Інститут телекомунікацій і глобального інформаційного простору
Національна академія наук України

Кваліфікаційна наукова праця
на правах рукопису

ТЕРЕНТЬЄВ ОЛЕКСАНДР МИКОЛАЙОВИЧ

Прим. № ____

УДК 004.89:519.22](043.3)

ДИСЕРТАЦІЯ
МОДЕЛІ, МЕТОДИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ В УМОВАХ НЕВИЗНАЧЕНОСТІ

05.13.06 – інформаційні технології

Подається на здобуття наукового ступеня доктора технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ О. М. Терентьєв
(підпис, ініціали та прізвище здобувача)

Науковий консультант
Трофимчук Олександр Миколайович,
член-кореспондент НАН України,
д.т.н., професор

Київ – 2021

АНОТАЦІЯ

Терентьєв О. М. Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності. – Кваліфікаційна праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора технічних наук за фахом 05.13.06 «Інформаційні технології». – Інститут телекомунікацій і глобального інформаційного простору, Національна академія наук України, Київ, 2021.

Дисертаційну роботу присвячено вирішенню актуальної науково-прикладної проблеми розроблення і використання моделей та методів прогнозування нелінійних нестаціонарних процесів (ННП) різної природи, призначених до використання у сучасних інформаційних системах підтримки прийняття рішень

У дисертації розроблено нову інформаційну технологію для реалізації та використання прогнозуючих моделей високого ступеня адекватності та якості, що дозволяє прогнозувати ННП різних типів. Розроблена інформаційна технологія ґрунтується на принципах багатомодельного підходу, інтеграції структурованої та неструктурованої інформації, системному використанні методів інтелектуального аналізу даних, моделювання, прогнозування та прийняття рішень. Науковими результатами досліджень є інформаційна технологія для розв’язання задач прогнозування ННП, удосконалені методи прогнозування ННП різної природи на основі використання ймовірнісних та комбінованих прогнозів, метод оцінювання параметрів математичних моделей та їх ансамблів, метод розкриття невизначеностей різних типів за допомогою ймовірнісно-статистичних методів з метою коректного розв’язання задач прогнозування розвитку обраних процесів,. підходи, що забезпечують адекватний опис причинно-наслідкових зав’язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних ННП, побудовано та реалізовано

інформаційну технологію для розв'язання задач моделювання та прогнозування ННП для її подальшого використання у системах підтримки прийняття рішень.

В роботі запропонований новий робастний непараметричний метод для аналізу ННП за їх подібністю. Суть методу полягає у пошуку процесів (часових рядів) досліджуваного типу зі схожою статистичною поведінкою, що в результаті допомагає зробити висновки стосовно наявність шаблонів поведінки досліджуваних процесів. Проведено аналіз складності алгоритму побудови оптимального шляху переміщення по матриці відстаней. Виконано огляд існуючих метрик для оцінювання якості побудови шляху переміщення, які забезпечують оптимізацію процесу аналізу подібності процесів.

У дисертаційному дослідженні запропоновано власну метрику, призначену для оцінювання ступення близькості досліджуваних часових рядів, яка підвищує коректність використання алгоритму аналізу подібності процесів.

Науковими результатами досліджень також є запропонована та реалізована методику побудови мережі Байєса за наявності прихованих вершин. Наведено приклад успішного застосування інформаційної технології аналізу даних на основі мереж Байєса для моделювання і прогнозування економічних процесів на регіональному рівні.

Запропоновано використання різних типів інформаційних технологій прогнозування, в основу яких покладено поетапне розкриття невизначеностей різної природи, уточнення результатів моделювання на кожному етапі дослідження. В основу запропонованої технології покладено багатомодельний підхід, який ґрунтується на використанні множини різнотипних моделей: регресійний аналіз, байєсівське моделювання і прогнозування, технологія на основі застосування методу аналізу подібності, аналізу неструктурованих даних. При цьому кожний тип моделей призначений для виконання поставленої конкретної задачі. Регресійне моделювання забезпечує формування моделей на основі множини

допустимих регресорів, їх оптимального вибору для конкретної моделі з наступним оцінюванням одно- або багатокрокового прогнозу.

У дисертаційному дослідженні запропонована методика моделювання, що має в основі пошук історичних аналогій поведінки досліджуваних процесів. В цьому випадку використовується історична вибірка даних та відрізок-шаблон для якого виконується пошук історичного аналогу, оскільки зазвичай для таких задач виміри прив'язані до часу.

Згідно з принципами багатомодельного підходу розроблено метод синтезу інформаційних технологій застосований для розв'язування задач прогнозування розвитку нелінійних нестаціонарних процесів різної природи, який ґрунтується на інтеграції різнотипної інформації й заснований на системному використанні методів аналізу даних, моделювання, методів прогнозування**.**

Запропонована інформаційна технологія реалізована із використанням спеціальних мов програмування SAS/Base та SAS/IML і має гнучку архітектуру. Методи, що були розроблені та реалізовані в роботі призначені насамперед для автоматизації процесу інтелектуального аналізу даних що описують досліджувані процеси.

Наукова новизна одержаних результатів визначається такими теоретичними і практичними результатами, отриманими автором:

Уперше:

– розроблено новий метод опрацювання невизначеностей, який ґрунтується на застосуванні теорії подібності процесів, який відрізняється робастністю результатів аналізу, що забезпечує отримання оцінок прогнозів високої якості за наявності неповних або спотворених даних;

– запропоновано новий метод побудови регресійних та ймовірнісно-статистичних моделей у формі мереж Байєса, який відрізняється можливістю врахування нестаціонарності і нелінійності стосовно змінних, що забезпечує високу адекватність моделей і якість прогнозів процесів досліджуваного типу;

– розроблено метод моделювання ННП різної природи в умовах невизначеності, який відрізняється від відомих урахуванням різних типів невизначеностей, що підвищує адекватність моделей і якість оцінок прогнозів за лінійними та нелінійними моделями;

– побудовано та досліджено ансамблі моделей для формального опису ННП, які відрізняються модифікованою комплексною структурою та високою адекватністю, що дозволяє підвищити якість оцінювання прогнозів розвитку досліджуваних процесів;

– запропоновано метод прогнозування, оснований на використанні адаптивного підходу до моделювання у поєднанні із статистичним та ймовірнісним моделюванням, що дає можливість урахувати структурно-параметричні невизначеності і забезпечує адекватний опис причинно-наслідкових зв'язків і можливих варіантів розвитку процесів різної природи під впливом груп внутрішніх та зовнішніх чинників;

– запропоновано нові моделі і методи створення інформаційних технологій розв'язування задач побудови математичних моделей для прогнозування ННП, які ґрунтуються на принципах багатомодельного та багатокритеріального підходів, інтеграції різнотипної інформації і засновані на системному використанні методів інтелектуального аналізу даних, ймовірнісно-статистичного моделювання, теорії подібності процесів, прогнозування і підтримки прийняття рішень, що підвищує обґрунтованість прийняття рішень в умовах наявності невизначеностей та ризиків різних типів;

– розроблено інформаційну технологію, в основу якої покладено поєднання принципів системного аналізу, методів обробки та оцінювання якості даних, прогнозного моделювання із використанням нових моделей та їх композицій, запропонованих критеріїв адекватності моделей, оцінок якості прогнозів, яка забезпечує високу якість проміжних та остаточних результатів дослідження ННП.

Удосконалено та розвинуто:

– інформаційну технологію розв’язання задач прогнозування розвитку досліджуваних процесів ННП, яка створює підґрунтя для прийняття ефективних рішень;

– метод оцінювання параметрів математичних моделей, який відрізняється комплексним застосуванням теорії оцінювання та байєсівського підходу, що забезпечує подолання проблеми зміщеності оцінок;

– інформаційну технологію, призначену для реалізації у системах підтримки прийняття рішень на основі системного підходу, множини методів ідентифікації і врахування невизначеностей, регресійного та інтелектуального аналізу даних, яка забезпечує побудову адекватних моделей досліджуваних процесів і обчислення високоякісних оцінок прогнозів.

Практичне значення отриманих результатів полягає у тому, що запропонована методики аналізу ННП перевірена експериментальним шляхом. Було побудовано множину довго-, середньо-, короткострокових та дуже короткострокових математичних моделей прогнозування навантаження енергосистеми в енергетиці. Методика була застосована під час торгівлі криптовалютою біткоїн в парі з доларами США, за вісім місяців отримано майже 43 центи прибутку на 1 долар, вкладений у біткоіни. Виконано застосування теорії подібності процесів для вирішення задачі прогнозування поширення коронавірусної хвороби, в рамках якого було підтверджено, що недостатнє охоплення населення тестуванням впливає на об’єктивну статистичну картину урахування кількості хворих та померлих внаслідок коронавірусної хвороби, тому для уникнення спотворення вхідних даних, в роботі використано підхід уточнення значень показників із урахуванням значень аналогів-країн. В рамках багатомодельного підходу успішно продемонстровано на прикладі розв’язання практичної задачі групування регіонів України за обсягом капітальних інвестицій, спрямовуваних на охорону навколишнього середовища. На основі запропонованої нового робастного непараметричного методу аналізу ННП за їх подібністю виконано

аналіз щодо продуктивності праці у сільськогосподарських підприємствах.

Запропоновані методи і моделі доведені до рівня практичної реалізації у вигляді інформаційної технології що дозволяє аналізувати та прогнозувати ННП в умовах невизначеності.

Отримані теоретичні та практичні результати дисертаційної роботи використовуються в науково-практичній діяльності наступних підприємств та установ – Державній службі України з лікарських засобів та контролю за наркотиками; ТОВ «Картезіан-Європа»; Smart Arbitrage Technologies Limited та у навчальному процесі Інституту прикладного системного аналізу Національного технічного університету «Київський політехнічний інститут імені Ігоря Сікорського», що підтверджується відповідними актами та довідками про впровадження.

Ключові слова: нелінійні нестаціонарні процеси, інформаційна технологія, прогнозуючі моделі, часові ряди, методи подібності, інтелектуальний аналіз даних, багатомодельний підхід.

Список публікацій здобувача за темою дисертації

I. Наукові праці, в яких опубліковані основні наукові результати дисертації

Монографії

1. Bidyuk P., Prosyankina-Zharova T., Terentiev O. Modelling Nonlinear Nonstationary Processes in Macroeconomy and Finances. *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing* / Ed. Hu Z., Petoukhov S., Dychka I., He M. Cham : Springer, 2019. Vol. 754. P. 735–745. URL: http://doi.org/10.1007/978-3-319-91008-6_72

2. Trofymchuk O. M, Bidiuk P. I., Prosiankina-Zharova T. I., Terentiev O. M. Decision support systems for modelling, forecasting and risk estimation: monography. Riga : LAP LAMBERT Academic Publishing. 2019. 176 p.

Статті у виданнях, індексованих у наукометричних базах

3. Terentyev A. N., Bidyuk P. I., Mironova A. V., Medin N. Y. Comparison of data mining methods while credit rating of natural persons. *Journal of Automation and Information Sciences*, 2009. № 41(10). P. 71–80. (Scopus)

4. Bidyuk P. I., Davidenko V. I.,Trofimenko D. V.,Terentyev A. N. Comparative analysis of estimation methods of vertices correlation while Bayesian networks construction. *Journal of Automation and Information Sciences*, 2010. №42(11). P. 36–45. (Scopus)

5. Prosyankina-Zharova T., Terentiev O., Bidyuk P., Makukha M. Features of SAS Enterprise Guide for probabilistic modeling system, macroeconomic analysis and forecasting. *Journal of Mathematics and System Science*. 2016. №6. P. 112–122. (Index of Copurnicus)

6. Bidiuk P. I., Prosiankina-Zharova T. I., Terentieev O. M., Lakhno V. A., Zhmud O. V. Intellectual technologies and decision support systems for the control of the economic and financial processes. *Journal of Theoretical and Applied Information Technology*. 2019. Vol. 96. No. 1. P. 71–87. (Scopus Q3)

7. Terentiev O. M., Prosiankina-Zharova T. I., Lahno V. A., Usatiuk Y. V. The features of the predictive computing modeling power system load in terms of reforming energy market. *Journal of Theoretical and Applied Information Technology*. 2020. Vol. 98. No. 2. P. 163–182. (Scopus Q3)

Статті у наукових фахових виданнях України

8. Терентьев А. Н., Бидюк П. И., Коршевнюк Л. А. Алгоритм вероятностного вывода в байесовских сетях. *Системні дослідження та інформаційні технології*. 2009. № 2. С. 107–111.

9. Бідюк П. І., Терентьєв О. М., Коновалюк М. М. Байєсівські мережі в технологіях інтелектуального аналізу даних. *Наукові праці [Чорноморського державного університету імені Петра Могили комплексу "Києво-Могилянська академія"]. Серія: Комп'ютерні технології*. 2010. Т. 134. Вип. 121. С. 6–16.

10. Кашкарьов А. О., Терентьєв О. М. Аналіз випадкових процесів за допомогою швидкого перетворення Фур'є. *Праці Таврійського державного агротехнологічного університету*, 2010. Вип. 10, Т. 10. С. 158–162.

11. Свиридюк Я. В., Дегтярьов А. О. Терентьєв О. М. Застосування нечітко-множинного методу для оцінювання ризику банкрутства корпорації. *Вісник Університету «Україна». Серія: Інформатика, обчислювальна техніка та кібернетика*. 2010. № 8. С. 105–108.

12. Терентьєв О. М., Бідюк П. І., Кузнєцова Н. В. Система підтримки прийняття рішень для аналізу фінансових даних. *Наукові вісті КПІ*. 2011. №1. С. 48–61.

13. Бідюк П. І., Коршевнюк Л. О., Терентьєв О. М., Просянкіна-Жарова Т. І. Аналіз інвестиційних і соціально-економічних процесів методами моделювання обмежених множин багатовимірних даних. *Наукові вісті КПІ*. 2012. №2. С. 87–93.

14. Терентьєв О. М., Кириченко В. Е., Связінська Н. О., Просянкіна-Жарова Т. І. Прогнозування фінансових ризиків з використанням наївного і доповненого деревом класифікаторів на основі байєсівських мереж. *Наукові вісті НТУУ КПІ*. 2016. №2. С. 60–68. URL: https://doi.org/10.20535/1810-0546.2016.2.63882

15. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України. *Реєстрація, зберігання і обробка даних*. 2016. Т. 18. № 3. С. 75–86.

16. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Застосування когнітивного та ймовірнісного моделювання у задачах формування сценаріїв розвитку соціально-економічних систем. *Наукові вісті НТУУ КПІ*. 2016. №5. С. 37–47. URL: https://doi.org/10.20535/1810-0546.2016.5.79876.

17. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І., Ефендієв В. В. Прогнозне моделювання нелінійних нестаціонарних процесів у рослинництві з використанням інструментів SAS Enterprise Miner. *Наукові вісті НТУУ КПІ*. 2017. №1. С. 24–36. URL: https://doi.org/10.20535/1810-0546.2017.1.87423.

18. Bidyuk P., Overmyer S., Prosyankina-Zharova T., Terentiev O. Methodology of Modelingand Forecasting Nonlinear Processesin Finances. *Наукові вісті НТУУ КПІ*. 2018. №1. С. 15–25. URL: https://doi.org/10.20535/1810-0546.2018.1.120361.

19. Bidyuk P., Prosyankina-Zharova T., Terentiev O., Medvedeva M. Adaptivemodellingforforecastingeconomicandfinancialrisksunderuncertaintyinter

msoftheeconomiccrisisandsocialthreats. *Технологічний аудит та резерви виробництва.* 2018. №4. С. 24–36. URL: https://doi.org/ 10.15587/2312-8372.2018.135483.

20. Modeling and forecasting financial and economic processes with decision support system / Bidyuk P. I. et al. *Наукові вісті КПІ.* 2019. № 5–6. С. 7–17. URL: https://doi.org/10.20535/kpi-sn.2019.5-6.176835.

II. Наукові праці, які засвідчують апробацію матеріалів дисертації

21. Бідюк П. І., Коршевнюк Л. О., Терентьєв О. М. Підтримка розв'язання слабкоструктурованих задач в органах державної влади. *Системный анализ и информационные технологии (САІТ-2012) :* матеріали 14 міжнар. наук.-техн. конф. (Київ, 24 квіт. 2012 р.). Київ : ННК "ІПСА" НТУУ "КПІ". 2012. С. 169–170.

22. Коршевнюк Л. О., Терентьєв О. М., Бідюк П. І. Методика побудови математичних моделей динамічних процесів.*Системний аналіз та інформаційні технології(SAIT 2013) :* матеріали 15-ї міжнар. наук.-техн. конф. (Київ, 27–31 трав. 2013 р.). Київ : ННК "ІПСА" НТУУ "КПІ", 2013. С. 288–289.

23. Prosyankina-Zharova T. I., Terentiev O. M., Bidyuk P. I., Gasanov A. Features of SAS Enterprise Guide for probabilistic modeling system, macroeconomic analysis and forecasting. *On control and optimization with industrial applications (COIA-2015*) : Proceedings of the fifth international conference (Baku, 27–29 August 2015), Baku : Institute of applied mathematics BSU, 2015. P. 365–368.

24. Терентьєв О. М., Связінська Н. О., Просянкіна-Жарова Т. І. Візуалізація типології регіонів України із використанням система SAS Enterprise Guide 7.1. *Геоинформационные системы и компьютерны етехнологии эколого-экономического мониторинга* : матеріали міжнар. наук.-

техн. конф. (Дніпропетровськ, 13–15 квітня 2016 р.). Днепропетровск : ГВУЗ «НГУ» МОН Украины, 2016. С. 21–25.

25. Терентьєв О. М., Просянкіна-Жарова Т. І., Бідюк П. І., Связинська Н. О., Кириченко В. Е. Text mining analysis of agriculture internet sources using SAS software. *Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMCI 2016)* : матеріали міжнар. конф. (м. Залізний порт, 24–28 трав. 2016 р.). Херсон : ХНТУ, 2016. С. 244–246.

26. Ivanova Y. V., Terentiev O. N., Korshevnyuk L. O., Prosyankina-Zharova T. I. Using modified logistic regression to increase customer responserate to marketing campaigns. *International conferenceon System Analysisand Information Technology (SAIT 2016) :* Proceedings of the 18-th conference(Kyiv, May 30 – June 2). Kyiv : Institutefor Applied System Analysis, NationalTechnical University of Ukraine “KPI”, 2016. С. 304–305.

27. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Методи заповнення пропусків даних у задачах прогнозного моделювання соціально-економічних процесів. *Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMCI 2016) :* матеріали міжнар. конф. (м. Залізний порт, 22–26 травня 2017 р.). Херсон : Вишемирський В. С., 2017. С. 185–187.

28. Бідюк П. І. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Застосування інструментів SAS BASE для дослідження ефективності методів обробки пропусків у вибірках даних з метою підвищення якості прогнозування показників продовольчої безпеки країни. *International conferenceon System Analysis and Information Technology (SAIT 2017*) : Proceedings of the 19-th. Conference (Kyiv, May 22–25 2017). Kyiv : National Technical University of Ukraine “KPI”, 2017. С. 253–255.

29. Трофимчук О. М., Бідюк П. І., Просянкіна-Жарова Т. І., Терентьєв О. М. Адаптивне моделювання і прогнозування у системах підтримки прийняття рішень сільськогосподарського призначення. *Сучасні*

інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях : колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2017. С. 21–28.

30. Trofymchuk O. M., Bidyuk P. I., Prosyankina-Zharova T. I., Terentiev O. M. For casting non stationary processesin demography, ecology, economy and finances. *Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях* : колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2018. С. 113–123.

31. Bidyuk P., Huskova V., Terentiev O. Client solvency estimation using intellectual data analysis approach, *Advanced Information Systems and Technologies (AIST 2018)* : Proceedings of the 6th International Conference (Sumy, 16–18 May 2018), Sumy : Sumy State University, 2018 . С. 12–16.

32. Bidyuk, P. Terentiev O., Prosyankina-Zharova T. Dynamic processes forecasting and risk estimation under uncertainty using decision support systems. IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) : Proceedings of the Proceedings of the conference (Kyiv, 29 May-2 June 2017). Kyiv : Igor Sikorsky Kyiv Polytechnic Institute. 2017. P. 795–800. (Scopus). URL: https://doi. 8100355

33. Шолохов О. В., Терентьєв О. М. Просянкіна-Жарова Т. І. Підвищення ефективності соціальних комунікацій на основі аналізу інтернет-джерел засобами text mining. *Прикладні системи та технології в інформаційному суспільстві :* матеріали міжнар. наук.-практ. конф. (Київ, 30 верес. 2020 р.). Київ : Київський нац. ун-т імені Тараса Шевченка, 2020. С. 23–44.

34. Trofymchuk O. M., Bidyuk P. I., Prosyankina-Zharova T. I., Terentiev O. M. Decision support system for modeling and forecasting ecological processes. *Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях*: колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2020. С. 23–44.

35. Korbicz J., Bidyuk P., Kuznietsova N., Terentiev O., Prosiankina-Zharova T. Multivariate distribution model for financial risks management. *The 9th International Conference "Information Control Systems & Technologies"(ICST2020)* (Odessa, Sept. 24–26, 2020). CEUR Workshop Proceedings, 2020. P. 416–429. (Scopus)

III. Наукові праці, які додатково відображають результати дисертації

Навчальні посібники

36. Згуровський М. З., Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Байєсівські мережі в системах підтримки прийняття рішень : навч. посіб. Київ : Едельвейс, 2015. 300 с.

37. Домрачев В. Н., Костецкий Р. И. Терентьев А. Н. SAS BASE: Основы программирования. Киев : Эдельвейс, 2014. 304 с.

38. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Прикладна статистика : навч. посіб. Вінниця : Едельвейс і К, 2013. 288 с.

Авторські свідоцтва

39. Трофименко Д. В., Давиденко В. І. Комп’ютерна програма «Bayesian Network Master BNetMaster» : авторське свідоцтво № 34443 України ; заявл. 09.06.2010; № 34657 ; опубл. 09.08.2010.

40. Колбасюк Д.А., Дегтярьов А. О., Терентьєв О. М. Комп’ютерна програма «Alpha»: авторське свідоцтво № 34191 України ; заявл. 21.05.2010; № 34419 ; опубл. 21.07.2010.

41. Бідюк П. І., Кириченко В. Е., Связінська Н. О. Комп’ютерна програма «IMLBayesNet» : авторське свідоцтво № 64117 України ; заявл. 17.12.2015 № 64562 ; опубл. 16.02.2016.

42. Щука Р. В., Іванов С. С., Терентьєв О. М., Бідюк П. І., Макуха М. П Комп’ютерна програма «SAS Similar Trajectories» № 72358 ; заявл.07.03.2017 ; опубл. 05.05.2017.

SUMMARY

Terentiev O. M. Models, methods and information technologies of forecasting nonlinear nonstationary processes in conditions of uncertainty. – Qualification scientific work on the rights of the manuscript.

The thesis for the degree of Doctor of Technical Sciences on specialty 05.13.06 «Information technologies». – Institute of Telecommunications and Global Information Sphere, National Academy of Sciences of Ukraine, Kyiv, 2021.

The dissertation is devoted to solving an actual scientific and applied problem of development and use of models and methods for forecasting future development of nonlinear non-stationary processes of different nature dedicated for using in modern information decision support systems.

The dissertation presents a new information technology for the implementation and use of predictive models of high degree of adequacy and quality, which allows predicting nonlinear nonstationary processes of different types. The developed information technology is based on the principles of a multi-model approach, integration of structured and unstructured information, systematic use of methods of data mining, modeling, forecasting and decision making. The scientific results of research are information technology for solving the problems of forecasting nonlinear non-stationary processes, improved methods for predicting nonlinear nonstationary processes of different nature based on the use of probabilistic and combined forecasts, improved methods for predicting nonlinear non-stationary processes of different nature based on the use of probabilistic and combined predictions, a method of estimating the parameters of mathematical models and their ensembles, a method of revealing uncertainties of different types using probabilistic and statistical methods in order to correctly solve problems of predicting the development of selected processes, approaches that provide an adequate description of the causal relationships of different groups of factors and identify possible options for the development of the studied nonlinear

nonstationary processes. It was built and implemented the information technology for solving problems of modeling and forecasting of nonlinear non-stationary processes for its further using in decision support systems.

It was proposed in the study a new robust nonparametric method for analysis of nonlinear non-stationary processes using their similarity. At the core of the method is the search for the processes (time series) under study with similar statistical behavior what provides the possibility for making conclusions regarding availability of reference behavior for the processes being studied. A complexity analysis for optimal path development algorithm across the matrix of distances was carried out. An analysis of existing metrics for estimating quality of development of the movement path that provides a possibility for optimization of the similarity analysis process. The dissertation studies propose original metric that should be used for estimating the degree of proximity of the processes under study. The metric enhances the correctness of application of the algorithm used for performing similarity analysis.

Among the scientific research results is proposed and implemented methodology for constructing Bayesian networks for the case of availability of hidden vertices. A successful example is provided illustrating successful application of the information technology for data analysis based on the Bayesian networks. The technology is used for modeling and forecasting economic processes at the regional level.

It was proposed an application of various types of the forecasting information technologies that are based upon step-by-step refinement of various type and nature uncertainties as well as improvement of modeling results at each step of the research. The technology proposed is founded upon so called multi-model approach that is based on the use of a set of various type models such as the following: regression analysis techniques, Bayesian modeling and forecasting, the technology on the basis of application of the similarity analysis method and analysis of unstructured data. Here each model type should be used for fulfilling specifically set problem. The regression modeling methodology provides the

possibility for constructing models on the basis of possible independent variables (regressors), their optimal selection for building specific model with subsequent estimation of single- or multistep forecasts.

The dissertation proposes the modeling methodology that is based upon the search of historical behavior analogies for the processes under study. In this case the historical data sample is used together with the reference-set for which the historical analogy is being searched because in solving such problems the measurements are usually linked to the time points.

According to the principles of multi-model approach the method of informational technology synthesis was developed and applied to solving the problem of forecasting future development for nonlinear non-stationary processes of various origin. The method is based on the integration of information of various type as well as on systemic usage of modern data analysis methods, mathematical modeling and forecasting techniques.

The information technology developed has been implemented with the use of specialized programming languages such as SAS/Base and SAS/IML, and has a flexible architecture. The methods developed and implemented in the frames of the research carried out are to be used for automating the processes of intellectual data analysis characterizing behavior of the processes being studied.

The scientific novelty of the results achieved is determined by the following theoretical and practical results developed by the author:

For the first time:

– it was developed a new method for processing the uncertainties that is based upon application of the processes similarity theory, and which is distinguished by the robustness with respect to the analysis results, what provides for generating high quality forecast estimates in conditions of incomplete and distorted data;

– a new method for constructing regression and probabilistic and statistical models in the form of Bayesian networks that is distinguished with the possibility of taking into account non-stationarity and nonlinearity with respect to the model variables, and provides for the high adequacy of the models constructed, and the

quality of the forecasts for the processes being studied;

– a new modeling method was developed for nonlinear non-stationary processes of various origin in conditions of uncertainty that is distinguished from the known ones with taking into consideration of various type uncertainties that results in higher adequacy of models and enhances quality of the forecast estimates for linear and non-linear models;

– it was constructed and studied the ensembles of models for the formal description of the nonlinear non-stationary processes that are distinguished by the modified complex structure and higher adequacy allowing for the enhancement of their future development forecast quality for the processes being studied;

– the forecasting method was proposed based upon application of the adaptive modeling approach together with the statistical and probabilistic modeling what provides the possibility for taking into consideration the structural and parametric uncertainties, and provides for the adequate description of causal relationships and possible future development alternatives for the process of various nature under influence of the groups of internal and external factors;

– the new models and methods were proposed for creating information technologies directed towards solving the problems of constructing mathematical models for forecasting nonlinear non-stationary processes that are based upon the principles of multi-model and multi-criteria approaches, integration of various type information and grounded on the systemic usage of the intellectual data analysis methods, probabilistic and statistical modeling, on the theory of processes similarity, theory of forecasting and decision support techniques what enhances substantiation of the decisions in conditions of availability of uncertainties and risks of various types;

– the information technology was developed the basis of which was created by the system analysis principles, the methods of processing and data quality estimation, forecast modeling with the use of the new models and their compositions, proposed criteria for the model adequacy analysis as well as forecasts quality estimation, that provide for the high quality of intermediate and

final data investigating results;

It was improved and further developed:

– the information technology for solving the problem of forecasting the nonlinear non-stationary processes being studied that creates the ground for making effective decisions;

– the method for parameter estimation of mathematical models that is distinguished with the complex application of estimation theory and Bayesian approach that provides for solving the problem of estimates bias;

– the information technology developed for application in the decision support systems based upon systemic approach, the set of the identification methods and taking into account possible uncertainties, application of the regression and intellectual data analysis methods that provides the possibility for constructing adequate models of the processes under study and estimation of high quality forecasts.

The practical value of the results achieved is in the following: the methodology for analysis of nonlinear non-stationary processes is proposed that was tested experimentally. The set of mathematical models was proposed for solving the problem of very short, short-, medium-, and long-term forecasting the load for energy supply system. The methodology was applied in practice for performing trading operations with bit-coin crypto-currency together with US dollars. Within the period of 8 months it was achieved the return of about 43 cents per one dollar invested into bit-coin.

Also the similarity theory was applied to solving the problem of forecasting the dissemination of corona-virus disease in the frames of which it was proved that insufficient number of testing the population influences objectively the whole situation characterized by the number of ill (infected) people and those who died due to the corona-virus. That is why the approach used in the study was directed towards avoidance of the data distortion by making use of the data refining with taking into consideration the data from the countries served as analogs.

It was demonstrated in the frames of the multi-model approach the

successful example of solving the practical problem of grouping the Ukrainian regions according to the volume of capital investments into environment conservation. On the basis of the proposed new nonparametric method for analysis the nonlinear non-stationary processes selected according to their similarity it was performed analysis of personnel productivity at the agricultural enterprises.

The methods and models proposed in the frames of the study performed are raised to the level of their practical implementation in the form of the information technology what creates realistic perspectives for analyzing and forecasting the nonlinear non-stationary processes functioning in conditions of uncertainty.

The theoretical and practical results produced in the frames of the dissertation studies are used in scientific and practical activities of the following enterprises: The State Service of Ukraine in the area of medical means and drug control; The Cartesian-Europe Ltd. Company; Smart Arbitrage Technologies Limited and in the scientific and educational processes of the Institute for Applied System Analysis at the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic”, that is proved by the appropriate official papers.

Keywords: nonlinear non-stationary processes, information technology, forecasting models, time series, similarity methods, intellectual data analysis, multi-model approach.

ЗМІСТ

СПИСОК СКОРОЧЕНЬ

BIC	-	Bayesian Information Criteria
MBR	-	Memory Base Reasoning - метод на основі виводу правил
MCMC	-	метод Монте-Карло з марковскими ланцюгами
MDL	-	міра опису мінімальної довжини (minimum description length)
MIT	-	Massachutetts Institute of Technology
MRE	-	міра Minimum Recognition Error
MSBs	-	mean squared biases – середньоквадратичне зміщення
NDVI	-	Normalized Difference Vegetation Index
OPTMODEL	-	модуль для вирішення задач лінійного та нелінійного програмування
PCA	-	Principal Component Analysis, метод головних компонент
PLS	-	partial least squares regression model - регресія часткових найменших квадратів
R2	-	коефіцієнт детермінації
RMSE	-	Root mean square error - середньоквадратична похибка
RSI	-	relative strength index - індекс відносної сили
SAS Base	-	мова програмування для середовища SAS
SAS EM	-	SAS Enterprise Miner
SAS/IML	-	SAS Interactive Matrix Language – інтерактивна матрична мова програмування у середовищі SAS
SBC	-	критерій Шварца-Байєса
SEMMA	-	Sample, Explore, Modify, Model, Assessment
SMA	-	simple moving average - просте ковзне середнє
SPSS	-	Statistical Package for the Social Sciences - статистичний пакет
SSE	-	sum square error - сумарна квадратична похибка
STEPWISE	-	методика поетапного включення регресорів
SWOE	-	smoothed weight of evidence – метод згладженого значення зваженої сукупності
Threshold	-	методи об’єднання рівнів атегоріальних змінних на основі значення порогу
USD	-	United States Dollar - долар США
VIF	-	Variance Inflation Factor
WOE	-	Weight of Evidence – вимірює статистичну значущість кожного класу змінної
АКФ	-	автокореляційна функція

АРКС - авторегресія із ковзним середнім
БД - база даних
ІТ - інформаційна технологія
КСС - кількість ступенів свободи
ЛФК - логарифмічна функція корисності
БМ - Мережа Байєса
ВВП - валовий внутрішній продукт
ВРП - валовий регіональний продукт
МГК - метод головних компонент
НКФ - нелінійна кореляційна функція
ННП - нелінійні нестаціонарні процеси
НФК - нелінійна кореляційна функція
ОМД - опис мінімлаьною довжиною
ОПР - особа що приймає рішення
ПК - персональний комп'ютер
СППР - система підтримки прийняття рішень
СУБД - система управляння базами даних
ФК - Фільтр Каламана
ЧАКФ - часткова автокореляційна функція
ЩРЙ - щільності розподілу ймовірностей

ВСТУП

Актуальність теми. Накопичення знать та розвиток сучасних інформаційних технологій дозволяють вирішувати актуальні задачі, пов'язані із прогнозування процесів різної природи, зменшуючи при цьому вплив невизначеностей на прийняття управлінських рішень. При цьому зростають вимоги до точності прогнозу, як на короткострокову, так і на довгострокову перспективу. Все це потребує нових методів, засобів та технологій. При цьому слід відзначити й те, що не зважаюч на накопичення значних обсягів інформації, невизначеність у вхідних даних не лише не зменшується, а й зростає. Інформація щодо динаміки розвитку досліджуваних процесів отримується з різних джерел, часто є неповною чи спотвореною, що значно ускладнює отримання прогнозів потрібної точності.

При цьому надмірне бажання деталізувати структуру, якнайточніше виявити закономірності, властивості досліджуваних явищ та процесів, зокрема, їх структури та взаємозв'язків, характер динаміки, тощо часо призводить до спотворення результатів. Тому актуальною є і залишається проблема розробки таких моделей, методів та інформаційних технологій прогнозування, за яких були б враховані основні властивості досліджуваного процесу для отримання прийнятного розв'язку даної задачі за конкремних умов.

Розвинута у роботі теорія вирішення проблеми прогнозування розвитку нелінійних нестаціонарних процесів в умовах наявності невизначеностей різного типу ґрунтується на дослідженнях та теоретичних результатах вітчизняних та закордонних вчених: О. М. Трофимчука, С. О. Довгого, О. Г. Наконечного, В. Є. Снитюка, М. З. Швиденка (*технології та системи підтримки прийняття рішень в умовах невизначеності*); В. С. Степашка, С. Г. Удовенка, П. І. Бідюка (*аналіз нелінійних систем та методи прогнозування*); М. М. Корабльова, Є. В. Бодянського., Ю. В. Крака, В. І. Литвиненка, С. А. Положаєнка (*методи інтелектуального аналізу даних*);

В. Я. Данилова, В. М. Азарскова (*теорія та прикладні методи системного аналізу*); О. І. Міхальова, Ю. П. Кондратенка, Ю. П. Зайченка, В. В. Кульби (*методи нечіткого аналізу і моделювання*); Т. Байєса, Дж. Перла, А. Л. Тулуп'єва (основи ймовірнісного підходу); Є. В. Малахова, С. Ф. Теленика, О. П. Гожого, О. С. Меняйленка, С. В. Цюцюри, Р. О. Коржа (*інформаційні технології*), С. Клайна, Дж. Краскала, А. А. Гухмана, М. Леонарда, Дж. Слоана, Т. Ли, Б. Ельшеймера, С. Шуберта (*теорія подібності процесів*), О. А. Павлова (*методи моделювання складних систем*), Ф. Бёрнштейна, С. Холсаппла, В. Г. Тоценка, О. Ф. Волошина, (*побудова систем підтримки прийняття рішень*), Д. В. Ланде, М. Бері, С. Аггарвала, Х. До Прадо, Е. Фемеди (*Text Mining*) та інших.

Незважаючи на наявність значної кількості досліджень з даної тематики, проблема розробки моделей, методів та інформаційних технологій прогнозування нелінійних нестаціонарних процесів залишається невирішеною повністю. Зокрема, слід відзначити відсутність загальних підходів до аналізу та попередньої обробки структурованих і неструктурованих даних, які описують нелінійні нестаціонарні процеси, в тому числі із застосуванням методів інтелектуального аналізу даних. Також нерозв'язаною залишається й низка задач, пов'язаних із виявленням характеру досліджуваних процесів, розробкою універсальних методик побудови моделей та їх ансамблів для отримання прогнозів нелінійних нестаціонарних процесів.

Вирішення вказаних задач неможливе без створення єдиної науково обґрунтованої інформаційної технології, яка б охоплювала весь комплекс задач побудови прогнозів високої якості, необхідних для прийняття ефективних управлінських рішень.

На підставі викладеного можна стверджувати, що проблема побудови інформаційних технологій аналізу та прогнозування нелінійних нестаціонарних процесів, які забезпечують високу обчислювальну ефективність, є актуальною.

Важливою і *актуальною науково-прикладною проблемою*, що розв'язується в дисертаційній роботі є підвищення ефективності підтримки прийняття рішень управління нелінійними нестаціонарними процесами в умовах невизначеностей та ризиків різної природи, на основі застосування сучасних методів системного аналізу, моделювання, прогнозування та інформаційних технологій.

Зв'язок роботи з науковими програмами, планами й темами. Тема дисертаційної роботи повністю відповідає науковим напрямам, за якими виконуються передові наукові дослідження в Інституті телекомунікацій і глобального інформаційного простору, зокрема, у сфері систем і методів прийняття рішень, прогнозування, побудови інтелектуальних інформаційних систем. Робота виконувалась в Інституті телекомунікацій і глобального інформаційного простору в рамках цільового проекту наукових досліджень НАН України 0120U000000 «Розробка інформаційної технології моделювання і прогнозування розвитку соціально-еколого-економічних систем в умовах невизначеності, нестаціонарності та ризику» (2020-2021 рр.) та наукової (науково-технічної) роботи 0121U109211 «Розробка інформаційно-аналітичної системи для дослідження спроможності та прогнозування розвитку територіальних громад із застосуванням методів інтелектуального аналізу даних» (2021-2023 рр.), а також в Інституті прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» в рамках науково-дослідницьких тем 0109U000300 «Побудова системи підтримки прийняття рішень на основі теорії байєсівських мереж для моделювання поведінки складних систем» (2009-2010 рр.), 0111U001241 «Розробка і реалізація методики інтелектуального аналізу даних із використанням теорії мереж Байєса та регресійного аналізу» (2011-2012 рр.), 0113U000650 «Розробка інформаційної технології моделювання та оцінювання фінансово-економічних ризиків із врахуванням невизначеностей різної природи (на основі байєсівських моделей)» (2013-2014 рр.).

Мета й завдання дисертаційного дослідження. Метою дослідження є створення нових моделей та методів виконання системних досліджень нелінійних нестаціонарних процесів різної природи, розроблення на їх основі методологічних засад розробки інформаційних технологій, що дозволить підвищити якість оцінок прогнозів та відповідних рішень в умовах невизначеності та ризику.

Для досягнення поставленої мети в роботі необхідно вирішити такі *завдання:*

– виконати огляд та аналіз сучасних методів моделювання і прогнозування розвитку нелінійних нестаціонарних процесів різної природи в умовах невизначеності;

– розробити системну методологію побудови математичних моделей для опису та прогнозування процесів, що характеризують розвиток нелінійних нестаціонарних процесів в умовах невизначеності;

– удосконалити методи прогнозування нелінійних нестаціонарних процесів різної природи на основі використання ймовірнісних та комбінованих прогнозів;

– запропонувати метод оцінювання параметрів математичних моделей та їх ансамблів, який дозволяє подолати зміщеність оцінок прогнозів;

– розробити метод розкриття невизначеностей різних типів з метою коректного розв’язання задач прогнозування розвитку обраних процесів;

– побудувати математичні моделі та ансамблі моделей обраних нелінійних нестаціонарних процесів для прогнозування їх розвитку з метою підтримки прийняття управлінських рішень в умовах інформаційної невизначеності;

– розробити та виконати апробацію запропонованих підходів до моделювання та прогнозування, запропонувати підходи, що забезпечують адекватний опис причинно-наслідкових зв’язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних нелінійних нестаціонарних процесів;

– виконати аналіз адекватності побудованих моделей та формування висновків на їх основі щодо можливості їх застосування для розв'язання задач оцінювання і прогнозування розвитку нелінійних нестаціонарних процесів;

– на основі системного підходу побудувати та реалізувати інформаційну технологію для розв'язання задач моделювання та прогнозування нелінійних нестаціонарних процесів у різних галузях досліджень для її подальшого використання у системах підтримки прийняття рішень.

Об'єкт дослідження – нелінійні нестаціонарні процеси різної природи.

Предмет дослідження – моделі, методи та інформаційні технології дослідження та прогнозування нелінійних нестаціонарних процесів і систем, прийняття рішень щодо їх розвитку в умовах наявності невизначеностей і ризиків.

Методи дослідження. *Теоретичну та методологічну основу дослідження* складають метод діалектичного пізнання стану та особливостей розвитку нелінійних нестаціонарних процесів, системний підхід до визначення факторів і механізмів їх розвитку. *При узагальненні теоретичних та методологічних засад* математичного моделювання та прогнозування розвитку нелінійних нестаціонарних процесів використані методи індукції та дедукції, аналізу і синтезу, аналогій і зіставлення, формалізації й моделювання, методи порівняння, дослідження часових рядів, кластерного та інтелектуального аналізу даних, сценарний підхід, елементи теорії прийняття рішень. *Для аналізу і обробки інформації* – методи системного, інтелектуального та статистичного аналізу даних, експертного оцінювання; *для моделювання нелінійних нестаціонарних процесів* – методи ймовірнісно-статистичного та економіко-математичного моделювання; *для побудови, аналізу і оцінювання прогнозів* – методи прогнозування на основі теорії подібності процесів, регресійного аналізу, ймовірнісно-статистичних моделей, дерев рішень та багатовимірних розподілів; *для розроблення інформаційної технології* використано методи системного аналізу, підтримки

прийняття рішень, багатокритеріального аналізу та байєсівського підходу; *для побудови практичних реалізацій* – методи та засоби проектування і реалізації систем підтримки прийняття рішень, прикладне програмування.

Наукова новизна отриманих результатів:

Вперше:

– розроблено новий метод опрацювання невизначеностей, який ґрунтується на застосуванні теорії подібності процесів, який відрізняється робастністю результатів аналізу, що забезпечує отримання оцінок прогнозів високої якості за наявності неповних або спотворених даних;

– запропоновано новий метод побудови регресійних та ймовірнісно-статистичних моделей у формі мереж Байєса, який відрізняється можливістю врахування нестаціонарності і нелінійності стосовно змінних, що забезпечує високу адекватність моделей і якість прогнозів процесів досліджуваного типу;

– розроблено метод моделювання нелінійних нестаціонарних процесів різної природи в умовах невизначеності, який відрізняється від відомих урахуванням різних типів невизначеностей, що підвищує адекватність моделей і якість оцінок прогнозів за лінійними та нелінійними моделями;

– побудовано та досліджено ансамблі моделей для формального опису нелінійних нестаціонарних процесів, які відрізняються модифікованою комплексною структурою та високою адекватністю, що дозволяє підвищити якість оцінювання прогнозів розвитку досліджуваних процесів;

– запропоновано метод прогнозування, оснований на використанні адаптивного підходу до моделювання у поєднанні із статистичним та ймовірнісним моделюванням, що дає можливість урахувати структурно-параметричні невизначеності і забезпечує адекватний опис причинно-наслідкових зв’язків і можливих варіантів розвитку процесів різної природи під впливом груп внутрішніх та зовнішніх чинників;

– запропоновано нові моделі і методи створення інформаційних технологій розв’язування задач побудови математичних моделей для

прогнозування нелінійних нестаціонарних процесів, які ґрунтуються на принципах багатомодельного та багатокритеріального підходів, інтеграції різнотипної інформації і засновані на системному використанні методів інтелектуального аналізу даних, ймовірнісно-статистичного моделювання, теорії подібності процесів, прогнозування і підтримки прийняття рішень, що підвищує обґрунтованість прийняття рішень в умовах наявності невизначеностей та ризиків різних типів;

– розроблено інформаційну технологію, в основу якої покладено поєднання принципів системного аналізу, методів обробки та оцінювання якості даних, прогнозного моделювання із використанням нових моделей та їх композицій, запропонованих критеріїв адекватності моделей, оцінок якості прогнозів, яка забезпечує високу якість проміжних та остаточних результатів дослідження нелінійних нестаціонарних процесів.

отримали подальший розвиток:

– метод аналізу інформації на основі засобів статистично-ймовірнісного моделювання, який дає змогу зменшити суб'єктивізм відбору найбільш значимих чинників в задачах прогнозування процесів різної природи, що створює передумови для прийняття обґрунтованих управлінських рішень;

– метод побудови моделей у формі байєсівської мережі для оцінювання розвитку динаміки процесів різної природи, який відрізняється коректністю формального опису за байєсівським інформаційним критерієм, що забезпечує обчислення високоякісних ймовірнісних оцінок прогнозів розвитку досліджуваних процесів;

– системна методологія побудови адаптивних моделей процесів різної природи, яка відрізняється удосконаленням існуючих методів моделювання в умовах невизначеності і нестаціонарності даних, методом структурно-параметричної адаптації, яка забезпечує підвищення адекватності моделей та якості оцінок прогнозів;

удосконалено:

– інформаційну технологію розв'язання задач прогнозування розвитку досліджуваних процесів нелінійних нестаціонарних процесів, яка створює підґрунтя для прийняття ефективних рішень;

– метод оцінювання параметрів математичних моделей, який відрізняється комплексним застосуванням теорії оцінювання та байєсівського підходу, що забезпечує подолання проблеми зміщеності оцінок;

– інформаційну технологію, призначену для реалізації у системах підтримки прийняття рішень на основі системного підходу, множини методів ідентифікації і врахування невизначеностей, регресійного та інтелектуального аналізу даних, яка забезпечує побудову адекватних моделей досліджуваних процесів і обчислення високоякісних оцінок прогнозів.

Практичне значення отриманих результатів полягає у тому, що результати теоретико-методологічного та емпіричного досліджень доведені до використання в науково-практичній діяльності, а також апробовані та впроваджені за безпосередньою участю автора в практичну діяльність підприємств та установ, що підтверджується відповідними довідками: Державної служби України з лікарських засобів та контролю за наркотиками; ТОВ «Картезіан-Європа» та Smart Arbitrage Technologies Limited. Результати використовуються у навчальному процесі Інституту прикладного системного аналізу Національного технічного університету «Київський політехнічний інститут імені Ігоря Сікорського»

Особистий внесок здобувача. Усі теоретичні та практичні результати, що виносяться на захист, отримані автором самостійно. Пошук та аналіз літературних джерел за тематикою дисертаційного дослідження виконано автором особисто. Наукові положення, висновки, рекомендації, що викладені в дисертації, розроблені особисто здобувачем. У працях, виконаних у співавторстві, автору дисертації належать: розроблено системну методологію моделювання і прогнозування нелінійних нестаціонарних процесів в енергетиці [6, 7]; запропоновано критеріальну базу для аналізу адекватності

лінійних і нелінійних моделей досліджуваних процесів [1-4, 6,7, 11, 12, 15, 16, 18, 19, 20]; розроблено метод моделювання і прогнозування нелінійних нестаціонарних процесів з використанням мереж Байєса [2-5, 8, 9, 12-14, 17,]; запропоновано метод ідентифікації невизначеностей ймовірнісно-статистичного типу [1, 2, 4-14, 17, 19, 20]; розроблено архітектуру системи підтримки прийняття рішень для аналізу та прогнозування нелінійних нестаціонарних процесів [6, 9, 21]; створено функціональну схему системи підтримки прийняття рішень із використанням запропонованої інформаційної технології [2, 6, 7, 12, 14, 16, 19, 20].

Апробація результатів дисертації. Основні положення дисертації було подано і обговорено більше, ніж на 17 міжнародних, всеукраїнських та науково-технічних конференціях: 12-й міжнародній науково-технічній конференції «Системний аналіз та інформаційні технології» (м. Київ, 24 квітня 2012 р.); 15-й міжнародній науково-технічній конференції «Системный анализ и информационные технологии» (м. Київ, 27–31 травня 2013 р.); 15-й міжнародній конференції «International conference on control and optimization with industrial applications» (м. Баку, 27-29 серпня 2015 р.); науково-технічній конференції «Геоинформационные системы и компьютерные технологии эколого-экономического мониторинга – 2016» (м. Дніпропетровськ, 13–15 квітня 2016 р.); міжнародній конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (м. Залізний порт, 24-28 травня 2016 р.); 18-й міжнародній науково-технічній конференції "Системний аналіз та інформаційні технології" (м. Київ, 30 травня – 2 червня 2016 р.); міжнародній конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (м. Залізний порт, 22-26 травня 2017 р.); 11-й всеукраїнській науково-практичній конференції «Моделювання та прогнозування економічних процесів» (м. Київ, 6-8 грудня 2017 р.); IEEE First Ukraine Conference on Electrical and Computer Engineering (м. Київ, 29 травня – 2 червня 2017 р.); 16-й міжнародній науково-практичній конференції «Сучасні

інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях» (м. Київ, Пуща-Водиця, 3-4 жовтня 2017 р.); 6-й міжнародній конференції «Control and Optimization with Industrial Applications» (м. Баку, 11-13 липня 2018 р.); 6-й міжнародній конференції «Advanced Information Systems and Technologies, AIST 2018» (Суми, 16-18 травня 2018 р.); 1st International Conference on Computer Science, Engineering and Education Applications (м. Київ, 18 -20 січня 2018 р.); 14-й міжнародній науково-практичній конференції «Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях: актуальні питання» (м. Київ, 06-07 жовтня 2020р.); 4-й міжнародній науково-практичній конференції «Прикладні системи та технології в інформаційному суспільстві» (м. Київ, 30 вересня 2020 р.); 14-й міжнародній конференції «Application of information and communication technologies» (м. Ташкент, 07-09 жовтня 2020 р.); 9-й міжнародній конференції «Information Control Systems & Technologies» (м. Одеса, 24–26 вересня 2020 р.).

Публікації. Основні наукові результати дисертації опубліковані у 42 наукових працях, у тому числі, 2 колективних монографіях, 18 статтях у фахових виданнях, в тому числі 5 статей у закордонних фахових виданнях, з них 4 – у виданнях, що індексуються у наукометричній базі Scopus, 15 – у збірниках наукових праць, матеріалах і тезах міжнародних і національних конференцій. Також опубліковано 7 робіт, що додатково відображають наукові результати дисертації: 3 навчальні посібники, 4 авторських свідоцтва на комп’ютерну програму.

Структура та обсяг роботи. Робота складається із вступу, 7 розділів, списку використаних джерел та додатків. Загальний обсяг роботи складає 293 сторінки, з яких основного тексту 285 сторінок, 77 ілюстрацій, 58 таблиць, 12 додатків. Список використаних джерел налічує 274 найменування на 32 сторінках.

РОЗДІЛ 1
СИСТЕМОЛОГІЧНИЙ АНАЛІЗ ПРОБЛЕМ ПРОГНОЗУВАННЯ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ

1.1 Загальна характеристика проблеми дослідження нелінійних нестаціонарних процесів (ННП)

Тенденції, характерні для сучасного етапу розвитку суспільства, концентрують увагу на необхідності вирішення фундаментальних проблем, пов'язаних із дослідженням та прогнозуванням нелінійних нестаціонарних процесів різної природи, розвиток яких триває в умовах невизначеності та ризику. Тобто зростає потреба у постійному удосконаленні методів прогнозування за рахунок формування принципово нових підходів до формалізації об'єкта прогнозування, побудови моделей (ансамблів моделей) та опрацювання результатів прогнозування, які слугуватимуть підґрунтям для прийняття коректних управлінських рішень.

Побудова моделей та удосконалення методів потребують розробки сучасних інформаційних технологій, які дозволяють автоматизувати вирішення комплексу задач, пов'язаних з аналізом значних обсягів інформації. Остання отримується з різних джерел, часто є неповною чи спотвореною. Існує необхідність виявлення в ній закономірностей, відображення властивостей досліджуваних явищ та процесів, зокрема, циклічності розвитку, сезонності, подібності процесів тощо.

Теоретична основою дослідження є роботи закордонних та вітчизняних вчених та фахівців в галузі інформаційних технологій, зокрема, О. М.Трофимчука, С. О. Довгого, О. Г. Наконечного, В. Є. Снитюка, М. З. Швиденка (технології та системи підтримки прийняття рішень в умовах невизначеності); В. С. Степашка, С. Г. Удовенка, П. І. Бідюка (аналіз нелінійних систем та методи прогнозування); М. М. Корабльова, Є. В. Бодянського., Ю. В. Крака, В. І. Литвиненка, С. А. Положаєнка (методи

інтелектуального аналізу даних); В. Я. Данилова, В. М. Азарскова (теорія та прикладні методи системного аналізу); О. І. Міхальова, Ю. П. Кондратенка, Ю. П. Зайченка, В. В. Кульби (методи нечіткого аналізу і моделювання); Т. Байєса, Дж. Перла, А. Л. Тулуп'єва (основи ймовірнісного підходу); Є. В. Малахова, С. Ф. Теленика, О. П. Гожого, О. С. Меняйленка, С. В. Цюцюри, Р. О. Коржа (інформаційні технології), С. Клайна, Дж. Краскала, А. А. Гухмана, М. Леонарда, Дж. Слоана, Т. Ли, Б. Ельшеймера, С. Шуберта (теорія подібності процесів), О. А. Павлова (методи моделювання складних систем), Ф. Бёрнштейна, С. Холсаппла, В. Г. Тоценка, О. Ф. Волошина, (побудова систем підтримки прийняття рішень), Д. В. Ланде, М. Бері, С. Аггарвала, Х. До Прадо, Е. Фемеди (Text Mining) та інших.

Незважаючи на наявність значної кількості досліджень з даної тематики, проблема розробки моделей, методів та інформаційних технологій прогнозування нелінійних нестаціонарних процесів залишається невирішеною повністю. Зокрема, слід відзначити відсутність загальних підходів до аналізу та попередньої обробки структурованих і неструктурованих даних, які описують нелінійні нестаціонарні процеси, в тому числі із застосуванням методів інтелектуального аналізу даних. Також нерозв'язаною залишається й низка задач, пов'язаних із виявленням характеру досліджуваних процесів, розробкою універсальних методик побудови моделей та їх ансамблів для отримання прогнозів нелінійних нестаціонарних процесів.

Наявність нестаціонарних процесів характерна для сфери фінансів, економіки, технологій і технічних систем, біології тощо [1-6, 20-27]. Нестаціонарність проявляється у формі детермінованого або стохастичного тренду, зміні в часі дисперсії та коваріації процесу (дисперсійно-коваріаційна нестаціонарність). Так, вузли машин і механізмів з часом зношуються, що призводить до зміни їх розмірів. Наприклад, тривале спрацювання підшипника призводить до появи поперечних та поздовжніх коливань осі; при цьому амплітуда і дисперсія цих коливань збільшуються з часом.

Зношуваність ріжучого інструменту призводить до погіршення якості оброблюваної поверхні, збільшення відхилень розмірів деталей від нормативу. У живих організмів спостерігається зростання вікових обмежень стосовно рухливості органів, можливості запам’ятовувати інформацію, тобто з’являється негативний тренд, який свідчить про погіршення фізичних та когнітивних характеристик індивідуума.

Напрям і характер розвитку тренду, дисперсія і стандартне відхилення – це основні статистичні характеристики, які використовуються в системах контролю якості на виробництві; вони утворюють основу для створення контрольних карт, які містять інформацію стосовно поточної якості продукції, тощо.

Розвиток всіх згаданих типів процесів має нестаціонарний характер, що зумовлено особливостями розвитку, функціонуванням в умовах множини випадкових збурень, не завжди непередбачуваною поведінкою складових та зовнішніми впливами т. ін. Досліджувані процеси досить часто містять детерміновані і стохастичні тренди, як індикатори довгострокових і середньострокових змін [28-30].

Ще одним типом нестаціонарності, що часто має місце на практиці, є змінна у часі дисперсія, тобто наявність гетероскедастичних процесів. Обидва типи нестаціонарності вимагають застосування спеціальних підходів до побудови математичних моделей та функцій прогнозування на їх основі.

Сучасні процеси у фінансах, економіці, екології та суспільсьві також характеризуються високою динамікою змінних, наявністю старших похідних, що призводить до додаткових труднощів при моделюванні і прогнозуванні [33-42].

Отже, нелінійні нестаціонарні процеси, характерні для різних сфер, відповідно, мають особливості розвитку та протікають під впливом багатьох специфічних факторів [37-51]. Тому кожний клас таких процесів потребує відповідного формального опису із використанням математичних моделей різних типів, методів інтелектуального аналізу даних тощо.

Таблиця 1.1 – Моделі, що застосовуються у інформаційній технології прогнозування нелінійних нестаціонарних процесів

Параметричні	Непараметричні
Авторегресійні моделі [52-56] - AR, ARMA, ARIMA [57] - ARCH / GARCH [57, 58]	Функціональної декомпозиції
Нелінійні регресійні моделі та їх різновиди [59] - GLM [60, 61, 62, 63]; - PLS [64, 65, 66]; - LARS [67, 68]; - логіт / пробіт бінарна та множинна регрессії [69, 70, 71, 68]	Дерева рішень [72-75]: - C4.5 [76]; - Кааса [77, 68]; - ID3; - CART [78].
Нейронні мереж [78, 79] - багатошаровий перцептрон; - згорткові.	Топологічні методи на основі метрик близькості [80]: - кластеризації; - класифікації.
Приховані ланцюги Маркова [81]	Непараметричні байєсівські підходи [82]
Машина опорних векторів [68, 83, 84]	Мережі Байєса [85-91]: - дискретні [92, 93]; - динамічні; - гібридні.
Адаптивного короткострокового прогнозування [94]: - модель дампфуючого тренду [95]; - модель Вінтерса з адитивною або мультиплікативною сезоностями [96].	Спеціалізовані методи - MBR [68, 97]; - аналізу подібності поведінки рядів значень [98-100].
Фільтр Калмана [101-103]: - нелінійний ФК	

Не зважаючи на те, що сучасні методи, моделі та їх комбінації, розроблені та широко застосовуються у прогнозному моделюванні, відсутній єдиний підхід до побудови інформаційних технологій, який би дозволив реалізувати різні підходи до прогнозного моделювання. Тому автором

запропоновано схему аналізу та прогнозування ННП, в основу якої покладений комбінованого використання методів аналізу даних, побудови моделей і прогнозів, прийняття рішень у задачах управління розвитком нелінійних нестаціонарних процесів рис. 1.1)[103].

Рисунок 1.1 – Загальна схема аналізу нелінійних нестаціонарних процесів, побудови моделей та оцінювання прогнозів

В роботі пропонується наступний метод моделювання ННП, який відрізняється тим, що формально описується окремо лінійною та нелінійною складовими досліджуваного процесу і забезпечує підвищення адекватності

моделі в цілому, а також підвищення якості оцінок прогнозів, що обчислюються за допомогою цієї моделі [57].

Загальну тенденцію розвитку (спаду чи зростання) того чи іншого процесу характеризують його трендом. Тренд можна трактувати, також, як поточне середнє значення відповідного часового ряду даних. Наприклад, зростання валового внутрішнього продукту (ВВП) розвинених країн на довгих часових інтервалах (десятки років) характеризується детермінованим позитивним трендом, а ВВП країн з перехідною економікою часто характеризується негативним (спадаючим) трендом [2, 5, 18, 36, 41, 105]. Також можна сказати, що детермінований тренд характеризує довгострокові зміни процесів [52, 56, 95, 106]. Якщо тренд швидко змінюється у часі в довільному напрямку, то його називають стохастичним. До таких процесів відносять, наприклад, процеси формування цін на біржах, значні коливання курсів валют, зумовлені такими випадковими процесами: загальна економічна нестабільність, світові фінансові кризи, локальні війни тощо.

Пропонована методика моделювання ННП складається з наступних етапів.

Етап 1. Виділяється нелінійна частина шляхом побудови авторегресії вищого порядку, наприклад, десятого.

Етап 2. Якщо залишки нелінійні, то підбирають нелінійну частину, у класі регресійних моделей та моделей на основі інтелектуального аналізу даних.

Етап 3. Якщо модель в результаті неадекватна, то продовжується ітераційний підбір.

Етап 4. В свою чергу лінійна модель AP(10) оптимізується з точки зору зменшення параметрів.

Етап 5. Останнім кроком є об’єднання лінійної і нелінійної складових у єдину модель, яка використовується для генерування оцінок прогнозів.

На рис. 1.2 схематично представлений пропонована методика моделювання ННП.

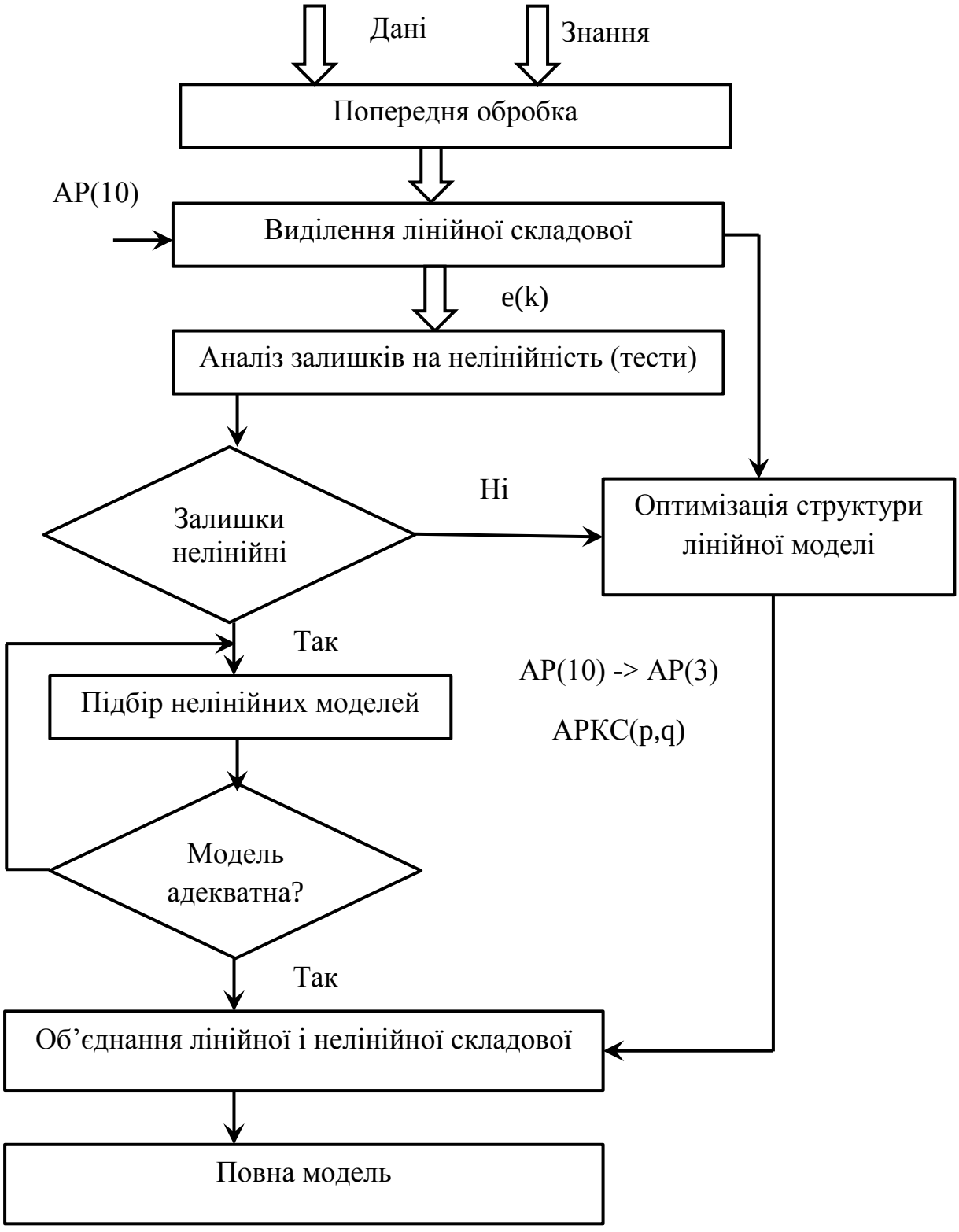


Рисунок 1.2 – Блок-схема методу моделювання ННП

Представлена на рис. 1.2. процедура складається з двох частин: лінійної та нелінійної. При цьому, для побудови моделей лінійної і нелінійної

складових використовуються ітераційні оптимізаційні процедури, які забезпечують адекватність моделі в цілому.

В загальному випадку для оцінювання параметрів моделі використовують різні методи: МСМС, МНК, рекурсивний МНК, нелінійний МНК, метод максимальної правдоподібності інші [56, 107, 108-120].

Наведена схема реалізації пропонованої методики може бути модифікована та доповнена залежно від потреб аналітика-користувача. Для пошуку нелінійної частини може бути використана нелінійна регресія (логістична регресія) або БМ. В свою чергу результати прогнозування нелінійної частини можуть бути вхідними змінними для побудови регресійної моделі (інших моделей), що уточнюють та доповнюють прогноз, отриманий для нелінійної частини.

Будь-яка прогнозна модель повинна вирішувати три необхідні завдання: надати правило перетворення вимірювання в прогноз, мати засіб для вибору важливих вхідних змінних з потенційно великої кількості кандидатів, бути придатною до регуляції складності для компенсації навчальних даних з шумами.

Під терміном «прогнозні моделі» [68, 69, 109] розуміють такі поняття як: прогнозна модель; контрольована класифікація; контрольоване навчання.

Однак, поняттю «прогноз» єдиного і загальноприйнятого визначення немає, хоча й існує значна кількість різноманітних визначень даного поняття.

Означення за Вікіпедією: прогноз (передбачення) – це науково обґрунтоване судження про можливі стани об'єкта в майбутньому і (або) про альтернативні шляхи і терміни їх здійснення [121].

У вузькому сенсі, це розподіл усіх суджень про майбутній стан об'єкта дослідження.

До основних методів прогнозування відносять:

- статистичні методи [110, 111];;
- експертні оцінки [112];;
- методи моделювання [21, 79];

– інтуїтивні (на основі досвіду фахівця, щодо застосування наукових методів в даному випадку).

В рамках даної роботи термін прогноз розглядається та використовується в рамках інформаційної технології прогнозного моделювання. нелінійних нестаціонарних процесів.

Під терміном «прогнозні моделі» [68, 69, 109] розуміють такі поняття:

– прогнозна модель;

– контрольована класифікація;

– контрольоване навчання.

Прогнозні моделі будуються на основі записів історичних подій і використовуються для прогнозування майбутніх повторень цих подій.

Прогнози – це найкращі оцінки для невідомих скорингових даних цільової змінної на основі відомих вхідних змінних і зв'язків між вхідними та цільової змінної.

Будь-яка прогнозна модель повинна вирішувати завдання [35, 108]:

– надати правило перетворення вимірювання в прогноз;

– мати засіб для вибору важливих вхідних змінних з потенційно великої кількості кандидатів;

– бути здатною до регуляції складності для компенсації навчальних даних з шумами.

В загальному випадку навчальні дані використовуються для створення моделі або правила, які пов'язують вхідні змінні з цільовою. В свою чергу прогнози класифікуються за трьома типами – рішення, рейтинги, оцінки [35, 108,].

Прогнози-рішення. В літературі замість терміну "прогноз-рішення" більш поширена назва "класифікація". Зазвичай будь-яке рішення пов'язане з якоюсь дією. Наприклад, класифікувати клієнта як «хорошого» або «поганого». Також класифікація використовується в задачах розпізнавання почерку, виявлення шахрайства, проведення рекламних акцій за допомогою прямих розсилок. Прогнози-рішення зазвичай відносять до категоріальної

цільової змінної [113]. Тому їх ідентифікують як первинне, вторинне та третинне рішення. Зазвичай, коли цільова змінна має категоріальний рівень виміру (бінарний, номінальний, порядковий), то оцінка моделі припускає прогнози-рішення.

Прогнози-рейтинги. Впорядковують спостереження на основі зав'язків вхідних змінних з цільовою. Модель намагається впорядковувати спостереження з більш високими значеннями вище спостережень з нижчими значеннями. Спостереження з більш високими значеннями мають великі скорингові бали. Фактично, обчислені бали несуттєві, важливий тільки відносний порядок. Приклад прогнозу-рейтингу – кредитний скоринг. Прогнози-рейтинги можуть бути перетворені в прогнози рішення, якщо прийняти первинне рішення для спостереження вище певного значення порогу, а вторинне і третинне рішення – для спостережень, нижчих відповідних значень порогу.

Прогнози-оцінки. Виконують апроксимацію очікуваного значення цільової змінної в залежності від вхідних значень. Для спостережень з чисельними цільовими змінними це число можна охарактеризувати як середнє значення цільової змінної за всіма спостереженнями, що мають поточні вхідні виміри. Для спостережень з категоріальними цільовими змінними це число може дорівнювати імовірності конкретного результату цільової змінної.

Прогнози-оцінки можуть бути перетворені як у прогнози-рішення, так і в прогнози рейтинги. Рішення можна оцінити по точності або проценту помилкової класифікації, рейтинги по узгодженості або неузгодженості, а оцінки за середньоквадратичною похибкою.

1.2 Формальне визначення нестаціонарності та нелінійності

Означення від MIT університету [30]. Нелінійна система – це система в якій зміна вихідних даних не пропорційна вхідних [30].

Означення за загальноекономічним і економіко-математичним словником Лопатнікова. Нелінійна система (процес) – динамічна система, в якій протікають процеси, що описуються нелінійними диференційними рівняннями [114].

Означення за Кембриджським словником: Нелінійний процес – це процес в якому один фактор нечітко або не безпосередньо впливає на інший [115].

Означення нелінійності. Процес називають нелінійним, якщо його неможливо описати з достатньою адекватністю лінійною моделлю. (Вікіпедія)

Означення за українською Вікіпедією [116]. Нелінійна система – динамічна система, в якій протікають процеси, описувані нелінійними диференціальними рівняннями.

Означення за англійською Вікіпедією [117]. Під динамічною системою розуміють систему будь-якої природи (фізичну, хімічну, біологічну, соціальну, економічну і так далі), стан якої змінюється (дискретно або неперервно) в часі. Нелінійна динаміка використовує при вивченні систем нелінійні моделі – частіше всього диференціальні рівняння і дискретні відображення.

Означення нестаціонарності. Дані носять нестаціонарний характер, якщо математичне очікування, дисперсія або коваріація змінюється з часом. Нестаціонарна поведінка може бути представлена у вигляді тренду, циклів, випадкового блукання або комбінації цих трьох факторів. Це означення можна зустріти в цій та близькій за саме цим сенсом і в [119, 52, 56, 95, 106, 118].

Означення нестаціонарності. Випадковий процес називають нестаціонарним, якщо хоча б один з трьох його основних статистичних параметрів – математичне сподівання, дисперсія або коваріація – змінюється у часі на інтервалі дослідження [17, 28, 29, 119].

Типи нелінійної динамічної поведінки:

– амплітудні осцилятори – будь-які коливання, присутні в системі, припиняються внаслідок якоїсь взаємодії з іншою системою або зворотного зв'язку тією ж системою [120].
– хаос – значення системи які неможна передбачити у далекому майбутньому, але при цьому їх флуктуації є не періодичними [121].
– мультистабільність – наявність двох і більше стабільних станів .
– солітони – одиночні хвилі, що само-підсилюється [122].
– граничні цикли – асимптотичні періодичні орбіти, до яких притягуються дестабілізовані нерухомі точки [24].
– автоколивання – коливання зворотного зв'язку, що відбуваються у відкритих дисипативних фізичних системах [123].

Нестаціонарні процеси можна розбити на дві групи [29, 28, 17]:

– детерміновані – характеризуються наявністю поліноміального тренду, нестаціонарної дисперсії, періодичністю.
– стохастичні – інтегрованого типу, тобто випадкове блукання, гетероскедостичність та інше.

1.3 Формально-математичне визначення та виявлення нестаціонарності процесів

Для виявлення нестаціонарності часових рядів використовуються наступні підходи [69, 37, 109, 68, 108]:

− візуальний аналіз графічного представлення часового ряду на наявність змінних математичного сподівання, дисперсії та коваріції;
− тести на відсутність сталості статистичних характеристик,

Процеси, представлені часовими рядами, називають стаціонарними, якщо три основні статистичні характеристики відповідного часового ряду не залежать від часу, а саме: математичне сподівання, дисперсія та коваріація [52, 56, 122, 95].

Формально стаціонарність процесу формулюють наступним чином:

– $\mu_x = E[x(k)] = const$ (постійне математичне сподівання);

– $\text{var}[x(k)] = E\left\{\left(x(k) - E[x(k)]\right)^2\right\} = E\left\{\left(x(k) - \mu_x\right)^2\right\} = const$ (постійна дисперсія);

– $\gamma_x = E\{[x(k) - \mu_x][x(k-s) - \mu_x]\} = const$, $s = 0,1,2,\ldots$ (постійна коваріація).

Якщо хоча б одна із наведених статистичних характеристик процесу змінюється в часі, то процес називають нестаціонарним. Існують деякі інші визначення стаціонарності, але наведене є самим поширеним і саме воно найчастіше використовується на практиці.

У випадку, коли $E[x(k)] \neq const$, тобто математичне сподівання змінюється в часі, то такий процес називають процесом з трендом або інтегрованим процесом (по аналогії із характером зміни сигналу на виході інтегратора) або процесом з одиничними коренями (відповідного характеристичного рівняння). Тренд (поточне середнє) може бути зростаючим або спадаючим, а за характером зміни в часі може бути детермінованим або стохастичним [126].

Для опису динаміки розвитку процесів з трендами застосуємо рівняння авторегресії з ковзним середнім (АРКС), яке має таку структуру [126]:

$$y(k) = \textit{дрейф} + \textit{детермінований тренд} + \textit{авторегресія} + \\ + \textit{сезонна компонента} + \textit{ковзне середнє}.$$

Наведена структура дає можливість описувати різноманітні детерміновані та випадкові ефекти, які наявні в динамічних процесах, у тому числі й тренди.

Детермінований тренд описують вибраною функцією, наприклад, поліномом від часу, сплайном, експонентою, комбінацією тригонометричних функцій і т. ін. Часто використовують поліноми від часу вигляду (1.1) [108]:

$$y(k) = a_0 + a_1 \cdot k + a_2 \cdot k^2 + \ldots + a_m \cdot k^m + \varepsilon(k), \qquad (1.1)$$

де k – дискретний час, який зв'язаний з неперервним реальним часом t через період реєстрації (дискретизації) даних: $t = kT_s$; $\varepsilon(k)$ – випадкова змінна, оцінку якої можна знайти після оцінювання рівняння: $\hat{\varepsilon}(k) = e(k)$, де $e(k)$ – похибка моделі. Очевидно, що після оцінювання моделі послідовність значень $\{e(k)\}$ буде містити всі коливання, що накладаються на тренд.

Випадкові тренди, тобто тренди, які не можна описати з необхідною точністю за допомогою детермінованих функцій, моделюють за допомогою випадкових процесів [108, 126].

Таким чином, описуючи тренд рівнянням (1.1), ми фактично видаляємо його з процесу і повна модель процесу буде складатись щонайменше з двох рівнянь: рівняння (1.1) для тренду і рівняння АРКС (*p*,*q*), яке описує коливання, що накладаються на тренд.

Тренд може бути видалений з процесу (даних) за допомогою різниць. Так, перші різниці видаляють тренд першого порядку (лінійний тренд), другі різниці видаляють квадратичний тренд і т.д. Наприклад, нехай $y(k) = a_0 + a_1 \cdot k$. Перші різниці цього процесу (1.2) [108]:

$$\Delta y(k) = y(k) - y(k-1) = a_0 + a_1 \cdot k - [a_0 + a_1 \cdot (k-1)] = a_1 \qquad (1.2)$$

призводять до видалення лінійного тренду.

Очевидно, що після видалення тренду ми вже не зможемо його спрогнозувати. Докладно задача моделювання процесів з трендом буде розглянута в подальшому.

Якщо процес містить сезонний ефект, то він враховується шляхом введення в праву частину рівняння залежної змінної із затримкою, що дорівнює періоду сезонного ефекту або введення складової ковзного

середнього з тією ж затримкою. Для зменшення дисперсії процесу з сезонним ефектом застосовують так звані “сезонні” різниці, тобто різниці вигляду [35]:

$$\Delta_4 y(k) = y(k) - y(k-4),$$

де “4” означає періодичність сезонного ефекту, в даному випадку “4” це лише приклад можливої ситуації..

Тест Дікі-Фуллера (Dickey-Fuller test, DF-тест) – це методика, яка використовується в прикладній статистиці і економетриці для аналізу часових рядів для перевірки на стаціонарність [125].

Даний тест заснований на оцінці параметру $\lambda = a_1 - 1$ рівняння

$$\Delta y_t = \lambda \cdot y_{t-1} + \varepsilon_t$$

еквівалентного рівнянню авторегресії

$$y_t = a_0 + a_1 \cdot y_{t-1} + \varepsilon_t$$

Також його називають тестом на одиничний корінь.

Нульова гіпотеза:

$$H_0: \lambda = 0;\ H_1: \lambda < 0.$$

Якщо значення t-статистики Стьюдента для параметру λ менше нижнього порогового значення DF-статистики, то нульову гіпотезу $\lambda = 0$ (про наявність одиничного кореня $a_1 = 1$) треба відхилити і прийняти альтернативну о стаціонарності процесу y_t.

Таблиці тесту Дікі-Фулера розраховані для рівнів значимості 1, 5 та 10%. Вказані в таблиці значення DF-тесту – від’ємні.

Таблиця 1.2 – Критичні значення статистики Дикі-Фулера при 1% рівні значимості

Розмір вибірки	AR-модель	AR-модель з константою	AR-модель з константою та трендом
25	-2,66	-3,75	-4,38
50	-2,62	-3,58	-4,15
100	-2,60	-3,51	-4,04
Нескінченність	-2,58	-3,43	-3,96

Є три основні версії тесту [125]:

1. Тест на одиничний корінь:

$$\Delta y_t = \lambda \cdot y_{t-1} + \varepsilon_t$$

2. Тест на одиничний корінь з самопливом:

$$\Delta y_t = a_0 + \lambda \cdot y_{t-1} + \varepsilon_t$$

3. Тест на одиничний корінь з самопливом і визначним відхиленням:

$$\Delta y_t = a_0 + a_1 \cdot t + \lambda \cdot y_{t-1} + \varepsilon_t$$

Кожна версія тесту має власні критичні значення, які залежать від розміру вибірки.

У всіх випадках нульовою гіпотезою є наявність одиничного кореня, $\lambda = 0$.

За своє суттю тест можна пояснити наступним чином. Якщо ряд є стаціонарним, то він намагається повернутися до свого середнього значення.

Якщо ж рід інтегрований то, як і у випадку моделей випадкового блукання, зміни значень ряду будуть відбуватися незалежно від поточного значення ряду.

Інтегровані нестаціонарні процеси – це процеси для яких за допомогою послідовного застосування операцій взяття послідовних різниць з нестаціонарних часових рядів можна отримати стаціонарні ряди [52, 108, 56].

Послідовні різниці стохастичного процесу визначаються співвідношеннями [108]:

$\Delta y_t = y_t - y_{t-1}$ – перші послідовні різниці.

$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$ – другі послідовні різниці і так далі.

Якщо перші різниці нестаціонарного ряду y_t стаціонарні, то ряд y_t називається інтегрованим першого порядку. Стаціонарний часовий ряд називається інтегрованим нульового порядку.

Якщо перші різниці настаціонарного ряду нестаціонарні, а другі різниці стаціонарні, то ряд y_t називається інтегрованим другого порядку. Якщо перший стаціонарний ряд отримують після k-кратного взяття різниць, то ряд y_t називається інтегрованим k −го порядку.

1.4 Аналіз наявності нелінійностей процесу

Однією з проблем при визначенні структури моделі є встановлення факту наявності нелінійностей в досліджуваному процесі та їх типу [126]. Для розв’язку цієї проблеми обов’язково використовують візуальний аналіз даних та формальні тести на наявність нелінійностей. Досвідченому фахівцю з моделювання візуальний аналіз дозволяє оперативно виявити наявність участків з лінійним або нелінійним трендом, в якійсь мірі наявність

гетероскедастичності та значних викидів (імпульсів), які можуть суттєво впливати на якість моделі. Необхідно зазначити, що існують навіть окремі навчальні курси з візуального аналізу даних. Це свідчить про те, що не варто нехтувати такою доступною, але ефективною можливістю дослідження даних.

Існує також ряд формальних тестів на наявність нелінійності. Розглянемо простий тест для визначення наявності нелінійності. Цей тест можна застосувати у випадку, коли можна набрати кілька груп (вибірок) спостережень для одного і того ж процесу [69, 17, 108, 111]:

$$\widehat{F} = \frac{\frac{1}{m-2}\sum_{i=1}^{m} n_i (\bar{y}_i - \widehat{y}_i)^2}{\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2},$$

де $\bar{y}_i$ – середнє значення для i –ї групи (вибірки або групи) даних; $\widehat{y}_i$ – середнє для лінійної апроксимації даних; m – число груп даних; n_i – число вимірів в i –й групі; n – загальне число вимірів.

Фактично, наведена статистика представляє собою відношення [126]:

$$\widehat{F} = \frac{Відхилення\ середніх\ значень\ від\ прямої\ регресії}{Відхилення\ значень\ y(k)\ від\ групових\ середніх}.$$

Якщо статистика $\widehat{F}$ з $v_1 = m-2$ та $v_2 = n-m$ степенями свободи досягає або перевищує рівень значимості, то гіпотезу щодо лінійності необхідно відхилити. Недоліком даного підходу є те, що для його застосування необхідно мати декілька (не менше трьох) груп даних для одного і того ж процесу, які можна отримати в результаті виконання повторних експериментів. Очевидно, що це не завжди можливо.

Наявність нелінійності можна встановити також за допомогою вибіркових нелінійних кореляційних функцій (НКФ), тобто, кореляційних функцій, розрахованих за вибірками експериментальних (статистичних) даних [106]. Наприклад, якщо дискретна НКФ

$$r_{yx^2}(s) = r_{y(k)x^2(k-s)} = \frac{1}{N}\frac{\sum_{k=s+1}^{N}\left\{[y(k)-\bar{y}][x(k-s)-\bar{x}]^2\right\}}{\sigma_y\sigma_x^2}, \quad s = 0,1,2,3,\ldots,$$

містить значення, які суттєво відрізняються від нуля в статистичному смислі, то процес містить квадратичну нелінійность відносно регресора x.

Наявність детермінованого тренду в процесі можна визначити шляхом оцінювання рівняння (поліном порядку m відносно часу) [106]:

$$y(k) = a_0 + c_1 k + c_2 k^2 + \ldots + c_m k^m.$$

Якщо хоча б один із коефіцієнтів $c_i,\ i = 1,\ldots,m$ є статистично значимим, то гіпотеза щодо відсутності тренду відхиляється. Якщо тренд відносно швидко змінює свій напрямок розвитку і для нього трудно знайти адекватне функціональне описання, то застосовують моделі випадкових трендів, які ґрунтуються на комбінаціях випадкових величин.

Найбільш часто при побудові моделей використовуються наступні аналітичні залежності [35, 106, 118]:

1) лінійна

$$y = a + b_1 \cdot x_1 + \cdots + b_n \cdot x_n + \varepsilon$$

де y – відгук, x_1 та x_n регресори; a, b_1 … b_n невідомі параметри, ε – похибка.

2) степенева $y = a \cdot x_1^{b_1} \cdot \ldots \cdot x_n^n \cdot \varepsilon$

3) напівлогарифмічна $y = a + b_1 \cdot \ln(x_1) + \cdots + b_n \cdot \ln(x_n) + \varepsilon$

4) гіперболічна $y = a + b_1 \frac{1}{x_1} + \cdots + b_n \frac{1}{x_n} + \varepsilon$

5) експоненційна $y = \exp(a + b_1 \cdot x_1 + \cdots + b_n \cdot x_n + \varepsilon)$

Також можуть використовуватися комбінації розглянутих залежностей.

Нелінійні регресійні рівняння можна розділити на два класи [1, 118]:

- рівняння які за допомогою заміни змінних можна звести до лінійного вигляду;
- рівняння для яких це неможливо.

В першому випадку рівняння регресії перетворюють до лінійного вигляду за допомогою введення нових лінеарізуючих змінних.

В таблиці 1.3 наведені лінеаризуючі перетворення для деяких нелінійних моделей [1, 118].

Таблиця 1.3 – Лінеаризуючі перетворення

Залежність	Формула	Перетворення	Залежність між параметрами
Гіперболічна	$y = a + \frac{b}{x}$	$y' = y$ $x' = \frac{1}{x}$	$a = a'$ $b = b'$
Логарифмічна	$y = a + b \cdot \ln(x)$	$y' = y$ $x' = \ln(x)$	$a = a'$ $b = b'$
Степенна	$y = a \cdot x^b$	$y' = \ln(y)$ $x' = \ln(x)$	$\ln(a) = a'$ $b = b'$
Експоненційна	$y = e^{a+bx}$	$y' = \ln(y)$ $x' = \mathrm{x}$	$a = a'$ $b = b'$
Показникова	$y = a \cdot b^x$	$y' = \ln(y)$ $x' = \mathrm{x}$	$\ln(a) = a'$ $\ln(b) = b'$

В подальшому, за допомогою оберненого перетворення можна отримати параметри початкового рівняння.

1.5 Методи моделювання нелінійних нестаціонарних процесів

Аналіз залишків моделі регресії дозволяє ідентифікувати наступні проблеми при побудові моделі [35, 52, 69, 108, 109]:

1) невірно задана структура моделі;

2) дисперсія не є постійною;

3) наявність корельованості похибок моделі;

Окрім цього, також можна виявити ще додаткові проблеми:

4) викиди в даних (екстремальні значення);

5) наявність колінеарності між регресорами;

Типові поліноміальні регресійні моделі:

1) Квадратична поліноміальна модель [69, 37, 109, 108, 106, 118]:

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j$$

2) Кубічна поліноміальна модель[4, 29, 64]:

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3 + \varepsilon_j$$

3) Поліноміальна модель з комбінованою змінною [69, 37, 109]:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j} X_{2j} + \varepsilon_j$$

Тест на відсутність відповідності моделі даним, порівнює варіацію навколо моделі з "чистою" варіацією в межах спостережень, що повторюються. На практиці, якщо є n_i спостережень відгука, що повторюються, $Y_{i1}, \dots, Y_{in_i}$, для тих саме значень регресорів x_i, то можна спрогнозувати реальне значення відгука за значенням x_i, або із

використанням значення прогнозу Y_i на основі моделі або із використанням середнього значення $\bar{Y}_i$ по значенням, що повторюються.

Цей тест розкладає залишкову похибку в першу компоненту через зміну повторів навколо їх середнього значення (так звана "чиста" похибка) і другу компоненту, обумовлену зміною середніх значень навколо прогнозу за моделлю (похибка "зміщення") [69, 37, 109].

$$\sum_i \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{Y}_i\right)^2 = \sum_i \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2 + \sum_i n_i \left(\bar{Y}_i - \hat{Y}_i\right)^2$$

Якщо модель адекватна, то обидві компоненти оцінюють номінальний рівень похибки. Однак, якщо другий компонент зміщення помилки набагато більший, аніж чиста помилка, то це є доказом того, що існує значна відсутність підгонки моделі. Значна відсутність підгонки вказує на те, що вказана модель є недостатньою для моделювання даних, і в цьому випадку треба доробити модель.

Класифікація типових нелінійних функції зв’язку між цільовою та регресорною змінною можна представити у вигляді дев’ятнадцяти функціональних залежностей наступного виду:

1) $\mathrm{Y} = \mathrm{a} \cdot \mathrm{X}^{\mathrm{b}}$

2) $\mathrm{Y} = \mathrm{a} + \mathrm{X}^{\mathrm{b}}$

3) $\mathrm{Y} = \mathrm{a} + \mathrm{b} \cdot \mathrm{X}^{\mathrm{c}}$

4) $\mathrm{Y} = \mathrm{a} - \mathrm{b} \cdot \mathrm{c}^{\mathrm{X}}$

5) $\mathrm{Y} = \log(X - a)$

6) $\mathrm{Y} = \log(a + b \cdot X)$

7) $\mathrm{Y} = 1/(1 + a \cdot X)$

8) $\mathrm{Y} = X/(a + b \cdot X)$

9) $\mathrm{Y} = a/(1 + b \cdot X)$

10) $\mathrm{Y} = 1/(a + b \cdot X + c \cdot X^2)$

11) $Y = e^{(X-a)}$

12) $Y = 1 - e^{-X^a}$

13) $Y = 1 - e^{-a \cdot X^a}$

14) $Y = 1 - e^{-\exp(a-b \cdot X)}$

15) $Y = a \cdot e^{-\exp(b-c \cdot X)}$

16) $Y = a \cdot e^{\frac{b}{X+c}}$

17) $Y = \frac{a}{1+e^{b+c \cdot X}}$

18) $Y = b \cdot e^{X-a}$

19) $Y = a + e^{\frac{b}{X+c}}$

Методика побудови нелінійної моделі [1, 35, 108]:

1. Для кожного регресору $x_1, \dots, x_n$ відбувається підбір найкращого, за відповідним критерієм, нелінійного зв'язку з цільовою змінною y .В таблиці вище наведені відповідні можливі зв'язки.

$\forall\, i = 1, \dots, n$ відбувається пошук зв'язку вигліду $y = f_j(x_i)$ такого щоб $\left(y - f_j(x_i)\right) \to min.$

2. На основі виявлених індивідуальних функцій зав'язків $f_1(x_1), f_2(x_2), \dots., f_n(x_n)$ будується загальна повна модель виду

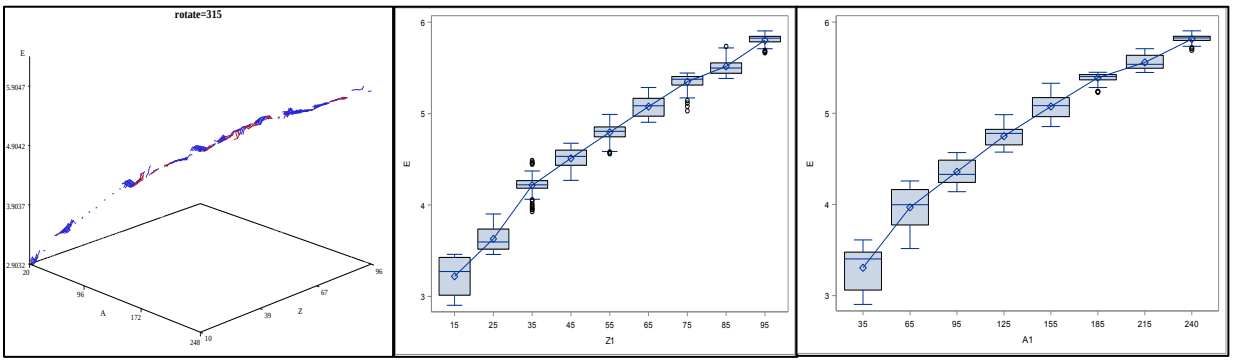
$$y = g(x_1, \dots, x_n) = g\big(f(x_1), \dots, f(x_n)\big)$$

В простому адитивному випадку функціональної залежності, формула може бути представлена у вигляді [108]:

$$y = f_1(x_1) + f_2(x_2) + \dots. + f_n(x_n)$$

Наведена формула відображає декомпозицію функції $y = g(x_1, \dots, x_n)$ як адитивне рівняння, складовими якого є індивідуальні нелінійні функції зв'язку між регресором за цільовою змінною.

В якості прикладу використання запропонованої методики розглядалася задача – моделювання залежності енергії зв’язку від зарядового числа ядра та атомного числа. Навчальний набір даних складався з 787 експериментальних спостережень. Дані включали значення: Z – заряд ядра; A – атомне число; E – енергія зв'язку ядра.



а) б) в)

Рисунок 1.2 – (а) Функціональна залежність *E=f(A,Z)*, (б) Графік box-plot залежності E від Z, (в) Графік box-plot залежності E від A,

За результатами відображеними на рис. 1.2, можна побачити залежність: при збільшенні A та Z збільшується і E. В рамках дослідження було створено п’ять різних груп наборів тренінгових та тестових даних.

На кожному тренінговому наборі даних обчислити значення оцінок коефіцієнтів моделі, а на відповідному тестовому наборі даних виконати тестування, після чого для значень прогнозів тренінгового та тестового наборів даних обчислити статистичні характеристики.

Описану вище методику було використано для моделювання спочатку індивідуальних нелінійних зав’язків між E та Z, між E та A, а потім загальних між E та Z, A.

Індивідуальні нелінійні зв’язки.

Найбільш перспективними моделями з урахуванням нелінійної залежності між $Y = E$ та $X = Z$ виявилися наступні:

$Y = a + b \cdot X^c$, для якої MAPE=1,74;

$Y = 1/(a + b \cdot X + c \cdot X^2)$, для якої MAPE=2,53.

Найбільш перспективними моделями з урахуванням нелінійної залежності між $Y = E$ та $X = A$ виявилися наступні:

$Y = a + b \cdot X^c$, для якої MAPE=2,89;

$Y = 1/(a + b \cdot X + c \cdot X^2)$, для якої MAPE=1,54.

$Y = a \cdot e^{-\exp(b - c \cdot X)}$, для якої MAPE=1,95.

Загальна нелінійна модель, з урахуванням індивідуальних нелінійних зав'язків.

В свою чергу, на основі виявлених індивідуальних функцій зав'язків було побудовано серію з шести нелінійних моделей, де $Y = E$, $X_1 = Z$, $X_2 = A$:

$Y = c_0 + c_1 \cdot X_1^{c_2} + c_3 \cdot X_2^{c_4}$, для якої MAPE=1,37;

$Y = c_0 + c_1 \cdot X_1^{c_2} + 1/(c_3 + c_4 \cdot X_2 + c_5 \cdot X_2^2)$, для якої MAPE=0,93;

$Y = 1/(c_0 + c_1 \cdot X_1 + c_2 \cdot X_1^2) + c_3 + c_4 \cdot X_2^{c_5}$, для якої MAPE=0,95;

$Y = c_0 + 1/(c_1 + c_2 \cdot X_1 + c_3 \cdot X_1^2) + 1/(c_4 + c_5 \cdot X_2 + c_5 \cdot X_2^2)$, для якої MAPE=1,51;

$Y = c_0 + 1/(c_1 + c_2 \cdot X_1 + c_3 \cdot X_1^2) + c_4 \cdot e^{-\exp(c_5 - c_6 \cdot X_2)}$, для якої а MAPE=1,88;

$Y = c_0 + c_1 \cdot X_1^{c_2} + c_3 \cdot e^{-\exp(c_4 - c_5 \cdot X_2)}$, для якої MAPE=0,91;

В той час як прості базові моделі у вигляді адитивних регресійних рівнянь

$Y = c_0 + c_1 \cdot X_1 + c_2 \cdot X_2$, для якої MAPE=2,15;

$Y = c_0 + c_1 \cdot X_1 + c_2 \cdot X_2 + c_3 \cdot X_1 \cdot X_2$, для якої MAPE=2,09;

Як можна побачити з отриманих результатів найменша похибка за критерієм мінімізації MAPE має модель у вигляді нелінійного рівняння:

$$Y = c_0 + c_1 \cdot X_1^{c_2} + c_3 \cdot e^{-\exp(c_4 - c_5 \cdot X_2)},$$

для якої MAPE=0,91 та в два рази нижча у порівнянні з регресійними моделями.

Модель регресії використовується у випадку коли відгук – інтервальна змінна і регресори – інтервального типу [69, 109, 68, 106].

Якщо ж відгук – номінальна змінна, а регресори – інтервального типу, то використовується модель логістичної регресії [108, 109, 68].

У випадку відгука дихотомічного типу (дві рівня), використовується бінарна логістична регресія [129]. Якщо ж відгук містить більше двох рівнів, то можливі наступні варіанти:

- Якщо відгук номінального типу, то використовують номінальну логістичну регресію;
- Якщо відгук ординарного типу, то це випадок ординарної логістичної регресії.

Логістичне перетворення

Модель логістичної регресії використовує logit-перетворення, що описується формулою:

$$logit(p_i) = ln\left(\frac{p_i}{(1-p_i)}\right)$$

де $i-$ номер спостереження; $\mathrm{p_i}-$ ймовірність того що в $i-$му спостережені відбудеться подія (наприклад, факт продажи товару, або неповернення кредиту); $ln-$натуральний логарифм (з базою $e = 2{,}7183$).

Модель логістичної регресії застосовує логіт-перетворення по відношенню до ймовірностей. Коли значення ймовірності p наближається до максимального значення 1, то значення $ln(p/(1-p))$ переходить до нескінченності, а коли p наближається до свого мінімального значення 0, то $p/(1-p)$ наближається до 0. Натуральний логарифм по аргументу, що наближається до 0, йде значення від'ємної нескінченності. Як наслідок логіт-перетворення не має верхньої або нижньої границь.

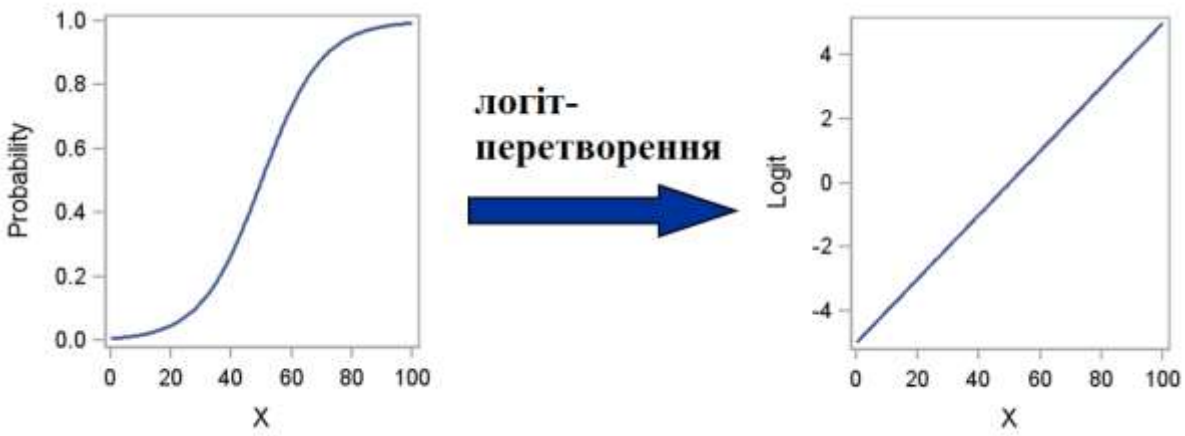


Рисунок 1.3 – Ілюстрація логіт-перетворення значень ймовірностей в логіт-бали

При використанні логістичної регресії робиться припущення про лінійну залежність логіт-балу від значення предиктора. Модель множинної логістичної регресії [1, 35, 129]:

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

де $logit(p_i) -$ логіт-бал для ймовірності події; $\beta_0 -$ вільний член регресійного рівняння; β_k –оцінка параметра k –го предиктора моделі.

На відміну від моделі лінійної регресії, логіт не є нормально розподіленою та дисперсія не є постійною. Таким чином, для оцінки параметрів логістична регресія вимагає більш складних методів обчислення оцінок параметрів, так званий Метод максимальної правдоподібності. Цей метод знаходить значення параметрів, які дозволяють зробити дані, що спостерігаються найбільш ймовірними. Це досягається шляхом максимізації функції правдоподібності, яка виражає ймовірність даних що спостерігаються, як функцію від невідомих параметрів.

На відміну від моделі лінійної регресії, логіт не є нормально розподіленою та дисперсія не є постійною [128]. Таким чином, для оцінки параметрів логістична регресія вимагає більш складних методів обчислення оцінок параметрів, так метод максимальної правдоподібності [129]. Цей

метод знаходить значення параметрів, які дозволяють зробити дані, що спостерігаються найбільш ймовірними. Це досягається шляхом максимізації функції правдоподібності, яка виражає ймовірність даних що спостерігаються, як функцію від невідомих параметрів.

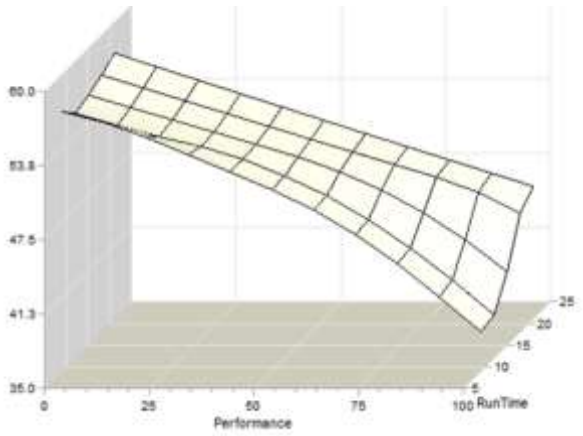


Рисунок 1.4 – Візуалізація нелінійної моделі з двома регресорами у тривимірному просторі, коли присутній нелінійний зв'язок між відгуком та регресорами

Для дослідження зв'язку між $k+1$ змінними (k регресорів плюс один відгук) використовують $k-$мірну поверхню для прогнозування. У загальному випадку в модель регресії може включати і більш складні зв'язки - квадратичну, кубічну, поліноміальну, перетин змінних [129].

Спочатку узагальнену авторегресійну модель гетероскедастичного процесу (УАРУГ). При використанні такої моделі умовна дисперсія процесу представляється за допомогою моделі авторегресії з ковзним середнім (АРКС). Нехай похибки моделі описуються рівнянням: де $\nu(k)-$ процес білого шуму з одиничною (для простоти) дисперсією $\sigma_\nu^2=1$; $h(k)-$умовна дисперсія процесу, яка дозволяє побудувати узагальнену умовно гетероскедастичну модель (УАРУГ) [52, 53, 108, 56]:

$$h(k)=a_0+\sum_{i=1}^{q}\alpha_i\varepsilon^2(k-i)+\sum_{i=1}^{p}\beta_i\,h(k-i) \quad . \qquad (1.3)$$

Оскільки процес $\{\nu(k)\}$ визначено в даному випадку як білий шум, то умовне і безумовне середнє процесу будуть дорівнювати нулю. Математичне сподівання для $\varepsilon(k)$ [129]: $E[\varepsilon(k)] = E\{\nu(k)\sqrt{h(k)}\} = 0$.

Важливим є те, що умовна дисперсія процесу $\{\varepsilon(k)\}$ є залежною від часу: $E_{k-1}[\varepsilon^2(k)] = h(k)$, оскільки $E_{k-1}[\nu^2(k)] = 1$. Для того щоб ця умовна дисперсія була скінченою, необхідно, щоб корені характеристичного рівняння, записаного для (1.3), знаходились всередині кола одиничного радіуса на комплексній площині. Таким чином, основною відмінною властивістю моделі УАРУГ є те, що умовна дисперсія збурень, які діють на процес $\{y(k)\}$, є процесом авторегресії з ковзним середнім.

Припустимо, що $\{y(k)\}$ – це процес авторегресії з ковзним середнім. При побудові моделі процесу можливі такі варіанти [108, 129]:

– якщо вдається побудувати адекватну модель АРКС, то похибки моделі будуть мати властивості білого шуму;

– якщо не вдається побудувати адекватну модель АРКС, то, використовуючи автокореляційну функцію для квадратів залишків, необхідно побудувати модель УАРУГ, яка дає можливість виконати аналіз поведінки дисперсії процесу; корелограма процесу $\{\varepsilon^2(k)\}$ надає можливість встановити наявність гетероскедастичності.

Оскільки $E_{k-1}[\varepsilon(k)] = \sqrt{h(k)}$, то рівняння (1.3) можна записати у формі:

$$E_{k-1}[\varepsilon^2(k)] = a_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i) \qquad (1.4)$$

За своєю структурою це рівняння схоже на рівняння АРКС(p,q) для послідовності $\{\varepsilon^2(k)\}$. Таким чином, методику побудови моделі гетероскедастичного процесу можна узагальнити наступним чином [108]:

Крок 1. За необхідності можна зробити попередню обробку (підготовку до моделювання) експериментальних даних (нормування, логарифмування,

аналіз екстремальних значень, заповнення пропусків даних) і застосувати до них тести на гетероскедастичність. Якщо досліджуваний процес містить тренд, то перед побудовою моделі дисперсії його необхідно видалити, наприклад, шляхом обчислення перших різниць або різниць вищих порядків. Досить часто візуальний аналіз даних надає можливість отримати суттєву інформацію щодо наявності у них гетероскедастичності. Разом із візуальним аналізом необхідно аналізувати параметри описової статистики, які полегшують оцінювання структури моделі.

Крок 2. Користуючись АКФ та ЧАКФ для експериментальних даних, побудувати модель АР(p) або АРКС(p,q) для процесу$\{ y(k) \}$ та обчислити ряд з квадратів залишків$\{ \hat{\varepsilon}^2(k) \}$, де $\hat{\varepsilon}^2(k) = e(k)$. Обчислити вибіркову дисперсію $\hat{\sigma}_{\hat{\varepsilon}}^2$ збурення $\hat{\varepsilon}(k)$: $\hat{\sigma}_{\hat{\varepsilon}}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\hat{\varepsilon}^2(k)$, де N число залишків після побудови моделі АР чи АРКС.

Крок 3. Обчислити та побудувати графік вибіркової автокореляційної функції для квадратів залишків [108]:

$$\rho(s) = \frac{\sum_{k=s+1}^{N}[\hat{\varepsilon}^2(k) - \hat{\sigma}_{\varepsilon}^2][\hat{\varepsilon}^2(k-s) - \hat{\sigma}_{\varepsilon}^2]}{\sum_{k=1}^{N}[\hat{\varepsilon}^2(k) - \hat{\sigma}_{\varepsilon}^2]}$$

Якщо встановлено, що існують такі значення $\rho(s)$, які відрізняються від нуля у статистичному смислі, то це свідчить про присутність процесу АРУГ або УАРУГ. Для того щоб переконатись у присутності гетероскедастичності, використовують Q статистику Люнга-Бокса, яка обчислюється за виразом [52, 108]: $Q = N(N+2)\sum_{i=1}^{n}\rho(i)/(N-i)$, де $n = N/4$ (емпіричне значення). Якщо значення $\hat{\varepsilon}^2(k)$ некорельовані, то Q- статистика повинна мати розподіл χ^2 з n ступенями свободи.

Крок 4. Побудувати модель УАРУГ (або іншу модифікацію)

$$h(k) = a_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i), \qquad (1.5)$$

використовуючи ряд значень $\hat{\varepsilon}^2(k)$ та значення умовної дисперсії $h(k)$. Якщо у цій моделі хоча б один із коефіцієнтів $\alpha_i, i \in 1,..q$ є статистично значущим, то процес є гетероскедастичним. Оскільки модель (1.4) описує залишки моделі з деяким наближенням, то у загальному випадку доцільно продовжити процес уточнення моделей, що описують вихідний процес у цілому. Тобто можна уточнити як початкову модель АР(p) чи АРКС(p,q), а також модель типу (1.5). Це виконується на наступному кроці виконання дослідження.

Крок 5. Далі можна скористатись моделлю (1.5) для того щоб отримати дійсні значення залишків, які описуються цією моделлю, тобто згенерувати ряд $\{\hat{\varepsilon}^2(k)\}$ і згенерувати ще один ряд $\{y_1(k)\}$, де $y_1(k) = y(k) - \hat{\varepsilon}_1(k)$. За допомогою отриманого ряду можна побудувати уточнену модель процесу типу АР(p) чи АРКС(p,q). За необхідності процес уточнення та ускладнення моделей може бути продовжений.

1.6 Типи невизначеностей, що характерні для різних підсистем складних систем

Невизначеність припускає наявність факторів, при яких результати дій не є детермінованими, а ступінь можливого впливу цих факторів на результати невідома [44, 130-143].

Аналіз умов наявності невизначеностей може виконуватись на абстрактному теоретичному рівні або з точки зору конкретної ситуації, наприклад, з точки зору виявлення можливості побудови математичної моделі чи з точки зору теорії інформації, де невизначеність виступає як характеристика ситуації вибору. Категорія невизначеності характеризується деякими змінними параметрами, які описують різні види невизначеностей:

глобальну невизначеність, ситуативну, політичну, соціальну і т.д. При вирішенні задач прийняття рішень в умовах наявності невизначеностей необхідно встановити рівень аналізу і типи невизначеностей, що розглядаються.

Необхідно зазначити, що часто невизначеність ототожнюють лише з відсутністю повної інформації про той чи інший об'єкт. Насправді незнання станів об'єкту не є єдиною невизначеністю. Поряд з цим іноді можна розглядати невизначеність цілей та невизначеність критеріїв вибору рішень. У багатьох реальних задачах складність прийняття рішення визначається насамперед числом альтернативних варіантів та числом і різнорідністю критеріїв оцінювання цих варіантів.

При вирішенні задач системного аналізу, прийняття рішень та дослідження операцій виділяють наступні основні типи невизначеностей [130-132]:

– невизначеність цілей – невизначеність вибору цілей в багатокритеріальних задачах;

– ситуаційна невизначеність – невизначеність впливу неконтрольованих факторів, що позначаються на процесах практичної діяльності;

– невизначеність природи – відсутність достатніх знань щодо оточення та зовнішні фактори;

– ненадійність очікувань – невизначеність розвитку певних подій у майбутньому;

– стратегічна невизначеність – невизначеність цілей і дій активного або пасивного партнера чи противника, так звана невизначеність конфліктів;

– інформаційна невизначеність – нечіткість та розпливчастість процесів і явищ та інформації про досліджувану систему, відсутність відомостей про достовірність інформації.

Додатково виділяють наступні види невизначеностей [133, 141]:

– структурна невизначеність – невизначеність структури моделі досліджуваної системи;

- параметрична невизначеність – апріорна невизначеність параметрів моделі системи, складність оцінювання і аналізу якості параметрів моделі;
- статистична невизначеність – невизначеність статистичних даних, що переважно випливає з об’єму даних, наявності пропусків, імпульсних викидів тощо;
- методична невизначеність – невизначеність методу обробки даних чи методу розв’язання задачі;
- Комбінаторна невизначеність – неможливість знання всіх можливих варіантів, вона пов’язана зі всіма іншими типами невизначеностей і найчастіше випливає з них.

Необхідно зазначити, що в реальних практичних задачах прийняття рішень і системного аналізу часто наявні різноманітні види невизначеностей, які разом складають деякий комплекс невизначеностей, так звану системну невизначеність [131].

Отже, в таких задачах дослідження розвитку соціальних, економічних, екологічних, суспільних процесів стикаються з наступними невизначеностями [132, 136] :

– інформаційна невизначеність, що полягає у відсутності, неповноті і неточності інформації про певні характеристики проектів, про ринки та сфери діяльності, на які розраховані інвестиційні проекти, тощо;

– невизначеність цілей, що зумовлена неясністю визначення раціональності способу розподілу ресурсів, вибору цілей при розв’язанні задачі та невизначеністю критеріїв вибору;

– ситуаційна невизначеність: невизначеність природи, що для задачі розподілу інвестиційних ресурсів виявляється у неповноті і неточності інформації про фінансовий стан інших учасників ринку, про параметри техніки і технології, динаміку цін, коливання ринкової кон’юнктури, валютних курсів, умов надання кредитів, тощо; ненадійність очікувань, що в сучасних умовах України зумовлена нестабільністю економічного

законодавства і поточної політичної ситуації, умов інвестування і використання прибутків й, наприклад, ризиком запровадження обмежень на торгові операції, тощо;

– стратегічна невизначеність, яка виявляється у неповноті і неточності інформації про цілі і наміри партнерів та конкурентів, про їх фінансовий стан, можливість утворення заборгованості, банкрутств, зривів договірних зобов'язань, можливість виводу на ринок іншими учасниками більш конкурентноспроможних товарів чи послуг, тощо.

Вочевидь, всі процеси, що відбуваються в соціально-економічних, фінансовій сфері, екології, управлінні протікають в умовах системної невизначеності, яка сполучає в собі всі розглянуті типи невизначеностей.

Таким чином, ефективність пошуку оптимальних рішень в практичних задачах залежить від методів аналізу, розкриття і врахування наявних в задачі невизначеностей, а також від того, наскільки адекватно ці методи зможуть відобразити реальну ситуацію.

Історично першими з'явилися імовірнісно-статистичні методи розкриття невизначеностей. Однак, незважаючи на розвиток імовірнісних методів, вони не можуть бути універсальним засобом в задачах системного аналізу і прийняття рішень для урахування усіх типів невизначеностей. Це перш за все відноситься до слабоструктурованих проблем [131] та задач з нечіткою вхідною інформацією [78, 183].

На даний час в теорії прийняття рішень (ТПР) накопичені наступні методи урахування невизначеностей при вирішенні задач [131, 141, 142]:

- метод коректив;
- аналіз чутливості;
- імовірнісний аналіз;
- дерево рішення;
- методи опису невизначених і неточних понять.

Узагальнення даних методів урахування і розкриття невизначеностей та аналіз їх придатності до застосування.

Необхідно зазначити, що розкриття невизначеностей постає задачею системно погодженого розкриття різнорідних невизначеностей на основі єдиних принципів, прийомів та критеріїв [131, 141].

Як зазначено вище, більшість з проаналізованих методів розкриття та урахування невизначеностей в задачах прийняття рішень мають суттєві специфічні недоліки, внаслідок чого оцінка їх результатів обмежена. Необхідно зазначити, що на практиці в деяких задачах намагаються користуватися комбінованим застосуванням декількох методів, що надає можливість отримувати додаткові висновки щодо допустимих рішень.

Для урахування інформаційної, ситуаційної невизначеності, ненадійності очікувань та стратегічної, комбінаторної невизначеності в процесах, що протікають у досліджуваних системах, де вплив людського фактору досить суттєвий, необхідною розробка сучасних інформаційних технологій, які б автоматизовували б процес формування варіантів рішень з урахуванням припущень щодо очікуваних значень. Вибір на користь впровадження інтелектуального та ймовірнісно-статистичного аналізу даних у інформаційних системах підтримки прийняття рішень ґрунтується на тому, що такий підхід до урахування невизначеностей позбавлений недоліків властивих іншим методам та постає придатним для опису та урахування системної невизначеності.

В загальному випадку модель багатокритеріальної задачі прийняття рішень в умовах невизначеностей може бути представлена у вигляді [141]:

$$\left(M, Z, A, C, Cs, S, R\right), \qquad (1.6)$$

де M – постановка задачі; Z – неконтрольовані невизначені фактори; A – множина можливих рішень; C – векторний критерій оцінювання рішень; Cs – множина шкал критеріїв; S – множина інтегральних оцінок рішень; R – правило розв’язання задачі (прийняття рішень).

Постановка задачі M у моделі (1.6) характеризує цілі ОПР та умови їх досягнення. Фактори Z зумовлюють наявність умов невизначеностей в задачі прийняття рішень, вони можуть бути представлені випадковими величинами. Множина допустимих рішень A становить сукупність рішень, що задовольняють наявним обмеженням з M, які розглядаються як можливі способи досягнення визначених цілей. Кожний варіант рішення оцінюється за сукупністю критеріїв, що представлені векторним критерієм $C=\{C_l\}$, $l=\overline{1,h}$. Для кожного критерію C_l задана шкала Cs_l, яка є множиною упорядкованих оцінок. Шкали Cs_l, $l=\overline{1,h}$ утворюють множину шкал критеріїв $Cs=\{Cs_l\}$. Кожному рішенню A_i з множини $A=\{A_i\}$, $i=\overline{1,n}$ надаються оцінки S_{il} за критеріями C_l, $l=\overline{1,h}$. Всі оцінки S_{il}, $i=\overline{1,n}$, $l=\overline{1,h}$ утворюють множину інтегральних оцінок рішень $S\subset Y$, де $Y=Cs_1\times Cs_2\times\ldots\times Cs_l$. R – правило розв'язання чи метод прийняття рішення, що можуть бути задані у вигляді аналітичного виразу, алгоритму чи вербального формулювання.

Вибір методу прийняття рішення R має вагоме значення в задачі прийняття рішень та суттєво впливає на якість остаточного результату. Як вже було зазначено, через концептуальні та методичні труднощі на даний час не існує єдиного методологічного підходу до вирішення задач прийняття рішень в умовах невизначеностей з класу задач розподілу ресурсів і вибору варіантів [130]. Однак, накопичена достатньо велика кількість методів формалізації постановки та розв'язання таких задач. При користуванні такими методами слід пам'ятати, що всі вони мають рекомендаційний характер, а вибір остаточного рішення завжди залишається за ОПР.

Розв'язок задачі прийняття рішення в умовах невизначеностей формулюється наступним чином [130, 139, 141, 142]:

- визначення типів невизначеностей, характерних для задачі;
- застосування обраного методу врахування невизначеностей;
- постановка детермінованої багатокритеріальної задачі ПР;
- розв'язання багатокритеріальної задачі прийняття рішень.

Формальна модель багатокритеріальної задачі прийняття рішення є прийнятною для задачі в постановці задачі дослідження та прогнозування розвитку нелінійних нестаціонарних процесів різної природи. Для постановки у відповідність розробленим вимогам пропонується розширити модель (1.6), і ввести до її складу групу ОПР: $D = \{D_t\}$, $t = \overline{1,k}$.

Зазначимо, що крім досліджених підходів до врахування і розкриття певних невизначеностей при розв'язанні задач прийняття рішень, доцільним є дослідження розроблених методів багатокритеріального прийняття рішень в умовах невизначеностей. Це надає можливість враховувати наявні невизначеності та зводити задачі до детермінованих постановок. Такі методи або їх елементи можуть бути використані в методології для усунення проблем, пов'язаних з факторами впливу системної невизначеності.

Детальне дослідження підходів до розв'язання багатокритеріальних задач розподілу ресурсів та методів прийняття рішень означеного класу, аналіз їх ефективності і можливостей застосування при розробці системної методології розподілу ресурсів. Дослідження охоплює основні сучасні напрямки з прийняття багатокритеріальних рішень в умовах невизначеностей: методи узагальненого критерію, методи послідовної оптимізації, методи теорії ігор, передові підходи і методи штучного інтелекту. За результатами аналізу до задачі розподілу ресурсів і вибору варіантів рекомендується застосовувати підхід, який ґрунтується на зведенні багатокритеріальної задачі до однокритеріальної. Цілком прийнятним також виявляється використання еволюційних алгоритмів пошуку кращого рішення.

1.7 Моделі та методи заповнення пропусків у структурованих даних

Складність побудови прогнозів нелінійних нестаціонарних процесів значною мірою зумовлена тим, що їх функціонування відбувається в умовах

невизначеностей різних типів, що практично унеможливлює отримання для побудови моделей повної та достовірної інформації, про їх стан та перебіг. Перш за все це пов’язано з технологічними аспектами збору та обробки інформації. А також з недосконалістю значної кількості методик, особливо таких, що базуються на опитування, зборі інформації у Інтернет.

Вирішенню проблеми забезпечення повноти вхідної інформації на етапі збору та попередньої обробки присвячено багато робіт вітчизняних та закордонних фахівців [143, 145, 146, 170], однак, вона залишається невирішеною. Відсутня також й ефективна інформаційна технологія аналізу та обробки такої інформації.

Більшість пропонованих варіантів класифікації невизначеностей [131,132, 136, 145] визначають пропуски як невизначеність даних, що, слід зазначити, є досить звуженим підходом. Даний тип невизначеності потребує окремої уваги, та має бути класифікований у окрему групу: невизначеностей, пов’язаних із життєвим циклом об’єкта дослідження, внутрішніми та зовнішніми факторами, що впливають на його розвиток, а також методами, що застосовуються під час збору та накопичення даних.

Тобто питання полягає не в усуненні невизначеності даних шляхом заповнення пропусків, але і в тому, щоб відтворити в повній мірі характер нелінійності. Адже й заповнення пропусків у даних може призвести до спотворення уяви про розвиток досліджуваного процесу, оскільки заповнення пропусків виконується з певною мірою припущення про характер досліджуваного процесу. Це в свою чергу впливатиме на якість прогнозу, отримуваного на основі моделей, вхідними даними для яких є часові ряди із заповненими пропущеними значеннями.

Дослідження різних методів заповнення пропусків показало, що існує багато різних підходів заповнення пропусків, в основу яких покладені різні методи [144, 145, 147].

В загальному вигляді задача заповнення пропусків в даних формулюється наступним чином. Нехай є навчальна вибірка даних

X={X_1,X_2,...,X_n}, Y - цільова змінна. Вхідні дані представлені у вигляді таблиці (табл.1.4), яка містить пропуски.

Таблиця 1.4 – Схематичний приклад таблиці даних, що містить цільову змінну та регресори

Цільова змінна	Регресори			
Y	X_1	X_2	…	X_n
Y_1	X_{11}	X_{21}	…	X_{2n}
…	…	…	…	…
Y_m	X_{1m}	X_{2m}	…	X_{nm}

При розв’язку задачі заповнення пропусків мінімізується функція:

$$f = \text{argmin}_*|Y - F(X)|,$$

де $F = F(X_1, X_2, \ldots, X_n)$ – функція, що визначає взаємозвязок між вихідною змінною Y та вектором X вхідних змінних.

Задача заповнення пропусків формулюється наступним чином:

$$Y_i = F_i(X_{i1}, X_{i2}, \ldots, X_{im}),\ i = \overline{1,n}$$

Вибір методу залежить, наприклад, від типу змінної - інтервальна або категоріальна; характеру досліджуваного процесу (наявність тренду і сезонності, тощо).

Видалення спостережень із пропусками ще називають аналізом повних спостережень [68, 69]. На практиці використання даного методу може призвести до того, що обсяг навчальної вибірки істотно зменшиться, і це спричинить необхідність додаткового опрацювання вхідних даних.

Наприклад, у задачах прогнозного моделювання цільової змінної з апріорно заданим співвідношенням результатів, видалення записів в навчальному наборі даних майже завжди призводить до зміни пропорції цільових результатів. Тому необхідно після виконання даної операції повторно формувати навчальну вибірку з заданими пропорціями результатів.

Іншим підходом заповнення пропусків є їх заміна константними значеннями [55, 145]. За своєю суттю, константні значення – це середнє значення по вибірці. Такий підхід використовується для інтервальних змінних. Для обчислень використовуватися як стандартна формула [68, 145, 147] (1.7)

$$\bar{x} = \frac{1}{m} \cdot \sum_{i=1}^{m} x_i, \tag{1.7}$$

де $\bar{x}$ – середнє значення для інтервального змінної X, яка приймає m значень $X = \{x_1, \dots, x_m\}$.

Варто зазначити, що спочатку виключаються пропуски з ряду даних, а тільки потім, за рештою значень обчислюється математичне сподівання.

Для випадку дискретного розподілу з відомим розподілом $P(X = x_i) = p_i$ для якого виконується умова $\sum_{i=1}^{\infty} p_i = 1$ використовується формула (1.8) [145]:

$$M[X] = \sum_{i=1}^{\infty} x_i \cdot p_i \tag{1.8}$$

Вказаний підхід (1.8) еквівалентний заміні пропусків на значення у вибірці, які найбільш часто зустрічається, розраховане на основі наявних значень змінної.

Мода, як середня величина, вживається частіше для категоріальних даних, що мають нечислову природу. Наприклад, колір об'єкта або

приналежність особи до політичної партії. Для інтервальних змінних замість моди часто вважають за краще використовувати медіанне значення [101, 145, 147].

Медіана є 50-й процентилем, 0,5-квантилем або другим квартилем вибірки або розподілу. Вона використовується для інтервальних змінних.

Якщо всі елементи вибірки різні, то медіана – це таке число вибірки, що рівно половина з елементів вибірки більше нього, а інша половина менше нього. У більш загальному випадку медіану можна знайти, упорядкувавши елементи вибірки за зростанням або спаданням і взявши середній елемент.

Обчислюється як середнє мінімального і максимального значень (1.9):

$$MIDRANGE = (MIN + MAX) / 2, \quad (1.9)$$

де MIN – мінімальне значення вибірки, MAX – максимальне значення вибірки.

Інший підхід в англомовній літературі позначається, як метод mid-minimum spacing. Для обчислення математичного очікування (середнього) використовується певний відсоток вибірки. В більшості статистичних пакетах це 90% вибірки. Інші 10% вибірки виключаються з розрахунків (з лівого і правого боку розподілу). За своєю суттю це вибіркова статистика за усіченою вибірці, очищена від викидів.

Метод заповнення пропусків із використанням вінзорірованого середнього схожий на методи середнього арифметичного і усіченого середнього. Розрахунок вінзорірованого середнього зводиться до того, що k% найбільших і k% найменших значень (зазвичай від 5% до 25%) замінюються найменшими і найбільшими значеннями з залишків масиву даних, після чого обчислюється середнє арифметичне [145, 147].

Часто використовуваним є метод заповнення пропусків із використанням середнього, розрахованого на основі метрики Мінковського.

В загальному випадку, відстань Мінковського між двома точками обчислюється за формулою (1.10) [188]:

$$\rho(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \tag{1.10}$$

де x та y точки у n-мірному просторі, $x = \{x_1, \ldots, x_n\}$, y=$\{y_1, \ldots, y_2\}$, p – числова константа, що приймає значення великі або рівні 1.

Для випадку однієї змінної формула (1.10) вироджується в формулу (1.11):

$$\rho(x) = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \tag{1.11}$$

де x точка у n-мірному просторі $x = \{x_1, \ldots, x_n\},\ p \geq 1.$

Формула (1.11) в подальшому використовується для обчислення середнього значення по вибірці.

Цей параметр задає ступінь відмінності порядку p, при обчисленні відстані Мінковського. В окремому випадку, при p = 1 значення еквівалентно медіанному, а для p = 2 зводиться до середньоквадратичного значенням (математичного сподівання).

Слід зазначити, що недоліком використання для заповнення пропусків методів усіченого середнього, вінзорізірованого виразу і інших подібних статистик, заснованих на заміні викидів або усікання вибірки, є їх мала ефективність при заповненні пропусків у коротких вибірках. У такому випадку заміна одних значень іншими не завжди обґрунтована.

М-оцінки [167] – широкий клас методів оцінки використовуваних в статистиці, заснований на мінімізації функціонала суми значень. Тьюки, Хубер та інші [8,166] запропонували більш стійкі статистики – математичне

сподівання, обчислене на усіченій вибірці, вінзорізоване середнє й інші схожі методи [8, 166].

Ідея застосування М-оцінок для заповнення пропусків в даних полягає в тому, що обчислюють більш робастні (стійкі) оцінки вибіркових статистик для аналізованої змінної, в першу чергу математичного очікування і дисперсії, таким чином, що б зменшити або повністю відсікти вплив викидів і аномальних значень. Реалізації алгоритмів засновані на різних підходах передбачають повне виключення небажаних значень з вибірки, заміні викидів на інші значення, які більше підходять під характеристики досліджуваної змінної, пошук оцінки вибіркової статистики на основі мінімізації функціоналу різниці між оцінкою і значеннями вибірки.

М-оцінка методом Хубера – підхід для побудови робастних регресійних моделей, менш чутливих до аномалій і викидів в даних. Одним з етапів запропонованої методики є попередня обробка даних і заповнення пропусків. Тобто, ідея застосування методу полягає у заповненні пропусків спеціальним чином обчисленим значенням математичного очікування по попередньо обробленій вибірці. Існує значна кількість модифікацій даної методики. Всі вони ґрунтуються на різних підходах, призначених для заміни або видалення викидів, пов'язаних з попередньо обробленими даними.

У загальному випадку значення математичного очікування є найкращою оцінкою параметра θ при мінімізації функціоналу суми квадратів $\sum(x_i - \theta)^2$, де x_i – це i-те значення з множини випадкових величин $X = \{x_1, \dots, x_m\}$, які в свою чергу є незалежні і однаково розподілені, тобто кожна з них має ту саму функцію розподілу, що і всі інші, і до того ж всі x_i незалежні в сукупності; θ – це значення оцінки, обчислене для кожної множини даних у просторі параметрів $\{\theta\}$ [166, 167].

У разі необхідності мінімізації функціоналу $\sum|x_i - \theta|$ найкращою оцінкою параметру θ є медіанне значення. В основі даного методу лежить та ж ідея, що й у методі найменших квадратів: мінімізація функції суми квадратів залишків $\sum(x_i - \theta)^2 \to min$.

У своїй роботі [166] Хубер запропонував загальну оцінку максимальної правдоподібності, яка мінімізує суму залишків $\sum_{i=1}^{n} \rho(x_i, \theta)$, де ρ – це функція втрат, яка може бути представлена в різній формі. В оригінальній роботі [166] Хубера запропонована така функція втрат (1.12) [166]:

$$\begin{gathered} \rho(\mathrm{x}, \theta) = \frac{(\mathrm{x}-\theta)^2}{2}, \text{ якщо } |\mathrm{x} - \theta| < \mathrm{k} \\ \rho(\mathrm{x}, \theta) = \mathrm{k} \cdot |x - \theta| - \frac{1}{2} \cdot k^2, \text{ якщо } |\mathrm{x} - \theta| \geq \mathrm{k} \end{gathered} \tag{1.12}$$

де ρ – це функція втрат, яка може бути представлена в різній формі, k – параметр настройки методу, який за своєю суттю вказує на те, який розмір оригінальної вибірки використовується для розрахунків стійкої оцінки.

В [166] рекомендують приймати значення k = 1,345, що дозволяє охопити 95% інтервалу вибірки. У спеціалізованих статистичних системах, як у випадку з системою SAS Enterprise Miner, за замовчуванням, використовується значення k = 1,5.

Для обчислення оцінки θ використовуються стандартні алгоритми оптимізації, наприклад, Ньютона-Рафсона (Newton-Raphson) або ітеративний метод найменших квадратів з повторним зважуванням ваг [167]. Як було згадано вище, для оцінки математичного очікування часто використовується функція $\rho(x, \theta) = \frac{(x-\theta)^2}{2}$, в той час як для медіанного значення $\rho(x, \theta) = -|x - \theta|$.

Алгоритм обчислення зваженої М-оцінки методом Тьюки, як і в випадку з методом Хубера, носить ітераційний характер і складається з наступних кроків [167]:

1. Обчислюється оцінка середнього значення вибірки;

2. Визначаються відстані від обчисленого середнього до кожного елемента вибірки. Відповідно до значення довжини відстані, елементам вибірки присвоюються різні ваги, з урахуванням яких перераховується середнє значення. Характер вагової функції такий, що спостереження,

віддалені від середнього досить далеко, не спричиняють значного впливу на значення зваженого середнього.

В якості опції втрат використовується наступна формула [167] (1.13):

$$\rho(x,\theta) = \frac{c^2}{6}\cdot\left(1-\left(1-\left(\frac{x-\theta}{c}\right)^2\right)^3\right), \text{якщо } |x-\theta| \le c$$
$$\rho(x,\theta) = \left(\frac{c^2}{6}\right), \text{якщо } |x-\theta| > c \quad (1.13)$$

де x – це випадкова величина з множини випадкових величин $x \in X$, $X = \{x_1, \dots, x_m\}$, які в свою чергу є незалежні і однаково розподілені, тобто кожна з них має ту саму функцію розподілу, що і всі інші, і до того ж всі x_i незалежні в сукупності; θ – це значення оцінки, обчислене для кожної множини даних у просторі параметрів $\{\theta\}$.

Рекомендованим [167] є значення c = 4,685, що дозволяє охопити 95% інтервал вибірки.

Хвиля Ендрюса (Andrews 'Wave) – це ще одна з різновидів стійких M-оцінок. В якості опції втрат використовується формула [168] (1.14):

$$\rho(x,\theta) = \frac{a^2}{\pi^2}\cdot\left(1-cos\left(\frac{\pi\cdot(x-\theta)}{a}\right)\right), \text{якщо } |x-\theta| \le a$$
$$\rho(x,\theta) = \frac{2\cdot a^2}{\pi^2}, \text{якщо } |x-\theta| > a \quad (1.14)$$

Рекомендованим [168] є значення a = 4,21 (фактично – число π помножене на 1,34), яке дозволяє охопити 95% інтервал вибірки.

За використання методу Hot-Deck заміна пропусків здійснюється на основі підстановки замість пропусків записів із схожими характеристиками або однаковими значеннями змінних. Перевагою даного методу є те, що його використання не потребує аналітичного апарату.

При роботі з часовими рядами даних також можуть використовуватися різні методи заміни пропущених значень, зокрема такі, що враховують

специфіку формування ряду, а саме змістовну складову прив'язки даних до часової шкали, що дозволяє розширити спектр застосовуваних методів.

Найпопулярнішими методами заповнення пропусків для даного класу даних є заповнення [68]:

- середнім значенням, розрахованим на основі відомих даних;
- мінімальним значенням, серед усіх відомих;
- максимальним значенням, серед усіх відомих;
- медіанним значенням, розрахованим на основі відомих даних;
- першим відомим значенням замість всіх пропусків;
- останнім відомим значенням замість всіх пропусків;
- попереднім значенням, що йде відразу перед пропуском;
- наступним значенням, що йде відразу після пропуску;
- константним значенням, обраним користувачем.

На увагу заслуговує і підхід, оснований на використанні індикаторних змінних передбачає заміну пропущених значень за допомогою будь-якого методу обробки пропусків, але при цьому створюється нова спеціальна змінна-індикатор, яка додається в список змінних-предикторів для подальшого аналізу. Така змінна-індикатор набуває нульових значень для записів, де дані з самого початку не містили пропусків і поодинокі значення там, де пропуски були.

До переваг даного методу належать [68]:

- можливість використання всього набору даних (репрезентативність вибірки не страждає),
- явне використання інформації про пропущені значення.

При роботі з геоданими, як правило, методи заповнення пропусків засновані на застосування підходу, що передбачає заповнення значеннями наявних даних, що розташовані якнайближче до пропуску, або тими, які пов'язані (асоційовані) з даним місцем розташування об’єкта. Як правило, це можуть бути розумно обґрунтовані значення з використанням найбільш значимих характеристик об’єкта, що має певну геолокацію.

Багаторазова заміна пропусків також є ефективним методом заповнення пропусків даних. Вона виконується у наступні три етапи:

Етап 1. Пропущені значення заповнюються значеннями *m* разів (*m* - параметр-лічильник багаторазового заповнення *)*, при цьому відповідно створюється *m* наборів даних.

Етап 2. *m* повністю заповнених наборів даних аналізуються за допомогою стандартних процедур.

Етап 3. Результати, отримані, на основі аналізу заповнених *m* наборів даних, об'єднуються в загальний результат. Наприклад, отримані значення оцінок параметрів (наприклад, коефіцієнти і стандартні помилки регресорів), отримані для кожного з *m* наборів даних (з етапу 1), об'єднуються в загальний результат.

Багаторазова заміна пропусків за своєю суттю є ітераційною формою стохастичного заповнення пропусків. Однак, замість того що б виконувати заповнення одного пропуску один раз, виконується багаторазове заповнення пропуску різними значеннями, які розташовані навколо коректного значення. Потім всі змодельовані значення разом з повними спостереженнями використовуються в звичайній регресійній моделі, після чого отримані результати об'єднуються. Кожне заповнене значення включає випадкову компоненту, значення якої відображає ступінь того, як зміни інших змінних моделі не можуть передбачити її істинні значення. Таким чином досягається ефект вбудовування рівня невизначеності навколо «правдивості» значень, що підставляються.

Досить поширеною помилкою при використанні різних методів заповнення пропусків, є припущення, що підставлені значення повинні відображати «реальні» значення. Метою заміни пропусків, є коректне відтворення матриці дисперсії (коваріації), яка спостерігалася б, якби в даних не було пропусків.

При виборі підходу для заповнення пропусків, потрібно визначити чи відповідає обраний підхід аналітичній моделі. Така узгодженість означає, що

модель заповнення пропусків включає (принаймні) ті ж змінні, які є в аналітичній або оціночній моделі. Також включаються будь-які перетворення по відношенню до змінних, які будуть потрібні для оцінки обраної гіпотези, що представляє інтерес. Досить часто використовується логарифмічне перетворення, створення комбінованих змінних, перетворення змінної інтервального типу в номінальну.

Зазвичай прагнуть відтворити коректну матрицю дисперсії (коваріації), що відображає ступінь залежності між змінними аналізу, тому, відповідно, оцінки пропусків повинні бути заповнені одночасно.

Це може призвести до недооцінювання зв'язку між змінними, які являють інтерес, і втраті потужності вибірки, використовуваної для знаходження зв’язків в даних, зокрема, таких як нелінійність і взаємні статистичні ефекти [145, 147, 158-171].

Наприклад, кожна група відображає підмножину спостережень в даних, які мають один і той же шаблон відсутності інформації. На першому етапі дослідження формується шаблон наявних пропусків. Приклад шаблону поданий на рисунку 1.5.

Шаблон пропусків даних

Група	Y	X_1	X_2	X_3	X_4	Частота	Процент
1	X	X	X	X	X	83	82.18
2	X	X	.	X	X	7	6.93
3	X	.	X	X	X	6	5.94
4	.	X	X	X	X	2	1.98
5	.	.	X	X	X	1	0.99
6	.	.	.	X	X	2	1.98

Рисунок 1.5 – Приклад шаблону наявності пропусків в даних

На другому етапі сформований шаблон «накладається» на дані із пропусками

Шаблон пропусків даних

Група	Математичне сподівання за групою				
	Y	X_1	X_2	X_3	X_4
1	700.0686	5115.0987	264.1963	97.6373	61.8493
2	776.3428	4645.5285	.	61.0571	60
3	535.0166	.	164.45	48.2666	36.3166
4	.	2971.15	82.55	1	19
5	.	.	7	14	11
6	.	.	.	45.4	28.85

Рисунок 1.6 – Приклад шаблону наявності пропусків в даних

Наприклад, група 1 представляє 73 спостереження набору даних, які містять заповнену інформацію для всіх 5 змінних. Окрім цього піддають аналізу значення математичних сподівань по кожній змінній з кожній з груп (всього шість груп).

Аналіз значень математичних очікувань в кожній з груп може дати додаткову інформацію щодо природи виникнення пропусків в даних або монотонності характеру їх появи. Наприклад, у випадку, якщо дані залежать від часу, може трапитись, коли в певний момент часу не буде інформації про досліджуваний процес, то це призведе до подальшого каскадного утворення пропусків. Окрім того, можуть бути виявлені шаблони пропусків даних, на які аналітик при первинному аналізі можливо не звернув увагу, але які повинні бути враховані для подальшого аналізу.

На третьому етапі здійснюється виявлення допоміжних змінних. Даний етап виконується за потребою. Допоміжні змінні – це змінні, що мають значення кореляції більше 0,4 із значеннями, що відсутні. Дослідження фахівців [145, 147] показали, що використання таких допоміжних змінних

покращує якість моделі множинного заповнення пропусків. Крім того, для отримання кращих результатів важливо знати і розуміти предметну область. Це дає можливість створювати допоміжні змінні і оптимально використовувати їх при створенні аналітичної моделі. А найпростішим способом пошуку допоміжних змінних є аналіз значень кореляції між змінними.

Підхід до заповнення пропусків з використанням методу Монте-Карло ланцюгами Маркова передбачає, що всі змінні моделі заповнення мають спільний багатовимірний нормальний розподіл.

Реалізація алгоритму описана в [68] під назвою – алгоритм розширення даних (DA – data augmentation algorithm). Вказаний алгоритм дозволяє заповнювати пропуски в даних, отримуючи значення для заповнення з умовного нормального розподілу (хоча, в загальному випадку розподіл може бути й іншого виду).

У більшості випадків обчислювальні експерименти із заповнення пропусків [68, 145, 147], показали, що припущення про багатовимірний нормальний розподіл, за достатнього обсягу вибірки, дозволяє отримати достовірні оцінки навіть в разі порушення гіпотези про нормальність розподілу. Однак, у випадку малого розміру вибірки і значної частки пропусків в даних зміщення оцінок буде спостерігатися.

Методи стохастичного заповнення пропусків використовуються для того, щоб додати мінливості значень прогнозів, які використовуються, для заміщення пропусків. З цією метою запропоновано підхід "add back lost variability" (повернення втраченої мінливості) [68]. Даний підхід полягає в тому, що до значення прогнозу додається флуктуаційне значення з математичним очікуванням рівним нулю і дисперсією рівною дисперсії залишків регресійної моделі.

При цьому вважається, що стандартні похибки, які виникають при оцінюванні регресії хоча і стають менш упередженими, ніж у випадку з єдиним заповненням, але все одно будуть менші [68, 145,147].

Виконуючи дослідження нелінійних нестаціонарних процесів різної природи було виявлено найбільш поширені причини виникнення пропусків в часових рядах даних. У проведених чисельних експериментах заповнення пропусків виконувалось у часових рядах соціально-економічних та еколого-економічних даних [171]. Часові ряди мали різну довжину та різну кількість пропусків. В результаті було запропоновано методику, яка дозволяє опрацьовувати часові ряди, пропуски у яких мають різні механізми формування та різну кількість пропущених значень (рис. 1.7).

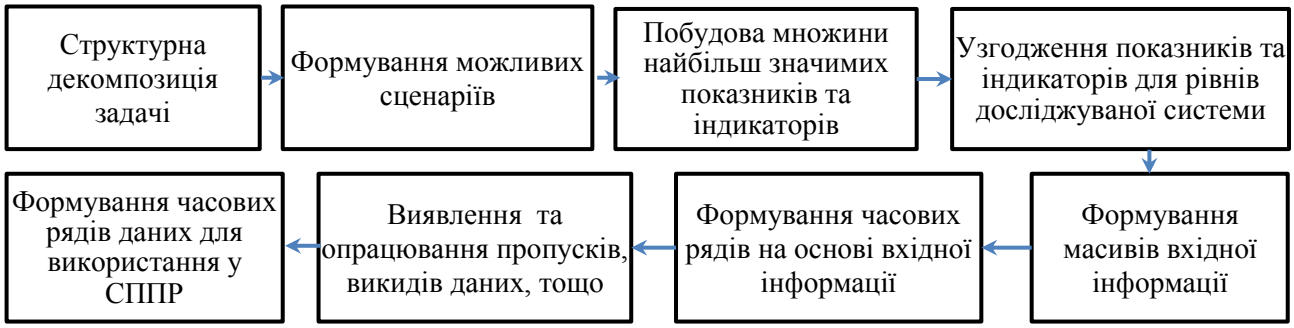


Рисунок 1.7 – Методика обробки вхідної інформації на етапі попереднього аналізу даних та заповнення пропусків

Як показали виконані чисельні експерименти, при виборі методу заповнення пропусків, потрібно визначити чи відповідає обраний метод конкретній аналітичній моделі досліджуваного процесу. Крім того, заповнення пропусків випадковими або детермінованими значеннями на основі використання, наприклад, аналітичної моделі або середнім за вибіркою, є штучним підходом, що не дозволяє виявити наявність шуму в даних та спотворює уяву про характер досліджуваного процесу. Тому для заповнення пропусків у часових рядах запропоновано використовувати методику на основі аналітичної моделі, побудова та реалізація якої здійснюється у два (за потреби у три) етапи. На першому етапі виконується аналіз наявності пропусків у досліджуваних часових рядах, на другому – аналіз характеру виникнення, кількість пропусків в даних та виконується

аналіз на наявність шаблонів, на третьому – виявлення допоміжних змінних за допомогою кореляційного аналізу даних та експертних оцінок стосовно області дослідження (даний етап виконується за потреби). Для визначення ефективних методів заповнення пропусків даних виконано порівняльний аналіз існуючих методів заповнення пропущених значень у довгих часових рядах даних та експериментально перевірено ефективність застосування різних алгоритмів за різної кількості пропусків.

Для таких бізнес-задач, як аналіз відтоку клієнтів, підвищення продажів, побудова скорингової карти при кредитуванні, які вирішуються методами прогнозного моделювання, найкращі результати показали методи заповнення пропусків з використанням прогнозних моделей. Так, при побудові скорингових моделей, підмішування даних про відхилені заявки клієнтів, на етапі верифікації інформації і схвалення заявки, з використанням методів нечіткої логіки і логістичної регресії для прогнозування (відновлення) значень відгуків, часто приводить до поліпшення якості фінальної скорингової моделі в середньому на 2-5% за статистикою Gini [16].

При обробці економетричних часових рядів для виявлення тренду і сезонності, досліджуваний часовий ряд повинен бути достатньо довгим [56].

Відновлення даних дозволяє подовжити ряд ретроспективних даних, однак, як показує практика не для всіх методів заповнення пропусків, подовження ретроспективної вибірки призводить до поліпшення прогнозних показників моделі. Найчастіше важливішою є наявність більшої кількості найбільш «свіжих» даних.

Одним із виконаних чисельних експериментів є заповнення пропусків даних у часовому ряду продажів квитків для прогнозування їх місячних обсягів.

Вибірка складалася з 144 значень місячних продажів квитків за період з січня 1959 по грудень 1960 року. Для побудови моделі використано відрізок даних з січня 1959 по червень 1960, для перевірки прогнозних показників – шість останніх місячних значень з липня по грудень 1960 року. 10%

навчальних даних з навчальної вибірки випадкових чином були заповнені пропусками.

Результати заповнення пропусків з наступною побудовою прогнозної моделі у вигляді адитивної моделі Вінтерса представлені в таблиці 1.7

Таблиця 1.7 – Результати прогнозування продажів квитків із використанням часових рядів, пропуски у яких були заповнені різними методами, кількість квитків, тис. шт.

Метод опрацювання пропусків	Місяць 1960 року						Середньо квадра-тична похибка прогнозу
	липень	серпень	вересень	жовтень	листопад	грудень	
Реальне значення	622	606	508	461	390	432	
Прогнозне значення без заміни пропуску	595	601	501	444	399	438	5,8
Прогноз з заміною пропуску на попереднє заповнене	596	603	503	447	401	434	5,35
Прогноз із заміною пропуску на медіанне значення	582	580	496	454	382	435	9
Прогноз з заміною пропуску на математичне сподівання	570	572	493	445	401	439	11,2
Прогноз з заміною пропуску за моделлю тренду	589	594	497	443	398	437	7
Прогноз з заміною пропуску на середнє між сусідніми	595	601	501	445	399	438	5,7

Чисельні експерименти були виконані також на одинадцяти найбільш довгих часових рядах (за 1940 – 2015 рр.), натуральних показників сільськогосподарського виробництва [171], які описують розвиток двох його підсистем: рослинництва та тваринництва [173, 174]. Слід зазначити, що виробництво валової продукції сільського господарства характеризується

позитивною динамікою. Такий же тренд мають і показники виробництва валової продукції рослинництва, а для тваринництва характерна негативна динаміка; показники, що описують дані процеси між собою корелюють достатньо слабко. Результати чисельних експериментів із заповнення пропусків подані у табл. 1.8.

Таблиця 1.8 – Значення похибки MAPE при заповненні пропусків даних різними методами для часових рядів за 1940-2015 рр. «Виробництво зернових та зернобобових в Україні, тис. т.» (зростаючий тренд) та «Виробництво молока в Україні, тис. т.» (спадаючий тренд)

Метод заповнення пропусків	Часовий ряд	Пропущено значень, %			
		5	10	15	20
Модель регресії	Виробництво зернових та зернобобових	17,59	15,06	14,69	13,84
	Виробництво молока	5,08	5,08	5,30	5,28
Середнє за вибіркою	Виробництво зернових та зернобобових	27,75	30,57	32,81	32,44
	Виробництво молока	26,82	30,87	35,01	35,11
Середнє між першим попереднім та першим наступним значеннями	Виробництво зернових та зернобобових	12,98	14,43	12,03	14,69
	Виробництво молока	4,91	4,34	4,03	5,76
Середнє п’яти попередніх значень, розрахованих з використанням вагових коефіцієнтів за принципом експоненційного згладжування	Виробництво зернових та зернобобових	19,64	19,59	20,24	19,2
	Виробництво молока	8,56	8,53	8,92	9,87
Модель авторегресії з урахуванням припущення про те, що сільськогосподарське виробництво має циклічність 3 роки	Виробництво зернових та зернобобових	20,84	20,03	19,49	23,39
	Виробництво молока	9,54	7,57	7,58	11,61

У експериментах, результати яких подані в таблиці 1.8 (та аналогічних, на інших часових рядах), використані часові ряди як з наявними пропусками даних, так і з примусовим формуванням пропусків у кількості від 5% до 20% від загальної кількості значень часового ряду. Як видно з таблиці 1.8, значення похибки MAPE суттєво відрізняється для досліджуваних часових рядів. Однак, найбільш ефективними виявилися методи заповнення пропусків значеннями, спрогнозованими із використанням моделі регресії, та значеннями, розрахованими як середнє між першим попереднім та першим наступним значеннями. Однак, доцільно заповнювати пропуски даних лише за наявності не більше 15% пропущених значень часового ряду. Практичну перевірку відновлюючих властивостей досліджених алгоритмів виконано при прогнозуванні цільової змінної – валової доданої вартості сільського господарства у фактичних цінах (основного показника який характеризує розвиток даної галузі) на основі використання моделі регресії на головних компонентах, у якій три перші головні компоненти описують 89% варіабельності всіх 14 змінних-регресорів, що були відібрані для побудови моделі (1.10):

$$Y(t)=a_0+a_1\cdot PRIN_1+a_2\cdot PRIN_2+a_3\cdot PRIN_3 \qquad (1.10)$$

де $PRIN_i$ – i-та головна компонента, отримана за методом головних компонент.

Значення показника валової доданої вартості сільського господарства спрогнозовано на три кроки вперед за моделлю як із заповненням пропусків, так і за фактичними даними. В ході експериментів із заповнення пропусків даних побудовано десять сценаріїв наявності до 20% пропусків даних у досліджуваних часових рядах, які були заповнені із використанням методів, поданих у таблиці 1.8.

Результати дослідження ефективності різних методів заповнення пропусків представлені у таблиці 1.9.

Таблиця 1.9 – Результати прогнозування валової доданої вартості сільського господарства України як із заповненням пропусків у часових рядах регресорів, так і за фактичними даними

Назва прогнозу	Валова додана вартість сільського господарства у фактичних цінах, млн. грн.			MSE	MAPE
	2013	2014	2015		
Фактичні значення	132245	161145	236003		
Прогнози на основі заповнення пропусків із використанням:					
- моделі лінійної регресії	197962,8	159993,8	170943,7	53394,3	24%
- середнього значення за вибіркою	111167,8	112288,1	106414,6	80879,4	61%
- методу розрахунку середнього між першим попереднім та першим наступним значеннями	120392	116399	112564	76113,0	52%
- методу розрахунку середнього п’яти попередніх значень, розрахованих з використанням вагових коефіцієнтів за принципом експоненційного згладжування	122973,6	116292,5	113925,1	75278,81	51%
- моделі авторегресії	115874,1	117063,5	100527,1	82794,8	62%
Прогноз без заповнення пропусків	101825,1	102443,7	94516,1	90166,0	78%

Як видно з таблиці 1.9, найкращі результати одержано з використанням методу заповнення пропусків на основі лінійної регресії (MAPE = 24%), а у випадку, коли заповнення пропусків не виконувалося – MAPE сягає 78%. Слід також відзначити, що за результатами обчислювальних експериментів, проведених на 15 часових рядах показників виробництва продукції сільського господарства за 1940 – 2015 рр., найкращі результати одержано при застосуванні методу заповнення пропусків середнім між першим попереднім та першим наступним значеннями (MAPE за всіма експериментами становить 8,44%), із використанням інших методів, зокрема,

моделі лінійної регресії, експоненційно середньозваженого значення за п’ятьма наявними вимірами, авторегресії, заповнення середнім за вибіркою, значення похибки MAPE становить відповідно 13,33, 15, 15,85, і 51,41%.

В ході виконання значної кількості експериментів було побудовано прогнози на основі часових рядів, які містили до 20% пропущених значень. Пропуски були заповнені із використанням різних методів за десятьма сценаріями наявності до 20% пропусків даних у досліджуваних часових рядах. Виявлено, що навіть, якщо часовий ряд містить 20% пропущених значень, прогноз, отриманий після їх заповнення значеннями, отриманими за пропонованою методикою, дав кращі результати. Зокрема, відхилення прогнозного значення показника відрізнялось від фактичного в межах 5-10%, а значення MAPE становило від 8,5 до 24%, MSE – від 5,4 до 11% (залежно від використаного методу заповнення пропусків), в той час у випадку, коли заповнення пропусків не виконувалось, значення вказаних показників становили 61-78%.

Як показало виконане дослідження, при виборі методу заповнення пропусків необхідно враховувати характер процесу, що аналізується, особливості формування часових рядів, кількість та щільність пропусків. Якщо кількість пропущених значень у часовому ряду перевищує 20% , то за таких довгих послідовностей пропусків їх доцільно заповнювати за допомогою моделей регресії та середніми значеннями по вибірці, оскільки інші є непрацездатними.

1.8 Постановка проблеми дисертаційної роботи

Як показало дослідження існуючих методів, моделей і технології моделювання та прогнозування ННП в умовах невизначеностей, дана проблема є актуальною. Однак, не зважаючи на значний доробок вітчизняних та закордонних науковців та фахівців-практиків, залишається не вирішеною. Мають місце проблеми та недоліки при практичному використанні існуючих

інформаційних технологій, і як наслідок, можна говорити про необхідність створення нових інформаційних технологій, що ґрунтуються на сучасних методах інтелектуального аналізу даних, багатомодельному підході та алгоритмах. Розділи 1.2 – 1.4 містять необхідні означення та математичні визначення щодо нестаціонарності та нелінійності процесів та їх ідентифікацію. Розділ 1.5 містить огляд методів моделювання ННП, а розділ 1.6 містить типи невизначеностей, що характерні для різних підсистем складних систем. Загалом проведений аналіз моделювання ННП в умовах невизначеностей показав необхідність вирішення наступних проблем:

1. Відсутність системної методології побудови математичних моделей для опису та прогнозування процесів, що характеризують розвиток нелінійних нестаціонарних процесів в умовах невизначеності.

2. Необхідність розробки нових та удосконалення вже існуючих методів прогнозування нелінійних нестаціонарних процесів різної природи на основі використання ймовірнісних та комбінованих прогнозів.

3. Недостатність існуючих метод оцінювання параметрів математичних моделей та їх ансамблів, що дозволяють подолати зміщеність оцінок прогнозів.

4. Розроблення метод розкриття невизначеностей різних типів з метою коректного розв’язання задач прогнозування розвитку обраних процесів.

5. Розробити математичні моделі та ансамблі моделей обраних нелінійних нестаціонарних процесів для прогнозування їх розвитку з метою підтримки прийняття управлінських рішень в умовах інформаційної невизначеності.

6. Провести апробацію запропонованих підходів до моделювання та прогнозування, запропонувати підходи, що забезпечать адекватний опис причинно-наслідкових зв’язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних нелінійних нестаціонарних процесів.

7. Зробити аналіз адекватності побудованих моделей та формування висновків на їх основі щодо можливості їх застосування для розв'язання задач оцінювання і прогнозування розвитку нелінійних нестаціонарних процесів;

8. Із використанням основ системного підходу побудувати та реалізувати інформаційну технологію для розв'язання задач моделювання та прогнозування нелінійних нестаціонарних процесів у різних галузях досліджень для її подальшого використання у системах підтримки прийняття рішень.

Отже аналіз виконаного огляду літератури та Інтернет джерел показав, що проблема розробки інформаційних технологій аналізу та прогнозування нелінійних нестаціонарних процесів, які забезпечують високу обчислювальну ефективність в умовах невизначеностей, є актуальною.

Задача дисертаційного дослідження полягає у створенні нових моделей та методів виконання системних досліджень нелінійних нестаціонарних процесів різної природи та розроблення на їх основі методологічних засад розробки інформаційних технологій, що дозволить підвищити якість оцінок прогнозів та відповідних рішень в умовах невизначеності та ризику.

Висновки до розділу 1

Окреслено задачі аналізу нелінійних нестаціонарних процесів, що мають місце у різних галузях досліджень, зокрема, в економіці, фінансах, екології, соціальній сфері та ін. Подано відповідні взаємодоповнюючі означення, які узагальнюють характеристики вказаних процесів і можуть бути використані при виконанні порівняльних досліджень. Виокремлено характеристики нелінійних процесів, що мають місце при аналізі статистичних/експериментальних даних у вказаних галузях.

Виконано огляд існуючих методів щодо заповнення пропусків даних, а також надано опис необхідних процедур та практичних рекомендацій.

Для вирішення окреслених проблем були виділені та застосовані

основні аспекти використання методології системного аналізу з метою розв’язання задач прогнозування нелінійних нестаціонарних процесів різної природи. Представлена загальна методика обробки інформації при розв’язанні задач формування інформаційної бази для їх дослідження, в тому числі ідентифікацію та опрацювання невизначеностей різних типів при формуванні часових рядів вхідних даних, визначено методи та моделі для прогнозування нелінійних нестаціонарних процесів та шляхи їх реалізації в рамках відповідної інформаційної технології.

З’ясовано, що метод аналізу повних спостережень, включає оцінку середнього і коваріації, ґрунтуючись на всіх доступних заповнених випадках. Коваріаційна (або кореляційний) матриця складається з елементів, кожен з яких розраховується для кожної пари змінних. Цей метод став популярний, тому що втрата потужності через недостатньої інформації не настільки значна, як при аналізі повних спостережень.

Залежно від аналізованої пари змінних, розмір вибірки змінюється, за рахунок обліку кількості пропусків у змінних. Даний підхід призводить до того, що оцінки параметрів будуть зміщені.

В той час як підхід на основі безумовного заповнення середнім значенням, включає заміну пропусків по кожній змінній, значенням середнього по відповідній змінній. Цей підхід відносно простий в застосуванні, тому що дозволяє заповнити пропуски центральними значеннями, що в результаті не змінює загального значення середнього змінної по вибірці.

При цьому, математичне очікування не змінюється, в той час як статистика розмаху розподілу (стандартне відхилення, дисперсія) зазнають змін, за рахунок появи концентрації спостережень навколо середнього вибірки, за рахунок заповнення пропусків середніми значеннями.

Більш того, на практиці інколи можна побачити, що значення кореляції між цільовою змінною і регресорами зменшуються, що свідчить про ослаблення впливу відповідних незалежних змінних.

Заповнення пропусків одиночними або детермінованими значеннями на основі використання моделі лінійної регресії або середнім за вибіркою, є штучним підходом, що не відображає можливість наявності шуму в даних. Прогнозні значення, отримані, наприклад, на основі лінійної регресії будуть відображати виключно лінійну залежність, яка графічно з себе представлятиме похилу пряму.

РОЗДІЛ 2
ТЕОРЕТИЧНІ ТА МЕТОДОЛОГІЧНІ ОСНОВИ ПОПЕРЕДНЬОГО АНАЛІЗУ ДАНИХ НА НАЯВНІСТЬ ЗВ'ЯЗКІВ ТА АНОМАЛІЙ

2.1 Основні етапи і задачі попереднього аналізу даних

В умовах, коли зростає потреба якісного інформаційного забезпечення, все більш актуальним стає питання створення ефективної інформаційної технології збору та попередньої підготовки даних до їх пдальшого використання. Тому значну увагу у роботі приділено саме розробці інформаційної технології попереднього опрацювання значних обсягів різнорідних даних, отриманих із різних інформаційних джерел. Вказана технологія заснована на системному використанні методів статистичного аналізу та автоматизованого інтегрування різнотипної інформації, методів моделювання, прогнозування та багатокритеріального прийняття рішень. Для подолання можливих невизначеностей статистичного, структурного та параметричного типів запропоновано множину методів, залежно від типу невизначеності. Запропонована методика попереднього опрацювання даних вирізняється тим, що при формуванні масивів вхідних даних їх інформативність, синхронність та коректність забезпечуються на етапі попередньої обробки, яка виконується для приведення їх до форми, що забезпечить можливість коректного застосування методів оцінювання параметрів моделі та отримання їх статистично значущих оцінок. Зокрема, розроблено методику корегування значних імпульсних (екстремальних) значень, нормування вимірів у заданих межах, логарифмування великих значень та фільтрації шумових складових, виявлення та усунення мультиколінеарності [149-164, 175-181].

Встановлено, що для побудови математичних моделей у задачах моделювання нелінійних нестаціонарних процесів доцільно використовувати підхід, що ґрунтується на розділенні всієї вибірки вхідних даних на

тренінгову, валідаційну та тестову. Проблемним питанням побудови математичних моделей нелінійних нестаціонарних процесів є так зване «прокляття розмірності», коли необхідно будувати модель на коротких часових вибірках даних. Тобто використання існуючих методів прогнозування в задачах прийняття рішень, які ґрунтуються на аналітичних процедурах, логічних правилах та раціональному експертному оцінюванні, у багатьох випадках не дає бажаного результату стосовно якості оцінок прогнозів, а тому в процесі дослідження зазвичай вирішується проблема системного використання альтернативних методів моделювання і прогнозування для підвищення якості оцінок прогнозів для кожного конкретного випадку. Зокрема, для моделей нелінійних нестаціонарних процесів важливо знайти підмножину вхідних змінних (які називаються ознаками або атрибутами), що якнайкраще характеризуватимуть предметну область. Для цього застосовуються три стратегії – стратегія фільтра (наприклад, накопичення ознак), стратегія обгортання (наприклад, пошук згідно заданої точності) і стратегія вкладення (вибираються ознаки для додавання або видалення в міру побудови моделі, заснованої на аналізі похибок прогнозування).

Для застосування техніки відбору ознак при побудові моделей нелінійних нестаціонарних процесів у роботі використано такі підходи: спрощення моделей, щоб зробити їх простішими для інтерпретації дослідниками / користувачами; зменшення часу на тренування (навчання) моделі; зменшення розмірності вхідних даних; покращене узагальнення шляхом уникнення перенавчання (формально, це зменшення дисперсії).

Пропонована методика попереднього аналізу даних, призначена для аналізу даних на наявність зв’язків та аномалій під час побудови моделей нелінійних нестаціонарних процесів, складається з таких етапів:

Етап 1. Графічний аналіз даних із використанням діаграм розсіювання включає використання графіків та кореляційного аналізу.

Етап 2. Діагностика екстремальних значень та викидів, включає аналіз

діаграм розсіювання та графіків на наявність викидів, за потреби, з’ясовуються причини появи викидів.

Етап 3. Аналіз впливу екстремальних значень, студентизовані залишки; D-статистика Кука; статистика DFFITS; статистика DFBETAS.

Етап 4 Дослідження даних на наявність колінеарності, включає використання VIF-статистики, індекс стану та пропорцію варіації. Перевірка пропорції варіації включає дослідження студентизованих залишків, D-статистика Кука та DFFITS статистика.

Етап 5. Аналіз категоріальних даних, включає дослідження їх розподілу, проведення попереднього дослідження наявності зв’язку між змінними (використовується статистика хі-квадрат Пірсона, V-статистика Крамера, співвідношення шансів, хі-квадрат тест Мантел-Гаензеля)

Етап 6. Стандартизація даних використовує методи стандартизації, узагальнення яких представлене в роботі.

2.2 Графічний аналіз даних діаграмами розсіювання

Для вивчення взаємозв'язку між двома неперервними змінними використовуються [179, 180]:

- графіки розсіювання (Scatter Plot)
- кореляція (найпопулярніший варіант – Пірсона).

Діаграма розсіювання – двовимірний графік, на якому кожна точка відповідає парі значень (X, Y), що описують об’єкт дослідження.

Даний графічний підхід використовується для:

- дослідження зв'язку між двома змінними;
- виявлення екстремальних значень і аномалій;
- ідентифікація наявних трендів;
- ідентифікація діапазону зміни X і Y;
- інтерпретації результатів аналізу.

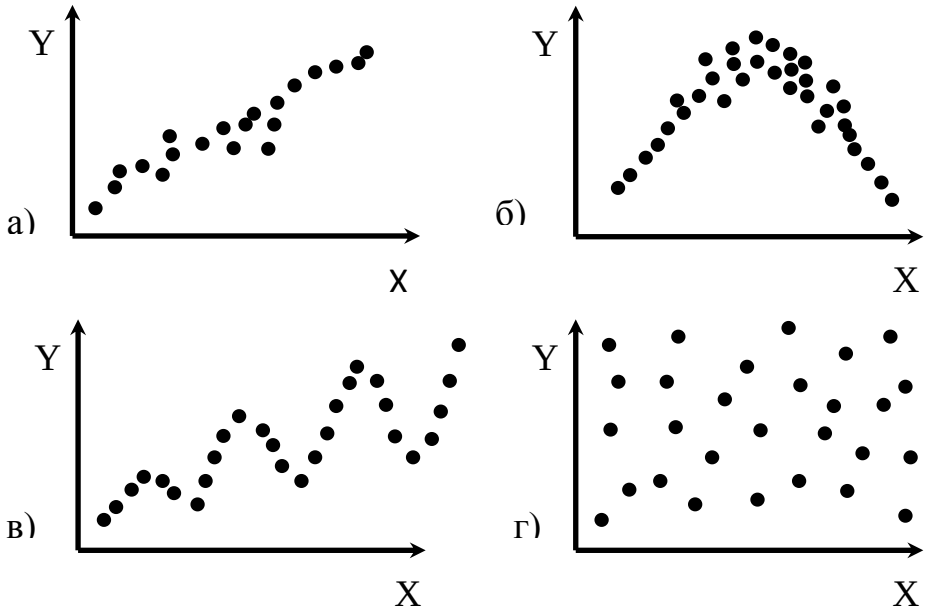


Рисунок 2.1 – Приклади можливих зв’язків між змінними *X* та *Y*

Опис зв'язку між двома неперервними змінними – це перший важливий крок у будь-якому статистичному аналізі. На рис. 2.1 показана графічна інтерпретація таких видів можливих зв’язків між змінними, як: наявність лінійного (а), нелінійної зв'язку(б), циклів(в), відсутність зв'язку між змінними (г).

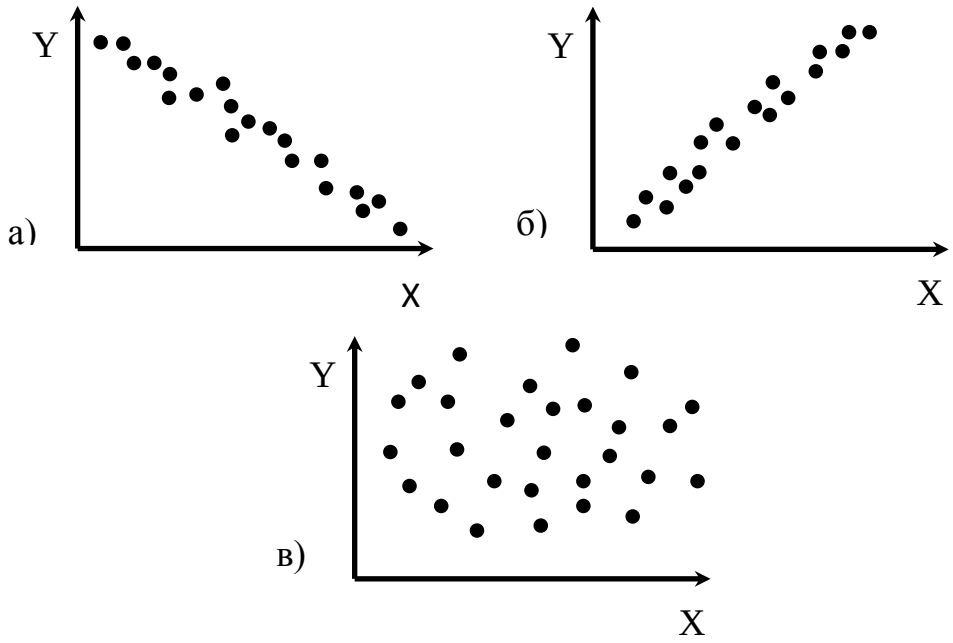


Рисунок 2.2 – Графічне представлення різних типів кореляції між змінними

На рис.2.2 наведені типи кореляції між змінними:

а) позитивна – при збільшенні значення X збільшується Y;

б) негативна – при збільшенні значення X зменшується Y;

в) відсутність зв'язку – відсутність зв'язку між X і Y.

При трактуванні результатів кореляції часто можуть виникати помилки. Наприклад, за результатами аналізу роблять висновок про наявність причинно-наслідкового зв'язку між змінними, але:

- сильна кореляція між двома змінними не означає, що зміна однієї змінної обов'язково призводить до зміни значення іншої змінної або навпаки;
- вибіркова кореляція може мати завищені значення тому що на змінну, яка аналізується, насправді впливають невраховані фактори;
- наявність кореляції не має на увазі причинності;
- кореляція Пірсона використовується для виявлення лінійного зв'язку;
- якщо значення коефіцієнта кореляції Пірсона близьке до нуля, то лінійного зв'язку немає;
- якщо значення коефіцієнта кореляції Пірсона близьке до нуля, то це не означає, що між змінними взагалі немає ніякого зв'язку.

На значення коефіцієнта кореляції досить сильно впливають екстремальні значення, тому при виникненні подібних ситуацій, пропонується:

- дослідити підозрілі значення, на предмет того чи є вони коректними;
- якщо підозріле значення виявилося коректним, то потрібно розрахувати статистики між підозрілим значенням і основною групою даних, щоб пересвідчитись у наявності запропонованого лінійного зв'язку;
- відтворити підозрілі значення, дані з фіксованим значенням інкрементації для того щоб зрозуміти, чи є спостереження дійсно екстремальним;
- обчислити два коефіцієнта кореляції: один з урахуванням підозрілого спостереження, а другий без нього, для того, щоб виявити, наскільки впливовим для поточного аналізу є підозріле значення.

Подібний аналіз даних використовується, як правило, при побудові моделі множинної регресії [175]. Будуються двомірні діаграми розсіювання і

обчислюються парні значення кореляцій між основною змінною і регресорами. Попередній аналіз при підготовці до побудови моделі множинної регресії часто включає розгляд двовимірних діаграм розсіювання і кореляцій між кожною із змінних-предикторів та відгуком.

Однак, на основі цих результатів не слід приймати рішення про виключення або включення регресорів. Мета полягає в тому, щоб дослідити форму зв'язку між відгуком і регресорами, а також виявлення екстремальних значень.

2.3 Діагностика екстремальних значень та викидів в моделі

В статті [179] наведена серія коректних прикладів, що демонструють важливість графічної інтерпретації даних аналізу. На рис. 2.3 наведена серія графічних прикладів діаграм розсіювання даних. Вигляд рівнянь та статистика R^2 для всіх чотирьох випадків, представлених на рис. 2.3, однакові, і описуються рівнянням $Y=3+0.5x$ та значенням статистики $R^2=0,67$:

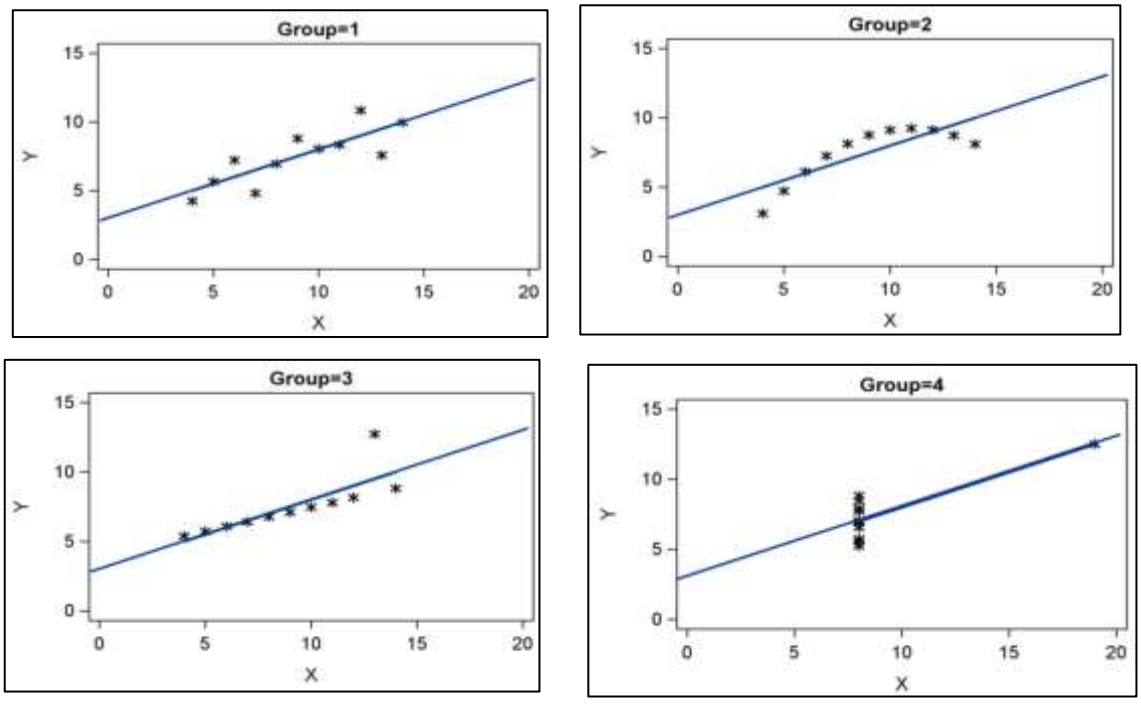


Рисунок 2.3 – Випадки діаграми розсіювання для значення статистики $R^2 = 0,67$

На діаграмі розсіювання (рис. 2.3 (а)) наведено типовий приклад адекватної лінійної регресійної моделі, що досить точно описує дані.

На діаграмі розсіювання (рис. 23 (б)) представлена підібрана лінійна залежність, яка зовсім не відповідає дійсному зв'язку між даними, тому що є явно виражена нелінійна залежність у вигляді поліному другого порядку.

На діаграмі розсіювання (рис. 2.3 (в)) яскраво виражене екстремальне значення в даних, що призводить до зміщення лінії регресії вгору. Це екстремальне значення суттєво пливає на зміну кута регресійного рівняння.

На діаграмі розсіювання (рис. 2.3 (г)) наведено приклад суттєвого впливу екстремального значення на зміну кута лінії регресії. Якби екстремальне значення було виключено з аналізу, кут нахилу лінії регресії дорівнював би дев'яноста градусам.

Наведені на рис. 2.3 приклади діаграм розсіювання показують, що взаємозв'язок, який описується регресійним рівнянням, може вводити в оману, при цьому, значення статистики R^2 для всіх випадків однакове, навіть незважаючи на суттєву різницю у характері зв'язків між даними.

Як підтверджує рис. 2.3, робити попередній візуальний аналіз даних, як підготовчий етап до побудови моделей досить корисно.

Для перевірки припущень щодо регресії можна використовувати залишки моделі , що обчислюються за формулою (2.1) 175:

$$r_i = Y_i - \hat{Y}_i, \tag{2.1}$$

де Y_i – реальне значення i –го спостереження, $\hat{Y}_i$ – значення прогнозу i –го спостереження.

Для перевірки припущень доцільно використовувати діаграми розсіювання залишків моделі в залежності від значень прогнозів; а також діаграми розсіювання залишків моделі в залежності від значень кожного з регресорів моделі [69, 175, 179].

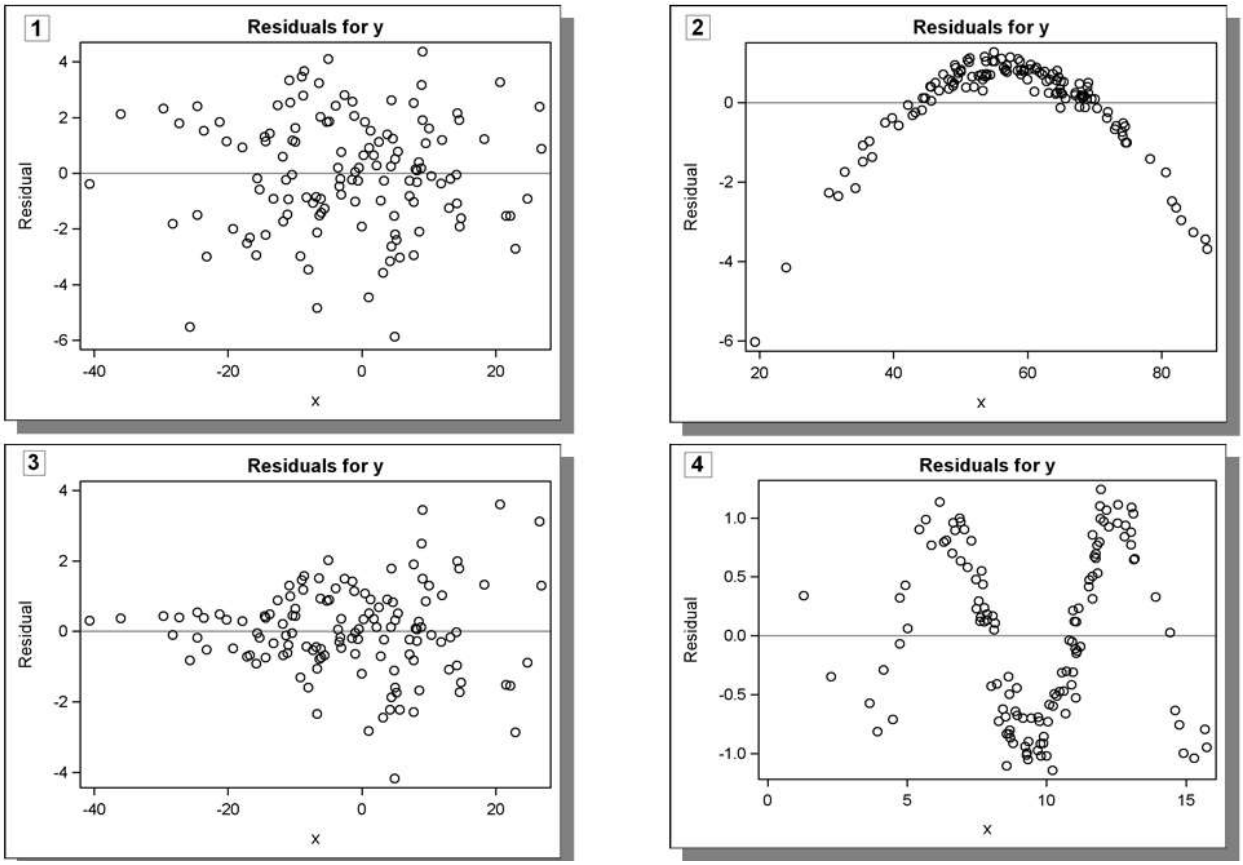


Рисунок 2.4 – Приклади діаграм розсіювання залишків моделі залежно від значень прогнозів

На рис 2.4 представлені чотири ситуації діаграм розсіювання залишків моделі у порівнянні з прогнозованими значеннями або значеннями змінної-предиктора. Якщо припущення регресійної моделі є дійсними, то залишки повинні бути випадковим чином розсіянні навколо лінії 0. Наявність будь-яких патернів або тенденцій в залишках свідчить про наявність проблем при побудові моделі, а саме:

– модель адекватна, тому що залишки випадковим чином розсіянні навколо значення 0, а також відсутні будь-які патерни;

– модель не відповідає даним. В даних наявний нелінійний зв'язок, як варіант вирішення проблеми – додати квадратичну залежність у рівняння регресії.

– зліва на право, як видно на рис. 2.4 (діаграмі розсіювання), дисперсія збільшується, у випадку, коли дисперсія не є постійною, можливі варіанти

вирішення проблеми: застосувати перетворення по відношенню до залежної змінної або, інший варіант – застосувати спеціальні процедури з наступним вибором моделі, що не передбачає рівної дисперсії;

– випадок, коли спостереження не є незалежними (наведений графік залишків містить автокореляцію) має місце, коли дані збиралися протягом часу, і тоді, щоб усунути цю проблему, необхідно використовувати спеціалізовані методи аналізу часових рядів.

2.4 Аналіз впливу екстремальних значень

Окрім перевірки припущень щодо регресії, також важливо перевірити наявність викидів в даних.

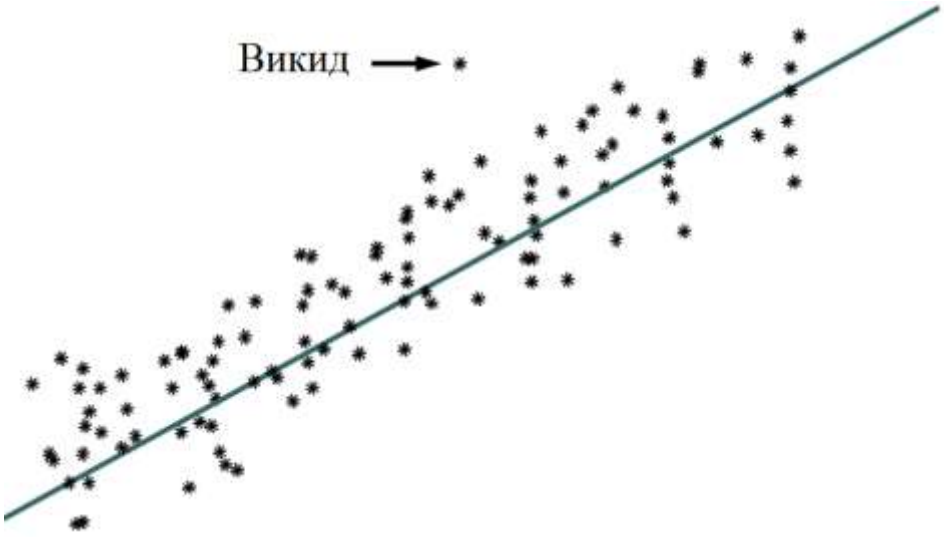


Рисунок 2.5 - Приклад викиду в даних

Спостереження, що знаходяться відчутно далеко від основної частини даних, є викидами. Ці спостереження часто є помилками даних або відображають незвичайні обставини, що виникли в процесі функціонування об’єкта. У будь-якому випадку, коректним є підхід, коли окремо з'ясовують причини виникнення викидів.

На рис. 2.3 представлено приклади, коли рівняння регресії та статистика R^2, були однакові, але аналіз діаграм розсіювання показав, що

необхідно будувати інші моделі. Ідентифікація екстремальних значень в даних є важливим етапом при дослідженні.

Для діагностики екстремальних значень використовуються такі статистики [68, 176]: студентизовані залишки; RSTUDENT-залишки; статистика Cook's D; статистика DFFITS; статистика DFBETAS.

Один із підходів, щодо перевірки викидів – використання студентизованих залишків. Вони обчислюються шляхом ділення залишків на значення стандартної похибки.

Для моделі, яка добре підходить до даних і не має відхилень, більша частина студентизованих залишків повинна бути близька до 0. В цілому, студентизовані залишки, що мають абсолютне значення менше 2,0, інколи виникають випадково. Студентизовані залишки, модуль яких знаходиться між 2,0 та 3,0, спостерігаються не часто і дійсно, можуть бути викидами. Модуль значення студентизованих залишків перевищує значення 3,0, зустрічаються ще рідше і підлягають більш глибокому вивченню та дослідженню.

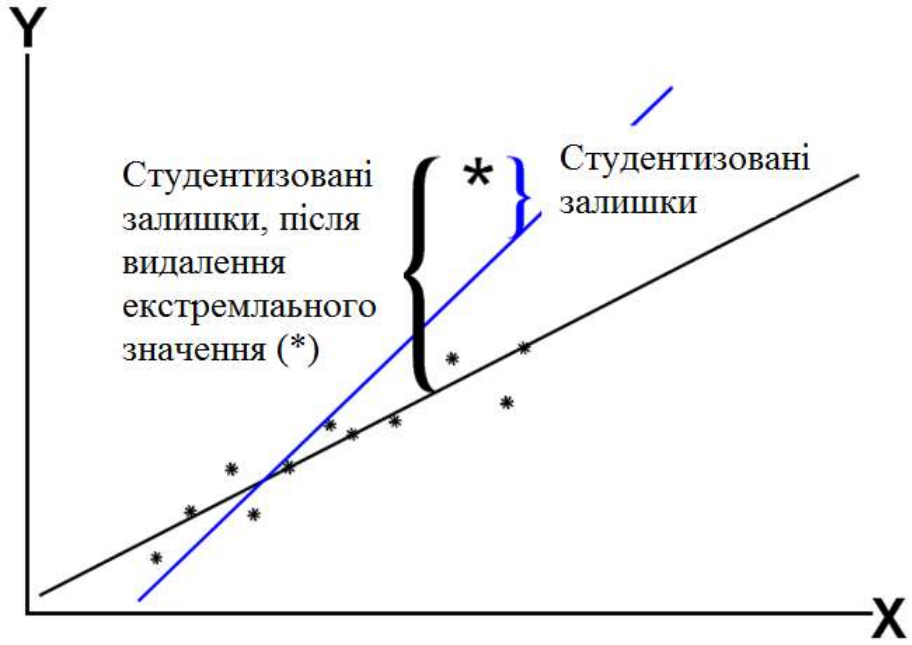


Рисунок 2.6 – Графічна ілюстрація студентизованих залишків до видалення екстремального значення, та після (в результаті отримуємо RSTUDENT залишки).

Якщо студентизовані залишки $|SR| > 2$, то є підозра на наявність викиду, якщо $|SR| > 3$ – однозначно, є викид, необхідно більш глибоке вивчення та дослідження даних.

Студентизовані залишки є звичайними залишками, розділеними на їх стандартні похибки. RSTUDENT залишки схожі на студентизовані залишки, за винятком того, що вони обчислюються після видалення i-го екстремального спостереження. Інший варіант пояснення, RSTUDENT залишки – це різниця між реальним значенням відгуку Y та значенням-прогнозу відгука $\hat{Y}$, за виключенням екстремального значення.

D- татистика Кука використовується для ідентифікації екстремальних значень. Ця статистика вимірює зміну оцінок параметрів, що виникає при видалені кожного спостереження (2.2) [68]:

$$\text{D} - \text{статистика Кука} = \left(\frac{1}{ps^2}\right)\left(b - b_{(i)}\right)'(X'X)\left(b - b_{(i)}\right), \qquad (2.2)$$

де $p -$ кількість параметрів регресії; $s^2 -$ середньоквадратична похибка регресійної моделі; $b -$ вектор оцінок параметрів; $b_{(i)} -$ вектор оцінок параметрів, отриманих після видалення i-го спостереження; $X'X -$ скоригована сума квадратів і матриці перехресних добутків.

В якості порогу відсікання, для ідентифікації екстремальних значень пропонується використовувати (2.3) [68]:

$$\text{D} - \text{статистика Кука}_i > \frac{4}{n}, \qquad (2.3)$$

де $n -$ кількість спостережень.

Статистика DFFITS (2.4) вимірює вплив, який має i-те спостереження на значення, що прогнозується [68].

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s(\hat{Y}_i)} \quad (2.4)$$

де $\hat{Y}_i -$ i −те значення прогнозу, $\hat{Y}_{(i)} - i$ −те значення прогнозу, при видаленні i −го спостереження, $s(\hat{Y}_i) -$ стандартна похибка i −го значення прогнозу.

Белсі Д., Кух Е. та Уелш Р. [183] запропонували формулу для обчислення порогового значення (2.5):

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}} \quad (2.5)$$

де $p -$ кількість регресорів в моделі, враховуючи вільний член, а $n -$ розмір навчальної вибірки.

Назва статистики DFBETAS походить від англійського Difference in Betas. Ця метрика містить стандартизовану різницю для кожної оцінки окремих коефіцієнтів, що виникає через відсутність i-го спостереження. Існують різні варіанти DFBETAS. Наприклад, варіант D-статистика Кука (2.6) [183]:

$$DFBETA_{ij} = \frac{b_j - b_{(i)j}}{s(b_j)}, \quad (2.6)$$

де $b_j -$ це оцінка параметра j −го регресора; $b_{(i)j} -$ це оцінка параметра j −го регресора, з вилученим i −м спостереженням; $s(b_j)$ −стандартна похиюбка b_j.

Великі значення статистики DFBETAS вказують на те які саме регресори можуть бути причиною екстремального впливу.

За рекомендацією Белсі Д., Куха Е. та Уелша Р. [183], прийнято значення порогу рівного 2, при перевищені якого вважається, що

спостереження екстремальне, або обчислювати поріг як $2\sqrt{\frac{1}{n}}$, тобто з поправкою на розмір вибірки.

На рис.2.7 представлені екстремальні спостереження, що знаходяться за межами двох стандартних похибок від середнього значення 0. В якості прикладу розглянуто аналіз залежності витрат часу на виконання фізичних вправ спортсменом в залежності від показників: X1 – якість виконання вправ, X2 – витрачений час на виконання вправ, X3 – вік спортмена, X4 – вага спортсмена, X5 – пульс у процесі виконання вправ, X6 – пульс у стані спокою, X7 – максимальний пульс.

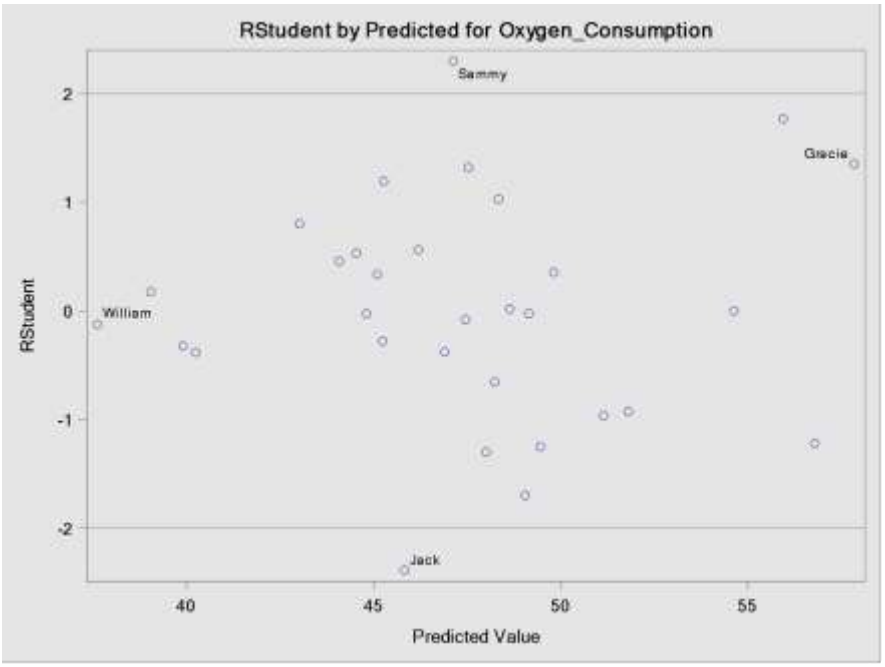


Рисунок 2.7 – RStudent залишки моделі; спортсмени Sammy та Jack – екстремальні спостереження

Як можна побачити з рис. 2.6, всього ідентифіковано два екстремальних спостереження – спортсмени Sammy та Jack

Оскільки допустимим вважається твердження, що 5% значень виходять за 95% довірчий інтервал, а в нашому випадку тільки два значення з 31, що становить 6% від вибірки, то ці екстремальні значення не викликають особливого занепокоєння, тому що 6% близько до 5%.

D-статистика Кука, вказує на те що спостереження Gracie є екстремальним (рис. 2.8).

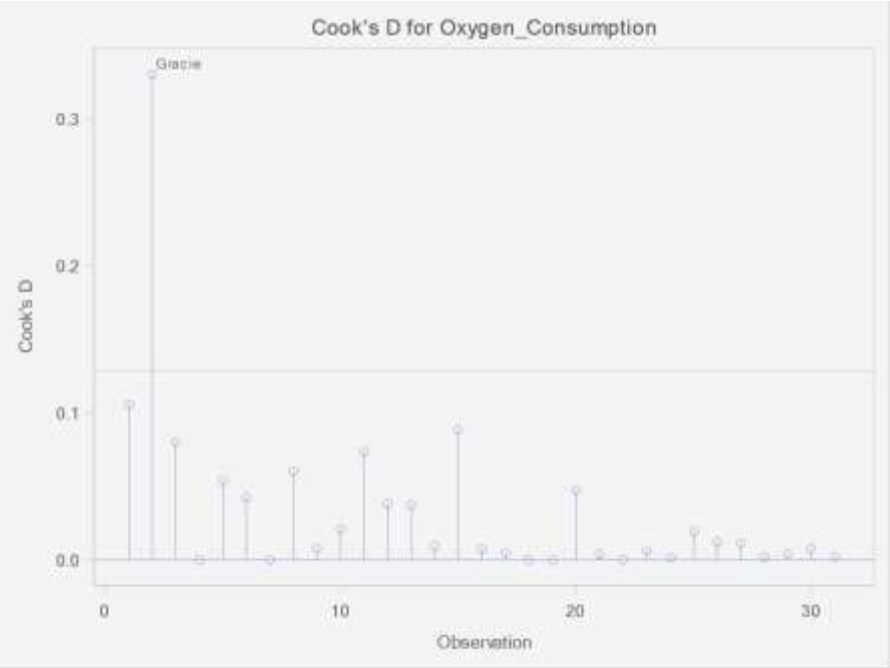


Рисунок 2.8 – Приклад D-статистика Кука

Статистика DFFITS також, вказує на те що спостереження спортсмену Gracie є екстремальним (рис. 2.9).

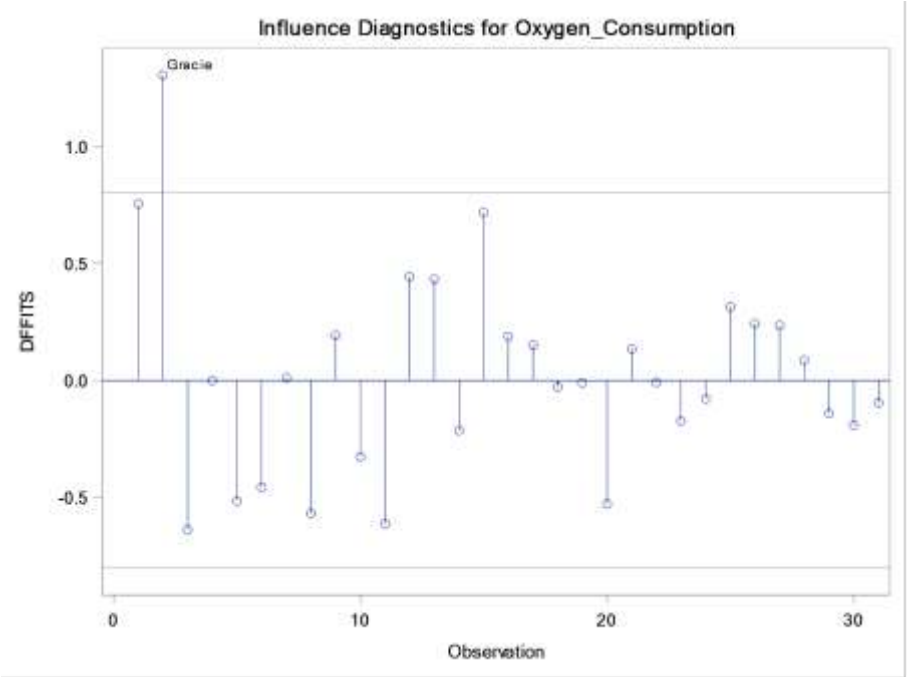


Рисунок 2.9 – Статистика DFFITS

На детальніше проаналізувати спостереження Gracie можна скориставшись статистики DFBETAS (рис. 2.10).

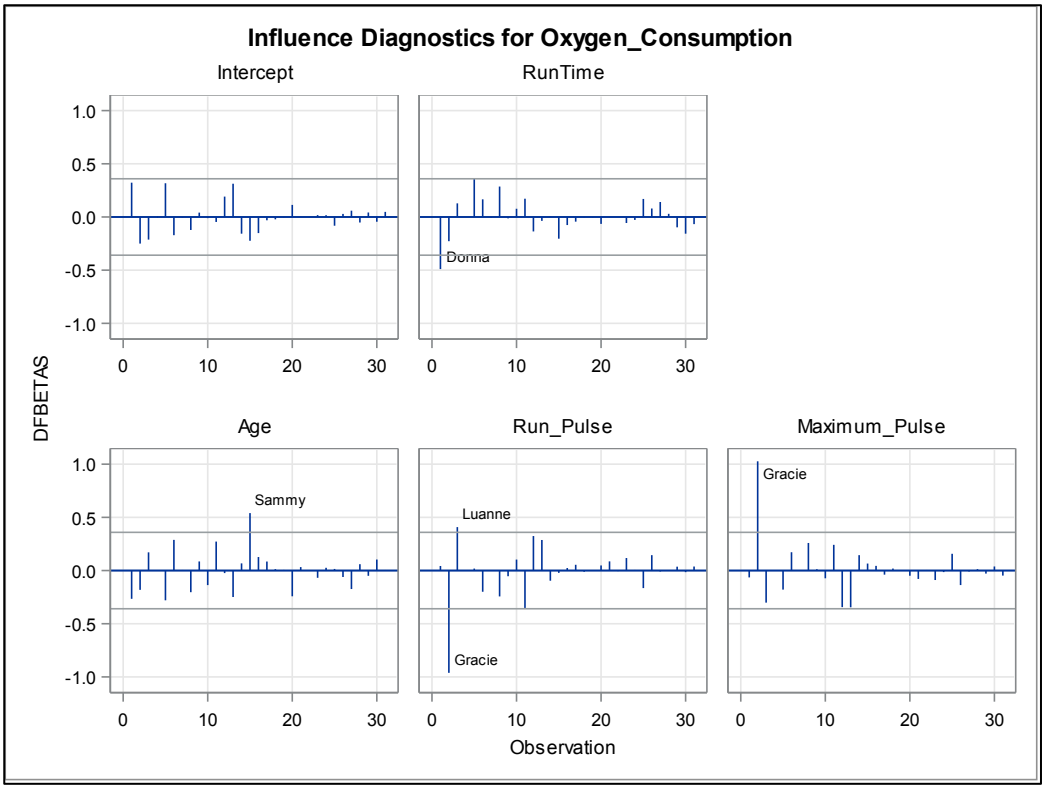


Рисунок 2.10 – Статистика DFBETAS

Як можна побачити з рис. 2.9, результати статистики DFBETAS, спостереження для об’єкту Gracie є аномальним за значеннями по змінними X5 (пульс у процесі виконання завдання) та X7 (максимальний пульс)

Виявлення викидів за графіками можливо на невеликих вибірках, але для великих наборів даних має сенс аналізувати детально саме тільки підмножини екстремальних значень.

2.5 Методичні підходи до розв’язання проблеми колінеарності в даних

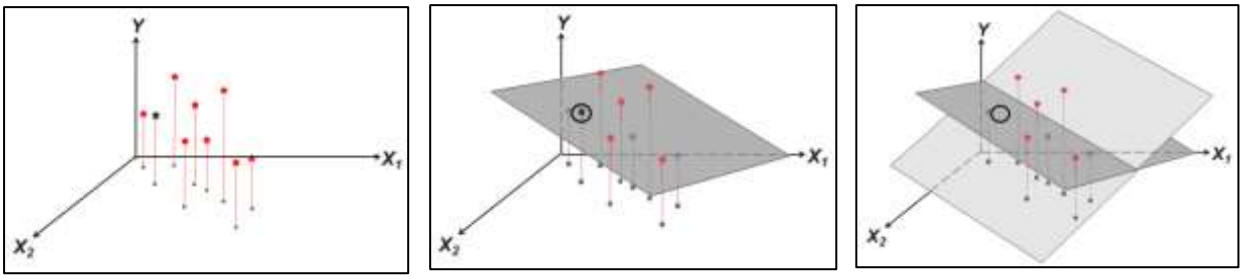
Метою побудови будь-якої математичної моделі є пошук найкращої площини, гіперплощини або огинаючих кривих, що проходять через дані для прогнозування відгуку.

Проблема колінеарності полягає в тому, що:

1. Колінеарність може збільшити уявну статистичну значущість ефектів, за рахунок повторного включення тієї ж порції інформації.

2. Колінеарність прагне збільшити дисперсію оцінок параметрів і, отже, збільшити похибка прогнозування.

На рис 2.10 наведена графічна інтерпретація проблеми колінеарності. В тривимірному просторі відображений відгук Y та два регресори X_1 і X_2. Кут нахилу лощини X_1 та X_2 відносно Y і є оцінкою коефіцієнта параметра моделі.як раз демонструє сильний лінійний зв'язок між відгуком та регресорами.



(а) (б) (в)

Рисунок 2.11 – Графічна інтерпретація проблеми колінеарності

(а) проблеми колінеарності; (б) побудова площини, що краще всього описує колінеарні дані; (в) видалення або переміщення точки даних, призводить до побудови нової площини, яка описує дані, що значно відрізняється від попередньої

При цьому, як показано на рис. 2.10(в), видалення тільки однієї точки даних (або переміщення цієї точки) призводить до появи іншої площини, що описує дані, при цьому кут цієї площини значно відрізняється від попереднього. Все це демонструє варіабельність оцінок параметрів при екстремальній колінеарності.

За наявності колінеарності оцінки коефіцієнтів моделі нестійкі, вони мають велику дисперсію. Більше того, реальний зв'язок між Y та X_S може

сильно відрізнятися від отриманих результатів. При цьому наявність колінеарності не порушує умов, що накладається на лінійну регресію.

Для діагностики колінеарності використовується набор інструментів для кількісного оцінювання проблеми колінеарності та визначення множини регресорів X_S, що містять колінеарність.

Метрика COLLIN аналізує матрицю $X'X$ на колінеарність, при цьому враховується також вільний член. Для аналізу матриці $X'X$ на колінеарність, застосовують метрику COLLINOINT [68], при цьому вільний член не враховується. Метрики COLLIN та COLLINOINT являють собою міру величини колінеарності та надають інформацію щодо множини регресорів X_S, що і є джерелом колінеарності.

Метрика VIF, скорочення від англійської Variance Inflation Factor, являє собою міру величини колінеарності.VIF – це відносна міра росту дисперсії внаслідок наявності колінеарності (2.7) [68, 176].

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.7}$$

де $R_i^2 -$ статистика R квадрат для регресуючого X_i на всіх регрессорах моделі X_S.

Наприклад, якщо модель для прогнозування відгуку Y включає регресори X_1, X_2, X_3, X_4 (тобто X_S для $i = 1, \dots, 4$), то для обчислення статистики R квадрат на регресорі X_3, будується модель, що включає регресори X_1, X_2, X_4. Обчислене значення статистики R^2 підставляють до формули 2.7:

$$VIF_3 = \frac{1}{1-R_3^2},$$

якщо значення $VIF_i > 10$, то присутня колінеарність.

Значення VIF_i обчислюється для кожного регресора моделі.

Таблиця 2.1 – Приклад оцінок коефіцієнтів параметрів моделі

Вільний член	Оцінка параметрів					
	КСС	оцінка параметра	стандартна похибка	t-стати стика	p-значення	Значення VIF-статистики
X1 (якість виконаня)	1	131.7825	72.20754	1.83	0.0810	0
X2 (час виконання)	1	-0.12619	0.30097	-0.42	0.6789	162.854
X3 (вік)	1	-3.86019	2.93659	-1.31	0.2016	88.86251
X4 (вага)	1	-0.46082	0.58660	-0.79	0.4401	51.01176
X5 (пульс у процесі виконання завдання)	1	-0.05812	0.06892	-0.84	0.4078	1.76383
X6 (пульс у спокої)	1	-0.36207	0.12324	-2.94	0.0074	8.54498
X7 (максимальний пульс)	1	-0.01512	0.06817	-0.22	0.8264	1.44425
Вільний член	1	0.30102	0.13981	2.15	0.0420	8.78755

З даних, наведених в таблиці 2.1 видно, що є значення VIF більші 10, а це свідчить про наявність проблеми колінеарності. Для початку має сенс звернутися до аналітиків за консультацією, і з’ясувати з якими змінними та з яких причин виникають проблеми.

Наведена ситуація трапляється часто, але в даному випадку це гарна демонстрація того, що коли змінна-агрегат включена в модель разом із своїми змінними-складовими, тоді обов’язково виникає проблема колінеарності. Якщо змінна-агрегат дійсно важлива для аналізу, то має сенс її залишити, а змінні-складові виключити з аналізу. Однак, використання змінних-агрегатів замість індивідуальних змінних має недолік у вигляді втрати інформації, тому в представлено прикладі має сенс виключити з аналізу саме змінну-агрегат Performance (табл. 2.2).

Таблиця 2.2 – Оцінки коефіцієнтів параметрів моделі після виключення змінної-агрегату Performance

Змінна	Оцінка параметрів					
	Змінна	КСС	оцінка параметра	стандартна похибка	t-статистика	p-значення
Вільний член	1	101.963	12.27174	8.31	<.0001	0
X2 (час виконання)	1	-2.63994	0.38532	-6.85	<.0001	1.58432
X3 (вік)	1	-0.21848	0.09850	-2.22	0.0363	1.48953
X4 (вага)	1	-0.07503	0.05492	-1.37	0.1845	1.15973
X5 (пульс у процесі виконання завдання)	1	-0.36721	0.12050	-3.05	0.0055	8.46034
X6 (пульс у спокої)	1	-0.01952	0.06619	-0.29	0.7706	1.41004
X7 (максимальний пульс)	1	0.30457	0.13714	2.22	0.0360	8.75535

Як можна побачити в табл. 2.2, значення VIF стали значно менші для моделі у порівнянні з повною моделлю. Однак для змінних X7 (максимальний пульс) та X5 (пульс у процесі виконання завдання) все ще спостерігається колінеарність.

Колінеарність може істотно вплинути на результат побудови регресії, що будується за методом STEPWISE (поеапного включення регресорів)[133]. Оскільки значимість важливих змінних може бути замаскована колінеарністю, як наслідок кінцева модель може не включати дуже важливі регресори. Тому рекомендується використовувати перевірку на колінеарність перед побудовою моделі в автоматичному режимі.

В загальному випадку існують інші підходи щодо вирішення проблеми колінеарності, наприклад, два методи – гребнева регресія та регресія на

головних компонентах. Крім того, центрування змінних-предикторів іноді може дозволити вирішити проблему колінеарності, особливо для випадків поліноміальної регресії в ANCOVA (Analysis of covariance) моделях.

Застосування методів та моделей для попереднього аналізу даних узагальнено на рис. 2.12.

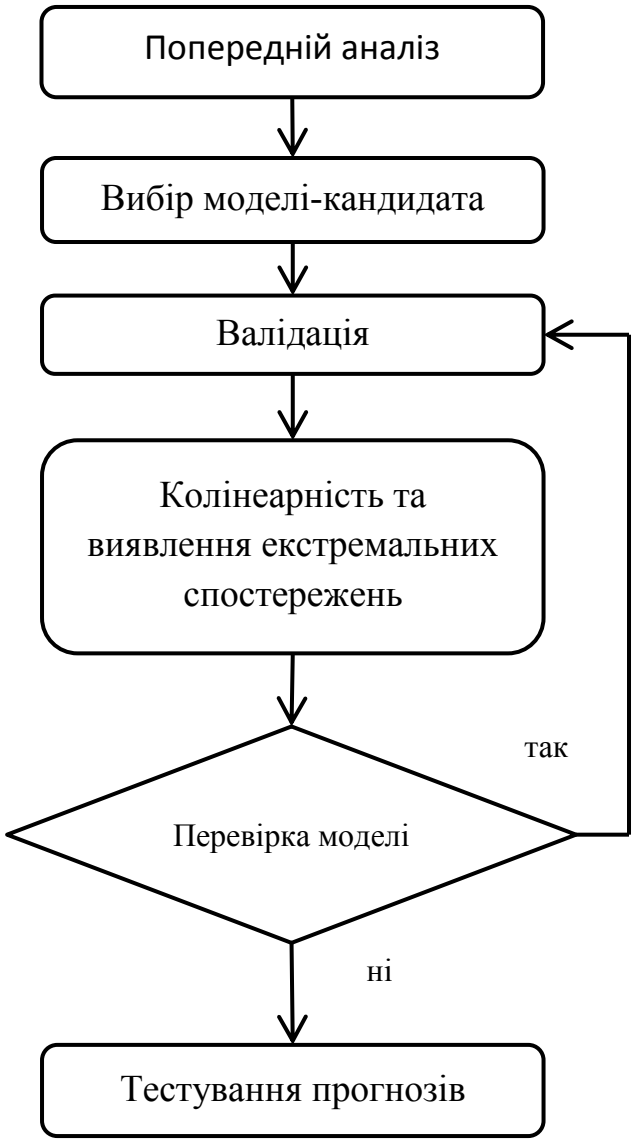


Рисунок 2.12 – Схема застосування моделей та методів попереднього аналізу даних під час побудови математичної моделі нелінійних нестаціонарних процесів

Пропонована методика включає виконання наступних етапів.

Етап 1. Попередній аналіз: використання описових статистик, графіків та кореляційного аналізу.

Етап 2. Вибір моделі-кандидата: ідентифікація однієї або декількох моделей-кандидатів.

Етап 3. Валідація: побудова та аналіз графіків та діаграм розсіювання залишків у порівнянні з значеннями прогнозів,перевірка рівності дисперсії.

Етап 4. Виявлення колінеарності та екстремальних спостережень: перевірка на колінеарність з використанням VIF статистики, індексу стану та пропорції варіації (обчислення студентизованих залишків, D-статистика Кука та DFFITS статистик).

Етап 5. Тестування моделі: якщо на етапах 3 та 4 очевидною є необхідність перевірки моделі, то будується нова модель із поверненням до цих кроків.

Етап 6. Тестування прогнозів: якщо є можливість, то модель перевіряється на даних, що не використовувалися для її побудови.

2.6 Аналіз категоріальних даних

При попередньому аналізі категоріальних даних, досліджують їх розподіл, обчислюють частоту появи відповідних значень змінних та наявність зв’язку між ними. Між двома категоріальними змінними існує зв'язок, якщо зміна розподілу однієї змінної відбувається разом із зміною значень інших змінних. Якщо розподіл першої змінної є однаковим, незалежно від значень іншої змінної, то зв'язок відсутній.

Кількість спостережень, що відбуваються в певних категоріях або інтервалах показує таблиця частот. У таблиці з односторонньою частотою розглядається одна змінна. Зазвичай однофакторна частотна таблиця, що будується із використанням спеціалізованого програмного забезпечення, містить такі стовпчики:

– частота – скільки разів значення змінної з'являється в наборі даних;

– процент – відсоток появи значення змінної в наборі даних;
– кумулятивна частота – накопичено значення частоти, до другого значення частоти додається перше і т. д.;
– кумулятивний процент – накопичене значення проценту: до другого значення відсотку додається перше і т. д.

Матриця перехресних значень (табл. 2.3) показує кількість спостережень для кожної комбінації змінних рядків і стовпців.

Таблиця 2.3 – Загальний вигляд матриці перехресних значень

	Стовпець 1	Стовпець 2	…	Стовпець c
Рядок 1	Статистика 11	Статистика 12	…	Статистика 1c
Рядок 2	Статистика 21	Статистика 22	…	Статистика 2c
…	…	…	…	…
Рядок r	Статистика r1	Статистика r2	…	Статистика rc

Кожна клітинка матриці перехресних значень містить наступні статистики:

– частота – кількість спостережень, що потрапляють у категорію, що утворюється комбінацією значеннями рядкової змінної та стовпчика;
– процент – кількість спостережень у відсотках від загальної кількості спостережень;
– процент по рядку – кількість спостережень у відсотках від загальної кількості спостережень у цьому рядку;
– процент по стовпчику – кількість спостережень у відсотках від загальної кількості спостережень у цьому стовпчику.

Зазвичай, для перевірки наявності зв’язку між двома категоріальними змінними [186] використовується статистика хі-квадрат Пірсона, яка

обчислює різницю між частотами що спостерігаються та емпіричними. Якщо статистика хі-квадрат значима, то це вагомий доказ того, що існує зв'язок між аналізованими змінними.

Тест хі-квадрат і p −значення, що йому відповідає, призначений для виявлення наявності зв'язку між змінними, не вимірює силу зв'язку між змінними і залежить від розміру вибірки.

Під нульовою гіпотезою щодо відсутності зв'язку між змінними Row та Column розуміють, що очікувана частота в будь-якій комірці R*C, значення комірок матриці по рядку (R / T) та стовпчику (C / T) будуть рівні. Очікувана кількість – це лише очікуваний відсоток від загального розміру вибірки, що обчислюється за формулою 2.8 [186]:

$$\text{Expected count} = \left(\frac{R}{T}\right) \cdot \left(\frac{C}{T}\right) \cdot T = \frac{(R \cdot C)}{T}, \tag{2.8}$$

де R – кількість спостережень по рядку, C – кількість спостережень по стовпчику, T – загальна кількцсть спостережень в матриці контингенції

p-значення для тесту хі-квадрат лише вказує, наскільки існує впевненість щодо нульової гіпотези відносно відсутності зв'язку. Цей тест не надає інформації щодо величини сили зв'язку. Значення статистики хі-квадрат також не говорить про це. Якщо навіть розмір вибірки буде збільшений, наприклад дублюючи кожне спостереження, подвоєння значення статистики хі-квадрат відбудеться, навіть якщо сила зв'язку не зміниться.

Наприклад, є набору даних людей частина з яких дожили до аступу пенсійного віку, а чатсина ні, при цьому є змінні Стать (жіноча або чоловіча стать) та Дожиття (Померти до пенсійнго віку=Так чи Ні). Обчислення статистики хі-квадрат виконується у три кроки.

Крок 1. Обчислити частотну матрицю значень (загальний вигляд частотної матриці представлений у табл. 2.4.

Таблиця 2.4 – Загальний вигляд частотної матриці

Показник	Померти до настання пенсійного віку=Ні	Померти до настання пенсійнго віку=Так	Загалом
Жінки	A	B	A+B
Чоловіки	C	D	C+D
Загалом	A+C	B+D	A+B+C+D

Крок 2. Обчислити частотну матрицю частот, що очікуються (приклад формул для розрахунків кожного елемента частотної матриці представлений в табл. 2.5)

Таблиця 2.5 – Загальний вигляд матриці очікуваних частот

Показник	Померти до пенсійного віку=Ні	Померти до пенсійного віку=Так
Жінки	(A+B)*(A+C) / (A+B+C+D)	A+B)*(B+D)/ (A+B+C+D)
Чоловіки	(C+D)*(A+C)/ (A+B+C+D)	(C+D)*(B+D)/ (A+B+C+D)

Крок 3. Обчислити значення хі-квадрат за формулою 2.9:

$$\chi^2 = \sum_{i=}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}, \qquad (2.9)$$

де i −й номер рядка (від 1 до r), j −й номер стовпчика (від 1 до c), O_{ij} − фактична кількість спостережень для (i, j) елемента матриці, E_{ij} − очікувана кількість спостережень для (i, j) елементу матриці.

Одним з показників сили зв’язку між двома номінальними змінними є V-статистика Крамера.

Ця статистика має діапазон від –1 до 1 для таблиць контингенції розміром 2 на 2 і від 0 до 1 для великих таблиць. Значення, що знаходяться далеко від 0 вказують на більш сильний зв’язок. V-статистика Крамера походить від статистики хі-квадрат Пірсона.

Співвідношення шансів вказує, наскільки більша ймовірність виникнення події в одній групі, відносно її виникнення в іншій.

Наприклад, порівняти шанси осіб різної статі дожити до пенсійного віку. Співвідношення шансів може бути використана як міра сили зв’язку для таблиць контингенції розміром 2 на 2. Шанси обчислюються зі значень ймовірностей (2.10):

$$Odds = \frac{P_{event}}{1 - P_{event}} \tag{2.10}$$

де P_{event} – ймовірність наступу події, наприклад, дожити до наступу пенсійного віку

Результати розрахунків представлені в табл. 2.6.

Таблиця 2.6 – Приклад таблиці контингенції 2 х 2

Група	Відгук		
	Так	Ні	Загалом
Група А	60	20	80
Група Б	90	10	100
Загалом	150	30	180

Як видно з таблиці 2.6, існує 90% ймовірність отримання значення «ТАК» для відгуку в групі Б, обчислюються як відношення: $\frac{90}{100} = 0{,}9$, а

ймовірність отримання значення «НІ» для відгука в групі Б, обчислюються як відношення: $\frac{10}{100} = 0{,}1$

Шанси в групі Б отримати значення відгуку «ТАК» по відношенню до значення НІ, дорівнюють: $\frac{0{,}9}{0{,}1} = 9$. Тобто в групі Б поява події «ТАК» в дев'ять разів більша (частіша) у порівнянні з появою події «НІ».

Аналогічно обчислюють значення для групи А. Ймовірність отримання значення «ТАК» для відгуку в групі А: $\frac{60}{80} = 0{,}75.$

Ймовірність отримання значення «НІ» для відгука в групі А: $\frac{20}{80} = 0{,}25.$ Шанси в групі А отримати значення відгуку «ТАК» по відношенню до значення «НІ», дорівнюють $\frac{0{,}75}{0{,}25} = 3.$

Відношення шансів групи А до групи Б дорівнює 1/3 або 0,3333, що вказує на те, що шанси на отримання відгуку зі значенням «ТАК» в групі А становлять одну третину по відношенню до групи Б.

Якщо цікавить співвідношення шансів у групі А по відношенню до групи Б, то треба зробити інверсію обчислень, $\left(\frac{1}{3}\right)^{-1} = 3.$

Аналіз співвідношення шансів засвідчує силу зв'язку між регресором-предиктором та змінною-відгуком. Якщо коефіцієнт шансів дорівнює 1, то зв'язку між регресором та відгуком немає. Якщо коефіцієнт шансів перевищує 1, то група А, що знаходиться в чисельнику формули, має більше шансів на те, що подія відбудеться.

Якщо коефіцієнт шансів менше 1, то група Б, що знаходиться в знаменнику, має більше шансів на відбуття події.

Якщо більше ніж 20% елементів матриці очікуваних частот містить значення менше п'яти, то в цьому випадку асимптотичний тест хі-квадрат Пірсона не використовується (табл. 2.7).

Це пов'язано з тим, що $p-$значення засновано на припущенні про значний розмір вибірки. Як наслідок, коли розмір вибірки замалий, то асимптотично $p-$значення не валідно.

Таблиця 2.7 – Приклад матриці очікуваних частот, для випадку коли 20% елементів матриці мають значення менше п’яти

Погода	Гарний настрій	Депресія
Сонячно	4	2
Дощі	2	3

Хі-квадрат тест Мантел-Гаензеля [186] – це тест для перевірки зв’язку між ординарними змінними (коли значення маються внутрішній порядок). Ординарний зв’язок передбачає, що при збільшенні однієї змінної інша змінна має тенденцію до збільшення або зменшення. Щоб результати тесту були значимими, для випадку коли змінна має більше двох рівнів, рівні повинні мати логічне впорядкування. Нульова гіпотеза: зв’язку між змінними по рядках та стовпчиках немає. Альтернативна гіпотеза: ординарний зв'язок між змінними по рядках та стовпчиках існує.

Хі-квадрат тест Мантел-Гаензеля [186]:

- дозволяє визначити, чи є ординарний зв'язок,
- при цьому він не вимірює силу ординарного зв’язку,
- залежить від розміру вибірки.

Статистика Хі-квадрат теста Мантел-Гаензеля є більш потужною, аніж загальна статистика хі-квадрат при використанні по відношенню до ординарних змінних.

Для виміру сили зв’язку між ординарними змінними, також можна використовувати кореляцію Спірмена. Ця статистика:

- має діапазон значень від –1 до 1,
- якщо близька до +1, то це означає що між змінними сильна додатна кореляція,
- якщо близька до -1, то між змінними сильна від’ємна кореляція,
- має сенс використовувати тільки для ординарних змінних, значення яких мають логічне упорядкування.

2.7 Стандартизація даних

В загальному випадку для стандартизації даних, в тому числі і при заповненні пропусків, саме стандартизованими значеннями, використовується наступна формула [183, 186] (2.11):

$$result = add + multiply \times \frac{(original - location)}{scale} \quad (2.11)$$

де result – кінцеве значення (стандартизоване), add – адитивна константа (зміщення), multiply – мультиплікативна константа, original – початкове (оригінальне) значення (до стандартизації), location – оцінка вибіркового середнього значення (як правило математичне сподівання), scale – оцінка вибіркового значення масштабу (як правило стандартне відхилення).

Середнє мінімального інтервалу - обчислюється за спеціальним алгоритмом [68, 106]. Для обчислення вибіркових статистик використовується 90% вибірки, 10% вибірки з лівого і правого боку розподілу виключається з розрахунків. За своєю суттю це вибіркова статистика за усіченою вибіркою, що очищена від викидів.

Відстань Евкліда між двома точками, в загальному випадку обчислюється за формулою:

$$\rho(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \ ,$$

але для випадку однієї змінної вироджується в формулу:

$$\rho(x) = \sqrt{\sum_{i=1}^{n} (x_i)^2} \ .$$

В загальному випадку, відстань Мінковського [188] між двома точками обчислюється за формулою:

$$\rho(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

але для випадку однієї змінної вироджується в:

$$\rho(x) = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

2.8 Особливості опрацювання неструктурованих даних

Враховуючи проблеми, з якими стикаються аналітики при побудові моделей нелінійних нестаціонарних процесів [3, 17, 35], особливо під час аналізу предметної області, все більше дослідників віддають перевагу використанню когнітивних, причинно-наслідкових моделей в поєднанні з математичними та економетричними моделями, хороші результати можна отримати й тоді, коли в якості регресорів додаються результати опрацювання неструктурованої інформації [130,189]. Це дозволяє урахувати альтернативи, суттєві для перебігу досліджуваного процесу, що підвищує якість інформаційного забезпечення процесу побудови моделей. Інформаційною основою для побудови таких «змішаних» моделей, є не лише статистичні показники та рейтингові оцінки, а й слабкоструктуровані Big Data.

Використання слабкоструктурованої інформації передбачає відбір та аналіз альтернатив, необхідних для побудові сценаріїв, виявлення чинників, які уточнюють уявлення про досліджувані процеси, а також при визначенні цільової установки дослідження.

Використання у моделях слабкоструктуровної інформації передбачає послідовну реалізацію набору кроків, які охоплюють попередню обробку,

зокрема, текстової інформації з Інтернет, формування набору правил для аналізу, відбір найбільш значимих критерії та цілей, так і візуалізацію одержаних результатів.

В даній роботі реалізовано метод застосування інформаційних технологій для розв'язування задач прогнозування нелінійних нестаціонарних процесів, який ґрунтується на інтеграції різнотипної інформації й заснований на системному використанні методів аналізу даних, моделювання, методів прогнозування і методів прийняття рішень**.**

Для вирішення цієї задачі запропоновано виконувати аналіз неструктурованої інформації, в тому числі, розміщеної в мережі Інтернет, використовуючи засоби SAS Textual Analytics. Методика використання засобів текстової аналітики, запропонована у роботі, передбачає послідовну реалізацію множини кроків, які охоплюють як попередню обробку текстової інформації, формування множини правил для аналізу, відбір найбільш значимих критерії та цілей, так і візуалізацію одержаних результатів. Використання засобів текстової аналітики, як це було запропоновано у роботі, дає можливість найбільш повно сформулювати альтернативи та визначити цільову настанову. У випадках, коли інформація, що описує досліджувані процеси має суттєві протиріччя, пропуски, похибки і затримки, може бути сумнівною щодо достовірності, доцільно застосовувати методику визначення подібності процесів.

Для реалізації пропонованої методики обробки та аналізу неструктурованої інформації, отриманої з мережі Інтернет, запропоновано алгоритм латентно-семантичного аналізу, реалізований засобами мови програмування SAS Base [99, 187]. Пропонована методика вирізняється тим, що на етапах інтерпретації результатів та виявлення найбільш значимих факторів можуть бути застосовані дерева рішень, кластерний аналіз [72, 188], ймовірнісно-статистичні моделі [142], експертні оцінки [90]. Узагальнена методика використання неструктурованих даних для побудови моделей нелінійних нестаціонарних процесів представлена на рис. 2.13.

Рисунок 2.13 – Методика опрацювання неструктурованих даних із використанням Text Mining

Проблему інтелектуального аналізу тексту можливо описати в рамках кроків, які необхідно виконати практично в будь-якій задачі інтелектуального аналізу даних. Першим кроком є отримання ознак з колекцій документів, щоб можна було виконувати обчислення і застосовувати статистичні методології. «Витягнуті» ознаки повинні якимось чином відображати зміст документів. В кращому випадку вони повинні фіксувати контент таким чином, щоб документи, які обговорюють схожі теми, але з іншою термінологією, мали аналогічні характеристики. Далі, необхідно сформулювати спосіб вимірювання відстаней між документами, де відстань будується, щоб вказати, як вони схожі за змістом. Як буде показано нижче, це не тільки дозволяє нам застосовувати методи класифікації і кластеризації, але також дозволяє формулювати стратегії для зменшення розмірності. З огляду на візуалізацію елементів в просторі з більш низькою розмірністю, ми можемо потім звернути нашу увагу на кластеризацію, дискримінантний аналіз [187,189].

Перша частина вилучення ознак – це попередня обробка словника або лексикону (сукупність усіх унікальних слів в колекції документів). Зазвичай це включає три частини: видалення стоп-слів, стемінг (визначення коренів слів) і визначення ваги термінів. До лексикону можна застосувати будь-яку,

всі або жодну з цих трьох частин. Застосування і корисність цих методів - відкрите питання дослідження в області інтелектуального аналізу текстових даних, яке повинно представляти інтерес для статистичного співтовариства[189, 190-193].

Стоп-слова – це слова, які не мають змістовного навантаження. Вони моужть бути видалені без спотворення тексту. Наприклади, сполучники, вигуки, тощо. Стоп-слова можуть бути задані у вигляді списку абоопрацьовуватись залежно від контексту документу.

Стемінг часто застосовується в області пошуку інформації, де метою є підвищення продуктивності системи та зменшення кількості унікальних слів. Стемінг - це процес видалення суфіксів і префіксів, залишаючи корінь або основу слова [187]. Наприклад, слова «захищений», «захисти», «підзахисний» і «захист» будуть скорочені до кореневого слова «захист». Це має сенс, оскільки слова мають схоже значення. Однак в деяких випадках скорочене слово має зовсім інше значення. Скоріш за все, це не впливає на результати пошуку інформації, але може мати деякі небажані наслідки при класифікації і кластеризації. Стемінг і видалення стоп-слів зменшить розмір словника, тим самим заощадить обчислювальні ресурси.

Один із способів кодування тексту – це підрахувати, скільки разів термін зустрічається в документі. Це називається частотним методом. Однак терміни з великою частотою не обов'язково більш важливі або володіють більш високою здатністю розпізнавання. Отже, ми могли б захотіти, щоб терміни мали вагові коефіцієнти по відношенню до локального контексту, документу або колекції документів.

Найпопулярнішим показником, мабуть, є IDF (обернена частота документа) - інверсія частоти, з якою слово зустрічається в документах відповідної колекції [187] (2.12):

$$IDF = \log(\frac{N}{df_i}) , \qquad (2.12)$$

де N – загальна кількість документів; df_i – кількість документів, що містять термін i.

Існує розширення цього показника, який позначається як TF-IDF. Точне формулювання TF-IDF приведено нижче [187] (2.13).

$$w_{ij} = f_{ij} * \log(\frac{N}{df_i}) , \quad (2.13)$$

де w_{ij} – ваговий коефіцієнт терміну i в документі j; f_{ij} – кількість входжень терміну i в документі j; N – загальна кількість документів; df_i – кількість документів, що містять термін i.

З огляду на те, що лексикон був попередньо оброблений (чи ні), вміст документів кодують. Один з найпростіших способів зробити – це використати підхід з використанням векторного простору набору слів. Цей підхід має перевагу, тому що він не вимагає будь-якої обробки природної мови, такої як ідентифікація частини мови або мовний переклад. Колекція документів кодується як терм-документа матриця X з n рядками і p стовпцями, де n представляє кількість слів у лексиконі (повний або попередньо оброблений словниковий запас), а p – кількість документів. Таким чином, елемент x_{ij} матриці X містить кількість разів, скільки і-й термін зустрічається в j-ому документі (частота терміна). Як було написано вище, це також може бути зважена частота. Кодування колекції документів у вигляді матриці дозволяє використовувати можливості лінійної алгебри для аналізу колекції документів [187].

Підхід векторного простору було розширено за рахунок включення порядку слів в сенсі пар або трійок слів. Замість однієї матриці кодується кожен документ як матрицю, що називається матрицею близькості, так що тепер у нас є p матриць для роботи. По-перше, потрібно виконати додаткову попередню обробку документів. Вся пунктуація видаляється (наприклад, коми, крапки з комою, двокрапки, тире і т. д.), а всі розділові знаки в кінці

речення перетворюються в крапку. Крапка вважається словом в лексиконі. Таким чином, кожна матриця близькості має *n* рядків і *n* стовпців [187].

Для вирішення завдань інтелектуального аналізу текстових даних, пов'язаних з кластеризацією, класифікацією і пошуком інформації, необхідно застосувати поняття відстані або подібності між документами. Найбільш використовуваною мірою при інтелектуальному аналізі текстових даних і пошуку інформації є косинус кута між векторами, що представляють документи [187]. Припустимо, у нас є два вектори документа $\vec{a}$ і $\vec{q}$, тоді косинус кута між ними, θ, визначається як (2.14):

$$\cos(\theta) = \frac{\vec{a}^T \vec{q}}{||\vec{a}||_2 \, ||\vec{q}||_2} \ , \tag{2.14}$$

де $||\vec{a}||_2 = \sqrt{\sum_{i=1}^{n} a_i^2}$ - L_2 норма вектора $\vec{a}$.

Слід зазначити, що великі значення цього показника вказують на близьке розташування документів, а менші значення – на те, що документи знаходяться далі один від одного.

Міра косинуса – це міра подібності, а не відстань. Зазвичай нам зручніше працювати з відстанями, але ми можемо легко перетворити міру подібності в відстані. По-перше, припустимо, що ми організували наші подібності в позитивно визначену матрицю C, де ij-й елемент цієї матриці вказує на схожість i-го і j-го документів. Тоді один із способів перетворити це значення в Евклідову відстань – використовувати наступну формулу [189] (2.15):

$$d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}, \tag{2.15}$$

Якщо два документи співпадають ($c_{ii} = c_{jj}$), то відстань між ними дорівнюватиме нулю.

Простір, в якому знаходяться документи, зазвичай має досить велику розмірність. З огляду на набір документів, поряд з відповідною матрицею відстаней, часто буває цікаво знайти зручний простір меншої розмірності для виконання подальшого аналізу. Це може бути вибрано для полегшення візуалізації, кластеризації та класифікації. Можна сподіватися, що, застосовуючи зменшення розмірності, можна видалити шум з даних і краще застосовувати методи статистичного аналізу даних для виявлення взаємозв'язків, які можуть існувати між документами [187].

Спочатку розглянемо особливо цікаву проекцію (спосіб зменшити кількість вимірювань), яку можна розрахувати безпосередньо з матриці термін-документ. Можна використовувати відому теорему лінійної алгебри, щоб отримати набір корисних проекцій за допомогою розкладання за сингулярними числами. Це стало відомо в інтелектуальному аналізі текстових даних та їх обробки природної мови як LSA (латентно-семантичний аналіз) [187, 193].

Нехай X позначає матрицю дійсних чисел розміром $m \times n$ і ранг дорівнює r, де $m \geq n$ і, отже, $r \leq n$.

Розкладання по сингулярних числах дозволяє нам записати матрицю як добуток трьох матриць [187] (2.16):

$$X = U \times S \times V^T, \tag{2.16}$$

де U – матриця розміру m × n; S – діагональна матриця n × n; V^T – матриця розміру n × n.

Стовпці U називаються лівими сингулярними векторами, $\{u_k\}$, і утворюють ортонормований базис, так що $u_i \cdot u_j = 1$ для $i = j$ і $u_i \cdot u_j = 0$ в іншому випадку. Рядки V^T містять елементи правих сингулярних векторів $\{v_k\}$. Елементи S відмінні від нуля тільки на діагоналі і називаються сингулярними значеннями. Таким чином, $S = diag\ (s_1, ..., s_n)$. Крім того, $s_k > 0$

для $1 \leq k \leq r$ і $s_k = 0$ для $(r + 1) \leq k \leq n$. Зазвичай, порядок сингулярних векторів визначається сортуванням сингулярних значень по спадаючій з найвищим сингулярним значенням у верхньому лівому індексі *S*-матриці [187]. Одним з важливих результатів SVD для матриці *X* є те, що (2.17)

$$X^{(l)} = \sum_{k=1}^{l} u_k \, s_k v_k^T \, , \tag{2.17}$$

– найближча до *X* матриця рангу l. Термін «найближча» означає, що $X^{(l)}$ мінімізує суму квадратів різниці елементів X і $X^{(l)}$ (2.18)

$$\sum_{ij} \left| x_{ij} - x_{ij}^{(l)} \right| \rightarrow min \, , \tag{2.18}$$

Один із способів розрахунку SVD – це спочатку обчислити V^T і S шляхом діагоналізації X^TX (2.19) [187]:

$$X^TX=VS^2V^T \, . \tag{2.19}$$

Потім обчислити U наступним чином (2.20):

$$U=XVS^{-1} \, , \tag{2.20}$$

де $(r + 1)$, ..., n стовпців *V*, для яких $s_k = 0$, ігноруються при матричному множенні. Вибір решти n-r сингулярних векторів в *V* або *U* може бути обчислений з використанням процесу ортогоналізації Грама-Шмідта.

Між PCA (метод головних компонент) і SVD існує прямий зв'язок у тому випадку, коли головні компоненти обчислюються з коваріаційної матриці. Якщо від кожного стовпчика матриці X відняти середнє значення цього стовпчика, то $X^TX = \Sigma_i g_i g_i^T$ пропорційно коваріаційній матриці змінних g_i [187].

Відповідно до рівняння 2.20 діагоналізація X^TX дає V^T, яка також дає головні компоненти $\{g_i\}$. Отже, праві власні вектори $\{v_k\}$ збігаються з головними компонентами $\{g_i\}$. Власні значення X^TX еквівалентні s_k^2, які пропорційні дисперсії головних компонентів.

Застосування SVD в аналізі даних має схожість з аналізом Фур'є. Як і у випадку з SVD, аналіз Фур'є включає розширення вихідних даних по ортогональному базису[187] (2.21):

$$x_{ij} = \sum_k c_{ik}\, e^{i2\pi jk/m} \,. \tag{2.21}$$

Зв'язок з SVD можна явно проілюструвати, нормалізував вектор $\{e^{i2\pi jk/m}\}$ і назвавши його v'$_k$ (2.22):

$$x_{ij} = \sum_k b_{ik}\, v'_{jk} = \sum_k u'_{ik} s'_k v'_{jk} \,, \tag{2.22}$$

яке породжує матричне рівняння (2.23):

$$X=U`S`V`^T \tag{2.23}$$

Це матричне рівняння аналогічно рівнянню 2.15. На відміну від SVD, однак, навіть якщо $\{v'_k\}$ є ортонормованим базисом, $\{u'_k\}$ в цілому не ортогональні.

Сингулярний розклад застосовується по відношенню до частотної матриці, що дозволяє спроектувати документи в простір меншої розмірності. SVD розмірність, що згенерована, за своїм змістом найкраще відображає підпростір термінів за критерієм найменших квадратів.

Таким чином, представлені методи попереднього аналізу даних формують цілісну методику підготовки статистичних даних для подальшого використання при побудові математичних моделей нелінійних

нестаціонарних процесів та використанні цих моделей для оцінювання прогнозів.

Висновки до розділу 2

У розділі розглянуто статистичні параметри (статистики), які використовуються для аналізу зв’язків між змінними, що використовуються у подальшому для побудови моделей досліджуваних процесів. Запропоновано процедуру аналізу досліджуваних процесів на наявність зв’язків різних типів.

Подано процедуру аналізу даних за допомогою діаграм розсіювання, які надають графічну інформацію стосовно величини відхилення даних від лінійності. Незважаючи на спрощення процедури аналізу такий підхід до аналізу надає корисну інформацію стосовно характеристик вимірів, яка може бути використана при оцінюванні типів даних, віднесення їх до конкретного класу та встановлення структури моделей.

Проаналізовано вплив екстремальних значень на статистичні характеристики вибірок даних і вказано на необхідність застосування спеціальних процедур обробки таких даних з метою покращення їх характеристик. Подано статистики, які необхідно використовувати у процедурах аналізу таких даних.

Розглянуто проблему мультиколінеарності, її вплив на обчислювальні характеристики алгоритмів моделювання та засоби боротьби з нею при побудові регресійних моделей.

Подано процедуру аналізу категоріальних даних, яка спрямована на підвищення якості даних і використана для побудови відповідних моделей. Також розглянуто процедури стандартизації даних перед їх використанням для побудови математичних моделей вибраних процесів. В подальшому аналізі даних наведені процедури використано при створенні системи підтримки прийняття рішень.

РОЗДІЛ 3
МЕТОДИКА АНАЛІЗУ ПОДІБНОСТІ ЧАСОВИХ РЯДІВ

3.1 Опис математичного апарату – методу виявлення подібності часових рядів

Одним із ефективних методів аналізу нелінійних нестаціонарних процесів різної природи є метод, оснований на використанні подібності часових рядів, що описують досліджувані процеси. Цей метод надає особливі переваги при вирішенні задач моделювання, що мають неповні або спотворені вхідні дані. У роботі пропонується новий робастний непараметричний метод на основі аналізу нелінійних нестаціонарних процесів за їх подібністю. Суть методу полягає у пошуку часових рядів (процесів) досліджуваного типу зі схожою статистичною поведінкою, що в результаті допомагає зробити висновки про наявність поведінкових шаблонів досліджуваних процесів. Для врахування різної поведінки рядів у часі за допомогою аналізу подібності розраховується відстань між першою часовою послідовністю та другою із урахуванням впорядкування. Зсув часу (ковзання) послідовно зміщує перший часовий ряд уздовж часової шкали з метою пошуку подібних зразків, які можуть виникати в довільні моменти часу. Перетворення часового ряду дозволяє «розтягувати» або «стискати» часовий вимір для випадків незначної відмінності у часі між схожими шаблонами.

Формально постановку задачі виявлення подібності часових рядів можна сформулювати таким чином. Нехай в наявності є два часових ряди Y та X, які математично записуються як послідовності значень $Y=\{y_1,\ldots,y_n\}$ та $X = \{x_1,\ldots, x_m\}$. Y – цільовий ряд значень, що містить n елементів, а X – вхідний ряд значень, що містить m елементів. При описі алгоритмів для послідовного перебору всіх значень цільового ряду Y використовується індекс j який змінюється від 1 до n, а для перебору всіх значень вхідного ряду X використовується індекс i який змінюється від 1 до m. В загальному

випадку вводиться деякий функціонал d_{ij}: $d_{ij} = d(x_i, y_j)$, що відображає відстань між двома точками x_i,та y_j.

Пошук часових рядів зі схожими статистичними поведінками може допомогти зробити висновки про наявність поведінкових шаблонів, що мають низку різноманітних застосувань, таких як:

– виявлення подібності матеріалів, захищених авторським правом;

– пошук та виявлення фінансових активів, що мають схожу поведінку протягом часу;

– ідентифікація поведінки шахраїв в фінансових операціях;

– розвиток епідеміологічної ситуації на основі захворювань, що відбувалися в минулі роки;

– схожі поведінки покупців протягом року, на вихідних та у періоди свят;

– типові проблеми із поставниками послуг, комплектуючих та ресурсів для виробництва.

Методи аналізу подібності часових рядів дозволяють порівняти два часові ряди з різною тривалістю, використовуючи метод динамічного перетворення, а також усереднення (згладжування) значень. Наприклад, сезонний цикл для фінансового ряду, що описує акцію компанії А може мати довжину хвилі у 5 місяців, тоді інший схожий фінансовий актив компанії Б може мати довжину сезонного циклу у 6 місяців. Спеціалізовані методи та підходи щодо перетворення часового ряду за часовою шкалою дозволяють розширювати та стискати цикли з метою виявлення схожих форм сезонного циклу, навіть при наявності незначних відмінностей. Загальний опис методу динамічного перетворення часового ряду в часі можна знайти в роботі [99].

Нехай в наявності є два часових ряди Y та X, які формально математично записуються як послідовності значень $Y = \{y_1, \dots, y_n\}$ та $X = \{x_1, \dots, x_m\}$. Y – цільовий ряд значень, що містить n елементів, а X – вхідний ряд значень, що містить m елементів.

При описі алгоритмів для послідовного перебору всіх значень цільового ряду Y використовується індекс j який змінюється від 1 до n, а для

перебору всіх значень вхідного ряду X використовується індекс i який змінюється від 1 до m.

В загальному випадку вводиться деякий функціонал d_{ij} що відображає відстань між двома точками x_i та y_j (3.1):

$$d_{ij} = d(x_i, y_j) \tag{3.1}$$

де x_i та y_j - відповідні і-е та j-е значення цільового та вхідного рядів даних.

Для обчислення відстані між часовими рядами для їх порівняння, з метою оцінювання подібності, можуть використовуватися різні формули. Найбільш відомими є модуль різниці, що зазвичай використовується в задачах бізнес-аналітики, та квадрат різниці відхилення, який використовується в задачах статистики.

Формула модуля різниці (Манхеттенська відстань або міських кварталів) (3.2) [188]:

$$d(x_i, y_j) = |x_i - y_j| \tag{3.2}$$

де x_i та y_j - відповідні і-е та j-е значення цільового та вхідного рядів даних.

Формула квадрата різниці (Евклідова міра) (3.3) [188]:

$$d(x_i, y_j) = (x_i - y_j)^2 \tag{3.3}$$

де x_i та y_j- відповідні і-е та j-е значення цільового та вхідного рядів даних.

Окрім наведених мір щодо обчислення відстані, абсолютної різниці та квадрату різниці, також можуть використовуватися будь-які інші, як то варіанти мір Мінковського (3.4) [99]:

$$d(x_i, y_j) = |x_i - y_j|^p \quad (3.4)$$

або

$$d(x_i, y_j) = |x_i - y_j|^{1/p}$$

де x_i та y_j - відповідні і-е та j-е значення цільового та вхідного рядів даних; , $p \geq 0$

Нормування даних бажано виконувати при проведені порівнянні часових рядів, в першу чергу з причини несумісність одиниць вимірювань змінних, що суттєво впливає на кінцевий результат. Наприклад, індекс споживчих цін у відсотках та валовий регіональний продукт у мільярдах гривень мають числові значення, що відрізняються на декілька порядків, тому їх нормують. Існує багато різних підходів до нормування даних, наведемо два найбільш популярні: за діапазоном та стандартне.

Нехай в наявності є вхідний вектор даних $X = \{x_1, \dots, x_m\}$

Перший підхід – нормування за діапазоном. Формула для обчислення нормованих значень має вигляд (3.5) [99]:

$$z_i = \frac{x_i - \min\limits_{i=1,..,m}\{x_i\}}{\max\limits_{i=1,..,m}\{x_i\} - \min\limits_{i=1,..,m}\{x_i\}} \quad (3.5)$$

де $\max_{i=1,..,m}\{x_i\}$ – максимальне значення, серед всіх значень X, а $\min_{i=1,..,m}\{x_i\}$ – відповідно мінімальне.

Як наслідок, область допустимих значень для $z_i \in [0; 1]$.

Відповідно для відображення результатів аналізу на схожість рядів та побудови графіків в кінці роботи методу та алгоритму комп'ютерної програми виконується зворотне перетворення за формулою (3.6) [99]:

$$x_i = z_i \cdot \left(\max_{i=1,..,m} \{x_i\} - \min_{i=1,..,m} \{x_i\} \right) + \min_{i=1,..,m} \{x_i\} \quad (3.6)$$

де $\max_{i=1,..,m}\{x_i\}$ – максимальне значення, серед всіх значень X, а $\min_{i=1,..,m}\{x_i\}$ – відповідно мінімальне.

Другий підхід – стандартизоване нормування. Використовується формула (3.7) [99]:

$$z_i = \frac{x_i - \bar{X}}{\sigma_X} \quad (3.7)$$

де $\bar{X}$ – математичне сподівання вектору даних X, а σ_X – стандартне відхилення.

Відповідно зворотне перетворення використовує формулу (3.8):

$$x_i = z_i \cdot \sigma_X + \bar{X} \quad (3.8)$$

де z_i – оцінка, σ_X – стандартне відхилення, $\bar{X}$ – математичне сподівання вектору даних X,

В реалізованому алгоритмі комп'ютерної програми [260] користувач може обрати один з трьох режимів роботи з даними: залишити дані без перетворень (не робити нормалізації); виконати нормалізацію за діапазоном; виконати стандартизовану нормалізацію.

Окрім нормування даних, з метою зменшення області допустимих значень як цільового, так і вхідного рядів даних, можуть бути використані інші різноманітні аналітичні підходи. Наприклад, такі як:

- логарифмування даних, натуральний логарифм (3.9):

$$w_i = ln(x_i) \quad (3.9)$$

- або десятковий (3.10)

$$w_i = log(x_i) \tag{3.10}$$

- квадратний корінь (3.11)

$$\mathrm{w_i} = \sqrt{\mathrm{x_i}} \tag{3.11}$$

- логістичне перетворення (3.12):

$$w_i = ln\left(\frac{\exp(x_i)}{1 + \exp(x_i)}\right) \tag{3.12}$$

- або перетворення Бокса-Кокса (Box-Cox) [124](3.13)

$$w_i = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda}, if\ \lambda \neq 0 \\ ln(x_i), if\ \lambda = 0 \end{cases} \tag{3.13}$$

де λ – параметр перетворення.

Позначимо як D матрицю відстаней, що має розмір $n \times m$, n рядків на m стовпчиків, кожен елемент матриці – це значення міри подібності d_{ji} між відповідними точками y_j та x_i. При чому $j -$ номер рядка (значення цільового часового ряду), а $i -$ номер стовпчика (значення вхідного часового ряду).

Матриця відстаней D містить усі можливі між елементами цільового та вхідного часових рядів попарні комбінації. Для аналізу подібності рядів цікава уся множина шляхів з матриці відстаней D, що проходять від елементу $D[1,1]$ до елементу $D[n,m]$. Зауважимо, що в цьому документі, для позначення номеру елементу матриці задається наступний формат (номер строки, номер стовпчика). Це обумовлено тим, що практична реалізація

алгоритму обчислення подібності, робилася на мові програмування SAS IML [99], яка підтримує саме такий синтаксис роботи з матрицями.

В табл. 3.1 наведено схематично матрицю відстаней D у загальному вигляді. Для наочності перший стовпчик матриці в табл. 3.1 містить значення цільового ряду, а перша строчка значення вхідного ряду. В подальшому будуть зустрічатися позначення для елементів матриці відстаней як у загально-математичному вигляді $d_{ji} = d(y_j, x_i)$, так і матрично-алгоритмічному $D[j,i] = d(y_j, x_i)$.

Таблиця 3.1 – Загальний вигляд матриці відстаней

	x_1	x_2		x_m
y_1	$d_{11} = d(y_1, x_1)$	$d_{12} = d(y_1, x_2)$	…	$d_{1m} = d(y_1, x_m)$
y_2	$d_{21} = d(y_2, x_1)$	$d_{22} = d(y_2, x_2)$	…	$d_{2m} = d(y_2, x_m)$
..				
y_n	$d_{n1} = d(y_n, x_1)$	$d_{n2} = d(y_n, x_2)$	…	$d_{nm} = d(y_n, x_m)$

Значення норми матриці відстаней D, може бути обчислене за різними формулами, найчастіше використовують звичайну суму всіх елементів, середнє значення за всіма елементами, мінімальне значення.

Звичайна сума всіх елементів розраховується за формулою (3.14):

$$Norm(D) = \sum_{j=1}^{n} \sum_{i=1}^{m} d_{ji} \tag{3.14}$$

де n – кількість рядків матриці відстаней, m – кількість стовпців матриці відстаней, i- номер стовця (значення вхідного ряду), j- номер рядка (значення цільового часового ряду), d_{ij}- значення міри подібності між відповідними точками.

Середнє значення за всіма елементами (3.15)

$$Norm(D) = \frac{1}{n \cdot m} \cdot \sum_{j=1}^{n} \sum_{i=1}^{m} d_{ji} \tag{3.15}$$

де де n – кількість рядків матриці відстаней, m – кількість стовпців матриці відстаней, і- номер стовця (значення вхідного ряду), j- номер рядка (значення цільового часового ряду), d_{ij}- значення міри подібності між відповідними точками.

Мінімальне значення (3.16)

$$Norm(D) = min_{ji}\{d_{ji}\} \tag{3.16}$$

де де n – кількість рядків матриці відстаней, m – кількість стовпців матриці відстаней, і- номер стовпця (значення вхідного ряду), j- номер рядка (значення цільового часового ряду), d_{ij}- значення міри подібності між відповідними точками.

Однак, на практиці для аналізу саме часових рядів ці формули не використовуються тому, що не враховують упорядкування значень часових рядів, що порівнюються.

Саме тому при роботі з часовими рядами, що мають часову шкалу, порівнюють різні шляхи, кожен елемент якого має послідовне хронологічне розміщення.

Строго формально-математично, шлях в матриці відстаней D – це переміщення по елементах матриці відстаней, починаючи з елементу $D[1,1]$ та закінчуючи елементом $D[n,m]$.

При цьому, переміщення по шляху повинно зберігати впорядкованість часової послідовності, а саме:

- неможна двічі використовувати один і той саме елемент матриці відстаней;

- недопустимо переміщення за шляхом у зворотному напрямі по відношенню до впорядкованої часової послідовності.

Введемо позначення П – це множина всіх можливих шляхів що існують для матриці відстаней D. В свою чергу π – це деякий шлях із множини усіх можливих, тобто $\pi \in$ П [3,4].

Найпростіший та очевидний шлях – прокладений по діагоналі матриці D, як показано на рис. 3.1а, але цього для вирішення задачі з аналізу подібності часових рядів не достатньо.

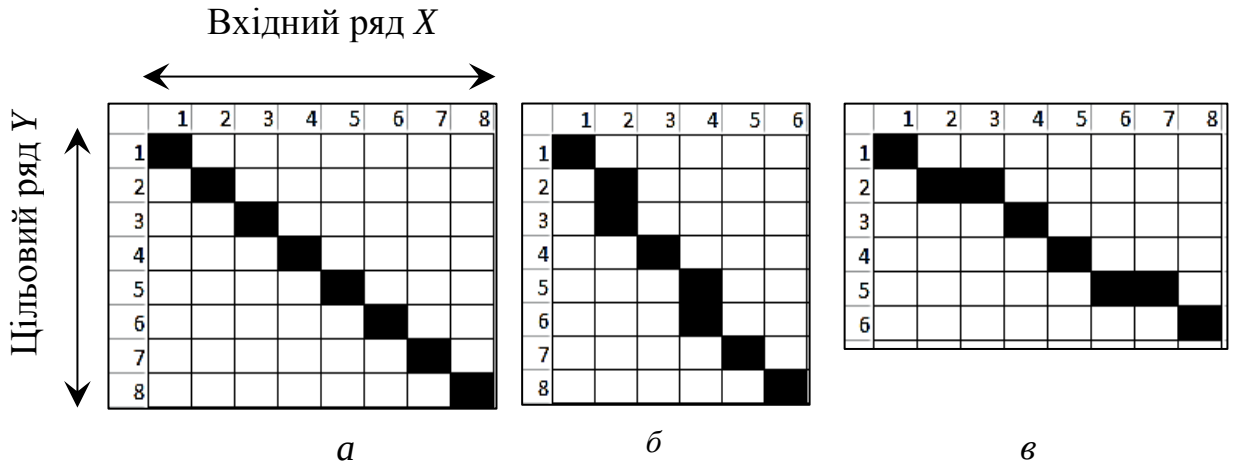


Рисунок 3.1 Приклади шляхів для матриці відстаней. ($а$) – випадок діагонального шляху, коли $n = m$. ($б$) – випадок наявності двох стиснень між елементами (2,2) і (3,2) та (5,4) і (6,4), зазвичай виникає коли $n > m$. ($в$) – випадок наявності двох розширень між елементами (2,2) і (2,3) та (5,6) і (5,7), зазвичай виникає коли $n < m$

Найкращим шляхом визнається той, який має найменше значенням функції вартості шляху, яка зазвичай обчислюється як сума вартостей всіх елементів з матриці відстаней, з яких складається шлях π, або середнє значення, тобто зазначена сума вартостей всіх елементів поділена на довжину шляху π, де довжина – це кількість елементів з яких складається шлях. В задачах оптимізації та комбінаторного аналізу значення матриці D відображають вартість переміщення за шляхом, тобто матриця відстаней

містить ціну за переміщення між відповідними елементами. В задачі аналізу подібності часових рядів, вартість – це відстань між значеннями цільового та вхідного часових рядів. Чим менша різниця між значеннями елементів рядів – тим менша вартість і навпаки.

Якщо деякі проміжки шляху прокладений по вертикалі, як показано на рис. 3.1.б, це інтерпретується як стиснення послідовності елементів цільового часового ряду. У випадках, коли деякі проміжки шляху прокладений по горизонталі, як показано на рис. 3.1.в, то це є розширенням послідовності елементів цільового часового ряду.

При побудові шляху в матриці відстаней зазвичай задають обмеження на розширення та стиснення цільового ряду.

RL – обмеження на стиснення цільового ряду (скорочення від анг. row limit). Зазвичай спостерігається, коли $n > m$. Використовується саме термін “стиснення” (compression) послідовності елементів цільового часового ряду, тому що у цьому випадку декілька значень цільової послідовності відповідає одному значенню вхідної, що візуально на графіку виглядає як скорочення цільового ряду або його стиснення (рис. 3.1 б).

CL – обмеження на розширення цільового ряду (скорочення від анг. column limit). Зазвичай спостерігається, коли $n < m$. Використовується саме термін “розширення” (expansion) послідовності елементів цільового часового ряду, тому що у цьому випадку одному значенню цільової послідовності відповідає декілька вхідної, що візуально на графіку виглядає як розтягування цільового ряду або його розширення (рис. 3.1в).

В залежності від налаштувань параметрів обмежень RL та CL виконується пошук оптимального шляху серед допустимих значень матриці відстаней.

За замовчанням, якщо не задані обмеження RL та CL, вони можуть визначатися за наступними правилами:

- якщо $n > m$ то $RL = n - m$ та $CL = 0$.
- якщо $n < m$ то $RL = 0$ та $CL = m - n$.

Для деякого шляху $\pi \in \Pi$ вводиться функція шляху, яка позначається як $Path_\pi$ (3.17).

$$Path_\pi = \{Path_\pi(p) : p = 1, \dots, P\} = \{Path_\pi(1), \dots, Path_\pi(P)\} \quad (3.17)$$

де аргумент p – це індекс функції шляху, який в загальному вигляді може приймати значення $p = 1, \dots, P$. P – це довжина шляху в матриці відстаней D, тобто кількість переходів між початковим та кінцевим елементами шляху π.

Діапазон допустимих значень довжина шляху визначається наступним інтервалом (3.18) [99]:

$$max(n, m) \leq P \leq (n + m - 1). \quad (3.18)$$

В свою чергу функцію шляху може бути представлена у вигляді множини пар компонент, цей вигляд зручний саме для реалізації алгоритмів комп'ютерних програм (3.19):

$$Path_\pi = \{(j_1, i_1), \dots, (j_p, i_p)\} \quad (3.19)$$

де $j_p \in \{1, \dots, n\}$ – індекс цільової послідовності часового ряду, а $i_p \in \{1, \dots, m\}$ – індекс вхідної послідовності часового ряду.

Для того, щоб функція шляху відповідала властивостям часового ряду, повинні виконуватися наступні умови (3.20) [99, 194]:

$$\text{Якщо } Path_\pi(p) = (j_p, i_p) \text{ та } Path_\pi(p+1) = (j_{p+1}, i_{p+1}), \text{ то } (j_{p+1} - j_p) \in \{0,1\}, (i_{p+1} - i_p) \in \{0,1\}. \quad (3.20)$$

де $j_p \in \{1, \dots, n\}$ – індекс цільової послідовності часового ряду, а $i_p \in \{1, \dots, m\}$ – індекс вхідної послідовності часового ряду.

Для функції шляху доступні наступні операції переміщення та обмеження.

Прямий перехід виконується, коли є переміщення по діагоналі матриці відстаней D (3.21) [99, 194]:

$$\left(j_{p+1} - j_p\right) = 1 \; ma \; \left(i_{p+1} - i_p\right) = 1 \tag{3.21}$$
$$Path_\pi(p+1) - Path_\pi(p) = (1,1)$$

Операція стиснення виконується, коли є вертикальне переміщення (номер строки збільшився на одиницю, а номер стовпчика залишився тим самим) (3.22):

$$\left(j_{p+1} - j_p\right) = 1 \text{ та } \left(i_{p+1} - i_p\right) = 0 \tag{3.22}$$
$$Path_\pi(p+1) - Path_\pi(p) = (1,0)$$

Операція розширення виконується, коли є горизонтальне переміщення (номер рядка залишився тим самим, а номер стовпчика збільшився на одиницю) (3.23):

$$\left(j_{p+1} - j_p\right) = 0 \text{ та } \left(i_{p+1} - i_p\right) = 1 \tag{3.23}$$
$$Path_\pi(p+1) - Path_\pi(p) = (0,1)$$

Недопустимий перехід виникає коли немає переміщення (залишилися на тому самому місці)(3.24):

$$\left(j_{p+1} - j_p\right) = 0 \text{ та } \left(i_{p+1} - i_p\right) = 0 \tag{3.24}$$
$$Path_\pi(p+1) - Path_\pi(p) = (0,0)$$

або з’являється ситуація переміщення в обернений напрям по відношенню до часової шкали (3.25):

$$\left(j_{p+1} - j_p\right) \notin \{0,1\} \text{ або } \left(i_{p+1} - i_p\right) \notin \{0,1\} \quad (3.25)$$

В загальному випадку, розширення послідовності цільового часового ряду – це теж саме, що і стиснення послідовності вхідного часового ряду, і навпаки, все залежить від того який ряд з яким порівнюється. Так саме стиснення послідовності цільового часового ряду – це теж саме, що і розширення послідовності вхідного часового ряду, і навпаки.

Зазвичай стиснення спостерігається коли довжина послідовності вхідного часового ряду менша за цільову, тобто $n > m$. як показано у прикладі на рис. 3.1б.

Область допустимих значень стиснення задається інтервалом [99] (3.26):

$$max\left(0, (n-m)\right) \leq RL \leq (n-1) \quad (3.26)$$

де $RL = 0$ – це відсутність стиснення, $RL = max\left(0, (n-m)\right)$ – мінімально допустимий розмір стиснення, а $RL = (n-1)$ – це максимально допустимий розмір стиснення.

В загальному випадку обмеження на стиснення можуть бути виражені рівнянням, що представляє кусочно-лінійні та нелінійні обмеження [99] (3.27):

$$(j - 1 - RL) = \frac{n-1}{m-1} \cdot (i-1) \quad (3.27)$$

де $\frac{n-1}{m-1}$ – це нахил лінії стиснення, для випадку геометричної інтерпретації, а RL враховується як частина зміщення лінії стиснення.

Описані лінійне обмеження можуть бути представлені у вигляді одного з наступних рівнянь [99, 194] (3.28):

$$(j-1-RL) \le \frac{n-1}{m-1}(i-1)$$
$$(j-1-RL)\cdot(m-1) \le (i-1)\cdot(n-1) \quad (3.28)$$
$$\frac{(j-1-RL)}{(i-1)} \le \frac{(n-1)}{(m-1)}$$

Якщо $i = 1$, то $1 \le j \le (RL+1)$. При цьому, якщо $RL = 0$, то $j = 1$, а у випадку коли $RL = n-1$, то $1 \le j \le n$.

Якщо $j = n$, то $i > \frac{n-1}{m-1}(n-1-RL)$. При цьому, якщо $RL = 0$, то $i = m$, а у випадку коли $RL = n-1$, то $1 \le i \le m$.

Зазвичай розширення відбувається, коли довжина послідовності вхідного часового ряду більша за цільову, тобто $n < m$, як показано на рис. 3.1в.

Область допустимих значень розширення відповідає інтервалу [99] (3.29):

$$max\big(0,(m-n)\big) \le CL \le (m-1) \quad (3.29)$$

де $CL = 0$ – це відсутність розширення, $CL = max\big(0,(m-n)\big)$ – мінімально допустимий розмір розширення, а $CL = (m-1)$ – це максимально допустимий розмір розширення.

В загальному випадку обмеження на розширення можуть бути виражені рівнянням, що представляє кусочно-лінійні та нелілійні обмеження [99, 194] (3.30):

$$(j-1) = \frac{n-1}{m-1}\cdot(i-1-CL) \quad (3.30)$$

де $\frac{n-1}{m-1}$ – це нахил лінії розширення, для випадку геометричної інтерпретації, а CL враховується як частина зміщення лінії розширення.

Відповідне лінійне обмеження може бути представлене у вигляді одного з наступних рівнянь [99] (3.31):

$$\begin{gathered}(j-1) \geq \frac{n-1}{m-1} \cdot (i-1-CL) \\ (i-1-CL)(n-1) \leq (j-1)(m-1) \\ \frac{n-1}{m-1} \leq \frac{j-1}{i-1-CL}\end{gathered} \tag{3.31}$$

Якщо $j = 1$ то $1 \leq i \leq (CL+1)$. При цьому, якщо $CL = 0$ то $i = 1$, а у випадку, коли $CL = m-1$, то $1 \leq i \leq m$.

Якщо $i = m$ то $j > \frac{n-1}{m-1} \cdot (m-1-CL)$. При цьому, якщо $CL = 0$, то $j = n$, а у випадку, коли $CL = (m-1)$, то $1 \leq j \leq n$.

Наведені рівняння (3.27) та (3.30) можна використати для побудови візуально графічної інтерпретації, як показано на рис. 3.2 [99].

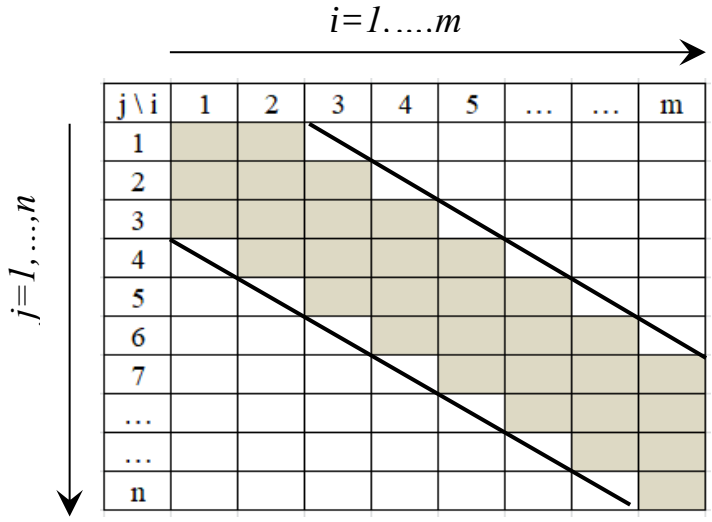


Рисунок 3.2 – Схематичне зображення обмежень на розширення та стиснення цільового ряду по відношенню до вхідного у вигляді матриці відстаней

Нелінійні обмеження на розширення та стиснення (можуть бути записані у вигляді сумісної комбінації нелінійних обмежень [99] (3.32-3.34):

- обмеження на стиснення:

$$(j-1-RL)(m-1) \leq (i-1)(n-1) \tag{3.32}$$

- обмеження на розширення (3.33):

$$(i-1-CL)(n-1) \leq (j-1)(m-1) \tag{3.33}$$

якщо $\mathrm{n} = \mathrm{m}$, то обмеження на перетворення спрощуються до

$$-\mathrm{RL} \leq (\mathrm{i}-\mathrm{j}) \leq \mathrm{CL} \tag{3.34}$$

Для часткових випадків виконуються наступні вирази [99]:

- якщо $m = 1$, то $i = 1$ та $1 \leq j \leq n$,
- якщо $n = 1$, то $j = 1$ та $1 \leq i \leq m$.

Для наведених обмежень на перетворення, в матриці відстаней D задаються значення за правилами [99]:

- якщо (j, i) належить допустимим обмеженням на перетворення – стиснення та розширення, що задаються формулами (5)-(7), то $d_{ji} = d(y_j, x_i)$;
- якщо (j, i) виходить за допустимі обмеження на перетворення, що задаються формулами (5)-(7), то $d_{ji} = \infty$.

Тобто в цьому випадку фактично задається значення штрафу, що дорівнює нескінченності, для того щоб при виборі наступного елементу шляху, відповідний елемент не був обраний алгоритмом, тому що це збільшить сумарне значення вартості шляху у нескінченість.

3.2 Алгоритм побудови шляху – переміщення по матриці відстаней

Практична реалізація алгоритму враховує ситуації: коли $n > m$ та $n < m$. У першому випадку повинно виконуватися $RL \geq (n-m)$, а у

другому $CL \geq (m-n)$. Якщо не ввести ці обмеження, то на практиці виникають некоректні ситуації, як показано на рис. 3.2 та 3.3:

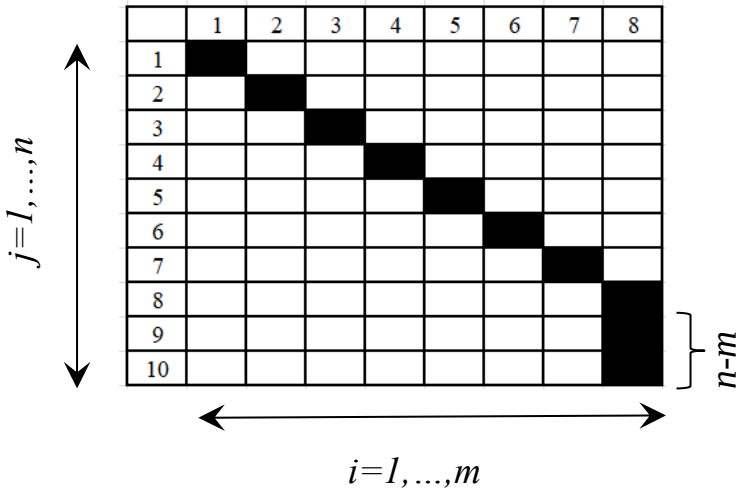


Рисунок 3.3 – Випадок $n > m$ оптимального шляху складається з діагональних елементів та останні (n-m) елементів m стовпчика матриці відстаней

Якщо в матриці відстаней D, кількість рядків більше, ніж кількість стовпчиків, тобто $n > m$, тоді в якості оптимального шляху обираються всі діагональні елементи матриці D та останні $(n-m)$ елементів m стовпчика матриці відстаней $D[m+1, m], \ldots, D[n, m]$ (рис. 3.4).

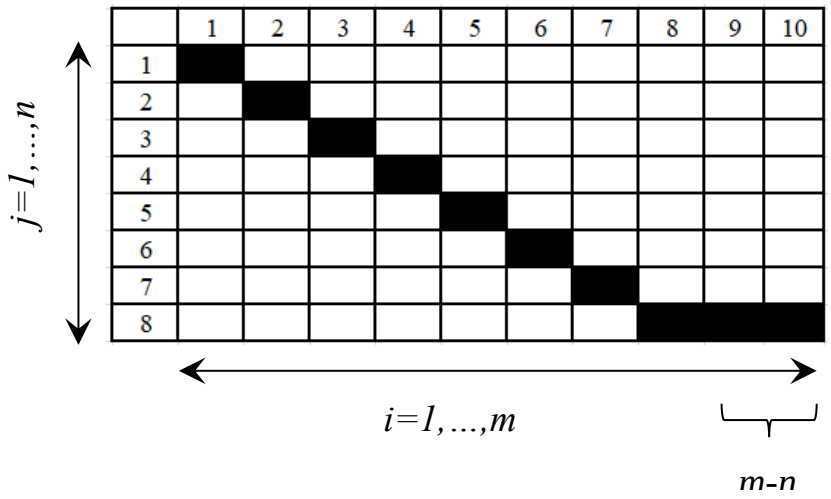


Рисунок 3.4 – Випадок $n < m$ оптимального шляху складається з діагональних елементів та останніх (m – n) елементів n рядка матриці відстаней

Якщо в матриці відстаней D, кількість строк менша кількості стовпчиків, тобто $n < m$, то в якості оптимального шляху обираються всі діагональні елементи матриці D та останні $(m - n)$ елементів n рядка матриці відстаней $D[n, n+1], \ldots, D[n, m]$ (рис. 3.3).

Схематично на рис. 3.5 наведені ітерації алгоритму, в якості обмежень на перетворення задані параметри $RL = 2$ та $CL = 1$. Сірим кольором позначені допустимі для побудови шляху елементи, побудований шлях позначений чорним кольором.

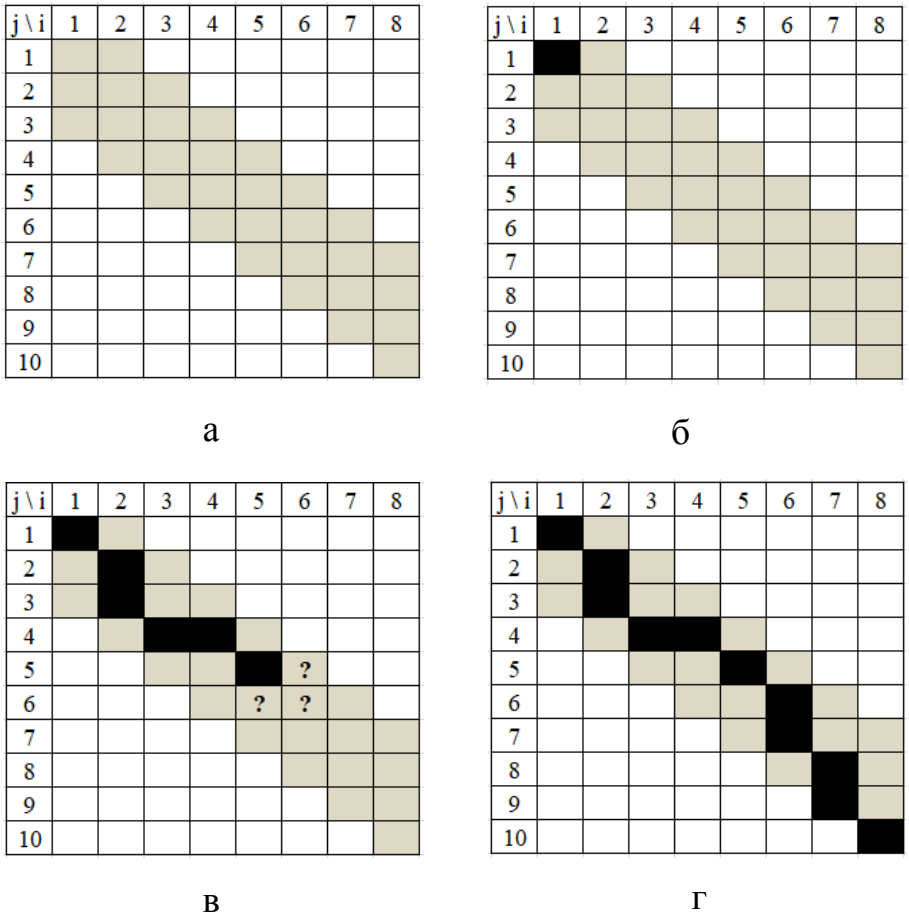


Рисунок 3.5 – Графічне представлення роботи алгоритму побудови шляху для матриці відстаней розміром 10 рядків та 8 стовпчиків.

На рисунку 3.5а зображені початкові значення матриці відстаней, з урахуванням обмежень на перетворення $RL = 2$ та $RC = 1$, допустимі для

аналізу елементи зображені сірим кольором. На рисунку 3.5б – перша ітерація алгоритму. Рисунок 3.4в відображує ітерацію алгоритму номер p, знаками питань позначені елементи – множина кандидатів для побудови шляху. Остання ітерація алгоритму представлена на рис. 3.5 г.

Алгоритм побудови шляху починається з елементу $\mathrm{D}[1,1]$, як показано на рис. 3.5 б. Ітерації виконуються послідовно, як це представлено на рисунку. Вирізняється передостання ітерація, коли на основі елементу отриманого на попередній p - 1 ітерації визначається множина кандидатів які можуть бути використані для побудови шляху на поточній ітерації алгоритму.

Якщо, на попередній $(p-1)$ ітерації, був обраний елемент $D[j,i]$, то множина кандидатів складається з елементів: $D[j+1,i]$, $D[j,i+1]$ та $D[j+1,i+1]$. Схематично це показано на рис. 3.4 в. При чому, кожен з елементів повинен задовольняти обмеженням представленим у вигляді формули (3.34).

В результаті обирається такий елемент (j_p, i_p) матриці D що $D[j_p, i_p] = min\{ D[j+1,i],\ D[j,i+1],\ D[j+1,i+1] \}$. Обраний елемент стає p - ланцюгом шляху π в матриці відстаней, тобто $Path_\pi(p) = (j_p, i_p)$.

Ітерація p повторюється доки не буде обрано останній елемент матриці відстаней $D[n,m]$.

На останній ітерації алгоритму повинен бути приєднаний останні елемент матриці відстаней $D[n,m]$, тобто $Path_\pi(P) = (n,m)$, p – останній елемент шляху (рис. 3.5 г).

3.3. Метрики оцінки якості побудованого шляху

Для оцінки якості побудованого шляху можна використовувати різні метрики та підходи або їхні комбінації. Нехай. є побудований шлях $\pi \in \Pi$, довжини P, що містить наступні значення (3.35) [99]:

$$Path_\pi = \{(j_1, i_1), \dots, (j_p, i_p)\} \quad (3.35)$$

де $p = 1, \dots, P$; $j_p \in \{1, \dots, n\}$ – індекс цільової послідовності часового ряду, а $i_p \in \{1, \dots, m\}$ – індекс вхідної послідовності часового ряду.

Значення функції вартості шляху (змінну позначено Score - в перекладі з англійської позначає виграш або бали) [99].

В базовому варіанті вартість шляху обчислюється за формулою (3.36):

$$Score = \sum_{p=1}^{P} d(j_p, i_p) \quad (3.36)$$

де $d(j_p, i_p)$ – значення елементу $D[j_p, i_p]$ матриці відстаней D.

Чим менше значення функції вартості шляху Score, тим краще.

На основі формули (3.36) можуть бути побудовані різноманітні модифікації розрахунку метрики. Наприклад, обчислення значення середньої вартості, яка за своєю суттю показує середню вартість одного переходу по шляху між елементами матриці відстаней (3.37) [99]:

$$ScoreAvg = \frac{Score}{P} = \frac{1}{P} \cdot \sum_{p=1}^{P} d(j_p, i_p) \quad (3.37)$$

де $d(j_p, i_p)$ – значення елементу $D[j_p, i_p]$ матриці відстаней D; P – довжина шляху; $Score$ – значення функції вартості шляху.

Слід зазначити, що чим меншою є довжина шляху P, тим краще.

На основі кількості операцій стиснень обчислюються статистики [99]:

- CN – Compression Number, кількість операцій стиснень (3.38):

$$CN = \left\| \{Path(p+1) - Path(p) = (0,1)\}_{p=1}^{P-1} \right\| \quad (3.38)$$

- процент співвідношення кількості стиснень до довжини шляху (3.39)

$$100\% \cdot \frac{CN}{P} \quad (3.39)$$

- процент співвідношення кількості стиснень до кількості елементів вхідної послідовності (3.40):

$$100\% \cdot \frac{CN}{m} \quad (3.40)$$

- процент співвідношення кількості стиснень до кількості елементів цільової послідовності (3.41):

$$100\% \cdot \frac{CN}{n} \quad (3.41)$$

- максимальна довжина стиснення (3.42):

$$\begin{aligned} C_{max} = \max_{m} \{m: Path(p) - (0,0) = \cdots \\ = Path(p+m) - (0,m), \quad 1 \le p \le P\} \end{aligned} \quad (3.42)$$

або максимальна кількість суміжних горизонтальних стиснень по матриці відстаней, що містить шлях.

Наступна група статистик обчислюється на основі кількості операцій розширень [99]:

- EN – Expansion Number, кількість операцій розширень(3.43):

$$EN = \left\| \{Path(p+1) - Path(p) = (1,0)\}_{p=1}^{P-1} \right\| \quad (3.43)$$

- процент співвідношення кількості розширень до довжини шляху (3.44):

$$100\% \cdot \frac{EN}{P}; \quad (3.44)$$

- процент співвідношення кількості розширень до кількості елементів вхідної послідовності (3.45):

$$100\% \cdot \frac{EN}{m} \quad (3.45)$$

- процент співвідношення кількості розширень до кількості елементів цільової послідовності (3.46):

$$100\% \cdot \frac{EN}{n} \quad (3.46)$$

- максимальна довжина розширення (3.47):

$$E_{max} = \max_{m}\{m: Path(p) - (0,0) = \cdots = Path(p+m) - (m,0), \; 1 \le p \le P\} \quad (3.47)$$

або максимальна кількість суміжних горизонтальних переміщень по матриці відстаней, що містить шлях.

На основі кількості операцій прямих переходів без розширень та стиснень розраховується DN (Diagonal Number), кількість прямих переходів без розширень та стиснень (3.48):

$$DN = \left\| \{Path(p+1) - Path(p) = (1,1)\}_{p=1}^{P-1} \right\| \quad (3.48)$$

або можна використовувати формулу:

$$DN = P - CN - EN$$

Відсоток співвідношення кількості прямих переходів до довжини шляху вираховується із $100\% \cdot \frac{DN}{P}$

Відсоток співвідношення кількості прямих переходів до кількості елементів вхідної послідовності - $100\% \cdot \frac{DN}{m}$

Відсоток співвідношення кількості прямих переходів до кількості елементів цільової послідовності – за формулою: $100\% \cdot \frac{DN}{n}$

Загальна кількість перетворень, обчислюється як сума кількості стиснень та розширень (3.49):

$$Warps = CN + EN \quad (3.49)$$

або можна використати формулу:

$$Warps = P - DN$$

Відповідні статистики шляху дозволяють оцінити кількість перетворень, які необхідно здійснити для переміщення по побудованому шляху в матриці відстаней. Окрім цього, подібні статистики можуть вказувати на різноманітні аномалії, наявні у вхідних даних.

На основі досвіду розв’язання практичних завдань, автором запропоновано власну метрику (*ScoreSim*), призначену для оцінювання ступеню близькості часових рядів Y та X.

Загальна постановка задачі розрахунку метрики для оцінювання ступеню близькості часових рядів має наступний вигляд. Нехай, для цільового ряду значень $Y = \{y_1, \dots, y_n\}$ формується додатковий вхідний ряд Y_0, що містить обернені (помножені на мінус одиницю) значення Y, тобто $Y_0 = \{-y_1, \dots, -y_n\}$.

Реалізація алгоритму відбувається наступним чином:

Крок 1. Побудувати матрицю відстаней D_0, розмірністю $n \times n$, на основі цільоового Y та вхідного Y_0 рядів даних;

Крок 2. Задатия обмеження $RL = 0$ та $CL = 0$,

Крок 3. Побудувати оптимальний для D_0 шлях π_0, який буде мати діагональний вигляд $\pi_0 = \{(1,1), \dots, (n,n)\}$,

Крок 4. Обчислити значення фнкції вартості шляху

$$ScoreAvg_0 = ScoreAvg(\pi_0) = \frac{1}{n} \cdot \sum_{j=1}^{n} D_0[j,j].$$

На основі розрахованого значення $ScoreAvg_0$, яке буде використовуватися в якості значення-нівеліру (значення, що описує максимально несхожі між собою ряди).

Наступний етап передбачає опрацювання цільового ряду Y та вхідного X:

Крок 1. Побудувати матрицю відстаней D_1,

Крок 2. Врахувати обмеження RL та CL,

Крок 3. Побудувати оптимальний для D_1 шлях π_1,

Крок 4. Обчислити значення фнкції вартості шляху:

$$ScoreAvg_1 = ScoreAvg(\pi_1).$$

На основі отриманих значень функцій вартостей обчислюється коефіцієнт подібності, запропонований автором (3.50):

$$ScoreSim(Y,X) = 100\% \cdot \left(1 - \frac{ScoreAvg_1}{ScoreAvg_0}\right) =$$

$$= 100\% - 100\% \times$$

$$\times \frac{\left(\frac{1}{\left(max(n,m) - \left(P_1 - max(n,m)\right)\right)} \cdot \sum_{k=1}^{P_1} D_1[j_p, i_p]\right)}{\left(\frac{1}{n} \cdot \sum_{k=1}^{P_0} D_0[j,j]\right)} \quad (3.50)$$

де P_0 – довжина шляху π_0, P_1 – довжина шляху π_1, $max(n,m)$ – повертає максимальне серед n та m значення.

За своєю суттю коефіцієнт $ScoreSim(Y,X)$ відображає ступінь схожості між часовими рядами Y та X виконуючи порівняння значення середньої вартості переходу по матриці відстаней між рядами Y та X, та середньої вартості максимально несхожих рядів Y та Y_0. Чим більше ряди Y та X несхожі один на одного, тим ближче відношення $\frac{ScoreAvg_1}{ScoreAvg_0}$ буде до одиниці. Тому, в (3.49) вираз $1 - \frac{ScoreAvg_1}{ScoreAvg_0}$, що дозволяє отримати значення, яке буде сходитися до нуля у випадку несхожості рядів Y та X, а у випадку максимальної близькості, збігатиметься до одиниці.

Вираз $\left(max(n,m) - \left(P_1 - max(n,m)\right)\right)$, використовується як значення штрафу шляху, побудованого по матриці D_1, має високу складність. А саме: чим довший шлях P_1, тим менше буде значення $\left(max(n,m) - \left(P_1 - max(n,m)\right)\right)$, що призведе до збільшення значення функції $ScoreAvg_1$. В той час як кращими вважаються шляхи, які мають менші значення функції $ScoreAvg_1$.

Зазначимо, що наведена метрика не буде працювати, як описано вище, для випадків, коли цільовий часовий ряд Y має однакові значення, $Y = \{y_1, \ldots, y_n\}$, такі що $y_1 = \ldots = y_n$, або близький до цього вигляд $y_1 \approx \cdots \approx y_n$ тому, що

за виняткового випадку немає сенсу формувати Y_0 та обчислювати $ScoreAvg_0$ тому, що матриці відстаней будуть містити нульові, або близькі до нуля значення. Такі випадки, коли константні або близькі до константних, значення цільового ряду вважаються як некоректні вироджені, і в цьому випадку рекомендується використовувати інші метрики оцінювання ступеня близькості цільового та вхідного часових рядів.

Наприклад, є ряд Y_0, що містить обернені (помножені на -1) значення Y, тобто $Y_0 = \{- y_1, \ldots, - y_n\}$. В таблиці 3.2 наведені оригінальні та нормовані (за діапазоном) дані щодо ефективності використання трудових ресурсів підприємства. Вхідні дані: T – рік, Y – індекс зростання продуктивності праці сільськогосподарських підприємств (відсотків до попереднього року), X – індекс продукції сільського господарства (відсотків до попереднього року).

Вхідний ряд даних X містить пропуски показників за 2000 та 2001 років. Ці пропуски створено штучно в рамках даного прикладу, з метою демонстрації роботи алгоритму на часових рядах з різною довжиною.

Таблиця 3.2 – Дані щодо ефективності використання трудових ресурсів підприємства

Рік, T	Цільова змінна *Y, %*	Обернена цільова змінна Y_0, %	Вхідний часовий ряд, *X, %*	Нормоване, %		
				Y	Y_0	*X*
1998	90,5	-90,5	90,4	0	0	0,0456
1999	94,6	-94,6	93,1	0,0537353	-0,0537	0,1335
2000	112,3	-112,3		0,2857143	-0,2857	0,3973
2001	127,7	-127,7		0,4875491	-0,4875	0,3973
2002	117,8	-117,8	101,2	0,3577982	-0,3577	0,3973
2003	93,3	-93,3	89	0,0366972	-0,0366	0
2004	166,8	-166,8	119,7	1	-1	1
2005	114,6	-114,6	100,1	0,3158585	-0,3158	0,3615
2006	115,5	-115,5	102,5	0,327654	-0,3276	0,4397
2007	105,6	-105,6	93,5	0,197903	-0,1979	0,1465

Графіки цільового, оберненого цільового та вхідного часових рядів представлені на рис. 3.6.

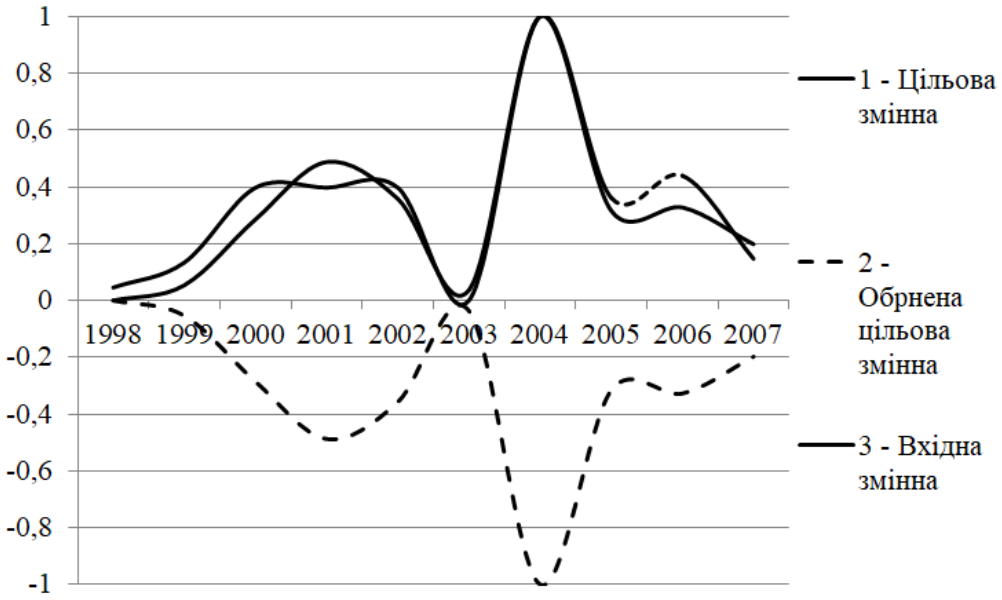


Рисунок 3.6 – Графіки цільового, оберненого цільового та вхідного часових рядів

В результаті проведених розрахунків отримано, що оптимальна довжина шляху для досліджуваних часових рядів дорівнює 10. Значення функції $ScoreAvg_0 = 0{,}6127$. Значення функції вартості шляху $ScoreAvg_1 = 0{,}0612$.

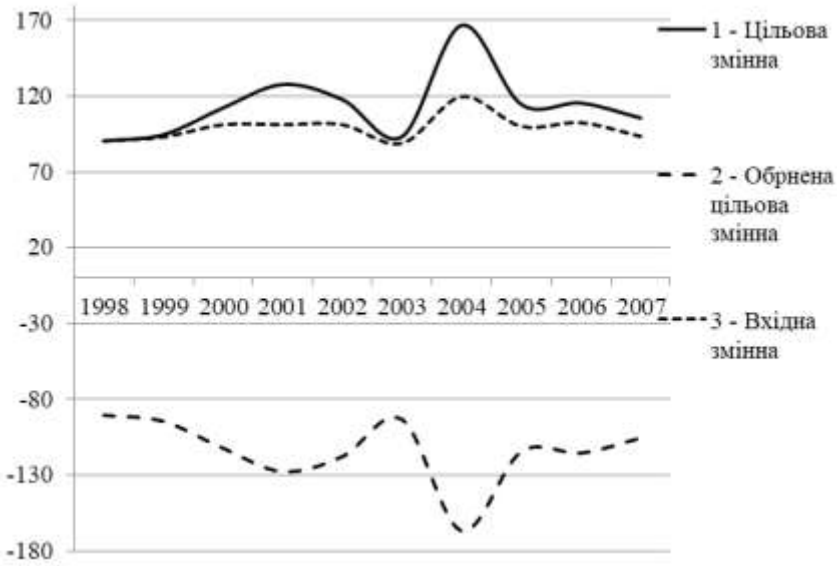


Рисунок 3.7 - Графіки нормалізованих значень (1) цільового, (2) оберненого цільового та (3) вхідних часових рядів

Обчислені значення $ScoreAvg_0$ та $ScoreAvg_1$ було отримано на нормалізованих даних прикладу.

Як наслідок значення міри подібності буде обчислене як:

$$ScoreSim(Y,X) = 100\% \cdot \left(1 - \frac{ScoreAvg_1}{ScoreAvg_0}\right) = 100\% \cdot \left(1 - \frac{0{,}0612}{0{,}6127}\right) = 89\%$$

3.4 . Аналіз складності алгоритму побудови оптимального шляху

Для випадку алгоритму перебору всіх комбінацій шляхів по матриці відстаней, кількість обчислень залежить від кількості елементів матриці відстаней, яка в свою чергу залежить від розміру цільового та вхідного рядів. Саме тому обчислювальна складність виражається як $O(n \cdot m)$, де n – розмір цільового ряду, а m – розмір вхідного ряду, якщо $n = m$ то відпоідно маємо квадратичну залежність $O(n^2)$ [195].

Зазвичай алгоритм працює із урахуванням обмежень на стиснення та розширення, як наслідок робота ведеться не з усіма елементами матриці відстаней, а тільки з тими, що входять до відповідної області.

В цьому випадку складність алгоритму може бути виражена як (3.51):

$$O\big((RL + CL + 1) \cdot \max(n,m) - (\min(RL,CL) + 1) \cdot \max(RL,CL)\big) \qquad (3.51)$$

де RL – обмеження на стиснення, а CL – обмеження на розширення.

У випадку, коли $m = n$, та $RL = CL = K$ формула (3.52) набуває вигляду [195]:

$$O\big((2 \cdot K + 1) \cdot n - (K + 1) \cdot K\big) \qquad (3.52)$$

де К – розмір обмеження (на стиснення або розширення).

Якщо $K \ll n$ (K значно менше за n) то формула (3.51) опису складності алгоритму, вироджується у формулу (3.52), що описує лінійну залежнсть складності алгоритму від розмірності цільового/вхідного часового ряду:

При проведенні дослідження складності алгоритму використано цільовий та вхідний часові ряди довжиною 280 точок, при цьому, розглядався винятковий випадок $n = m$ та $RL = CL$. Для проведення обчислювальних експериментів використовувався персональний комп'ютер на базі процесору Intel I3-7100U, тактова частота процесора - 2.40 GHz, ємність оперативного запам'ятовуючого пристрою 300 Мб. Програмним середовищем для написання коду та виконання експериментів обрано SAS Enterprise Guide 7.1.

В таблиці 3.3 наведені результати часу, с, витраченого на виконання обчислень залежно під параметру обмеження K.

Таблиця 3.3 – Результати вимірювання витраченого часу, необхідного для виконання обчислень, залежності під параметру обмеження K, с

K	1	2	3	4	5	6	7	8	9	10
Час, с	60	120	160	210	260	306	360	400	450	500

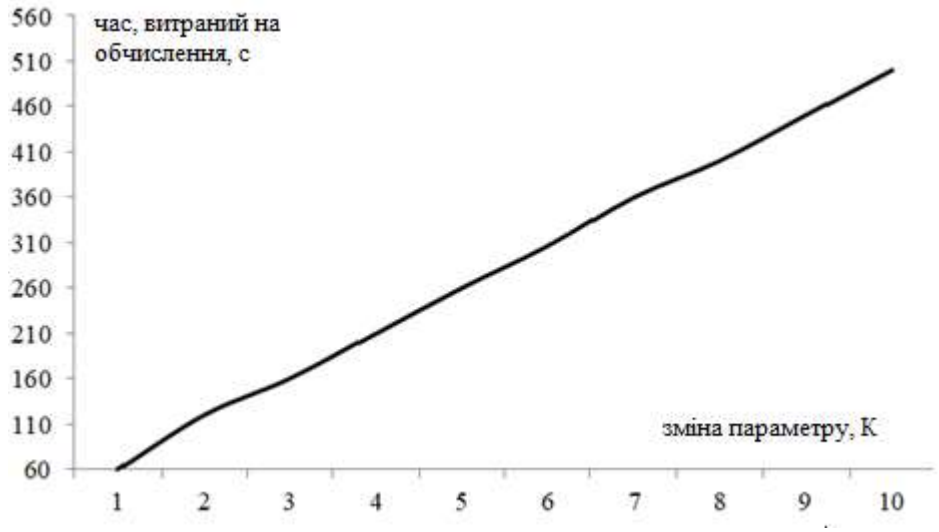


Рисунок 3.9 – Залежність між часом на обчислення та параметром обмеження K

Як можна побачити з отриманих результатів обчислювального експерименту (табл. 3.3 та рис. 3.9), при збільшені значення параметру обмеження *K* на одиницю, час обчислень збільшується лінійно. Для випадку аналізу часових рядів довжиною у 280 точок в середньому на 50 секунд.

3.5 Приклад застосування методики визначення подібності процесів

В якості прикладу застосування методики визначення подібності нелінійних нестаціонарних процесів використано часовий ряд даних щодо продуктивності праці у сільськогосподарських підприємствах.

Таблиця 3.3 – Зміна індексу сільськогосподарської продукції (X) залежно від зміни продуктивності праці (Y)

Рік (T)	Індекс продуктивності праці в сільськогосподарських підприємствах, % (Y)	Індекс сільськогосподарської продукції, X
1998	90,5	90,4
1999	94,6	93,1
2000	112,3	
2001	127,7	
2002	117,8	101,2
2003	93,3	89
2004	166,8	119,7
2005	114,6	100,1
2006	115,5	102,5
2007	105,6	93,5

Вхідний ряд даних змінні *X* має штучно створені пропуски значень показника у 2000 та 2001 рр. Пропуски було створено з метою демонстрації роботи алгоритму на часових рядах з різною довжиною.

Відповідно, отримаємо:

$$Y = \{90,5;\ 112,3;\ 127,7;\ 117,8;\ 93,3;\ 166,8;\ 114,6;\ 115,5;\ 105,6\}$$

$$X = \{90,4;\ 93,1;\ 101,2;\ 89;\ 119,7;\ 100,1;\ 102,5;\ 93,5\}$$

$$n = 10;\ m = 8;\ j = 1, \ldots, 10;\ i = 1, \ldots, 8$$

На рис. 3.10 наведено графіки цільового ряду даних Y та вхідного ряду X.

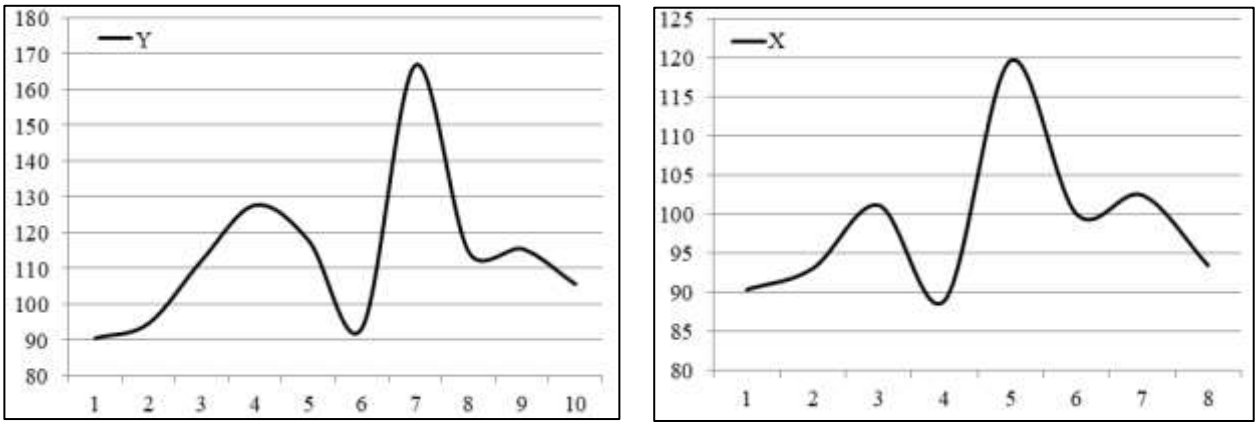


Рисунок 3.10 – Графіки цільового ряду даних *Y* та вхідного ряду - *X*

Як видно з рис. 3.10, в загальному випадку довжини досліджуваних часових рядів можуть бути різними, тобто не співпадати навіть за часовою шкалою. Процеси є нелінійними: мають три фази підйому та три фази спаду.

В рамках реалізованого методу, спочатку видаляються пропуски по зміним, і тільки після цього виконується операція нормалізації даних.

Для наведених вхідних рядів часових даних виконано стандартизацію за діапазоном [108]:

$$z_i = \frac{x_i - \min\limits_{i=1,..,m}\{x_i\}}{\max\limits_{i=1,..,m}\{x_i\} - \min\limits_{i=1,..,m}\{x_i\}}$$

де $\max_{i=1,..,m}\{x_i\}$ – максимальне значення, серед всіх значень X, а $\min_{i=1,..,m}\{x_i\}$ – відповідно мінімальне.

Нормалізовані значення цільової та вхідної змінних представлені у табл. 3.4.

Таблиця 3.4 – Значення цільового та вхідного часових рядів після нормування за діапазоном

Номер індексу часу (T)	Значення цільового часового ряду (Y)	Значення вхідного часового ряду (X)
1	0	0,0456026
2	0,0537353	0,1335505
3	0,2857143	0,3973941
4	0,4875491	0
5	0,3577982	1
6	0,0366972	0,3615635
7	1	0,4397394
8	0,3158585	0,1465798
9	0,327654	
10	0,197903	

Для розрахунку значень матриці відстаней було застосовано нормування значень. В якості метрики визначення відстані використовувався модуль різниці:

$$d(y_j, x_i) = |y_j - x_i|$$

де x_i та y_j - відповідні і-е та j-е значення цільового та вхідного рядів даних, $p \geq 1$.

В табл. 3.5 представлені значення матриці відстаней, обчислені на основі нормованих значень.

Таблиця 3.5 – Значення матриці відстаней

Індекс часу		Нормовані значення, індекс часу вхідного ряду (i)							
Нормовані значення, індекс часу вхідного ряду (j)		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
	j=1	0,0456	0,1336	0,3973	0	1	0,3616	0,4397	0,1465
	j=2	0,0081	0,0799	0,3436	0,0537	0,9462	0,3078	0,3861	0,0928
	j=3	0,2401	0,1522	0,1116	0,2857	0,7142	0,0758	0,1541	0,1391
	j=4	0,4419	0,3539	0,0901	0,4875	0,5125	0,1259	0,0478	0,3409
	j=5	0,3122	0,2242	0,0395	0,3578	0,6422	0,0038	0,0819	0,2112
	j=6	0,0089	0,0969	0,3607	0,0367	0,9633	0,3248	0,4031	0,1098
	j=7	0,9543	0,8664	0,6026	1	0	0,6384	0,5603	0,8534
	j=8	0,2703	0,1823	0,0815	0,3159	0,6841	0,0457	0,1239	0,1692
	j=9	0,2821	0,1941	0,0697	0,3276	0,6723	0,0339	0,1121	0,1811
	j=10	0,1523	0,0643	0,1995	0,1979	0,8021	0,1637	0,2418	0,0513

Для побудови шляху необхідно задати обмеження на перетворення. Нехай в рамках даного прикладу RL = 2 – обмеження на стиснення цільової змінної, а CL=1 обмеження на розширення цільової змінної. Шлях по матриці відстаней був побудований за 11 ітерацій.

В табл. 3.6 наведені результати обчислень – значення матриці відстаней D з урахуванням наведених обмежень на перетворення цільового ряду.

Таблиця 3.6 - Значення матриці відстаней D з урахуванням обмежень на перетворення цільового ряду $RL = 2$ та $CL = 1$.

Індекс часу	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
j=1	0,0456	0.1335						
j=2	0,0081	0,0799	0.3436					
j=3	0,2401	0,1522	0,1116	0.2857				
j=4		0,3539	0,0901	0,4875	0.5124			
j=5			0,0395	0,3578	0,6422	0.0037		
j=6				0,0367	0,9633	0,3248	0.4031	
j=7					0	0,6384	0,5603	0.8534
j=8						0,0457	0,1239	0,1692
j=9							0,1121	0,1811
j=10								0,0513

Результати виконання ітерацій алгоритму представлені в табл. 3.7.

Таблиця 3.7 – Ітерація 11

Індекс часу	*i* = 1	*i* = 2	*i* = 3	*i* = 4	*i* = 5	*i* = 6	*i* = 7	*i* = 8
j=1	0,0456	0.1335						
j=2	0,0081	0,0799	0.3436					
j=3	0,2401	0,1522	0,1116	0.2857				
j=4		0,3539	0,0901	0,4875	0.5124			
j=5			0,0395	0,3578	0,6422	0.0037		
j=6				0,0367	0,9633	0,3248	0.4031	
j=7					0	0,6384	0,5603	0.8534
j=8						0,0457	0,1239	0,1692
j=9							0,1121	0,1811
j=10								0,0513

Як можна побачити з (табл. 3.7), за 11 ітерацій алгоритму було побудовано шлях, але він все ще не оптимальний тому, що присутній зайвий перехід від D [1,1] до D [2,1] якого можна позбавитися. Це призведе до скорочення довжини шляху з 11 до 10 та зменшення значення функції вартості шляху на 0,0081 одиниць. Для розв'язку цієї задачі виконується проходження по матриці вартостей D з метою виявлення конфігурацій шляху та видалення зайвих переходів (рис. 3.11).

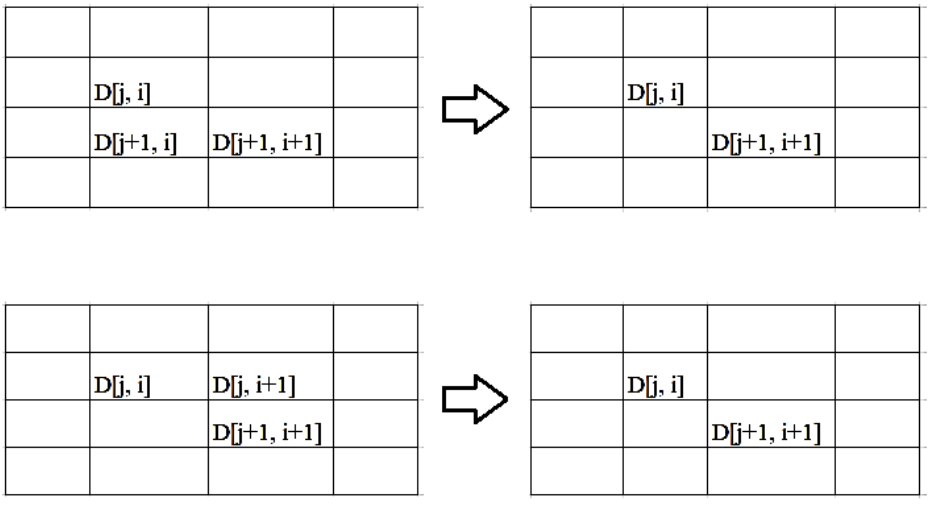


Рисунок 3.11 – Дві типові конфігурації наявності зайвих переходів шляху в матриці відстаней, в першому випадку видаляється D [j+1, i], а в другому D [j, i+1]

Результати розрахунку оптимального шляху по матриці відстаней представлені в табл. 3.8.

Таблиця 3.8 – Оптимальний шлях по матриці відстаней

Індекс часу	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
j=1	**0,0456**							
j=2	0,0081	**0,0799**						
j=3	0,2401	0,1522	**0,1116**					
j=4		0,3539	**0,0901**	0,4875				
j=5			**0,0395**	0,3578	0,6422			
j=6				**0,0367**	0,9633	0,3248		
j=7					**0**	0,6384	0,5603	
j=8						**0,0457**	0,1239	0,1692
j=9							**0,1121**	0,1811
j=10								**0,0513**

Таблиця 3.9 – Результати побудови шляху π, представлені у вигляді таблиці переходів між елементами матриці відстаней D

Параметр p	Індексні значення (j_p,i_p)	Відстань, d_{ji} $(D[j,i]$
1	(1,1)	0.0456
2	(2,2)	0.0798
3	(3,3)	0.1116
4	(4,3)	0.0901
5	(5,3)	0.0395
6	(6,4)	0.0366
7	(7,5)	0
8	(8,6)	0.0457
9	(9,7)	0.1121
10	(10,8)	0.0513

В табл. 3.9 наведені індексні значення (j_p,i_p) для оптимального шляху π, що відповідає матриці відстаней D, з урахуванням обмежень на перетворення

RL = 2 та *CL* = 1. При цьому довжина побудованого шляху π дорівнює P=10, а значення функції вартості Score= 0,6123.

У додатку І наведено ітераційний процес побудови шляху в матриці відстаней, що відповідає значенням, представленим у табл. 3.10, а рис. 3.12 відображає схему співставлення точок цільового та вхідного часових рядів.

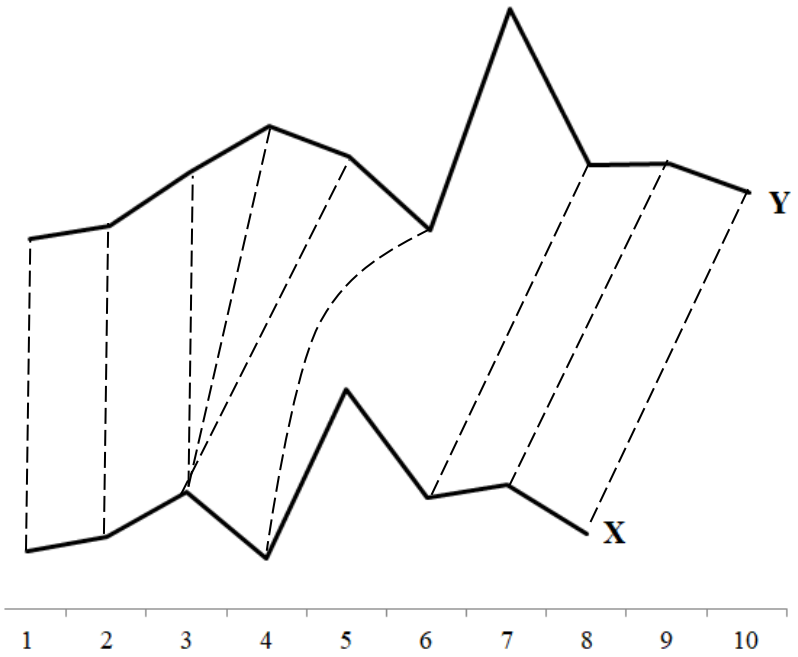


Рисунок 3.12 – Схема співставлення точок цільового та вхідного часових рядів.

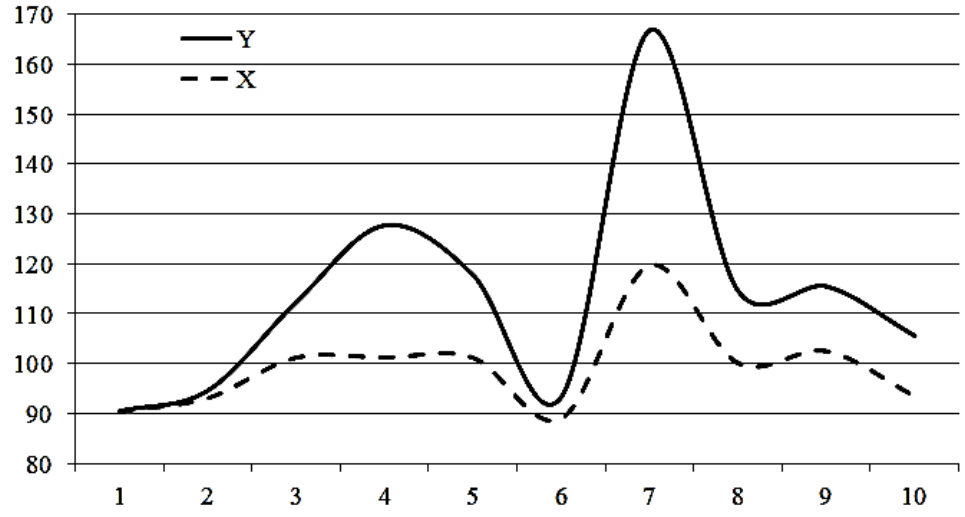
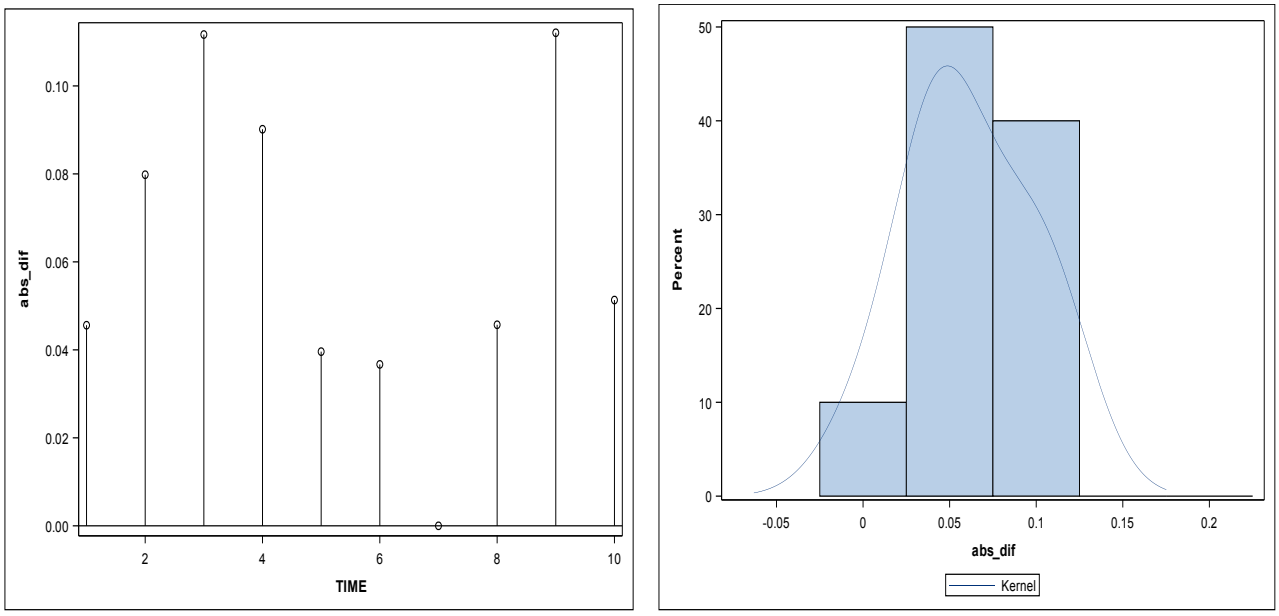


Рисунок 3.13 – Графіки цільового та вхідного рядів даних після виконання операцій перетворення з розширення та стиснення

Як можна побачити з рис. 3.12 та 3.13 для вхідного часового ряду фактично було виконано операцію розширення значень для вхідного ряду значень для співставлення зі значеннями часу 4 та 5, на основі попереднього значення x_3.

На рис. 3.14 наведені графіки значень абсолютних різниць, між нормованими значеннями цільового та вхідного рядів після перетворень.



а) *б*)

Рисунок 3.14 (а) – Графік абсолютних різниць між нормованими значеннями цільового та вхідного рядів після перетворень (б) - Гістограма та графік розподілів абсолютних різниць між нормованими значеннями цільового та вхідного рядів після перетворень

Як видно з рис. 3.14, що похибки нормально розподілені, а максимальне відхилення не перевищує 12% (для 3 та 9 значень індексів часу).

В табл. 3.10 наведені значення статистик порівняння цільового та вхідного рядів.

Таблиця 3.10 - Статистики порівняння цільового та вхідного рядів

Статистика (укр/анг. назва)	Статистика (укр. назва)	Значення
Score	Значення функції Score, за формулою (8)	0,6127
Score_Avg	Значення функції ScoreAvg, за формулою (9)	0,0612
Score_Sim	Значення функції ScoreSim, за формулою (21)	89%
Довжина шляху	Довжина шляху	10
Кількість прямих переходів (без розширень та стиснень)	Кількість прямих переходів (без розширень та стиснень), за формулою (20)	8
Кількість стиснень	Кількість стиснень, за формулою (10)	2
Кількість розширень	Кількість розширень, за формулою (15)	0
Compression maximum	Максимальна довжина стиснень, за формулою (14)	2
Compression path percent (%)	Процент співвідношення кількості стиснень до довжини шляху, за формулою (11)	20
Compression input percent (%)	Процент співвідношення кількості стиснень до кількості елементів вхідної послідовності, за формулою (12)	25
Compression target percent (%)	Процент співвідношення кількості стиснень до кількості елементів цільової послідовності, за формулою (13)	20
Expansion maximum	Максимальна довжина розширення, за формулою (19)	0
Expansion path percent (%)	Процент співвідношення кількості розширень до довжини шляху, за формулою (16)	0
Expansion input percent (%)	Процент співвідношення кількості розширень до кількості елементів вхідної послідовності, за формулою (17)	0
Expansion target percent (%)	Процент співвідношення кількості розширень до кількості елементів цільової послідовності, за формулою (18)	0

Як можна побачити ряди подібні на 89% за метрикою ScoreSim.

Висновки до розділу 3

У розділі запропоновано новий робастний непараметричний метод аналізу нелінійних нестаціонарних процесів за їх подібністю. Суть методу полягає у пошуку процесів (часових рядів) досліджуваного типу зі схожою статистичною поведінкою, що в результаті допомагає зробити висновки стосовно наявність шаблонів поведінки досліджуваних процесів. Для врахування різної поведінки рядів у часі за допомогою аналізу подібності розраховується відстань між першою часовою послідовністю та другою із урахуванням впорядкування.

Розроблено алгоритм побудови шляху переміщення по матриці відстаней. Запропоновано аналітичне і графічне представлення функціонування алгоритму побудови шляху для матриці відстаней між елементами часових послідовностей. Запропоновано метрики для оцінювання якості побудови шляху переміщення, які забезпечують оптимізацію процесу аналізу подібності процесів. Автором запропоновано власну метрику (*ScoreSim*), призначену для оцінювання ступеня близькості досліджуваних часових рядів Y та X, яка підвищує коректність використання алгоритму аналізу подібності процесів.

Виконано аналіз складності алгоритму побудови оптимального шляху переміщення по матриці відстаней. Встановлено, що для випадку перебору всіх комбінацій шляхів по матриці відстаней, кількість обчислень залежить від кількості елементів матриці відстаней, яка в свою чергу залежить від розміру цільового та вхідного рядів.

Подано приклад успішного застосування методики визначення подібності нелінійних процесів з використанням часових рядів даних щодо продуктивності праці у сільськогосподарських підприємствах.

РОЗДІЛ 4
ЗАСТОСУВАННЯ МЕРЕЖ БАЙЄСА У ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ ПРОГНОЗУВАННЯ

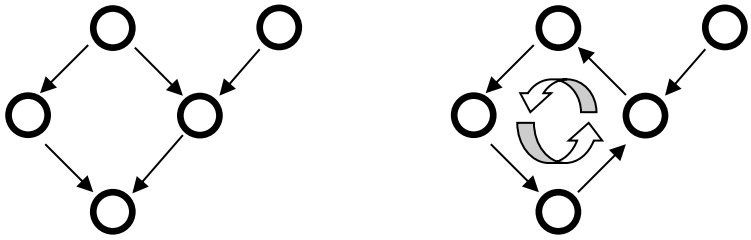
4.1 Особливості використання байєсових мереж для прогнозування нелінійних нестаціонарних процесів

Використання байєсових мереж (БМ) для аналізу процесів різної природи, діяльності людини та функціонування технічних систем дозволяє враховувати та використовувати будь-які вхідні дані – експертні оцінки і статистичну інформацію. Змінні, що використовуються в БМ, можуть бути як дискретними, так і неперервними, а характер їх надходження при аналізі та прийнятті рішення може бути і в режимі реального часу, і у вигляді статистичних масивів інформації та баз даних. Завдяки представленню взаємодії між факторами процесу у вигляді причино-наслідкових зв'язків в мережі досягаються максимально високий рівень візуалізації та, як наслідок, чітке розуміння суті взаємодії факторів процесу між собою. Саме це відрізняє БМ від інших методів ІАД. Також перевагами БМ є можливості врахування невизначеностей статистичного, структурного і параметричного характеру, побудова моделей за наявності прихованих вершин і при неповних спостереженнях, а також формування висновку за допомогою різних методів – наближених і точних. Загалом, можна сказати, що БМ – це високоресурсний метод ймовірнісного моделювання процесів довільної природи з невизначеностями різних типів, який забезпечує можливість достатньо точного опису їх функціонування, оцінювати прогнози та будувати системи управління [47, 196-237].

БМ являють собою графи з деякими спеціальними властивостями. В загальному, граф – це сукупність вузлів (вершин), з'єднаних дугами. Дуга вказує на існування причинного зв'язку між вершинами. Якщо дуга не направлена, то вона свідчить лише про наявність зв'язку між вузлами.

Направлені ж дуги вказують додатково ще й напрям залежності – від причини до наслідку. Відповідно графи, що містять лише дуги без напрямку називаються ненаправленими, лише направлені дуги – направленими.

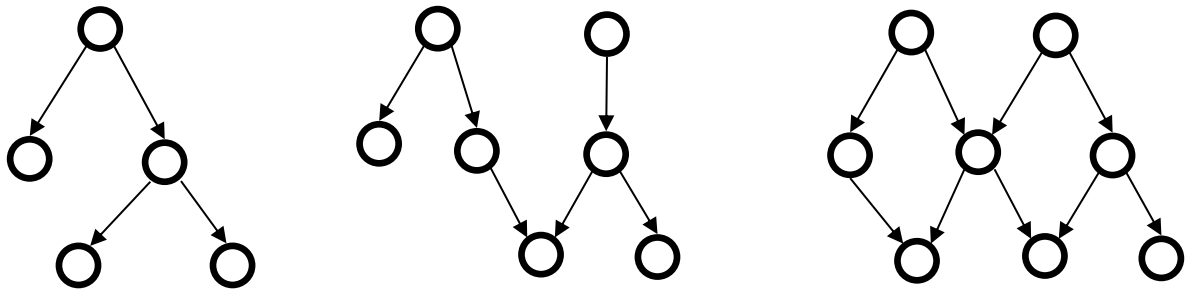
Направлений граф є ациклічним, якщо для будь-якої вершини графу не існує направленого шляху, що починається і закінчується в цій вершині (рис. 4.1, а). В протилежному випадку направлений граф називається циклічним (рис. 4.1, б).



а) б)

Рисунок 4.1 – Приклад циклічних та ациклічних графів

Розрізняють наступні типи структур графів БМ [228] (рис. 4.2)*:*



а) Дерево б) Полідерево в) Багатозв’язна мережа

Рисунок 4.2 – Структури графів БМ

Дерево – це структура, в якій кожна вершина може мати не більше однієї батьківської (рис.4.2, а).

Однозв’язна мережа (полідерево) – структура, в якій кожна вершина може мати більше одного батька, але існує лише один (ненаправлений) шлях між будь-якими двома вершинами (рис. 4.2, б) [228].

В багатозв’язних мережах між будь-якими двома вершинами може існувати декілька шляхів (рис. 4.2, в). Саме такі мережі зустрічаються найчастіше при вирішенні різноманітних задач.

У направлених графах виділяють батьків і нащадків, а також кореневі вершини та листки. Нащадки – вершини, які залежать від однієї або більше інших вершин. Відповідно батьківські вузли мають одного або більше нащадків. Одна і та ж вершина може бути як батьківською, так і нащадком. Кореневі вузли – вузли, які не мають жодного батька, а листки – вузли, що не мають нащадків.

БМ представляє собою пару [228] $<G,B>$, де G – це направлений ациклічний граф, а B – це множина параметрів, що визначають мережу, або множина таблиць умовних ймовірностей вершин (ТУЙ, CPT - conditional probability table): $P = P\left(X^{(i)} \middle| pa\left(X^{(i)}\right)\right), i = 1...N$ для кожного можливого значення $x^{(i)} \in X^{(i)}$ та $pa(X^{(i)}) \in Pa(X^{(i)})$, де $Pa(X^{(i)})$ - множина батьків змінної $X^{(i)} \in G$.

Кожна змінна $X^{(i)} \in G$ є вершиною мережі Байєса. Повна спільна ймовірність БМ обчислюється за формулою (4.1) [228]:

$$P(X^{(1)},...,X^{(N)}) = \prod_{i=1}^{N} P(X^{(i)} \middle| Pa(X^{(i)})). \qquad (4.1)$$

З математичної точки зору БМ – це модель подання існуючих і відсутніх імовірнісних залежностей. При цьому зв'язок $A \to B$ являється причинним, коли подія A – причина виникнення B, тобто коли існує механізм, відповідно до якого значення, прийняте A, впливає на значення, прийняте B.

Формула Байєса являє собою модель раціонального вибору в умовах неточної або неповної інформації. Процес прийняття рішень може ґрунтуватися на двох протилежних методах – дедукції й індукції. Обираючи

шлях дедукції, виходять із певної гіпотези або припущення про сутність події й припускають, що відбудеться згодом, якщо ця гіпотеза достовірна. Оскільки між гіпотезою й передбачуваним розвитком подій існує стійкий зв'язок, дедукція вважається об'єктивним методом, але вона лише підтверджує або спростовує запропоновану гіпотезу й не здатна створити нові знання.

Індуктивні міркування мають протилежну спрямованість: гіпотези будуються й оцінюються на підставі даних досвіду. В основі доказу лежить процес індукції, що відображає перехід від спостережень до закономірностей, що є їх основою. Переваги індуктивного методу аналізу полягають у тому, що за допомогою даних досвіду одержують додаткову інформацію про раніше невідомі явища, будують нові гіпотези й поглиблюють свої знання про зовнішній світ. На жаль, при цьому неминуче виникає так звана проблема індукції: ми не можемо бути повністю впевненими в правильності доводів.

Значно складніше індуктивна диференціальна діагностика, що реалізується у визначенні ймовірності різних визначень виходячи з наявної інформації й результатів проведених досліджень. Таким чином, дедуктивний метод більше надійний і об'єктивний, але менш продуктивний, чим індуктивний.

Нехай Ω - випадковий простір (множина) подій випадкових експериментів, що містить всі можливі значення випадкової змінної. Розглянемо дві події $E \in \Omega$ і $H \in \Omega$, що відповідно є спостереженням та гіпотезою. Формула Байєса використовується для обчислення ймовірності того, що подія E відбудеться за умови, що відбулась подія H. Тобто для обчислення $P(E \mid H)$ - умовної ймовірності E при заданій події H [228].

Умовна ймовірність $P(E \mid H)$ дорівнює відношенню сукупної ймовірності подій E та H $P(E \cap H)$ до ймовірності події H, за умови що вона не дорівнює нулю (4.2) [228]:

$$P(E \mid H) = \frac{P(E \cap H)}{P(H)}. \tag{4.2}$$

Аналогічно для події H (4.3):

$$P(H \mid E) = \frac{P(H \cap E)}{P(E)}. \tag{4.3}$$

Звідси можна записати правило Байєса (враховуючи властивість комутативності сукупної ймовірності) [228] (4.4):

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} \tag{4.4}$$

Дана формула показує причинно-наслідкові зв'язки між спостереженнями та гіпотезами. Ймовірності $p(H)$ та $p(E \mid H)$ є апріорними, тобто задаються до початку спостережень. Ймовірність $P(H \mid E)$ є апостеріорною. Перевага байєсівського методу полягає в тому, що апріорні ймовірності можна уточнювати (оновлювати) у відповідності до реалій протікання процесу, що досліджується. Це дозволяє уточнювати ймовірності подій при надходженні додаткової інформації.

Головне припущення теорії побудови БМ полягає в тому, що події є вичерпними ($\cup_{i=1}^{n} H_i = \Omega$) і не перетинаються (4.5). При виконанні цих умов ймовірність події E можна обчислити за допомогою умовних ймовірностей.

$$p(E) = \sum_{i=1}^{n} p(E \cap H_i) = \sum_{i=1}^{n} p(E \mid H_i) \cdot p(H_i). \tag{4.5}$$

Підставивши даний вираз в формулу (2.4) отримаємо вираз (4.6) [228]:

$$p(H_k \mid E) = \frac{p(E \mid H_k) \cdot p(H_k)}{\sum_{i=1}^{n} p(E \mid H_i) \cdot p(H_i)}, \quad (4.6)$$

На основі цієї формули будуються БМ.

Формула Байєса з успіхом була апробована в різних сферах людської діяльності в системах прийняття управлінських рішень для вирішення складних задач у реальних умовах [196-204, 212]. Наприклад, унікальна властивість класифікувати об'єкти завдяки їй широко використовується в комп'ютерному програмуванні як спам-фільтри, в медицині для встановлення діагнозу, програмуванні рухів тіла, в проектуванні аерокосмічних засобів, для оцінки ризиків, при оцінці дорожніх ситуацій, для вибору стратегій управління, транспортними засобами, для оцінки стану шахт та мереж електропостачання, для побудови механізму логічного висновку експертних систем, систем прийняття рішень, для оцінки катастроф тощо.

В залежності від типів станів вершин та властивостей самих вершин виділяють дискретні, динамічні, неперервні, гібридні мережі Байєса.

4.2 Опрацювання невизначеностей із використанням БМ

З глобальним накопиченням бізнес-інформації багатьма компаніями все більш необхідним стає інтелектуальний аналіз даних (Data Mining) – технологія виявлення прихованих взаємозв'язків всередині великих баз даних. Більшість інструментів інтелектуального аналізу даних ґрунтується на двох технологіях: машинне навчання (machine learning) і візуалізація (візуальне подання інформації). Байєсівські мережі поєднують у собі ці дві технології.

Прихованими називаються змінні, значення яких не є спостережуваними. Не зважаючи на активні розробки методів побудови

мереж Байєса за навчальними даними [228], побудова мереж з прихованими вершинами залишається однією з центральних проблем в дослідженні БМ.

Переважна більшість досліджень проводиться з припущенням, що дані є повними, тобто значення всіх змінних є відомими для кожного випадку в базі даних. Проте насправді це не є реалістичним, адже приховані змінні грають ключову роль в багатьох проблемах реального життя: невідомий регулюючий механізм може бути ключем до складних біологічних систем; зв'язок симптомів може натякнути на приховану фундаментальну проблему в діагностичних системах; умисно замаскована певна економічна сила може бути причиною пов'язаних фінансових феноменів [92, 93, 196]. Насправді ж приховані змінні просто служать підсумковим механізмом, що «захоплює» інформацію з деяких спостережуваних змінних та «проводить» її до деякої іншої частини мережі. Тому, приховані змінні можуть значно спростити структуру БМ та в результаті привести до кращого загального ймовірнісного виводу.

Побудова мереж Байєса фактично означає вибір оптимальної структури мережі. Оптимальна вказує на те, що обрана структура буде максимально правдоподібно відповідати навчальним даним або процесу, який моделюється. Термін «побудова БМ» в широкому сенсі означає розв'язання задач: пошук оптимальної структури БМ, тобто направленого ациклічного графа, що найбільш адекватно відповідає навчальним даним або досліджуваному процесу; обчислення значень таблиць умовних ймовірностей БМ для кожної вершини цього графа.

Переважна більшість існуючих методів зустрічаються з наступними проблемами:

1. Наявність впорядкованої множини вершин. У більшості методів, особливо старих, вважається, що ця множина задана, але на практиці при роботі з реальними даними це дуже часто не відповідає дійсності.

2. Низька обчислювальна ефективність. Деякі сучасні методи працюють без використання впорядкованої множини вершин, замість неї використовується тест на умовну незалежність. Але в даному випадку необхідно виконати експоненціальну кількість таких тестів, що в свою чергу призводить до зменшення ефективності роботи методу у зв'язку із зростанням об'єму обчислень.

3. Проблема побудови великих БМ. Існує декілька методів здатних побудувати структуру БМ з декількома сотнями вершин, використовуючи навчальну вибірку, що складається з мільйонів записів. До таких методів відносяться Tetrad II [224] та SopLeq [209].

Отож, для опису БМ необхідно визначити топологію графа (структуру) і параметри кожної вершини. Таку інформацію можна отримати з навчальних даних, але отримання правильної топології мережі є більш складної проблемою, аніж визначення параметрів вершин. Особливий підхід необхідний у випадку, коли деякі вершини приховані або коли маємо справу з некоректними чи неповними даними. Тому дослідники виділяють 4 випадки навчання БМ. Вони наведені в таблиці 4.1 [204, 209, 228].

Таблиця 4.1 – Випадки навчання мереж Байєса

Структура	Спостереження	Метод
Відома	Повні	Максимальна Оцінка Правдоподібності
Відома	Часткові	Градієнтні методи, ЕМ-алгоритм (максимізація математичного сподівання), застосування вибірки Гіббса
Невідома	Повні	Пошук в просторі моделей
Невідома	Часткові	Структурний ЕМ-алгоритм, алгоритм стиснення границь

Розглянемо випадки, коли спостереження повні, тобто БМ не має прихованих вершин.

В даному випадку обчислюються значення параметрів кожного умовного ймовірнісного розподілу, які максимізують правдоподібність навчальних даних, що містять M записів (припускають, що ці записи незалежні). Нормалізоване логарифмічне рівняння правдоподібності для навчальної множини $D = \{D_1, ..., D_M\}$ має вигляд (4.7) [228]:

$$L = \frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} \log P(X_i \mid Pa(X_i), D_m), \tag{4.7}$$

де $Pa(X_i)$ - множина предків вершини X_i, D - навчальна вибірка, а D_m - це m-й елемент навчальної вибірки, M - кількість подій в навчальній вибірці.

Розглянемо, наприклад, просту мережу фіктивного медичного Розглянемо, наприклад, просту мережу медичного прикладу, що описує смертність пацієнтів хворих на коронавірусну хворобу, залежно від виявлених в нього супутніх хвороб та проблем із здоров'ям.

Схема мережі зображена на рис. 4.3.

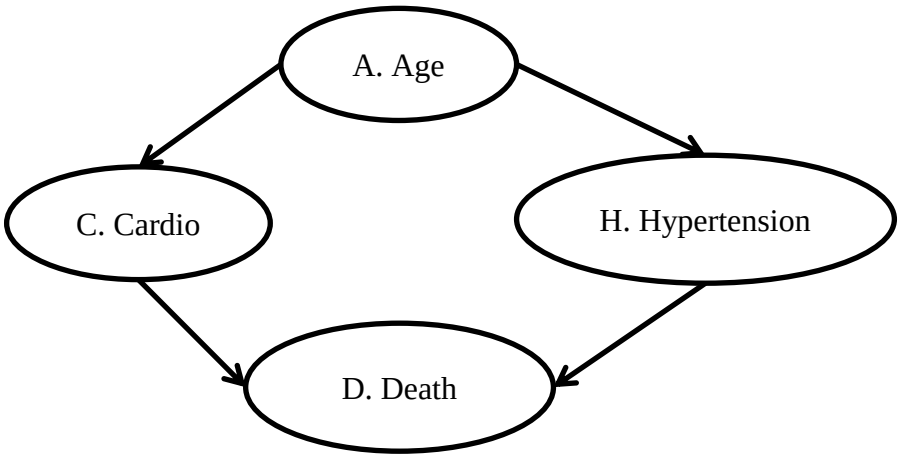


Рисунок 4.3 – Фрагмент побудованої мережі Байєса для прогнозування ймовірності смерті пацієнта хворого на коронавірусну хворобу залежно від наявності інших проблем із здоров'ям та факторів ризику

Вершина Age має сім станів – приналежність до відповідної вікової групи пацієнта: 0-9; 10-39; 40-49, 50-59; 60-69; 70-79; ≥80. Інші вершини мають по два стани «присутній» і «відсутній»: Age, Cardio, Hypertension.

Таблиця 4.2 – Опис показників, що використовувалися як вершини БМ

Назва вершини	Опис вершини
Death	Цільова змінна – ймовірність смерті пацієнта хворого на коронавірсуну хворобу
Age	Вік пацієнта
Cardio	Чи має пацієнт серцево-судинні захворювання
Hypertension	Чи хворий пацієнт на гіпертонію

Для цього дискретного випадку оцінка вузла D обчислюється наступним чином. Нехай є дані для навчання, тобто інформація про те скільки випадків смерті пацієнтів було зафіксовано від коронавірсуної хвороби, а пов'язані причинами із цим були наявність серцево-судинних захворювань та гіпертонії $N(D=1, C=1, H=1)$, скільки було випадків смерті від коронавірусної хвороби, спостерігалися серцево-судинні захворювання, але не було гіпертонії $N(D=1, C=1, H=0)$ і так далі. Використовуючи ці дані, максимальна оцінка правдоподібності вершини D обчислюється як:

$$P(D=d|C=c, H=h) = \frac{N(D=d, C=c, H=h)}{N(C=c, H=h)} = \\ = \frac{N(D=d, C=c, H=h)}{N(D=0, C=c, H=h) + N(D=1, C=c, H=h)}$$

Якщо вище наведені вершини описуються функцією Гауса, то можна обчислити вибіркове середнє і дисперсію, а після цього за допомогою

лінійної регресії обчислити матрицю вагів. Для інших видів розподілу використовують більш складні процедури.

Найбільш ймовірною моделлю в даному випадку є повний граф, тому що в цьому випадку буде задіяно найбільшу кількість параметрів, отже така модель найбільше відповідатиме даним.

Формула Байєса має вигляд (4.8):

$$P(G \mid D) = \frac{P(D \mid G) \cdot P(G)}{P(D)}, \tag{4.8}$$

де G - направлений ациклічний граф, що відповідає випадковим змінним, а $D = \{x^1, ..., x^N\}$ - множина даних.

Прологарифмуємо цю формулу (4.9):

$$\log(P(G \mid D)) = \log(P(D \mid G)) + \log(P(G)) + (-\log(P(D))), \tag{4.9}$$

У формулі (4.9) доданок $-\log(P(D))$ грає роль штрафуючої компоненти за надмірно складні моделі. Складова $P(D \mid G)$ також може виступати штрафуючою компонентою за надмірно складні моделі (цей випадок відомий як бритва Оккама [68, 176]).

Для виконання точних розрахунків пов'язаних з вибором моделі необхідно обчислити $P(D) = \sum_G P(D \mid G)$, що є завданням експоненційної складності.

Замість цього можна використовувати Байєсовий інформаційний критерій, який визначається як (4.10) [228]:

$$\log(P(G \mid D)) \approx \log(P(D \mid G, \hat{\theta}_G)) - \frac{\log(N)}{2} \cdot \dim(G), \tag{4.10}$$

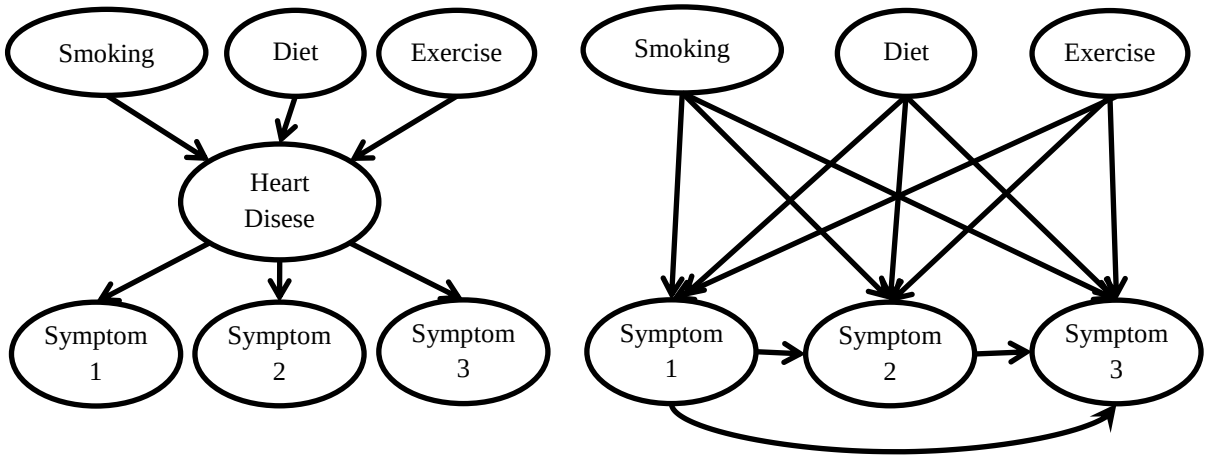
де N - число моделей, $\dim(G)$ – розмір моделі (кількість вільних параметрів), $\hat{\theta}_G$ - максимально правдоподібна оцінка параметрів, $-\frac{\log(N)}{2}\dim(G)$ - штрафуючи компонента за надмірно складні моделі.

Наступним кроком після вибору структури є навчання структури, так що б направлений ациклічний граф найкраще задовольняв даним. Це завдання є NP-важкою задачею, тобто задачею з нелінійною поліноміальною оцінкою числа ітерацій. Тому зазвичай використовують локальні алгоритми пошуку такі, як, наприклад, жадібний метод пошуку екстремуму (greedy hill climbing method) або метод гілок і меж для пошуку в просторі графів.

Всі існуючі сучасні методи побудови структури мереж Байєса можна умовно розділити на дві категорії [205, 206, 211] на основі оціночних функцій (search & scoring) та застосовуючи тест на умовну незалежність (dependency analysis).

Багато реальних задач характеризуються наявністю прихованих змінних (так званих латентних змінних), які не є спостережуваними і доступними для навчання [219]. Наприклад, історії хвороби часто включають описи спостережуваних симптомів, лікування, що застосовується, і, можливо, результату лікування, але рідко містять відомості про безпосереднє спостереження за розвитком самого захворювання. Очевидне питання: "Якщо хід розвитку захворювання не спостерігається, то чому не побудувати б модель без нього"? Відповідь на це питання можна знайти на рис. 4.4, де показана невелика фіктивна діагностична модель для сердечного захворювання. Існують три спостережувані фактори, що сприяють або перешкоджають розвитку захворювання, і три спостережувані симптоми. Передбачається, що кожна змінна має три можливі стани (наприклад, none (відсутній), moderate (помірний) і severe (серйозний)). Видалення прихованої змінної з мережі, приведеної на рис. 4.4а), спричиняє створення мережі, показаної на рис. 4.4б) і при цьому загальна кількість параметрів

збільшується з 78 до 708. Таким чином, приховані змінні дозволяють різко зменшити кількість параметрів, потрібних для визначення байєсової мережі. А даний факт, у свою чергу, дозволяє різко зменшити об'єм даних, необхідних для визначення в процесі навчання даних параметрів.



а) *б*)

Рис. 4.4 Ілюстрація необхідності використання прихованих змінних:

Приховані змінні важливі, але їх введення призводить до ускладнення завдання навчання. Наприклад, на рис. 4.4, а) не показано, як знайти в процесі навчання розподіл умовних ймовірностей для змінної HeartDisease, якщо задані її батьківські змінні, оскільки значення змінної HeartDisease в кожному випадку невідоме; така ж проблема виникає при пошуку в процесі навчання розподілів ймовірностей для симптомів.

Таким чином, перевагами використання прихованих вершин в мережах Байєса є:

– значне спрощення структури мережі;

– зменшення кількості параметрів, необхідних для визначення БМ, а це в свою чергу дозволяє різко зменшити об’єм даних, необхідних для визначення в процесі навчання цих параметрів.

Недоліки використання прихованих вершин в мережах Байєса:

– ускладнення задачі навчання БМ;

– у разі невідомої структури мережі невирішеною проблемою залишається визначення кількості прихованих вершин, місце їх розташування та кількість станів кожної з них.

Іноді ми маємо інформацію щодо прихованих вершин в мережах Байєса. Проте в загальному випадку ми не маємо ніяких даних щодо прихованої вершини. Відповідно постає дві основні проблеми навчання БМ з прихованими вершинами:

1. Скільки станів повинна мати прихована вершина?
2. Як умовні ймовірності можуть бути знайдені без даних?

Число станів може бути визначене експериментально. Так як прихована вершина насправді діє як випадкова причина, то очікувано, що число станів буде відповідним кількості станів вершин, які вона розділяє. Насправді, хорошою відправна точкою можна встановити число станів як максимальне серед числа станів предків і нащадків і потім видаляти стани, якщо вони матимуть в остаточному випадку досить низькі ймовірності.

Для урахування структурно-параметричних невизначеностей і адекватного опису причинно-наслідкових зв’язків та можливих варіантів розвитку подій під впливом різних груп детермінованих та випадкових чинників моделі у формі мереж Байєса, запропонована методика, представлена у табл. 4.3.

Особливістю методики є використання ймовірнісно-статистичних методів для створення та застосування уніфікованих за структурою моделей у просторі станів. Для цього спочатку вирішується задача оцінювання елементів структури і параметрів моделі. Структура моделі оцінюється на підставі дослідження закономірностей протікання процесу, застосування статистичних тестів для перевірки наявності нелінійності, інтегрованості, гетероскедастичності, аналізу кореляційних функцій, візуального аналізу даних. При цьому вибирається кілька найбільш ймовірних структур моделей-кандидатів. Потім обчислюються оцінки параметрів моделей-кандидатів і

обирається краща з них, використовуючи відповідні статистичні характеристики адекватності моделей.

Таблиця 4.3 – Методика застосування МБ для опрацювання коротких часових рядів та таких, що містять неповні чи спотворені дані

Етап	Пропоновані методичні підходи
Попередня підготовка даних	Збір інформації, попередня обробка даних, підготовка даних до аналізу
	Обробка пропусків, пошук аномалій
	Виявлення закономірностей у досліджуваних процесах, зв’язків у даних
	Зменшення простору вхідних змінних
Моделювання причинно-наслідкових зв’язків	Етап 1. Скорочення розмірності задачі моделювання. Етап 2. Масштабування розподілів даних з метою їх приведення до зручної для подальшого використання форми й дискретизація неперервних змінних. Етап 3. Формулювання семантичних обмежень. Етап 4. Пошук структур моделей-кандидатів. На цьому кроці з множини можливих структур моделей вибирається декілька кращих моделей-кандидатів за допомогою відповідних критеріїв якості та оцінюються їх параметри. Етап 5. На цьому етапі виконується порівняння характеристик ймовірнісних моделей-кандидатів з метою вибору кращої для опису досліджуваного процесу
Етап	Пропоновані методичні підходи
Прогнозування на основі адаптивного підходу з комбінованим використанням регресійних та ймовірнісно-статистичних моделей у формі мереж Байєса	Етап 1. Представлення адитивної мережної моделі: Етап 2. Представлення вимірів незалежних змінних у вигляді вектора Етап 3. Представлення адитивної моделі із використанням довільних функцій Етап 4. Формування динамічної мережевої моделі, за якою буде обчислюватись прогноз. Головною особливістю двокрокової методики моделювання є те, що параметри декомпозиції обчислюються повторно після отримання нових вимірів. За даної методики змінна *Y(k)* залежить від вимірів незалежних змінних у часі. Етап 5. Обчислення умовної ймовірності змінної *Y(k)* з використанням адитивної декомпозиції ймовірнісної моделі Етап 6. Отримання остаточного результату. На основі моделі розглянутого типу за узагальненим алгоритмом формується ймовірнісний висновок.
Алгоритм ймовірнісного висновку	1 Виконується адитивна декомпозиція мережі Байєса на окремі складові загальної мережі. 2 Вивід окремих підмножин вузлів основної моделі виконується за L–S алгоритмом. 3 Для кожної підмножини (кліки) вузлів обчислюється спільний розподіл ймовірностей. З цією метою обчислюються ймовірності кількість значень категорійної змінної.

Крім того, застосування пропонованого підходу доцільно застосовувати для оцінювання невідомих параметрів моделей, при побудові яких опрацьовуються короткі часові ряди даних.

Проблема знаходження умовних ймовірностей або навчання параметрів прихованих вершин, маючи відому топологію самої мережі, була досліджена в наукових роботах [203, 208]. Так, однією з перших робіт в даному напрямку є робота Голмарда Дж. та Малета А. [226], в якій описано алгоритм побудови мереж деревовидної структури. Пізніше був досліджений загальний випадок для направлених ациклічних графів [208]. В даних роботах описується алгоритм максимізації математичного очікування (EM-алгоритм [197]) для ймовірнісних мереж. EM-алгоритм, як і градієнтний спуск, знаходить локальний максимум на множині правдоподібності, що визначена параметрами множини. Деякі дослідники зазначали, що існують певні проблеми із застосуванням алгоритму EM для вирішення цієї проблеми, і запропонували градієнтний спуск як можливу альтернативу. В даний час проводяться активні дослідження щодо для прямого порівняння обох методів. Серед інших методів розв'язання проблеми навчання БМ з прихованими вершинами та відомою структурою варто виділити застосування вибірки Гіббса [225] та послідовного оновлення [214].

Процес навчання БМ складається із зовнішнього циклу, що є процесом пошуку структури, і внутрішнього циклу, що є процесом оптимізації параметрів. У разі повних даних внутрішній цикл виконується дуже швидко, оскільки при цьому досить лише витягнути інформацію про умовні частоти з набору даних. Якщо ж є приховані змінні, то для виконання внутрішнього циклу може потрібно застосувати багато ітерацій алгоритму EM або алгоритму на основі градієнта, а кожна ітерація буде обчислювати розподіли апостеріорних ймовірностей в байєсовій мережі, що є NP–складною задачею.

Коли деякі вершини БМ приховані а структура мережі відома, то можна застосовувати EM (expectation maximization) алгоритм для

знаходження локальної оптимальної оцінки максимальної правдоподібності параметрів. EM-алгоритм або алгоритм максимізації математичного очікування був запропонований в 1977 році в роботі [197] і адаптований для навчання БМ з прихованими вершинами та відомою структурою [197, 201].

Алгоритм складається з двох базових кроків E (expectation) та M (maximization) . Головна ідея алгоритму полягає в тому що, якби були відомі значення всіх вершин, то навчання (на кроці M) було б простим, оскільки ми мали би всю необхідну інформацію. Тому спочатку на кроці E робиться обчислення значення очікуваної правдоподібності (expectation of the likelihood) включаючи латентні змінні, так ніби ми мали можливість їх спостерігати. Фактично робиться "доповнення" даних для прихованих змінних на основі поточних оцінок параметрів мережі $\theta^{(i)}$, тобто знаходиться розподіл апостеріорних ймовірностей по прихованих змінних за наявності отриманих даних. На кроці M робиться обчислення значення максимальної правдоподібності (maximum likelihood estimates) параметрів, використовуючи максимізацію значень очікуваної правдоподібності отриманих на кроці E. Далі алгоритм знову виконує крок E з використанням параметрів $\theta^{(i+1)}$ отриманих на кроці M і так далі [228].

Нехай $X = \{X_1,...,X_n\}$ - вершини БМ, $\Theta = \{\theta_{ijk}\}, 1 \le i \le n, 1 \le j \le q_i, 1 \le k \le r_i$ - множина таблиць умовних ймовірностей $P(X_i = x_k \mid Pa(X_i) = pa_j)$ (де q_i та r_i кількість станів вершини X_i та її предків $Pa(X_i)$ відповідно), $D = \{X_i^l\}, 1 \le i \le n, 1 \le l \le N$ - множина даних, а D_m - дані прихованих змінних, які відсутні в навчальній вибірці. Нехай $\log P(D \mid \Theta)$ - логарифмічна функція правдоподібності параметрів при заданому наборі даних. Встановивши певні початкові значення Θ^* можна обчислити ймовірнісний розподіл для прихованих змінних $P(D_m \mid \Theta^*)$ і потім обчислити $Q(\Theta : \Theta^*)$ - математичне очікування попередньої логарифмічної функції правдоподібності [228] (4.11):

$$Q(\Theta : \Theta^*) = E_{\Theta^*} \log P(D \mid \Theta), \qquad (4.11)$$

Рівняння (4.11) може бути записане в наступному вигляді (4.12) [228]:

$$Q(\Theta:\Theta^*)=\sum_i\sum_{X_i=x_k}\sum_{P_i=pa_j}N^*_{ijk}\log P_{\theta_{ijk}}(D), \tag{4.12}$$

де $N^*_{ijk}=E_{\Theta^*}[N_{ijk}]=N\cdot P(X_i=x_k,P_i=pa_j\mid\Theta^*)$ отримується за допомогою ймовірнісного висновку в мережі Байєса.

Принцип ЕМ-алгоритму зображений на рис. 4.5.

$1: i=0$

вибір Θ^0

$2:$ *repeat*

$3:$ $\quad i=i+1$

$4:$ $\quad \Theta^i=\arg\max_{\theta} Q(\Theta:\Theta^{i-1})$

$5:$ *until* $\quad Q(\Theta^i:\Theta^{i-1})\le Q(\Theta^{i-1}:\Theta^{i-1})$

Рисунок 4.5 – Схема алгоритму ЕМ

Таким чином, на кроці E оцінюється значення N^* на основі параметрів Θ^i, а на кроці M вибирається найкраще значення параметрів Θ^{i+1}, максимізуючи Q.

В роботі [197] було доведено, що алгоритм сходиться, а також факт того, що не обов'язково буде знайдено глобальний оптимум функції $Q(\Theta:\Theta^*)$. Наявність локальних оптимумів є серйозним недоліком даного алгоритму. Одне з рішень полягає в накладенні розподілів апріорної вірогідності на параметри моделі і застосуванні версії MAP (Maximum a posteriori) алгоритму ЕМ. Ще одне рішення полягає в перезапуску обчислень з новими випадково вибраними параметрами. Іноді має також сенс вибирати для ініціалізації параметрів більш обґрунтовані значення.

На основі даного розглянутого принципу ЕМ-алгоритму можливо вивести певні його варіанти і удосконалення. Наприклад, дуже часто Е-крок

(обчислення розподілів апостеріорних ймовірностей по прихованих змінних) є трудомістким при використанні великих байєсових мереж. В цьому випадку можна застосувати наближений Е-крок і алгоритм навчання буде залишатись ефективним. При використанні алгоритму формування вибірки МСМС процес навчання стає повністю інтуїтивно зрозумілим: кожен стан (конфігурація прихованих і спостережуваних змінних), який відвідав алгоритм МСМС, розглядається точно так, ніби він був повним спостереженням. Тому параметри можуть оновлюватися безпосередньо після кожного переходу із стану в стан в алгоритмі МСМС. Для визначення за допомогою навчання параметрів дуже великих мереж показали свою ефективність і інші форми наближеного ймовірнісного висновку, такі як варіаційні і циклічні методи.

Випадок, коли розглядається БМ з прихованими вершинами та невідомою структурою - є найскладнішим випадком навчання БМ.

Структурне навчання полягає в знаходженні правильних зв’язків між вершинами. Більш цікавою проблемою є створення прихованих вершин. Приховані вершини можуть зробити модель більш компактною. Стандартний підхід полягає в додаванні щоразу по одній прихованій вершині до певної частини мережі, виконуючи структурне навчання на кожному кроці, поки оцінка моделі не перестане зменшуватись. Ще одна проблема – це вибір можливої кількості прихованих вершин та виду розподілу їх умовних ймовірностей. Разом з цим стоїть задача вибору місця створення прихованої вершини. Немає сенсу робити її нащадком, так як приховані нащадки завжди можуть бути скорочені. Тому необхідно знайти існуючу вершину, яка потребує нового предка у випадку, коли поточна множина можливих предків не є адекватною.

Одним з підходів може бути дослідження умовної ентропії $H(X \mid Pa(X))$: якщо вона дуже велика, то це значить, що поточна множина не може «пояснити» залишкову ентропію; якщо $Pa(X)$ найкраща множина предків (в значенні BIC (Bayesian Information Criteria) або хі-квадрат), яку

можливо знайти в поточній моделі, то це передбачає необхідність створення нової вершини і включення її до множини $Pa(X)$.

Для розв'язання завдання структурного навчання БМ з прихованими вершинами використовують структурний ЕМ-алгоритм або алгоритм стиснення границь. Одним з удосконалень вищеописаного принципу максимізації математичного очікування є структурний ЕМ-алгоритм [197, 228] (SEM - structural EM algorithm), який відрізняється від параметричного алгоритму ЕМ тим, що цей алгоритм може оновлювати не лише параметри, але і структуру. Схему алгоритму SEM зображено на рис. 4.6.

$1: i = 0$

вибір початкової мережі (G^0, Θ^0)

$2:$ *repeat*

$3: \quad i = i + 1$

$4: \quad (G^i, \Theta^i) = \arg\max_{G,\Theta} Q(G, \Theta : G^{i-1}, \Theta^{i-1})$

$5:$ *until* $Q(G^i, \Theta^i : G^{i-1}, \Theta^{i-1}) \le Q(G^{i-1}, \Theta^{i-1} : G^{i-1}, \Theta^{i-1})$

Рис. 4.6 Принцип структурного ЕМ-алгоритму

Даний алгоритм навчає мережі на основі штрафних ймовірнісних значень, які включають значення, отримані за допомогою байєсівського інформаційного критерію (BIC), принципу мінімальної довжини (MDL) [288], а також інших критеріїв. Як і в звичайному алгоритмі ЕМ використовуються поточні параметри для обчислення очікуваних кількостей в Е-кроці, а потім ці кількості застосовуються в М-кроці для вибору нових параметрів, так і в алгоритмі структурного ЕМ використовується структура для обчислення очікуваних кількостей, після чого ці кількості застосовуються в М-кроці для оцінки правдоподібності потенційних нових структур (у цьому полягає відмінність цього методу від методу зовнішнього циклу/внутрішнього циклу, в якому обчислюються нові очікувані кількості для кожної потенційної структури). Крок 4 даного алгоритму передбачає

пошук найкращої структури та найкращого набору параметрів для даної структури. На практиці ці кроки чітко відрізняються(4.13):

$$G^i = \arg\max_G Q(G, \Theta^i : G^{i-1}, \Theta^{i-1}), \tag{4.13}$$

$$\Theta^i = \arg\max_\Theta Q(G^i, \Theta : G^{i-1}, \Theta^{i-1}), \tag{4.14}$$

Перший пошук (4.13) в просторі можливих графів повертає нас до вихідної проблеми пошуку оптимальної структури БМ. Щодо пошуку в просторі параметрів (4.14), в роботі [197] було запропоновано повторювати цю процедуру декілька разів з використанням інтелектуальної ініціалізації. Даний крок означає запуск параметричного ЕМ-алгоритму для кожної структури G^i, починаючи зі структури G^0.

Таким чином, структурний алгоритм ЕМ дозволяє вносити до мережі декілька структурних змін без яких-небудь повторних обчислень очікуваних кількостей і має здатність визначати в процесі навчання нетривіальні структури байєсових мереж. Проте необхідно виконати ще дуже багато роботи, перш ніж можна буде стверджувати, що завдання визначення структури в процесі навчання остаточно вирішене.

Структурний ЕМ-алгоритм дає можливість навчання параметрів і структури БМ з визначеним числом прихованих змінних. На жаль, проблема визначення кількості прихованих вершин, які необхідно додавати до структури мережі, на сьогодні не є остаточно вирішеною. Тому існує гостра необхідність в створенні додаткового механізму для розв'язання цієї задачі.

На основі вищерозглянутого алгоритму максимізації математичного очікування автором було розроблено методику обчислення параметрів БМ за умови неповної вхідної інформації і відомої топології мережі. Отримана методика є комплексною системою, яка повністю описує весь процес

знаходження невідомих параметрів прихованих вершин: Дана методика передбачає обчислення параметрів мережі Байєса за умови неповної вхідної інформації і відомої топології мережі, є комплексною системою, яка повністю описує весь процес знаходження невідомих параметрів прихованих вершин і складається з таких кроків:

Крок 1. Побудова мережі Байєса за навчальними даними або «вручну». Необхідно визначити топологію та знайти значення параметрів всієї мережі. Можливі два варіанти: коли структура мережі нам відома, і коли є лише навчальні дані. В другому випадку побудова структури здійснюється за два кроки: на першому кроці будується топологія мережі з використанням, наприклад, евристичного алгоритму, а на другому – визначаються параметри мережі, які максимально правдоподібні до навчальних даних.

Крок 2. Генерування вибірки за заданою структурою мережі (застосовується у випадку відсутності навчальних даних). У випадку, коли навчальні дані не задані або їх недостатньо, відбувається псевдовипадкове генерування вибірки за побудованою на першому етапі структурою мережі. Генерація відбувається наступним чином. Спочатку формується ймовірнісний висновок у мережі без інстанційованих вершин $P(S_{ij})$, $i = 1,\ldots, N$; $j = 1,\ldots,S$. Далі обирається вершина N_{i^*} і інстанціюється один із її станів S_{ij} із ймовірністю цього стану $P(S_{ij})$. Перераховуються ймовірності станів вершин після інстанціювання $P(S_{ij} \mid S_{i^*} = S_{i^* j^*})$. Далі обирається наступна вершина. Ця операція повторюється доки не залишаться неінстанційовані вершини. Інстанційовані стани утворюють запис у вибірці $(S_{i_1} \ldots S_{i_N}), i_k \in (1,\ldots,S)$. Дії повторюється доки не буде створено необхідної кількості записів у вибірці.

Крок 3. Додавання прихованих вершин до структури мережі. Додавання прихованої вершини до мережі. У випадку, коли вершина вставляється між декількома існуючими, то видаляються попередні дуги між ними і створюються нові, які пов’язані з прихованим вузлом. Якщо

необхідно додати приховану батьківську вершину, то просто створюється вершина і відповіді дуги.

Крок 4. Початкова ініціалізація невідомих параметрів мережі. Параметри прихованих вершин ініціалізуються початковими значеннями. Це можуть бути як випадково згенеровані значення, так і значення, що вказані експертами.

Крок 5. Обчислення параметрів мережі на основі згенерованих даних з використанням ЕМ-алгоритму. На даному етапі запускається ітераційний процес ЕМ-алгоритму, який використовує попередньо згенеровану вибірку даних і результатом якого є оцінки невідомих параметрів прихованих вершин мереж Байєса.

Отже, на першому етапі необхідно побудувати БМ, тобто визначити топологію та знайти значення параметрів всієї мережі. На цьому етапі можливі два варіанти, коли структура мережі нам відома і тоді залишається її перенести в програмне середовище і заповнити ТУЙ, або коли є лише навчальні дані. В другому випадку побудова структури здійснюється в два кроки: на першому кроці будується топологія мережі з використанням, наприклад евристичного алгоритму, а на другому – знаходяться параметри мережі, які максимально правдоподібні до навчальних даних.

На другому етапі у випадку, коли навчальні дані не задані або їх недостатньо, відбувається псевдовипадкова генерація вибірки за побудованою на першому етапі структурою мережі.

Генерація відбувається наступним чином. Спочатку робиться ймовірнісний висновок у мережі без інстанційованих вершин $P(S_{ij}), i = 1,...,N; j = 1,...,S$. Далі обирається вершина N_{i^*} і інстанціюється один із її станів $S_{i^*j^*}$ із ймовірністю цього стану $P(S_{i^*j^*})$. Перераховуються ймовірності станів вершин після інстанціювання $P(S_{ij} \mid S_{i^*} = S_{i^*j^*})$. Далі обирається наступна вершина. Ця операція повторюється доки не залишиться неінстанційованних вершин. Інстанційовані стани утворюють запис у вибірці $(S_{i_1}..S_{i_N}), i_k \in (1,...,S)$.

Алгоритм повторюється доки не буде створено необхідної кількості записів у вибірці [228].

На третьому етапі відбувається додавання прихованої вершини до мережі. Якщо ця вершина вставляється між декількома існуючими, то видаляються попередні дуги між ними і створюються нові, які пов'язані з прихованим вузлом. Якщо необхідно додати приховану батьківську вершину, то просто створюється вершина і відповіді дуги.

На наступному етапі параметри прихованих вершин ініціалізуються початковими значеннями. Це можуть бути як випадково згенеровані значення, так і значення, що вказані експертами.

На завершальному етапі запускається ітераційний процес алгоритму EM, який використовує попередньо згенеровану вибірку даних і результатом якого є оцінки невідомих параметрів прихованих вершин мереж Байєса.

4.3 Методи оцінювання якості побудови ймовірнісного висновку

Для оцінювання якості роботи будь-якого методу побудови ймовірнісного висновку можуть скористатися такими величинами: середньоквадратична похибка, KL-відстань або квадратична відстань Хелліндже́ра (Hellinger) [204, 228].

Середньо квадратична похибка MSE (4.15):

$$MSE = \frac{1}{\sum_{X^{(i)} \in X \setminus E} card\left(A^{(i)}\right)} \cdot \sum_{X^{(i)} \in X \setminus E} \left(\sum_{\forall x^{(i)} \in X^{(i)}} \left(P(x^{(i)} | e) - \hat{P}(x^{(i)} | e) \right)^2 \right) \quad (4.15)$$

KL-відстань D_K між значенням ймовірності $P(X^{(i)}|e)$ та оцінкою $\hat{P}(X^{(i)}|e)$ (4.16) [204, 228]:

$$D_K\left(P(X^{(i)}|e), \hat{P}(X^{(i)}|e)\right) = \sum_{\forall x^{(i)} \in A^{(i)}} \left(P(x^{(i)}|e) \cdot \log\left(\frac{P(x^{(i)}|e)}{\hat{P}(x^{(i)}|e)} \right) \right). \quad (4.16)$$

KL-відстань D_K всієї БМ (4.17) [204, 228]:

$$D_K\left(P,\hat{P}\right)=\frac{1}{card\left(X\setminus E\right)}\cdot\sum_{X^{(i)}\in X\setminus E}D_K\left(P(X^{(i)}|e),\hat{P}(X^{(i)}|e)\right). \quad (4.17)$$

Квадратична відстань Хелліндже ра D_H між значенням ймовірності $P(X^{(i)}|e)$ та оцінкою $\hat{P}(X^{(i)}|e)$ (4.18) [204, 228]:

$$D_H\left(P(X^{(i)}|e),\hat{P}(X^{(i)}|e)\right)=\sum_{\forall x^{(i)}\in A^{(i)}}\left(\sqrt{P(x^{(i)}|e)}-\sqrt{\hat{P}(x^{(i)}|e)}\right)^2 \quad (4.18)$$

Квадратична відстань Хелліндже ра D_H всієї БМ (4.19) [204, 228]:

$$D_H\left(P,\hat{P}\right)=\frac{1}{card\left(X\setminus E\right)}\cdot\sum_{X^{(i)}\in X\setminus E}D_H\left(P(X^{(i)}|e),\hat{P}(X^{(i)}|e)\right). \quad (4.19)$$

В наведених формулах $X=\{X^{(1)},...,X^{(N)}\}-$ множина всіх вершин графа БМ G, де кожна j-а вершина мережі ($j=1,...,N$) має $A^{(j)}=\{0,1,...\alpha^{(j)}-1\}$ ($\alpha^{(j)}\geq 2$) станів; запис $card(A^{(j)})$ означає потужність множини $A^{(j)}$ (кількість елементів з яких складається множина); $E\subset X, E=e-$ множина подій (інстанційовані вершини); $P(X^{(i)}|e)-$ значення ймовірності вершини $X^{(j)}$ за умови відбуття події $E=e$, $\hat{P}(X^{(i)}|e)-$ значення оцінки ймовірності.

4.4 Огляд програмних засобів, що можуть бути використані при створенні інформаційної технології прогнозування із використанням мереж Байєса

Використання БМ для аналізу ННП дозволяє враховувати та використовувати будь-які вхідні дані – експертні оцінки і статистичну інформацію. Перевагами використання БМ є представлення взаємодії між факторами процесу у вигляді причино-наслідкових зв’язків, завдяки чому

досягається максимально високий рівень візуалізації і, як наслідок візуалізації, чітке розуміння суті взаємодії факторів процесу між собою.

Саме це відрізняє БМ від інших методів, що застосовуються для поудови моделей ННП. Також перевагами БМ є можливості врахування невизначеностей статистичного, структурного і параметричного характеру, побудова моделей за наявності прихованих вершин і при неповних спостереженнях, а також формування висновку за допомогою різних методів: наближених і точних.

BayesiaLab [228] дуже зручна у використанні програма, розроблена французькою компанією Bayesia SA. В програму інтегровано декілька методів для автоматичної побудови БМ та ймовірнісного висновку, особливий інтерес викликає швидкий та зручний метод SopLeq для побудови БМ. Ціна програми залежить від типу ліцензії та комплектації, коливається від 190 € за найдешевшу академічну версію до 88000 € за повну ліцензію на професійну версію. Існує можливість користування програмою безкоштовно протягом місяця, але в цьому випадку немає можливості зберігання результатів роботи.

Програмний продукт Netica [228] розроблений компанією Norsys Software Corporation, Ванкувер (Канада). Norsys спеціалізується в розробці програмного забезпечення для байєсових мереж. Програма Netica – основне досягнення компанії, розробляється з 1992 року і стала комерційно доступною в 1995 році. Нині Netica є одним з найширше використовуваних інструментів для розробки БМ. Використовується для побудови БМ експертами та імовірнісного виводу, але в програмі відсутня можливість автоматичної побудови БМ за навчальними даними. Вартість академічної версії програми складає $285, а комерційної $685. Також доступна безкоштовна версія програми, але в цьому випадку максимально можливий розмір БМ для роботи та зберігання обмежується 15 вершинами, а максимально можливий розмір вибірки даних – 999 записами.

Hugin-Expert [228] розроблена компанією Hugin Expert A/S, Ольсборг (Данія). Їх основний продукт Hugin почав створюватися під час робіт за проектом ESPRIT для проблеми діагностики нервово-м'язових захворювань. Потім почалася комерціалізація результатів проекту і основного інструменту - програми Hugin. До теперішнього моменту Hugin адаптована в багатьох дослідницьких центрах компанії в 25 різних країнах, вона використовується у ряді різних областей, пов'язаних з аналізом рішень, підтримкою прийняття рішень, прогнозированням, діагностикою, управлінням ризиками та оцінками безпеки технологій.

Hugin є програмною реалізацією системи підтримки прийняття рішень на основі БМ і дві версії Pro і Explorer. Система використовує два основні режими роботи: режим редагування і побудови мережі Байєса, а також заповнення таблиць умовних ймовірностей (ТУЙ) та побудови ймовірнісного висновку.

Ціна програми в залежності від типу ліцензії від 6426 до 50999 датських крон. Також є можливість користування безкоштовно версією програми, але в цьому разі немає можливості автоматичної побудови БМ.

Програма Bayes Net Toolbox for Matlab [228] розроблена професором Кевіном Мерфі для системи Matlab; програма надається безкоштовно. Всі модулі програми доступні користувачу. Існує можливість доповнювати програму своїми методами та модулями. Але платою за цю свободу та можливості є необхідність глибокого знання пакету Matlab та повне розуміння теорій БМ. Серед недоліків також треба відзначити відсутність зручного графічного інтерфейсу та можливсоті візуалізації результатів.

MSBNx [228] - ,безкоштовна програма розроблена компанією MicroSoft. Основний недолік відсутність можливості автоматичної побудови БМ за навчальними даними.

GeNIe [228] - безкоштовна програма розроблена лабораторією СППР Пітсбургського Університету. Дуже зручна в використанні, є можливість

автоматичної побудови БМ, а також підключення власних модулів з алгоритмами написаними на мові програмування С.

Програмне забезпечення Bayesware Discoverer [228] розроблене англійською компанією Bayesware. Серед особливостей варто відмітити автоматичну побудову БМ за навчальними даними, симуляційне порівняння декількох моделей на основі БМ, використання алгоритму Bound and Collapse для випадку неповних даних, заповнення ТУЙ та побудова ймовірнісного висновку, робота з категоріальними даними та неперервними змінними. Програма є платною. Вартість ліцензії коштує від $3000 до $8500. Існує можливість безкоштовного використання програми протягом 30 днів без будь-яких обмежень функціональності.

4.5 Інформаційна технологія застосування мереж Байєса у прогнозному моделюванні

В роботі представлено практичну реалізацію інформаційної технології вирішення задачі прогнозування нестаціонарного процесу – інвестиційної діяльності в економіці регіону. Актуальність задачі полягає в тому, що інвестиційне забезпечення є одним із важливих джерел розвитку економіки регіону [237, 174, 173] .

Дослідження проведене на матеріалах Черкаської області, яка є лідером з виробництва продукції сільського господарства (7% від валової продукції сільського господарства України), але за рейтингом інвестиційної привабливості тривалий час не піднімається вище 13 місця (у 2004 рік) серед інших областей. Аналогічна ситуація спостерігається і в таких традиційно аграрних регіонах, як Вінницька, Полтавська, Сумська, Хмельницька, Херсонська областях [173].

Для побудови математичних моделей використано соціально-економічні показники Черкаської області за 2000 – 2011 рр. [173, 174].

Таблиця 4.3 – Змінні що використовувалися для моделювання та їх опис

Змінна	Опис змінної	Одиниця виміру
Y	валовий регіональний продукт (у фактичних цінах)	млн. грн.
x01	інвестиції в основний капітал (у фактичних цінах)	млн. грн.
x02	основні засоби (у фактичних цінах, на кінець року)	млн. грн.
x03	обсяг інноваційних витрат в промисловості	тис. грн.
x04	індекс споживчих цін (грудень до грудня попереднього року)	відсотків
x05	обсяг реалізованої промислової продукції (робіт, послуг) (у фактичних цінах)	млн. грн.
x06	продукція сільського господарства (у порівняних цінах 2005 року)	млн. грн.
x07	обсяг реалізованої будівельної продукції (у фактичних цінах)	млн. грн.
x08	доходи населення	млн. грн.
x09	рівень зареєстрованого безробіття (на кінець року)	відсотків
x10	середньорічна кількість найманих працівників	тис.
x11	середньорічна номінальна заробітна плата	грн.
x12	Зовнішня торгівля товарами, експорт	млн. дол. США
x13	Зовнішня торгівля товарами, імпорт	млн. дол. США
x14	прямі іноземні інвестиції в область	тис. дол. США
x15	роздрібний товарообіг підприємств (у фактичних цінах)	млн. грн.
x16	відправлення (перевезення) вантажів усіма видами транспорту	млн. т.
x17	відправлення (перевезення) пасажирів транспортом загального користування	млн. осіб.

Для побудови моделей також буде використовуватися змінна $Y1$, що являє собою валовий регіональний продукт (ВРП) за попередній рік, тобто $Y1(k) = Y(k-1)$.

Одна з головних проблем при побудові математичних моделей соціально-економічних процесів – це наявність достатнього об’єму

статистичних даних для виявлення закономірностей, виявлення адекватної структури та обчислення оцінок параметрів моделі. Існуюча система збору та накопичення статистичної інформації, обмеження щодо конфіденційності окремих показників, адаптація статистичної методології до міжнародних стандартів ускладнюють формування достатньо великих наборів рядів часових даних відповідних показників. Це призводить до того, що при побудові адекватних математичних моделей для регіонального рівня, необхідно ширше застосовувати сучасні інтелектуальні засоби, спроможні працювати та виявляти закономірності на коротких часових вибірках даних.

Проблема побудови математичних моделей на коротких наборах даних в статистиці більш відома під назвою – "прокляття розмірності". Для вирішення цієї проблеми в загальному випадку використовують статистичні параметри і методи попереднього відбору найбільш значущих змінних процесу, а саме: критерієм R-квадрат, мережею Байєса та методом головних компонент. У зв'язку з тим, що навчальна вибірка даних складається лише з десяти точок, а змінних аналізу, що можуть приймати участь в побудові моделі сімнадцять для відбору найбільш значимих пропонується обирати ті які надають кращі значення за критерієм R-квадрат [35, 126]:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)},$$

де $\mathrm{var}(\hat{y})$ – дисперсія залежної змінної, оціненої за допомогою побудованої математичної моделі, а $\mathrm{var}(y)$ – вибіркова дисперсія залежної змінної, оціненої за допомогою її фактичних значень; для адекватної моделі $R^2 \to 1$.

Побудована за цим підходом модель множинної регресії має вигляд:

$$y(k) = 1215 - 2{,}8428 \cdot x_{06}(k-1) + 3{,}8722 \cdot x_{07}(k-1) - 0{,}000923 \cdot x_{08}(k-1) + \\ + 23{,}2503 \cdot x_{10}(k-1) + 52{,}8486 \cdot x_{11}(k-1) - 12{,}1045 \cdot x_{15}(k-1)$$

.Для побудованої моделі середня абсолютна похибка в процентах (MAPE) дорівнює 3,23%. Формула для обчислення *MAPE* має вигляд:

$$MAPE = \frac{100}{n} \cdot \sum_{i=1}^{n} \left| \frac{y - \hat{y}}{y} \right|,$$

де n – це кількість вимірів, y – реальне значення, а $\hat{y}$ – прогнозне значення.

Одним з досить корисних підходів є використання методу головних компонентів (МГК), що був запропонований К. Пірсоном у 1901 році [108]. Суть методу полягає у тому, що визначається набір ортогональних векторів, які краще всього описують напрям зміни в даних у порівнянні з вхідними даними у початковому вигляді.

У загальному випадку фахівці радять використовувати для аналізу таку кількість компонент, щоб сума їх власних чисел була більшою за 80% або 90% кількості вхідних змінних. Тобто для заданої множини даних достатньо перших трьох компонент тому, що їх сума дорівнює 15,02 і пояснює 83,7% змін у даних.

В рамках виконаних досліджень пропонується використовувати таку кількість компонентів, що мінімізують середньоквадратичну похибку (RMSE) моделі у вигляді регресії при включенні відповідного компоненту.

Таблиця 4.4 Список компонентів для методу головних компонент

Номер компоненти	Власне значення компоненти	Процент пояснення змін в даних відповідним компонентом	Кумулятивний процент пояснення змін в даних	Значення RMSE при включенні компоненти в регресійну модель
PCA1	11,27	62,61%	62,61%	1677
PCA 2	2,14	11,9%	74,52%	1809
PCA 3	1,65	9,18%	83,7%	1842
PCA 4	0,98	5,48%	89,17%	1915
PCA 5	0,86	4,82%	94%	2192
PCA 6	0,71	3,98%	97,98%	2660
PCA 7	0,26	1,47%	99,44%	3247
PCA 8	0,1003	0,56%	100%	3247

В табл. 4.4 наведено побудовані компоненти; як можна побачити, найменша похибка прогнозування ВРП за критерієм RMSE досягається при використанні лише першого компоненту.

Побудована за цим підходом модель має вигляд:

$$y(k) = 11832 + 1833{,}3 \cdot PCA_1(k).$$

Для побудованої моделі MAPE дорівнює 7,21%.

Мережі Байєса досить широко використовуються для визначення причинно-наслідкових зв'язків при моделюванні процесів, що описуються великою кількістю факторів [213, 218, 220, 223]. Саме тому для пошуку зв'язків між ВРП та іншими економічними показниками було використано мережу Байєса. Для подальшого використання при прогнозуванні в рамках виконаного дослідження запропонована методика, що поєднує у собі такі математичні підходи – мережі Байєса і регресійний аналіз.

Запропонована двоетапна методика складається з таких кроків.

Етап 1. Будується топологія мережі Байєса, що надає інформацію, необхідну для виявлення причинно-наслідкових зв'язків між змінними та силу зв'язків між ними.

Крок 1. Обчислюються відхилення між поточним та попереднім значенням для всіх змінних. Це робиться для того щоб позбутися трендової складової в змінних, оскільки ця операція в грубому наближенні еквівалентна знаходженню першої похідної.

Крок 2. Знайдені на попередньому кроці відхилення перетворюються у номінальні змінні завдяки використанню квантильного перетворення. Для даної задачі було задано чотири квантилі.

Крок 3. Для побудови топології мережі використано програмне забезпечення SAS/IML з такими налаштуваннями [99]:

– метод побудови MDL [99];

– максимальна кількість предків – 5;

– критерій оптимізації – мінімальна ентропія.

Етап 2. Прогнозування на основі побудованої причинно-наслідкової структури мережі Байєса.

На основі побудованої топології мережі Байєса визначаються найбільш значущі змінні, що впливають на цільову, після чого будується рівняння множинної регресії із примусовим включенням до моделі виявлених змінних. Оцінка параметрів моделі виконується на основі рекурсивного методу найменших квадратів.

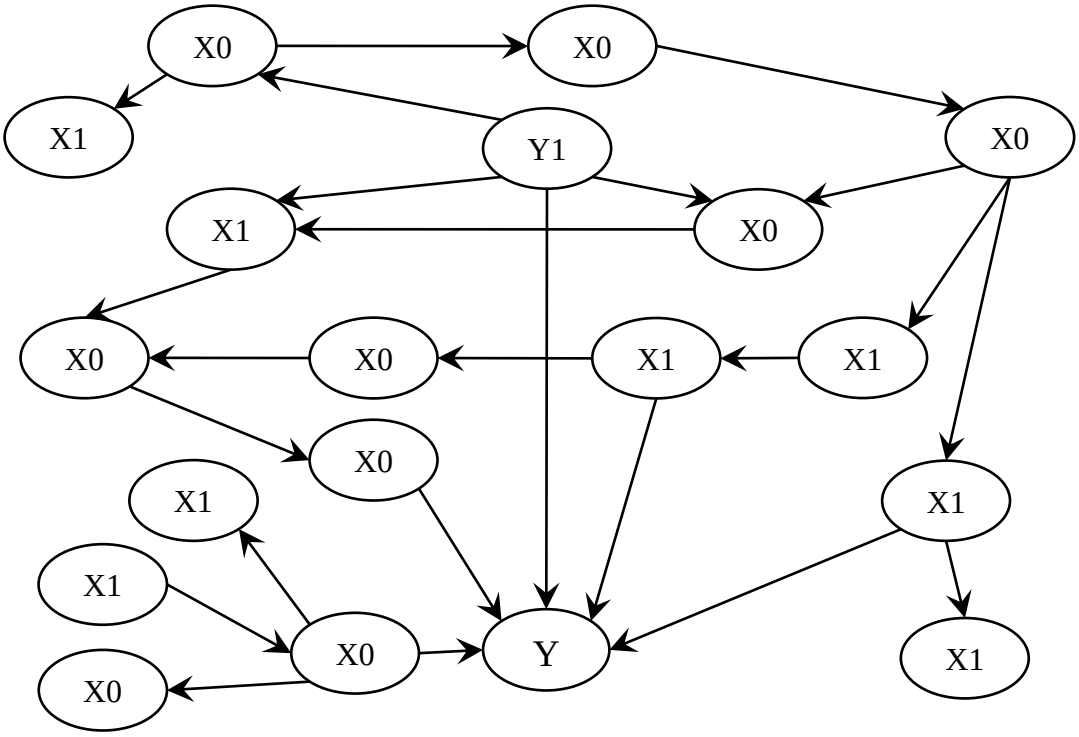


Рисунок 4.7 – Топологія побудованої мережі Байєса для дослідження розвитку системи інвестиційної діяльності Черкаської області

Найбільш значущими для подальшого прогнозування виявилися показники обсягу реалізованої промислової продукції, прямих іноземних інвестицій, відправлення (перевезення) вантажів, валової продукції сільського господарства та значення валового регіонального продукту за попередній рік.

Рівняння для прогнозування має вигляд:

$$y(k) = 11429{,}9 + 2{,}7017 \cdot y(k-1) - 1{,}3027 \cdot x_{05}(\text{k-1}) - 3{,}9693 \cdot x_{06}(\text{k-1}) + \\ + 38{,}8244 \cdot x_{14}(\text{k-1})$$

Для побудованої моделі MAPE дорівнює 6,4%.

В табл. 4.5 наведено прогнозні значення, оцінені на основі запропонованих математичних моделей,

Таблиця 4.5 – Реальні та прогнозні значення, отримані за побудованими моделями

Рік	Реальні історичні значення	Множинна лагова регресія за критерієм R-квадрат	Лінійна регресія на основі методу головних компонентів	Комбінована модель мережі Байєса та множинної регресії
2002	3852	3867,18	4182,49	4216,53
2003	4565	4544,54	4914,24	5357,96
2004	6623	6663,65	6771,29	6262,34
2005	9014	9046,33	9939,39	9339,17
2006	10957	10872,25	10595,96	11028,1
2007	13656	13656,16	13400,55	13336,96
2008	19101	19104,44	16591,6	16264,23
2009	18707	18668,16	21755,59	18725,04
2010	20016	20068,79	18340,37	21961,15
2011	21417	21361,75	21314.26	21304,72

В табл. 4.6 представлені статистичні характеристики моделей.

Таблиця 4.6 Статистичні характеристики математичних моделей

Статистична характеристика	Множинна лагова регресія за критерієм R-квадрат	Лінійна регресія на основі методу головних компонентів	Комбінована модель мережі Байєса та множинної регресії
AIC	80,64	135,4	110,34
MSE	85,98	1677,83	395,43
BIC	82,02	135,79	111,33
SSE	14785,91	19705502,04	625446,49

Аналіз часового ряду даних ВРП свідчить про наявність таких статистичних характеристик:

- наявність лагу першого порядку у даних (табл. 4.7);

Таблиця 4.7 Значення ЧАКФ для часового ряду ВРП

Порядок запізнення	Значення ЧАКФ
1	0,798
2	-0,136
3	-0,111
4	-0,147
5	0,006

- наявність нестаціонарності за математичним сподіванням та дисперсією.

Для урахування вище зазначених особливостей ряду побудована авторегресійна модель довгострокового прогнозування, що складається з двох рівнянь:

– рівняння тренду:

$$trend(k) == -1059{,}24 + 1867.61 \cdot \mathrm{k}$$

де k - це змінна що приймає значення від 1 до 21 (значення $\mathrm{k} = 1$ відповідає 2000 року, $\mathrm{k} = 2$ відповідає 2001 і так далі, до $\mathrm{k} = 21$, що відповідає. 2020 року).

– рівняння прогнозування відхилень від тренду:

$$dy(k) = -523.6 + 0.537 \cdot dy(k-1) + 0.00016698 \cdot x_{01}(k-1) +$$
$$+ 6.001318 \cdot 10^{-5} \cdot x_{01}(k-2) - 0.0001721 \cdot x_{01}(k-3)$$

де dy – відхилення від тренду, а x_{01} – рівень інвестицій в основний капітал.

На основі наведених рівнянь остаточний прогноз обчислюється так:

$$y(k) = trend(k) + dy(k).$$

Із використанням описаних рівнянь побудовано три сценарії зростання ВРП Черкащини – оптимістичний, помірний та песимістичний. Результати наведено на рис. 4.8.

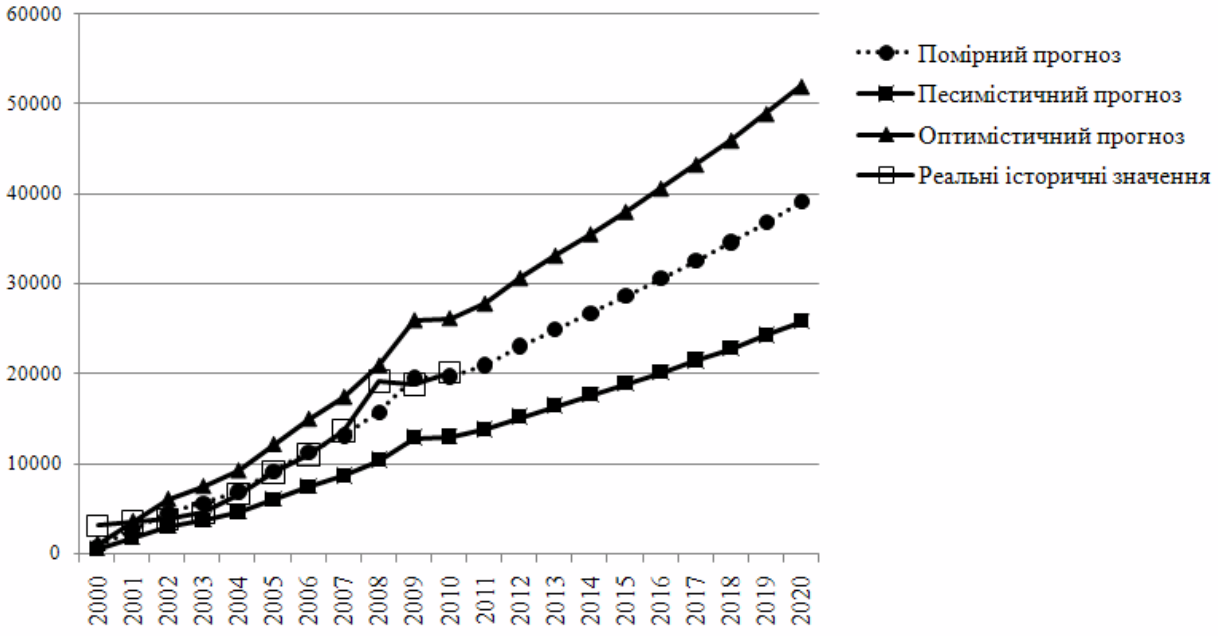


Рисунок 4.8 – Сценарії розвитку динаміки ВПР Черкаської області на період до 2020 року, млн. грн.

Як показали виконані розрахунки, побудована економіко-математична модель для довгострокового прогнозування обсягів інвестицій в основний капітал має прийнятні статистичні характеристики та забезпечує належну якість прогнозу і може бути використана під час прогнозування обсягів інвестицій в основний капітал на довгострокову перспективу [229-237].

Саме по собі зростання обсягів інвестицій не забезпечує економічне зростання, а тому, спрогнозувавши обсяги інвестицій в основний капітал, необхідно перевірити, чи знаходяться прогнозовані значення на рівні, що відповідає необхідному для простого та розширеного відтворення. Крім того, при розробці державних програм та стратегій важливим є окреслення строків, коли даний результат може бути досягнуто. Скориставшись твердженням про те, що для досягнення простого відтворення достатньо, щоб

інвестиції складали 30% від обсягу валового регіонального продукту, а для розширеного відтворення – до 40% [238].

За основу взято прогноз за помірним варіантом розвитку подій. Динаміка даного показника, співвідношення валового регіонального продукту та інвестицій в основний капітал представлені у табл.4.8.

Таблиця 4.8 – Динаміка співвідношення обсягу інвестицій та валового регіонального продукту Черкаської області у 2000-2019 р. та її прогноз

Роки	Валовий регіональний продукт, млн грн	Інвестиції в основний капітал (розраховані за моделлю), млн грн	Співвідношення інвестицій та валового регіонального продукту, %
2000	3 179	330	10
2001	3 590	431	12
2002	3 852	706, 8	18
2003	4 565	1142, 7	25
2004	6 623	2690, 6	41
2005	9 014	2408,2	27
2006	10 957	3734,8	34
2007	13 656	4808,3	35
2008	19 101	5814, 6	30
2009	18 707	3616,3	19
2010	22354	3800,18	17
2011	27012	5672,52	21
2012	31265	6878,3	22
2013	33087	8271,75	25
2014	38466	10385,82	27
2015	50843	15252,9	30
2016	59412	18417,72	31
2017	73176	24148,08	33
2018	93315	32660,25	35
2019	103514	38300,18	37
2020*	139080	55632	40

Як видно з табл. 4.8, навіть за помірним сценарієм, на період до 2015 року в область спрямовуватиметься обсяг інвестицій в основний капітал, достатній для забезпечення простого відтворення, а у 2020 році є можливим досягнення рівня розширеного відтворення за існуючої їх структури.

Вибрано змінні для аналізу регіонального розвитку Черкаської області України і побудовано прогнозуючі математичні моделі деяких процесів на основі зібраних статистичних даних. Для редукції кількості вхідних змінних використано метод головних компонент і байєсівську мережу. Зокрема, побудовано множинну лагову регресію, яка забезпечує отримання оцінок короткострокового прогнозу з MAPE = 3,23%.

З використанням методу головних компонент побудовано регресійну модель і показано, що найменша похибка прогнозування валового регіонального продукту за критерієм RMSE досягається при використанні лише першої компоненти (RMSE = 1677; MAPE = 7,21). За допомогою створеної байєсівської мережі встановлено, що найбільш значущими для подальшого прогнозування є показники обсягу реалізованої промислової продукції, прямих іноземних інвестицій, відправлення (перевезення) вантажів, валової продукції сільського господарства та значення валового регіонального продукту за попередній рік. Модель, побудована з використанням для редукції кількості змінних байєсівської мережі, має хороші прогнозуючі характеристики, зокрема MAPE = 6,4%.

Побудована авторегресійна модель для довгострокового прогнозування, що складається з двох рівнянь: рівняння тренду і рівняння для прогнозування відхилень від тренду. Отримана модель має прийнятні прогнозуючі характеристики. Використовуючи створені моделі, побудовано три сценарії зростання ВРП Черкащини – оптимістичний, помірний та песимістичний. Показано, що навіть за помірним сценарієм на період до 2015 року в область спрямовуватиметься обсяг інвестицій в основний капітал, достатній для забезпечення простого відтворення, а у 2020 році є можливим досягнення рівня розширеного відтворення.

Висновки до розділу 4

У розділі розглянуто особливості застосування ймовірнісно-статистичних моделей у формі байєсівських мереж для моделювання нелінійних нестаціонарних процесів та прогнозування їх розвитку. Перевагами таких моделей є їх можлива висока розмірність, об'єднання в одній моделі дискретних і неперервних змінних, можливість боротьби із невизначеностями ймовірнісно-статистичного характеру та ін. Також мережі Байєса надають можливість обчислювати альтернативні оцінки прогнозів у вигляді умовних ймовірностей попадання значень оцінок прогнозів у деякий інтервал, заданий користувачем.

Показано, яким чином опрацьовуються невизначеності різних типів за допомогою байєсівського підходу до моделювання. Зокрема, це невизначеності, пов'язані з неповними даними, наявністю прихованих вершин (змінних) мережі і т. ін. Запропонована методика побудови мережі за наявності прихованих вершин, яка успішно апробована на практичних прикладах аналізу фактичних даних.

Для оцінювання якості побудови ймовірнісного висновку запропонована множина відповідних критеріїв, використання яких забезпечує підвищення якості моделі та оцінок прогнозів. Сформульована відповідна методика використання запропонованих критеріїв.

Виконано огляд сучасних програмних засобів для побудови ймовірнісно-статистичних моделей у формі байєсівських мереж, які можуть бути успішно застосовані у якості елементів інформаційних технологій аналізу даних у формі нелінійних нестаціонарних процесів.

У розділі подано успішне застосування інформаційної технології аналізу даних на основі байєсівських мереж до моделювання і прогнозування економічних процесів на регіональному рівні.

РОЗДІЛ 5
ОСОБЛИВОСТІ ЗАСТОСУВАННЯ БАГАТОМОДЕЛЬНОГО ПІДХОДУ ДО ПРОГНОЗУВАННЯ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ РІЗНОЇ ПРИРОДИ

5.1 Багатомодельний підхід у прогнозуванні нелінійних нестаціонарних процесів

Зростаючі вимоги до сучасних інформаційних технологій з боку різних груп користувачів, і особливо з боку осіб, зацікавлених у підвищені ефективності управління та зниженні ризику прийняття невірного управлінського рішення, спонукають до удосконалення існуючих методів та інформаційних технологій аналізу даних, пошуку нових рішень, що можуть бути швидко адаптовані до нових умов функціонування досліджуваних процесів.

У даній роботі пропонуються різні інформаційні технології прогнозування, в основу яких покладено поетапне розкриття невизначеностей різної природи, уточнення результатів моделювання на кожному етапі дослідження. Особливістю такого підходу є те, що особа, яка приймає рішення, має можливість отримати не лише «готове» рішення, а й брати участь у формуванні моделей, пропонувати власні сценарії та горизонти прогнозування для опрацювання з використанням фактичних даних. Інформаційна технологія в даному випадку повинна бути зручним і зрозумілим інструментом користувача, робити процес прогнозування більш прозорим та обґрунтованим, приховуючи при цьому всю складність застосовуваного математичного та технічного інструментарію.

Процеси, характерні для розвитку більшості реальних складних систем та об’єктів, визначаються наявністю невизначеностей різних типів, спричинених як зовнішнім середовищем функціонування, так і їх складністю, нелінійністю, нестаціонарністю, високою динамічністю тощо. Все це

приводить до того, що кожне рішення оцінюється за одним або декількома критеріями в рамках однієї або декількох моделей – тобто виникає потреба створення і застосування багатомодельного підходу до розв'язання складних задач.

Крім того, для виявлення закономірностей, що характеризують поведінку досліджуваних процесів, розроблено та широко застосовується сучасний математичний апарат аналізу часових рядів. Однак, у випадку наявності нелінійностей, виникають проблеми із практичним застосування теоретичних розробок. Значною мірою це пов'язано з тим, що такі часові ряди, крім складності структури ряду, часто містять пропуски, шумові складові, аномальні значення, є нестаціонарними, можуть містити локальні особливості, які можуть бути досить складними, неперіодичними ефектами. Все це значно ускладнює процес оцінювання структури і параметрів адекватної моделі для прогнозування.

Тому на перший план виходять методи, які не завжди можуть мати теоретичне обґрунтування, а переважно засновані на практичному досвіді аналітика і розробника, а також вхідних умовах задач прогнозування і подальшого прийняття рішень з урахуванням отриманих результатів прогнозування.

Добре зарекомендували себе й потужні аналітичні платформи, в рамках яких реалізовані сучасні інформаційні технології, розраховані на використання досить значних обчислювальних потужностей та які вже містять широке коло можливих структур математичних моделей, потужний апарат інтелектуального аналізу структурованих та неструктурованих даних. [31, 35, 243, 254]. Це надає переваги у застосуванні адаптивного підходу при побудові моделей, а також неформалізованого підходу при побудові допустимих ансамблів математичних моделей.

З метою створення єдиної концептуальної основи для вирішення задач побудови інформаційних технологій прогнозування запропоновано системну методологію побудови моделей нелінійних нестаціонарних процесів,

розроблену на основі адаптивного підходу до моделювання з комбінованим використанням сценарного моделювання, регресійних та ймовірнісно-статистичних моделей у формі мереж Байєса, багатовимірних розподілів та узагальнених лінійних моделей [78]. Її перевагами є комплексне застосування множини ймовірнісно-статистичних методів виявлення та врахування невизначеностей (цифрова і оптимальна фільтрація, альтернативні методи оцінювання структури і параметрів математичних моделей), і забезпечення підвищення якості проміжних і остаточних результатів обробки даних, побудови математичних моделей і оцінювання прогнозів. Загальна схема застосування пропонованого підходу подана на рис. 5.1.

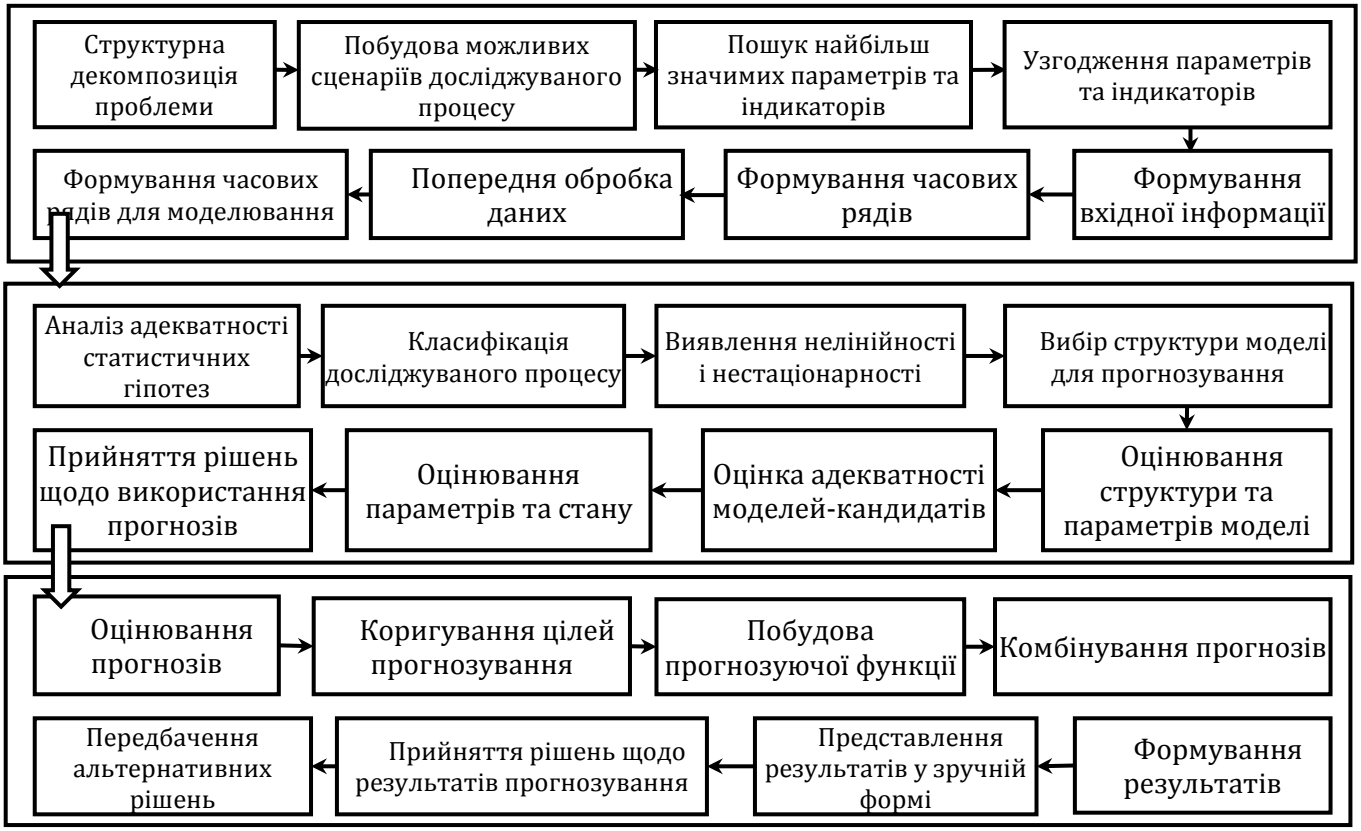


Рисунок 5.1 – Загальна схема застосування пропонованого багатомодельного підходу

Як видно з рис. 5.1, пропонований підхід може бути застосований для прогнозування процесів, що мають місце у системах різного типу, зокрема, технічних, соціально-економічних, екологічних тощо. Перевагою розробки є

те, що її легко «перелаштувати» на вирішення різних задач за різних вхідних умов: урахувати можливі варіанти розвитку ситуації, опрацювати короткі вибірки значної кількості різнорідних показників, експертних оцінок та можливих варіантів рішень. Важливе місце у розробленій методиці застосування багатомодельного підходу приділено саме всебічному дослідженню предметної області задачі, огляду варіантів розвитку подій, а також передбаченню їх наслідків на перспективу. Дана проблема вирішується шляхом розробки інформаційної технології, яка поєднує в собі процедури синтезу оптимального рішення, використання інтелектуального аналізу даних, прогнозного моделювання, а також доступних інструментів підтримки прийняття рішень [243, 254].

Як видно з рис. 5.1, значна увага приділена забезпеченню системності використання методів аналізу та автоматизованого інтегрування різнотипної інформації, методів моделювання, прогнозування та багатокритеріального прийняття рішень. Для подолання невизначеностей запропоновано множину різних методів, залежно від типу невизначеності.

Очевидно, що за пропонованого підходу значна увага приділена методиці попереднього опрацювання значних обсягів різнорідних даних, отриманих із різних інформаційних джерел. Створена методика вирізняється тим, що при формуванні масивів вхідних даних їх інформативність, синхронність та коректність забезпечуються на етапі попередньої обробки даних, виконуваної для приведення їх до форми, яка забезпечить можливість коректного застосування методів оцінювання параметрів моделі та отримання їх статистично значущих оцінок. Крім того, на особливу увагу заслуговує проблема опрацювання пропусків в даних, оскільки у реальних часових рядах зазвичай наявні пропуски, спричинені різними факторами, і уникнути їх утворення не можливо, і, часто трапляється так, що вилучити такі дані з аналізу не можливо – це спотворить уяву про досліджуваний процес. Тому в даній методиці пропонується використовувати різні підходи до заповнення пропусків із різною кількістю пропущених значень

(представлені у першому розділі). Опрацювання вхідних даних має бути системним та комплексним, і що головне, адаптовано до потреб користувача: поряд із заповненням пропусків даних, повинно виконуватись корегування значних імпульсних (екстремальних) значень, нормування вимірів у заданих межах, логарифмування великих значень та фільтрації шумових складових, виявлення та усунення мультиколінеарності.

Передбачено також, що вхідні дані можуть бути використані у багатоваріантних розрахунках, коли розглядаються їх можливі зміни за різних сценаріїв розвитку досліджуваних процесів.

Впровадження технологій інтелектуального аналізу даних у інформаційну технологію прогнозування та підготовки варіантів рішень представлено на рис. 5.2.

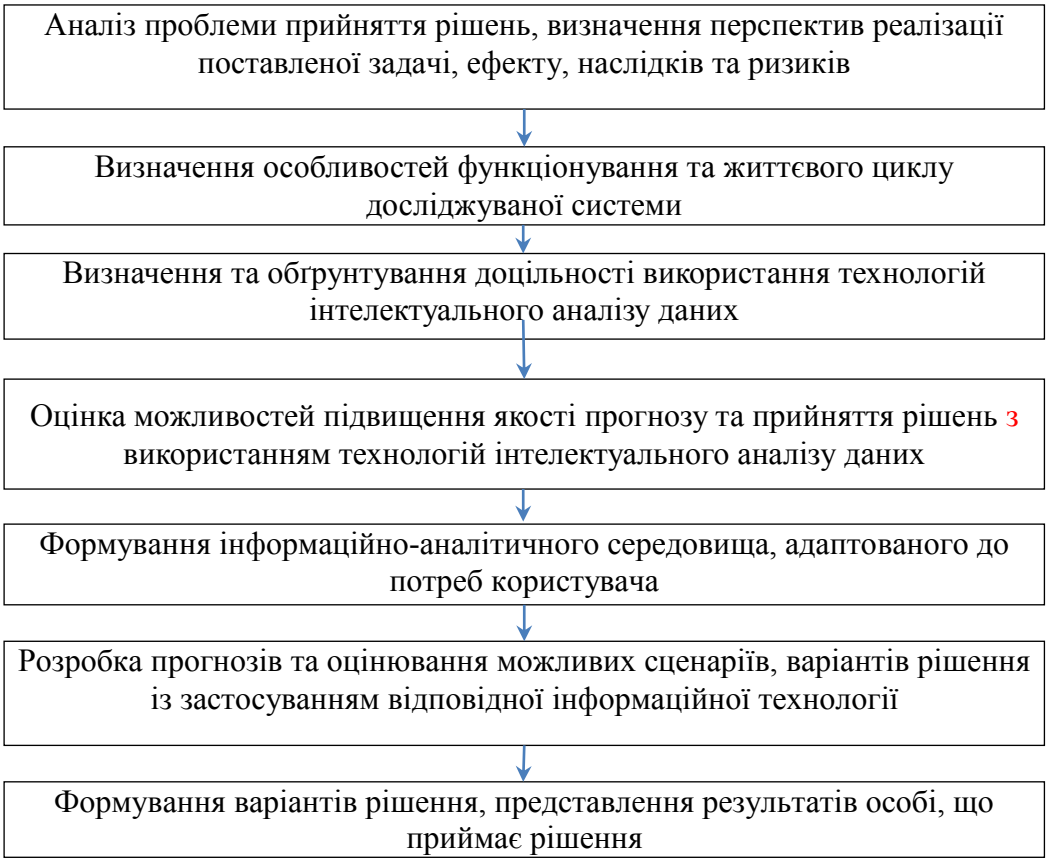


Рисунок 5.2 – Схема використання інтелектуального аналізу даних у інформаційній технології прогнозування для подальшого використання у інформаційній технології прийняття рішень

Запровадження у інформаційну технологію прогнозування сценарного підходу [257], невід'ємною складовою якого є попередній аналіз предметної області, відбір альтернатив і формулювання цільових установок можливих варіантів розвитку подій, дозволяє розробляти прогнози навіть в умовах нечіткої, неповної інформації, оцінювати можливі ризики та невизначеності різних типів на основі процедури оцінювання і ймовірних варіантів розвитку ситуації, що забезпечує підвищення якості остаточних результатів [256].

Інформаційна технологія прогнозування, пропонована в роботі, поєднує в собі різноманітні методичні підходи до використання окремих методів та моделей а також їх комбінацій для отримання кінцевого результату – оптимального управлінського рішення – обґрунтованого прогнозними розрахунками [245, 254]. У таблиці 5.1 узагальнено основні методичні підходи, пропоновані до застосування на різних етапах.

Таблиця 5.1 – Методичні підходи, що реалізовані у інформаційній технології прогнозування

Етап	Методичні підходи
1. Збір інформації, формування моделі діагностики	Емпіричні методи дослідження, метод статистичних спостережень, аналізу та синтезу, системного аналізу
2. Огляд стану та динаміки об'єкта дослідження, виявлення характерних ознак	Статистичний аналіз, data-mining, text-mining, факторний аналіз, багатовимірний аналіз даних, метод головних компонент, кореляційно-регресійний аналіз, типологічні та структурні групування, RFM-аналіз, кластерний аналіз, когнітивний аналіз
3. Причинно-наслідковий аналіз	Кореляційно-регресійний аналіз, ймовірнісне моделювання (мережі Байєса)
4. Розробка сценаріїв розвитку подій	Експертних оцінок, морфологічний аналіз, сценарне моделювання
5. Обґрунтування варіантів управлінських рішень	Data-mining, нейронні мережі, економетричне моделювання, SWOT-аналіз, когнітивне моделювання
6. Аналіз результатів	Метод експертних оцінок, метод Делфі, графічний

Як видно з таблиці 5.1, варіантів композицій застосування методичних підходів до аналізу та прогнозування розвитку є досить багато, всі вони опрацьовані .

Доповнюючи підхід, представлений в [244, 256] у загальному сенсі, завдання багатокритеріального вибору кращого рішення на множині математичних моделей, можна представити так (5.1):

$$M = \{\mathrm{M_0}\ (\mathrm{Y, I}), \mathrm{M_E}\ (\mathrm{X}), \mathrm{M_{OE}}, \mathrm{M_D(Q)}, \mathrm{M_{MO}}, \mathrm{M_{ME}}, \mathrm{M_U}, \mathrm{A}, \mathrm{M_H}, \mathrm{M_{RS}}, M_v\ \}, \quad (5.1)$$

де $\mathrm{M_0}(Y,I,P)$ – ідентифікуюча модель системи; Y – ендогенні змінні; I – вектор керованих змінних; ME(X) – модель навколишнього середовища; X – екзогенні змінні; MOE – модель взаємодії об’єкта та навколишнього середовища; MD(Q) – модель поведінки системи; Q – збурюючі впливи; Mv – модель взаємодії з підсистемами інших рівнів; MMO – модель зміни стану системи; MME – модель зміни стану навколишнього середовища; MU – модель керуючої системи; A – правило вибору процесів зміни об’єкта; MH – модель впливу особи, що приймає рішення, на систему та результати дослідження; MRS – модель системних ризиків.

Проблема формування оптимальної методики прогнозного моделювання нелінійних нестаціонарних процесів значною мірою зумовлена необхідністю одночасного подолання цілої низки невизначеностей різних типів: ситуаційної, статистичної, структурної, ймовірнісної тощо. Хоча, більшість існуючих інформаційних систем прогнозування та підтримки прийняття рішень мають відповідні засоби, однак, увага зосереджується на опрацюванні невизначеностей лише окремих типів. У даному дослідженні для вирішення цієї проблеми запропоновано інформаційну технологію застосування інтелектуального аналізу великих масивів структурованих та неструктурованих даних. Важливою перевагою пропонованого підходу є можливість розробки моделей для різних горизонтів прогнозування, адже від того, який горизонт прогнозування обрано, залежить не лише вибір типу моделі, а й підбір вхідних даних, зокрема, консолідація найбільш вагомих

чинників. Крім того, необхідним є урахування тенденцій розвитку досліджуваних процесів, формально-математичний опис алгоритмів та методів генерування набору кандидатів-сценаріїв розвитку процесу (рис. 5.3) [258].

Етап	**Характеристика етапу**			
Завантаження вхідних даних	Завантаження часових ряди історичних даних			
Діагностика даних	Визначення аномалій та їх обробка (виключення, згладжування).			
Обробка пропусків	Заповнення пропусків			
Формування набору додаткових чинників	Додаються чинники, суттєві для аналізу досліджуваних процесів (дані, що надаються за бажанням користувача, довідкових дані, даних щодо екзогенних параметрів процесу, експертні оцінки, дані, що доповнюють відомості про аналізований процес)			
Строк	Дуже короткострокове	Короткострокове	Середньострокове	Довгострокове
Особливості побудови прогнозу	на кожну годину, враховуються цикли всередині доби.	на добу (тиждень) вперед, врахо-вуючи цикли та закономірності в середині тижня.	Трендова складова, тренд та комбінований фактор з іншими регресорами, похідні показники	Враховуються додаткові чинники, користувача (на довгострокову перспективу)
Включення додаткових чинників	Додаткові регресори, що описують процес	Додаткові регресори, що описують процес.	Враховуються додаткові чинники, користувача	В залежності від предметної області
Горизонт прогнозування	від 1 до 24 годин	від доби до місяця	від 1 місяця до 3 років	більш 3 років
Сценарій	оптимістичний, реалістичний, песимістичний			
Побудова моделей-кандидатів	На кожному горизонті прогнозування будується модель за історичними даними та за даними користувача			
Етап 1	Одноступеневі моделі (експоненційні, регресійні, авторегресійні, узагальнені лінійні моделі)			
Етап 2	Двоступеневі моделі (регресійні та авторегресійні моделі. моделі із включенням трендової складової та урахуванням залишків (різниця між реальним та прогнозним значенням), що включаються до моделі у вигляді ковзного середнього, за умови, що між залишками та цільовою змінною є кореляція (автокореляція). Моделі класу експоненційного згладжування – Хольта, Тейла-Вейджа, Брауна, Вінтерса (з адитивною або мультиплікативною сезонною складовою), з урахуванням демпфуючого тренду та інші модифікації, Узагальнені лінійні моделі, нейронні мережі, ймовірнісні моделі та нечіткі методи.			
Вибір кращої моделі-кандидата	Вибір за результатами виконання етапу 1 та етапу 2, на основі: середньої абсолютної процентної похибки (MAPE), максимального MAPE, коефіцієнт детермінації (R^2), середньоквадратична похибка (RMSE).			
Побудова прогнозу	Використовуються кращі моделі-кандидати, вхідні дані: часові ряди досліджуваних процесів та додаткові чинники, що є значимими на даному горизонті прогнозування.			

Рисунок 5.3 – Методика застосування моделей різного типу та їх комбінацій залежно від горизонту прогнозування

Як видно з рис. 5.3, оцінювання якості побудованих сценаріїв, обґрунтування вибору найкращого та найімовірнішого з них виконано з використанням ймовірнісно-статистичного моделювання. Для побудови сценаріїв можуть бути застосовані різні методики, що передбачають використання як кількісних, так і якісних показників, зокрема, соціально-демографічних, економічних, а в окремих випадках і суспільно-політичних. Для вирішення проблеми відсутності достатньо довгих часових рядів статистичних даних, розроблена методика формального опису причинно-наслідкових зв'язків. Особливістю методики є використання ймовірнісно-статистичних методів для створення та застосування уніфікованих за структурою моделей в просторі станів. Для цього спочатку вирішується задача оцінювання елементів структури і параметрів моделі. Структура моделі оцінюється на підставі дослідження закономірностей протікання процесу, застосування статистичних тестів для перевірки наявності нелінійності, інтегрованості, гетероскедастичності, аналізу кореляційних функцій та візуального аналізу даних. При цьому вибирається кілька найбільш ймовірних структур моделей-кандидатів. Потім обчислюються оцінки параметрів моделей-кандидатів, обирається краща з них, використовуючи відповідні статистичні характеристики адекватності моделей. Ефективність застосування даного методу перевірено під час розв'язування практичних задач прогнозування розвитку складних соціально-еколого-економічних систем. Також, для прогнозування фінансових та економічних процесів запропоновано метод на основі адаптивно-оптимізаційного підходу до моделювання та прогнозування з використанням регресійних, узагальнених лінійних та ймовірнісно-статистичних моделей у формі мереж Байєса [237, 254, 256, 258].

Застосування багатомодельного підходу виправдане і у випадках, коли дані, які описують досліджувані процеси, неповні або є сумніви щодо їх достовірності. Для розв'язання цієї задачі запропоновано багатомодельний підхід; а в основу цього методу покладено інтегроване використання теорії

подібності процесів та ймовірнісно-статистичного моделювання (рис. 5.6).

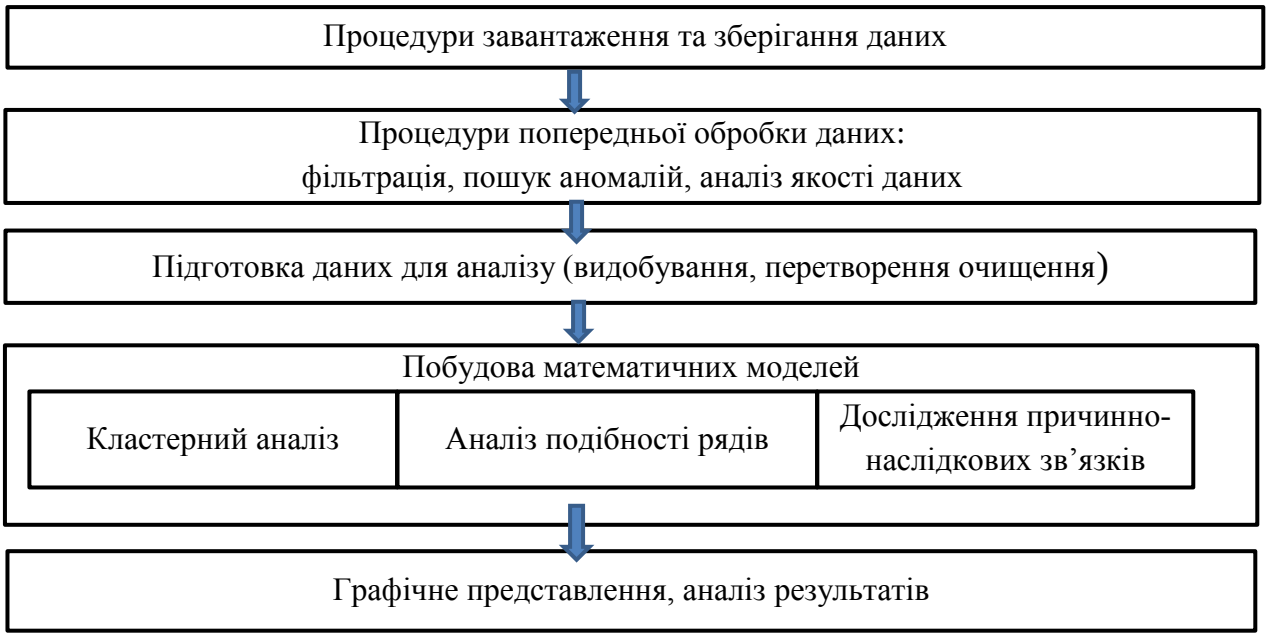


Рисунок 5.4 – Схема багатомодельного підходу із використанням теорії подібності процесів та ймовірнісно-статистичного моделювання

Для оцінювання якості прогнозів, обчислених за побудованими моделями, запропоновано процедуру автоматизованого вибору кращої прогнозної моделі, у якій для вибору було використано інтегральний критерій якості, який включає 2-5 окремих статистичних критеріїв якості. Процедура автоматизації забезпечує можливість побудови, аналізу та вибору кращої із множини можливих моделей, кількість яких може сягати кількох сотень. Витрати часу залишаються при цьому цілком прийнятними для практичного використання пропонованої методики у автоматизованій системі, призначеній для прогнозного моделювання нелінійних нестаціонарних процесів.

Отже, для розв'язання задач прогнозування нелінійних нестаціонарних процесів необхідно створити такі інформаційні технології:

- аналізу та попередньої підготовки інформації;
- побудови моделей;
- прогнозування;

- опрацювання результатів прогнозування та представлення їх вигляді, придатному для використання у процесі прийняття рішень.

Враховуючи те, що результати прогнозування створюють основу для прийняття ефективних рішень, проблема вибору оптимального рішення може вирішуватися за використання багатомодельного підходу, який передбачає, що для обґрунтування вибору необхідно дослідити результати прогнозування, отримані із використанням множини різнорідних моделей, в тому числі і математичних.

Для створення інформаційної технології побудови моделей використано різні класи моделей, вибір яких здійснюється із застосуванням методології системного аналізу, ієрархічного підходу до створення та аналізу процесу моделювання, адаптації структури і параметрів моделей до особливостей досліджуваних процесів та на основі порівняння оцінок прогнозів за числовими критеріями їх якості.

5.2 Особливості застосування регресійних моделей

Не зважаючи на те, що дана група методів регресійного аналізу вважається однією з найбільш опрацьованих у дослідженнях науковців та аналітиків-практиків та широко застосовується у інформаційних технологіях, їх застосування у прогнозування нелінійних нестаціонарних процесів заслуговує на увагу.

Зазвичай, моделі включають нелінійні, поліноміальні зв'язки, а також комбінації змінних. При цьому розрізняють два загальні підходи – прогнозування та аналітичний аналіз. Основною метою прогнозування є вибір параметрів моделі, що є найбільш значимі. В той час як аналітичний підхід, або як його інколи називають, пояснюючий аналіз, націлений на розуміння зв’язків між регресорами та відгуком. До уваги приймається не лише сила зв’язку, а й напрям (додатній або від’ємний коефіцієнт моделі) [1, 69, 261].

Для дослідження зв'язку між $k+1$ змінними (k регресорів плюс один відгук) використовують k −мірну поверхню для прогнозування. У загальному випадку в модель регресії може включати і більш складні зв'язки - квадратичну, кубічну, поліноміальну, перетин змінних.

Приклад формули регресії із нелінійним зв’язком відгуку та регресорів.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_1^2 + \beta_3 \cdot X_2 + \beta_4 \cdot X_2^2 + \varepsilon$$

Модель множинної лінійної регресії у порівнянні із моделлю простої лінійної регресії, має наступну основну перевагу – множинна регресія дозволяє досліджувати зв'язок не з однією, а з декількома регресорами.

При цьому є і деякі недоліки, а саме – збільшення складності при виборі кращої моделі та подальшій інтерпретації моделі.

Парні (прості) лінійні регресійні моделі встановлюють лінійну залежність між двома змінними [106]. У загальному випадку парна вибіркова регресійна модель [108] має вигляд:

$$Z = a_0 + a_1 X + E, \tag{5.1}$$

де Z – вектор значень залежної змінної, $Z = [z_1, z_2, \ldots, z_n]$; X – вектор значень незалежної змінної (регресора) $X = [x_1, x_2, \ldots, x_n]$; a_0, a_1 – невідомі параметри регресійної моделі; E – вектор значень випадкових величин $E = [e_1, e_2, \ldots, e_n]$, поява яких у моделі зумовлена наявністю випадкових збурень, недосконалою структурою моделі, похибками вимірів та обчислень оцінок параметрів моделі.

Слід зазначити, що при побудові ансамблів моделей (багатомодельному підході) нестаціонарних процесів, процесів, що характеризуються нелінійністю, кращі результати вдалось отримати використовуючи регресійні моделі в комбінації із мережами Байєса, деревами рішень, когнітивними моделям.

Переваги використання регресійних моделей за пропонованою багатомодельною методикою представлені у задачі прогнозного

моделювання урожайності соняшника. Інформаційною основою дослідження є дані Державної служби статистики України у Черкаській області щодо урожайність сільськогосподарських культур [254] та індекси біомаси, обчислених на основі показників вимірювання спектрального відбиття, відповідно, у червоній та ближній інфрачервоній областях, отримані в результаті обробки супутникових знімків. Для вирішення даної задачі пропонується застосування статистичних, графічних та регресійних процедур у інформаційній технології прогнозування, в основу якої покладено багатомодельний підхід

Набір вхідних даних 1 (GRID_50KM_COORDINATES) містить GPS координати сільськогосподарських полів. Таблиця даних містить 297 рядків та 3 стовпчики: ID поля (Grid), широта (Lat), довгота (Long).

Набір вхідних даних 2 (GRID_SUNFLOWER) містить значення обчислених індексів для 219 полів з різних регіонів України з 2003 по 2017 роки. Таблиця даних містить 4215 рядків та 9 стовпців: номер поля; рік; місяць; потенційна біомаса (PYB); потенційна біомаса продуктивних органів (PYS); лімітована вологою біомаса (WLIB); лімітована вологою біомаса продуктивних органів (WLIS); потенційний LAI (PLAI); лімітований вологою LAI (WLAI).

Набір вхідних даних 3 –урожайність соняшника за період з 1913-2017 рр. у Черкаській області (Productivity) [173].

В якості платформи для інформаційної технології використано спеціалізоване програмне забезпечення компанії SAS Institute. Побудова прогнозів здійснювалась у декілька етапів.

Спочатку були побудовані діаграми розсіювання по координатах досліджуваних полів, розташованих в Черкаській області. З метою створення «прив’язки» до геоданих [47]. Дана процедура автоматизована за рахунок створення відповідного програмного коду:

```
data agro.map_id;
set agro.GRID_50KM_COORDINATES;
```

```
Group=1;
IF lat >48 AND lat <50 AND
long > 30 AND long < 33 THEN
  DO;
    Group=2;
    lat2=lat;
         long2=long;
         lat=.;
    long=.;
  END;
run;
SYMBOL1 VALUE=CIRCLE CV=BLACK;
SYMBOL2 VALUE=TRIANGLE CV=RED;
PROC GPLOT DATA=AGRO.MAP_ID;
PLOT lat*long lat2 * long2 / OVERLAY;
        RUN; QUIT;
```

В результаті отримано діаграму розсіювання, що відповідає розташуванню полів.

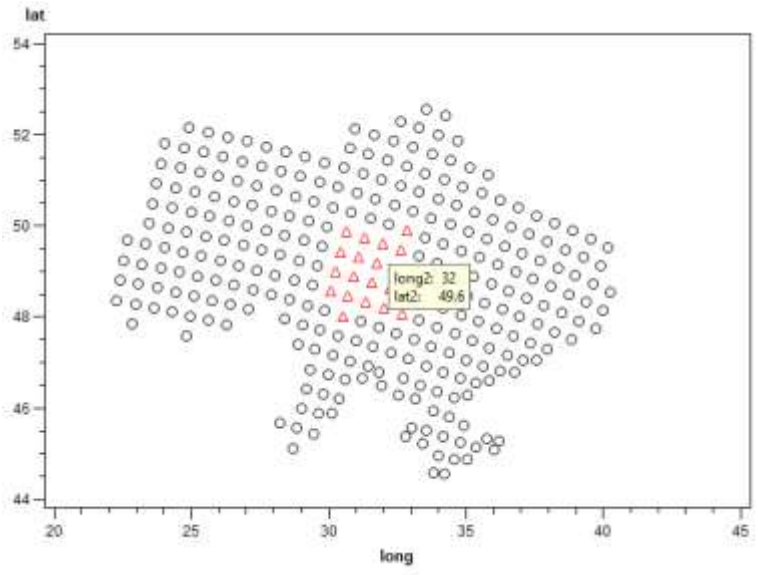


Рисунок 5.5 – Побудована карта полів у вигляді діаграми розсіювання. Червоний трикутник – поля Черкащини

На наступному етапі реалізовано інформаційну технологію для опрацювання таблиць даних, призначену для відбору показників біомаси

досліджуваних полів із загальної інформації щодо показників біомаси всіх полів Черкаської області.

Реалізоване програмне забезпечення яке дозволяє відібрати із загального набору даних ті поля, на яких досліджуватиметься урожайність обраної сільськогосподарської культури (соняшника).

Представлений код автоматично формує таблицю з ID цих полів, а потім виконує лівостороннє об'єднання з таблицею GRID_SUNFLOWER, в результаті утворюється фінальна таблиця CHERKASY_GRID_SUNFLOWER, яка містить відібрані показники.

Фрагмент коду попередньої підготовки даних:

```
data agro.cherkasy_id;
set agro.GRID_50KM_COORDINATES;
IF lat >48 AND lat <50 AND
long > 30 AND long < 33;
run;

data CHERKASY_GRID_SUNFLOWER;
MERGE agro.cherkasy_id (in=D1)
    agro.GRID_SUNFLOWER (in=D2);
BY Grid;
IF D1=D2;
        run;
```

Для обчислення агрегованих (середніх) значень показників: потенційна біомаса (PYB); потенційна біомаса продуктивних органів (PYS); лімітована вологою біомаса (WLIB); лімітована вологою біомаса продуктивних органів (WLIS); потенційний LAI (PLAI); лімітований вологою LAI (WLAI) використовується код, реалізований мовою SAS SQL.

```
 PROC SQL;
```

```
  CREATE TABLE CHERKASY_GRID_SUNFLOWE_MEAN AS
  SELECT t1.year,
        (MEAN(t1.PYB)) FORMAT=BEST12. AS MEAN_of_PYB,
        (MEAN(t1.PYS)) FORMAT=BEST10. AS MEAN_of_PYS,
        (MEAN(t1.WLYB)) FORMAT=BEST12. AS MEAN_of_WLYB,
        (MEAN(t1.WLYS)) FORMAT=BEST14. AS MEAN_of_WLYS,
        (MEAN(t1.PLAI)) FORMAT=BEST14. AS MEAN_of_PLAI,
        (MEAN(t1.WLAI)) FORMAT=BEST14. AS MEAN_of_WLAI
    FROM WORK.CHERKASY_GRID_SUNFLOWER t1
    GROUP BY t1.year;
QUIT;
        RUN;
```

Для того, щоб вхідні дані, крім потенційних регресорів моделі, також містили цільову змінну, було використано представлений нижче код програми.

```
data AGRO.CHERKASY_GRID_SUNFLOWE_MEAN;
MERGE
WORK.CHERKASY_GRID_SUNFLOWE_MEAN (IN=D1)
AGRO.PRODUCTIVITY (IN=D2);
BY Year;
IF D1=D2;
        run;
```

В результаті виконання даного коду для набору даних CHERKASY_GRID_SUNFLOWER, дописується стовпець з цільовою змінною – урожайністю соняшника у Черкаській області за відповідний рік. В результаті виконання процедур попередньої обробки даних отримано набір вхідних даних AGRO.CHERKASY_GRID_SUNFLOWER, який буде використаний у подальшому аналізі.

Таблиця 5.1 – Результати обчислення агрегованих значень показників вегетації із використанням розробленої програми

Рік	MEAN_of_PYB	MEAN_of_PYS	MEAN_of_WLYB	MEAN_of_WLYS	MEAN_of_PLAI	MEAN_of_WLAI	Урожайність
2005	3028,096	1054,964	1234,887	310,7368	1,170533	0,461965	15
2006	3325,307	1160,805	1478,952	238,9974	1,336329	0,611068	15,8
2007	2151,01	942,6025	119,7957	31,58617	0,642753	0,033493	17,1
2008	3179,55	1032,712	952,4107	211,9878	1,373521	0,382127	19,3
2009	3179,899	1223,231	592,9779	138,7237	1,157377	0,236591	22,8
2010	2842,099	1191,817	1266,751	475,4378	0,939179	0,422371	20,9
2011	3626,157	1526,457	1386,22	600,8449	1,192168	0,434943	22,1
2012	2954,812	1428,786	397,3948	59,47271	0,820713	0,134636	26,7
2013	1636,692	964,7109	412,947	165,4768	0,473071	0,161681	31,1
2014	3904,658	1578,541	3391,957	1142,598	1,313567	1,149661	28
2015	3550,627	1439,618	3174,403	1174,937	1,258188	1,137391	28,5
2016	3091,028	1253,368	3075,16	1237,499	1,063946	1,062462	28,3
2017	3306,358	1132,661	2016,587	566,3511	1,307088	0,818077	24,8

Наступним етапом дослідження є аналіз регресорів моделі на мультиколінеарність з метою уникнення ситуації, коли один і той же ефект-фактор, присутній в різних регресорах включається в модель декілька разів. Це призводить до зміщення оцінок коефіцієнтів моделі, а в подальшому – і прогнозних значень.

Проблеми спричинені коллінеарністю:

– колінеарність може збільшити уявну статистичну значущість ефектів за рахунок повторного включення тієї ж порції інформації.
– колінеарність прагне збільшити дисперсію оцінок параметрів і отже, збільшити похибку передбачення.

За колінеарності, оцінки коефіцієнтів моделі нестійкі, мають велику дисперсію. Отже, справжній зв'язок між урожайністю та регресором моделі може значно відрізнятися від отриманих результатів.

При цьому колінеарність не є порушенням припущень для лінійної регресії. Для кількісної оцінки проблеми колінеарності та визначення підмножини регресорів, що містять колінеарність, доцільно скористатися показником VIF, який надає оцінку сили колінеарності у вигляді значення Variance Inflation Factor. За своєю суттю VIF – це міра росту дисперсії за наявності колінеарності. Якщо $VIF_i > 10$, то присутня колінеарність.

Значення VIF обчислюється для кожного регресора моделі за формулою:

$$VIF_i = \frac{1}{(1 - R_i^2)}$$

де R_i^2 – це значення R^2 для регресора X_i по відношенню для всіх інших регресорів моделі.

Код програми:

```
PROC REG DATA=AGRO.CHERKASY_GRID_SUNFLOWE_MEAN;
MODEL Productivity=
MEAN_of_PYB MEAN_of_PYS
MEAN_of_WLYB MEAN_of_WLYS
MEAN_of_PLAI MEAN_of_WLAI / VIF;
    RUN;
```

Таблиця 5.2 Оцінки коефіцієнтів регресійної моделі та значення коефіцієнта VIF

Variable	Parameter Estimates					
	DF	Parameter Estimate	Standard Error	tValue	Pr>|t|	Variance Inflation
Intercept	1	11,0853	7,63149	1,45	0,1966	0
MEAN_of_PYB	1	-0,0259	0,01429	-1,81	0,1201	89,05
MEAN_of_PYS	1	0,04981	0,01874	2,66	0,0376	18,71
MEAN_of_WLYB	1	-0,0224	0,01798	-1,24	0,2602	476,38
MEAN_of_WLYS	1	0,01471	0,01414	1,04	0,3385	45,51
MEAN_of_PLAI	1	24,7491	23,139	1,07	0,3259	51,73
MEAN_of_WLAI	1	54,9146	39,4301	1,39	0,2131	280,05

Як видно з таблиці 5.2, всі регресори колінеарні, що пояснюється тим, що всі вони обчислювалися на основі індексів RED та NIR (вимірювання спектрального відбиття, отримані відповідно у червоній та ближній інфрачервоній областях відповідно), розрахованих за даними супутникової зйомки.

Для усунення мультоколінеарності в даному випадку запропоновано зменшити простір вхідних змінних за методом головних компонент, тобто замість шести регресорів використовувати декілька базисних, які в свою чергу, є комбінацією вхідних.

Код програми для створення головних компонент та виводу статистик для аналізу:

```
ODS GRAPHICS ON;
PROC PRINCOMP DATA =
AGRO.CHERKASY_GRID_SUNFLOWE_MEAN
OUT=WORK.CHERKASY_GRID_SUNFLOWE_PCA
OUTSTAT=WORK.OUTSTATS
PREFIX='PRIN'n SINGULAR=1E-08 VARDEF=DF STD
   PLOTS(ONLY)=SCREE;
VAR MEAN_of_PYB MEAN_of_PYS MEAN_of_WLYB
MEAN_of_WLYS MEAN_of_PLAI MEAN_of_WLAI;
RUN;
      ODS GRAPHICS OFF;
```

В таблиці 5.3 наведені значення власних чисел отриманих головних компонент.

Таблиця 5.3 – Власні числа отриманих головних компонент

№	Eigenvalue	Difference	Proportion	Cumulative
1	4,4045735	3,4769514	0,7341	0,7341
2	0,9276221	0,301488	0,1546	0,8887
3	0,6261341	0,5933163	0,1044	0,9931
4	0,0328178	0,0252207	0,0055	0,9985
5	0,0075971	0,0063415	0,0013	0,9998
6	0,0012556		0,0002	1

На рис. 5.4, представлені результати застосування методу головних компонент.

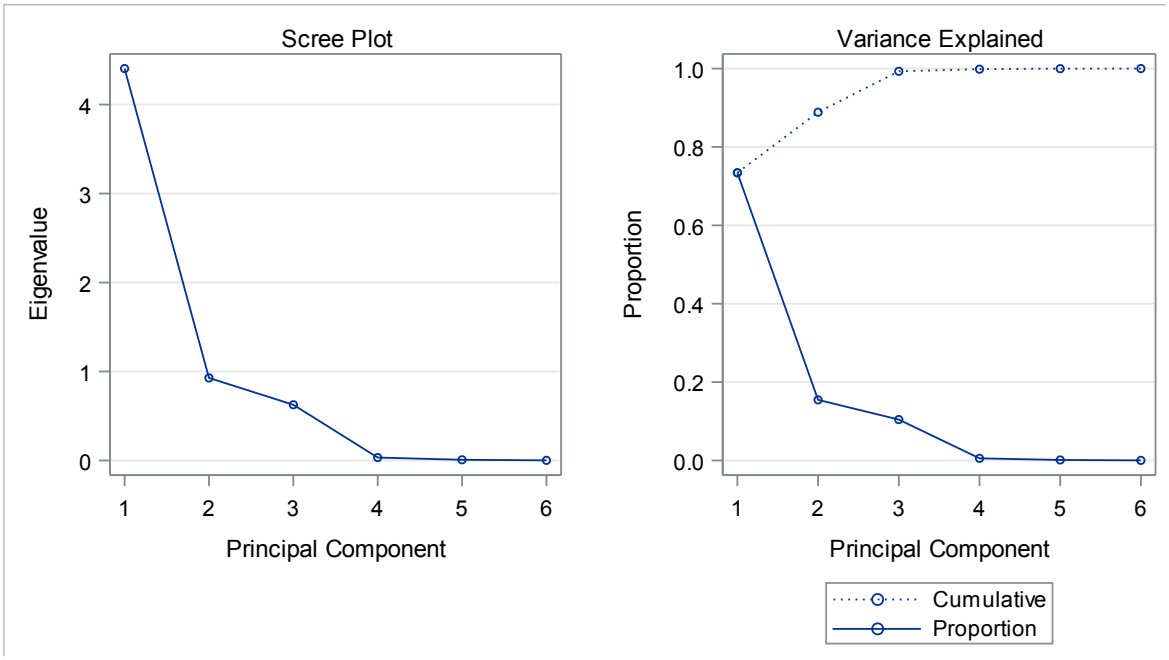


Рисунок 5.4 – Графічне представлення результатів застосування методу головних компонент

Як можна побачити з рис. 5.4, достатньо використовувати дві головні компоненти замість шести вхідних, що дозволило враховувати 88% варіабельності вхідних регресорів.

Для ідентифікації наявності лагових ефектів, обчислимо значення часткової автокореляційної функції.

Код програми має вигляд:

```
ODS GRAPHICS ON;
proc arima data=CHERKASY_GRID_SUNFLOWE_PCA ;
identify var=Productivity nlag=6;
run;
    ODS GRAPHICS OFF;
```

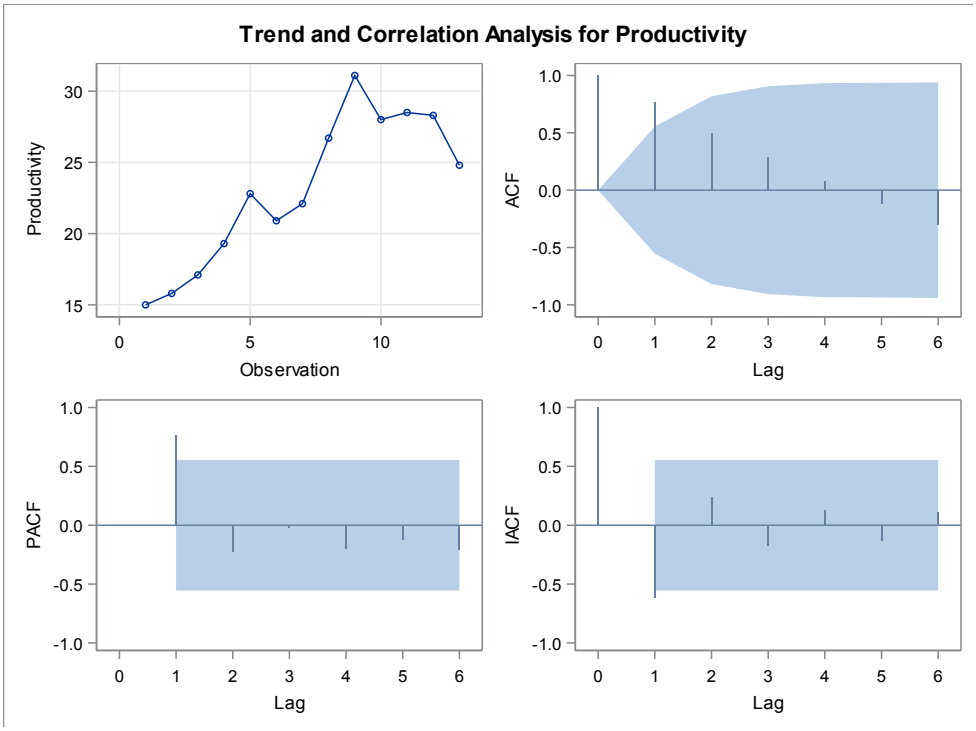


Рисунок 5.5 – Результати роботи процедури ARIMA

Як видно з рис. 5.5, графік PACF показує, що в модель необхідно включити лаг першого порядку.

Код програми для побудови математичної моделі прогнозування урожайності на основі лагової змінної та двох головних базисних компонентів.

```
data AGRO.CHERKASY_GRID_SUNFLOWE_PCA_LAG;
set CHERKASY_GRID_SUNFLOWE_PCA;
Productivity_LAG=LAG(Productivity);
run;
PROC REG DATA=AGRO.CHERKASY_GRID_SUNFLOWE_PCA_LAG;
MODEL Productivity=
Productivity_LAG
PRIN1
PRIN2 / VIF;
OUTPUT OUT=PREDICTION
PREDICTED=predicted_Productivity ;
```

RUN;

В результаті виконання даного коду отримано регресійну модель для прогнозування урожайності соняшника (табл. 5.4)

Таблиця 5.4 – Оцінки параметрів регресійної моделі прогнозування урожайності соняшнику та значення коефіцієнта колінеарності VIF

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1,2613	5,32576	0,24	0,8187	0
Productivity_LAG	1	0,9826	0,23158	4,24	0,0028	3,11586
PRIN1	1	-1,58276	0,98105	-1,61	0,1453	2,00662
PRIN2	1	0,31397	1,01308	0,31	0,7645	2,12831

На підставі даних, представлених у таблиці 5.4 можназробити висновок, що проблема мультиколінеарності була подолана. Адже, немає жодної змінної для якої показник VIF більший за 10.

Статистичні характеристики побудованої моделі, призначеної для прогнозування урожайності соняшника представлені в таблиці 5.5.

Таблиця 5.5 – Статистичні характеристики побудованої моделі прогнозування продуктивності вирощування соняшнику

Назва статистичної характеристики	Значення
Root MSE	2.39126
Dependent Mean	23.78333
Coeff Var	10.05437
R-Square	0.8268
Adj R-Sq	0.7618

Для візуалізації результатів прогнозування, а саме, порівняння реальних та прогнозних значень розроблено код програми:

```
SYMBOL1 VALUE=CIRCLE INTERPOL=SPLINES LINE=1 WIDTH=2
CV=BLACK;
SYMBOL2 VALUE=CIRCLE INTERPOL=SPLINES LINE=3 WIDTH=2
CV=RED;
Legend1 FRAME;
PROC GPLOT DATA=PREDICTION;
PLOT Productivity*Year predicted_Productivity * Year /
LEGEND=LEGEND1 OVERLAY FRAME;
    RUN; QUIT;
```

Результат прогнозування урожайності соняшника, отримані за використання пропонованої інформаційної технології прогнозування, основаної на багатомодельному підході програми показано на рис. 5.8.

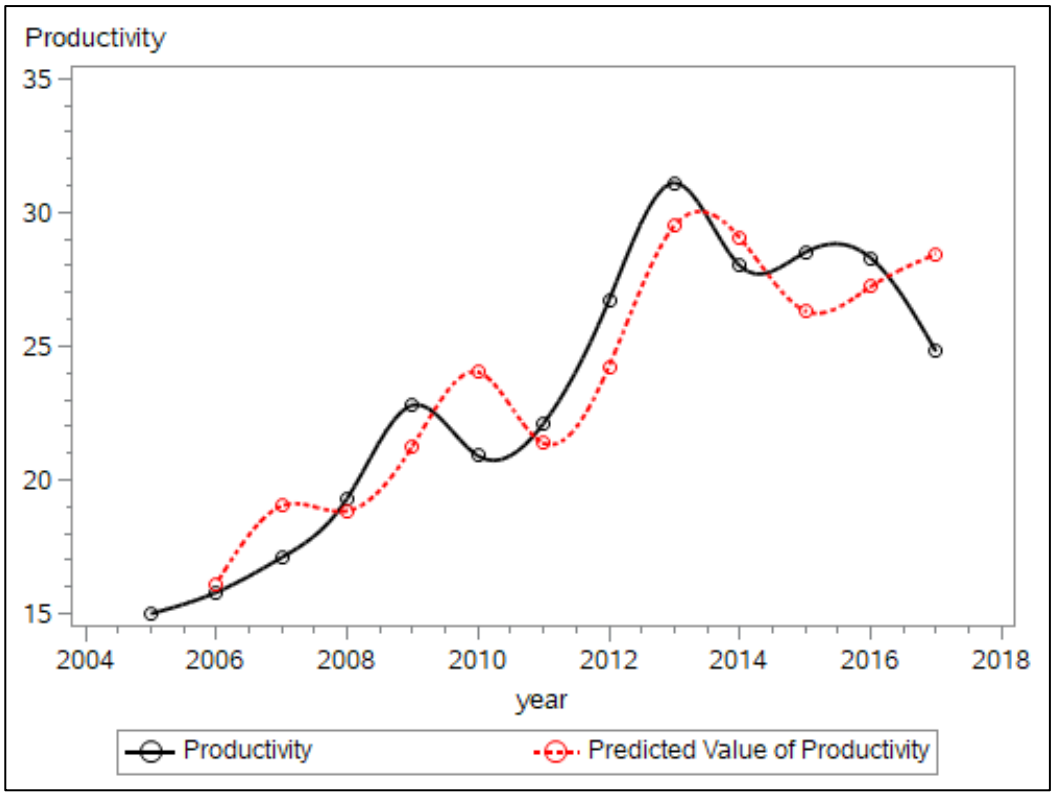


Рисунок 5.8 – Порівняння реальних і прогнозних значень продуктивності отриманих за побудованою моделлю

Таблиця 5.6 – Результати моделювання, за отриманою моделлю (на основі цих значень будувався рис. 5.8)

Рік	Урожайність	
	Статистичний показник	Прогноз
2006	15,8	16,53
2007	17,1	20,64
2008	19,3	18,29
2009	22,8	20,54
2010	20,9	24,59
2011	22,1	21,31
2012	26,7	23,88
2013	31,1	30,46
2014	28	28,98
2015	28,5	27,22
2016	28,3	28,88
2017	24,8	27,34
2018	31,5	27,69
2019	33,1	33,65

Статистичні характеристики побудованої моделі:

- середня абсолютна похибка (MAPE) отриманої моделі дорівнює 5,8%;
- коефіцієнт детермінації (R2) дорівнює 82,6%;
- середня квадратична похибка (RMSE) становить 2,39.

Для прогнозування стану поверхні Землі залежно від погодних умов на основі метеорологічних даних та індексу NDVI, розроблено методику, що реалізується за таких кроків:

Крок 1.Завантаження даних, їх попередня обробка.

Крок 2. Розділення даних випадковим чином на 2 підвибірки: навчальну (70%) та перевірочну (30%).

Крок 3. Побудова моделей.

Крок 4. Порівняння моделей за статистичними критеріями.

Крок 5. Побудова загального звіту.

В результаті, було побудовано два набори мереж Байєса. Для побудови першого набору мереж Байєса вихідні дані були переведені в дискретну форму за методом OptBining (дискретизація проводилась за допомогою програмного забезпечення SAS Enterprise Miner).

Для побудови другого набору мереж Байєса, робились такі перетворення:

1. Обчислено перші різниці даних показників NDVI, температури та кількості опадів.
2. Обчислені показники переведені в дискретну форму за методом рівних проміжків.

На основі дискретизованих даних побудована мережа Байєса. Для побудови мережі використано метод Greedy Thick Thinning (побудова відбувалась автоматизовано засобами програмного забезпечення GeNIe).

На рис. 5.12 зображено мережу Байєса, побудовану за перши набором даних для однієї з мозаїк. Точність класифікації становить 55,36%.

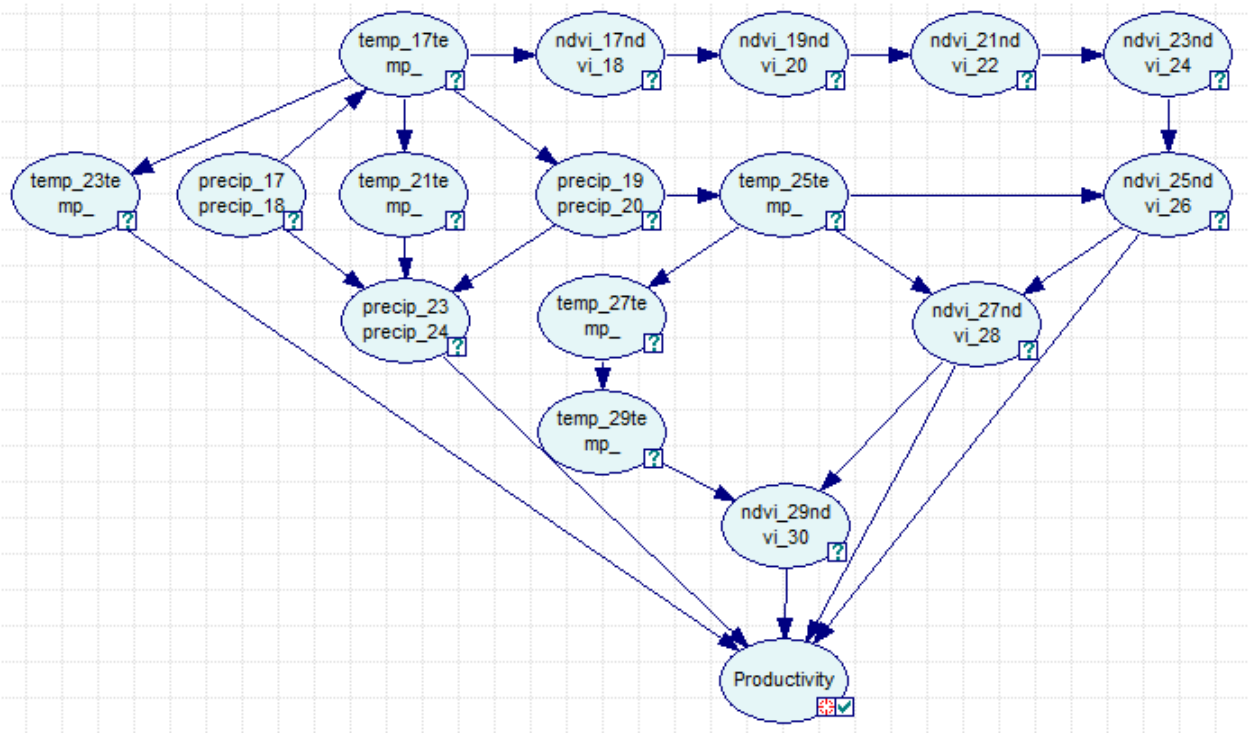


Рисунок 5.9 – Мережа Байєса (мозаїка 1)

На основі результатів, отриманих за побудованою мережею, стан «зеленої маси» покриву поверхні безпосередньо визначається метеорологічними факторами та впливом навколишнього середовища (індекс NDVI) на різних термінах вегетації рослин [250].

Точність класифікації мережі Байєса для мозаїки 2, побудованої за другим набором даних становить 70,98%.

У табл. 5.7 представлена зведена інформація щодо показників якості усіх побудованих моделей залежності стану зеленого покриву ділянок (мозаїки_1) від NDVI і погодних факторів.

Таблиця 5.7 – Значення показників якості моделей залежності стану зеленої маси поверхні від NDVI і погодних факторів

Модель	Вибірка					
	Навчальна			Перевірочна		
	RASE	SSE	MAX E	RASE	SSE	MAX E
Gradient Boosting	1,836	185,452	8,562	2,041	95,785	4,724
LARS Validation	3,050	511,625	11,534	2,626	158,562	7,062
LARS SBC	3,059	514,795	11,591	2,642	160,522	7,129
Partial Least Squares	2,810	434,427	10,140	2,842	185,792	8,389
Regression after DTree	2,931	472,452	11,143	3,007	207,943	8,437
Regression	2,896	461,165	11,146	2,755	174,534	7,469

Як видно за таблиці 5.7, найкращою виявилась модель, побудована за методом Gradient Boosting.

Отже, пропонована інформаційна технологія, основана на використанні багатомодельного підходу із використанням БМ дає можливість отримувати прогнози високої якості. Коло застосовуваних моделей може бути розширене. Крім того, апарат БМ добре зарекомендував себе і в задача, коли необхідно виявити причинно-наслідкові зв’язки під час розв’язку слабкоформалізованих задач, коли складно виявити фактори, які необхідно

використовувати при побудові моделей для прогнозування досліджуваних процесів чи систем.

5.3 Інформаційна технологія виявлення причинно-наслідкових зв'язків з використанням мереж Байєса

Досить часто під час дослідження процесів різної природи доводиться стикатися з проблемою необхідності побудови адекватної моделі в умовах невизначеності, зокрема коли відсутня повна та достовірна інформація про досліджуваний процес, його характер, умови протікання, взаємодію з іншими процесами та системами, чинники, що впливають на перебіг процесу тощо.

За таких умов складно не лише формалізувати поставлену задачу, а й виявити найбільш значимі фактори, чітко окреслити цілі та визначити причинно-наслідкові зв'язки, що існують між вхідними та вихідними показниками. Такі проблеми характерні для соціально-економічних, еколого-економічних та суспільно-політичних процесів та систем. Звичайно, можна було б розглядати їх окремі складові, будуючи «прості» моделі, зокрема, моделі множинної регресії, авторегресії, тощо. Однак, прогнози розвитку соціально-економічних та суспільно-політичних процесів закладають підґрунтя для прийняття важливих управлінських рішень, наслідки яких можуть відобразитися на різних аспектах людської життєдіяльності, тому інформаційна технологія прогнозування повинна передбачати ретельний аналіз предметної області, виявлення найбільш вагомих чинників, що впливають на розвиток досліджуваних системи чи процесу та пропонувати особі, що приймає рішення низку варіантів можливого перебігу подій.

Саме тому при вирішенні таких слабоформалізованих задач як прийняття управлінських рішень процесами та системами та процесами, особливо на середньо- та довгострокову перспективу, як показує практика, можна одержати з використанням методів, що ґрунтуються на різних ідеях та підходах, а саме, композиціях методів морфологічного, когнітивного та

сценарного аналізу, ймовірнісного моделювання, дослідження часових рядів, нейронних мереж тощо.

Наприклад, під час створення інформаційної технології прогнозування у задачі прийняття рішень щодо розвитку аграрного сектору регіону використано багатомодельний підхід, в рамках якого застосовано сценарний аналіз, когнітивне моделювання, ймовірнісно-статистичне моделювання, математичне моделювання тощо. Крім того, для обґрунтування цільових установок побудови сценаріїв було використано не лише статистичні дані, експертні оцінки, а й використано аналіз неструктурованої інформації з Інтернет-джерел. Для аналізу корпусу текстів було використано програмне забезпечення SAS Textual Analytics Suit, що складається з SAS Content Categorization та SAS Sentimental Analysis Studio. Була розроблена відповідна таксономія із 604 правил, опрацьовано більше 180 Інтернет-джерел. В результаті дослідження виявлені, ті чинники, які є суттєвими для розвитку аграрного сектору регіону і які необхідно включити у модель. Застосування сукупності вказаних підходів дозволило розглядати не лише з точки зору загальної характеристики окремого фактору, а й в контексті конкретної системи, процесу, їх особливостей та цільового спрямування. Наприклад, показник зростання обсягів житлового будівництва вважається позитивним фактором, оскільки свідчить про розвиток даної галузі, залучення інвестицій у економіку регіону. Однак, враховуючи те, що житлове будівництво швидкими темпами зростає у містах, і особливо у обласному центрі, це свідчить про те, що змінюється структура населення: відбувається відтік населення працездатного віку із сільської місцевості спричиняє додаткове соціальне навантаження на бюджети сільських громад, а в містах посилює конкуренцію на ринку праці. Тобто, як показує даний приклад, самі по собі показники обсягу інвестицій не можна розглядати як виключно позитивний чинник, треба розглядати його проти в місце у системі, досліджувати причинно-наслідкові зв’язки з іншими складовими в контексті конкретної задачі. Розвиток інвестиційної діяльності в агропромисловому виробництві є

фактором пливу на активізацію розвитку економіки та підвищення соціально-економічного рівня населення, з другого – зростання економіки та покращення життя населення сприяють притоку інвестиційних ресурсів, адже основним інвестором у сільській місцевості є жителі регіону. Це, як показало дослідження, є особливістю соціально-економічної системи регіону, в економіці якого переважає агропромислове виробництво. При виборі цільової установки було враховано двонаправленість факторів впливу на розвиток соціально-економічної системи регіону та визначено, що цільовою установкою є розвиток агропромислового виробництва в регіоні [252].

Дослідження виконане за наступною методикою, яка передбачає реалізацію інформаційної технології прогнозування і включає такі етапи:

Етап 1. Збір інформації, попередня обробка, аналіз предметної області.

Етап 2. Аналіз задачі, формулювання цілей, очікуваних результатів.

Етап 3. Відбір найбільш значимих факторів для кожного горизонту прогнозування.

Етап 4. Визначення найбільш значимих змінних, що впливають на цільову змінну.

Етап 5. Побудова моделі, оцінка параметрів моделі.

Етап 6. Розробка сценаріїв.

Етап 7. Розробка прогнозу для кожного сценарію.

Етап 8. Виявлення причинно-наслідкових зв’язків у кожному з сценаріїв.

Етап 9. Формування висновку.

Найбільш значимі фактори, що впливають на розвиток агропромислового виробництва регіону, відібрані в результаті проведеного аналізу, згруповані у Додатку Д. Для виявлення змінних, найбільш значимих для дослідження виконано SWOT-аналіз, опрацьовано інформацію з Інтернет-джерел, побудовано когнітивні моделі.

Використання когнітивного моделювання (як побудова карт, так і побудова матриць) та ймовірнісного моделювання у процесі формування

сценаріїв розвитку соціально-економічних систем передбачає виконання наступних кроків:

Крок 1. Визначення цільових установок побудови сценаріїв;

Крок 2. Визначення основних факторів, що впливають на розвиток системи.

Крок 3. Відбір із набору факторів найбільш значимих

Крок 4. Визначення сильних та слабких характеристик системи на основі відібраних найбільш важливих змінних

Крок 5. Формування множини вхідних концептів для побудови когнітивної карти

Крок 6. Побудова когнітивної карти на основі відібраних на кроці 5 концептів

Крок 7. Аналіз когнітивної карти

Крок 8. Оцінювання системного ризику

Крок 9. Формулювання можливих сценаріїв розвитку подій

Крок 10. Оцінювання якості побудованих сценаріїв, обґрунтування вибору найкращого та найімовірнішого

Для дослідження використано такі фактори, як природно-кліматичні та екологічні умови, ресурсний потенціал, соціально-економічний розвиток, ринкова інфраструктура тощо. Всього, було опрацьовано набори вхідних даних, що складаються із 56 показників соціально-економічного розвитку України та Черкаської області за 2010–2013 рр., а окремі показники і за 2003-2013 рр.

Для аналізу когнітивної карти використано показники: консонансу (c_i)[257]:

$$c_{ij} = \frac{|v_{ij}+\overline{v_{ij}}|}{|v_{ij}|+|\overline{v_{ij}}|}$$

де v_{ij}, $\overline{v_{ij}}$ – пара зв’язків у транзитивно замкненій когнітивній матриці; дисонансу (d_i) [257],

$$d_{ij} = 1 - c_{ij}$$

впливу системи на концепт ($\overline{P_i}$) [157]

$$\overrightarrow{P_i} = \frac{1}{n}\sum_{j=1}^{n} p_{ij}$$

та концепту на систему ($\overline{P_j}$) [157]

$$\overleftarrow{P_j} = \frac{1}{n}\sum_{i=1}^{n} d_{ij}$$

Показники консонансу впливу для більшості концептів є достатньо високими і знаходяться в межах від 0,64 до 0,91. Максимальний показник дисонансу достатньо низький – його значення становить 0,3. Тобто між відібраними концептами та системою існує суттєвий взаємовплив і вони можуть бути використані при побудові сценаріїв розвитку системи.

Когнітивна карта розвитку агропромислового виробництва в регіоні представлена на рис. 5.10.

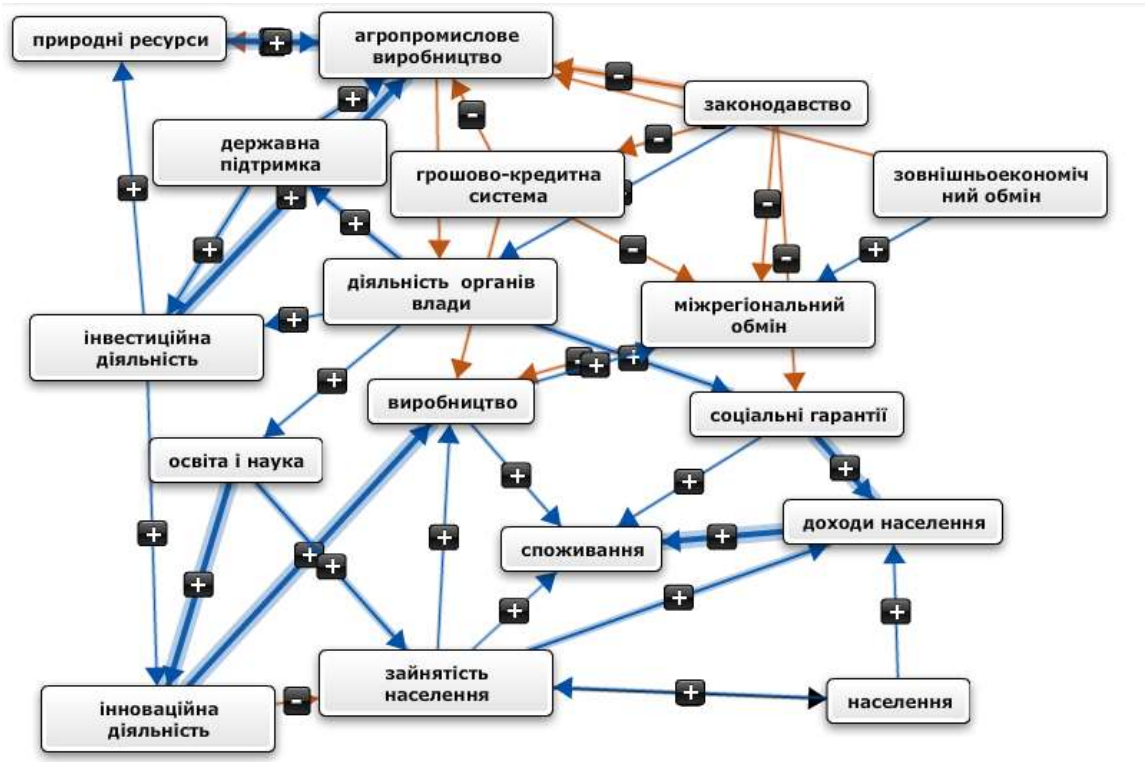


Рисунок 5.10 – Когнітивна карта розвитку агропромислового виробництва в регіоні

Отриманий набір найбільш значимих чинників використовується для побудови мережі Байєса, за допомогою якої визначаються змінні, що найбільше впливають на цільову.

Після цього будується рівняння множинної регресії із примусовим включенням до моделі виявлених змінних. Оцінка параметрів моделі виконується на основі рекурсивного методу найменших квадратів.

На основі одержаних результатів аналізу, сформовано можливі сценарії розвитку агропромислового виробництва регіону.

Оцінювання якості побудованих сценаріїв, обґрунтування вибору найкращого та найімовірнішого виконано з використанням ймовірнісного моделювання. Топологія побудованої мережі Байєса для одного із сценаріїв представлена на рис. 5.11

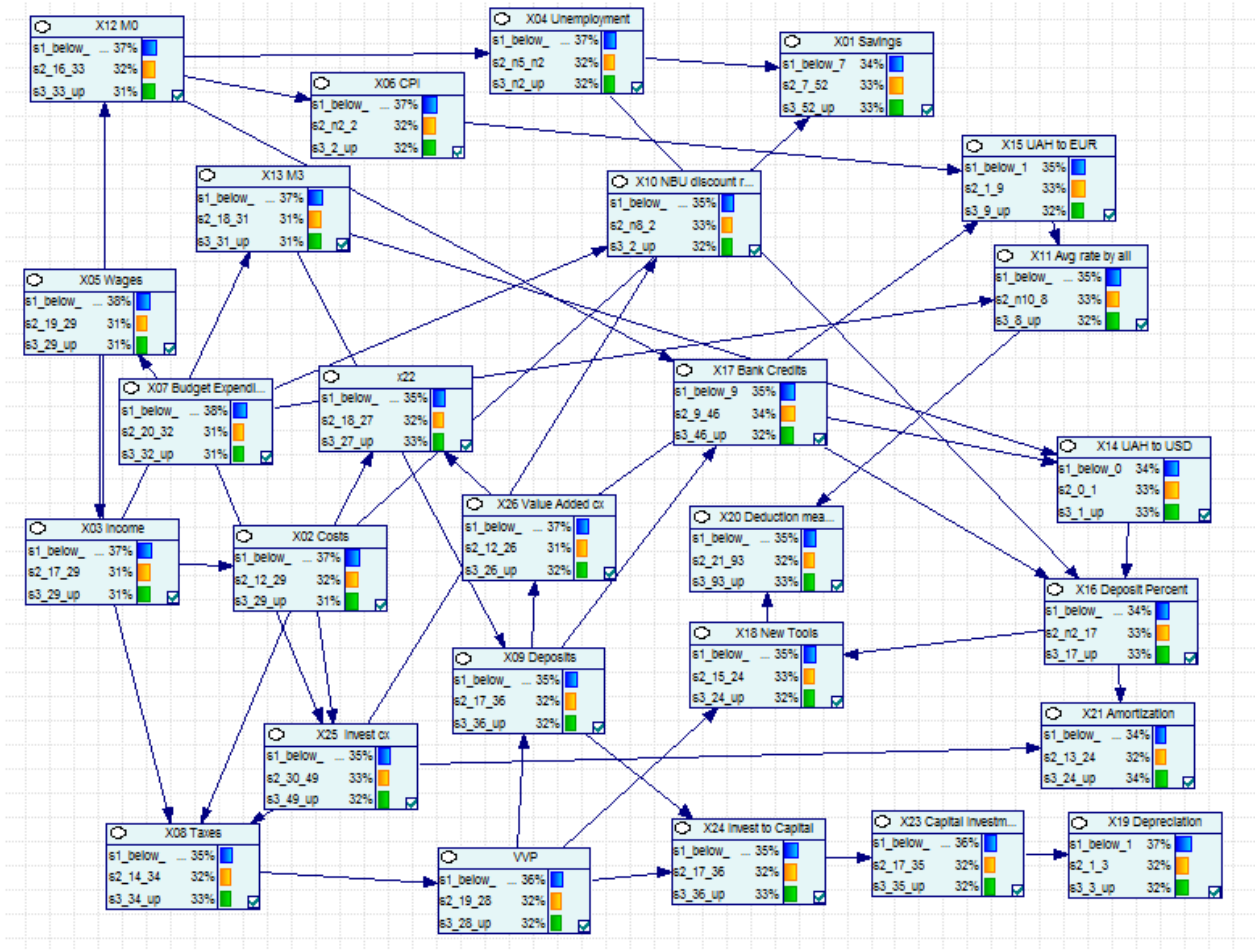


Рисунок 5.11 – Топологія мережі Байєса для сценарію S3

Для побудови топології мережі Байєса, розробленої для сценарію S3 використано програмне забезпечення GeNIe 2.0 з такими налаштуваннями:

– метод побудови K2;

– максимальна кількість предків – 5;

– критерій оптимізації – мінімальна ентропія.

Оцінка параметрів моделі виконана на основі рекурсивного методу найменших квадратів

Таблиця 5.8 – Значення статистичних характеристик тестової вибірки наївної мережі Байєса та за допомогою методу K2 в системі GeNIe 2.0.

Назва моделі	Відсоток неправильно класифікованих спостережень	ROC індекс	Коефіцієнт GINI
Наївна мережа	21,3	0,72	0,44
K2 GeNIe 2.0	12,5	0,81	0,62

Як можна побачити з отриманих результатів моделювання, наведених в таблиці 5.8, більш складний підхід із використанням спеціалізованого програмного забезпечення GeNIe 2.0, на основі методу K2, у порівнянні із найпростішим випадком використання мереж Байєса у вигляді наївної мережі, дає більш якісні прогнозуючі характеристики. В якості тестової вибірки для навчання та тестування використовувалися статистичні дані за період 2003 – 2013 роки, що описуються економічні та соціальні процеси в сільському господарстві.

На основі побудованої топології виявлено, що за реалізації сценарію S3 найбільшого приросту валового регіонального продукту (2 %) можна досягти за умови зростання таких чинників, як податкові надходження (більше ніж на 30 %), обсяг депозитів в економіці (більше ніж на 20 %), капітальних інвестицій в сільське господарство (більше, ніж на 36 %), зростання обсягів валової доданої вартості продукції сільського господарства (більше ніж на 25 %), зростання обсягів капітальних інвестиції в харчову промисловість

(більше ніж на 20 %). Ймовірність реалізації такого сценарію досить низька – 26 %. Якщо ж зростуть податкові надходження (більше ніж на 14 %), обсяг депозитів в економіці (не менше ніж на 17 %) та інвестицій в основний капітал (більше, ніж на 36 %), валова додана вартість продукції сільського господарства (більше ніж на 15 %), обсяги капітальних інвестицій в харчову промисловість (більше ніж на 10 %), то валовий внутрішній продукт зросте на 2,8 %. Ймовірність реалізації такого розвитку подій – 51 %.

Використання когнітивного та ймовірнісного моделювання у задачах формування сценаріїв розвитку соціально-економічних систем дозволяє дослідити виявити причинно-наслідкові зв'язки, що характеризують стан системи та процеси, що мають місце, змоделювати їх можливі варіанти, визначити шляхи покращення соціально-економічної ситуації за умов як позитивного так і негативного розвитку подій.

Дана інформаційна технологія призначена для удосконалення розробки планів та програм соціально-економічного розвитку регіонів України та окремих галузей, може бути використана у системах підтримки прийняття рішень, призначених для органів місцевого самоврядування.

Перевагами пропонованого підходу, який ґрунтується на використанні композиції методів когнітивного та ймовірнісного моделювання є можливість отримувати прогнози та формувати обґрунтовані варіанти рішень навіть в умовах коли чітке формулювання задачі ускладнене або не можливе.

Для визначення ефективності даного методу розроблено прогноз валового регіонального продукту Черкаської області за розробленим сценарієм на 2015-2020 рр. Відхилення прогнозованого значення та відповідного статистичного показника у 2019 р. становило менше 10%.

5.4 Застосування методу подібності процесів за багатомодельного підходу

Підхід щодо обчислення міри подібності часових рядів, може бути застосований і в складі ансамблів моделей [44]. Зокрема, у випадку, коли дані

неповні або є сумніви щодо їх достовірності, доцільно застосовувати методику, що має в основі пошук історичних аналогій поведінки досліджуваних процесів. В цьому випадку необхідна історична вибірка даних та відрізок-шаблон для якого виконується пошук історичного аналогу. Зазвичай для таких задач виміри прив'язані до часу.

Реалізація пропонованої методики виконується після попередньої обробки даних із виконанням описаних у попередніх розділах процедур. Подальші дії виконуються у наступній послідовності:

Крок 1. Робота з вхідними даними: часовим рядом історичних даних та часовим рядом шаблону, який буде виконувати роль цільового ряду, і з яким буде виконуватися порівняння часового ряду історичних даних.

Крок 2. Визначення часового інтервалу на якому буде виконуватися аналіз та агрегування даних. (Виконується для випадку, коли дані мають вигляд транзакцій. Тоді їх необхідно агрегувати до заданого часового інтервалу, наприклад, секундні виміри привести до хвилинних агрегованих значень, або до погодинних і так далі).

Крок 3. Ідентифікація або задання користувачем (аналітиком) значення тривалості сезонного циклу (більшість часових рядів, що описують реальні процеси характеризуються наявністю сезонних складових).

Крок 4. Розбиття вхідного часового історичний ряду даних на відрізки відповідної довжини. В результаті буде отримано K вхідних історичних часових рядів.

Крок 5. Обчислення значень міри подібності між кожним з вхідних часових рядів, отриманих на попередньому кроці (крок 4), та цільовим часовим рядом. Результати будуть мати вигляд ковектору значень подібностей:

$$Sim = \{Sim_1, \ldots, Sim_k\}$$

де Sim_i - міра подібності між i-м вхідним інтервалом та цільовим рядом, що обчислюється за формулою (21).

Крок 6. Пошук оптимального значення подібності, серед усіх значень ковектору отриманого на попередньому кроці (6):

$$max\ \{Sim_i : i = 1,\ldots, K\}$$

Слід зазначити, що у загальному випадку, замість міри подібності, формула (21), може біти використана будь-яка інша метрика – наприклад стандартна ScoreAvg (9), тоді відповідно на кроці 6 вирішується задача мінімізації – тобто визначається пара рядів для яких значення ScoreAvg найменше:

$$min\{ScoreAvg_i : i = 1, \ldots, K\}$$

$$ScoreSim(Y, X) = 100\% \cdot \left(1 - \frac{ScoreAvg_1}{ScoreAvg_0}\right)$$

Після того, як результати сформовані, аналітик на основі побудованих графіків та обчислених статистик, приймає рішення, щодо подальшого використання результатів, або введення додаткової інформації та уточнень в задачу.

В якості прикладу використання запропонованого підходу, можна навести практичний приклад прогнозування погодинне споживання електроенергії на 24 годинному інтервалі, споживачами однієї з енергетичних компаній України [258].

Вхідний набір даних містить 26112 погодинних спостережень за період з 2 січня 2015 року по 24 грудня 2017 року. В якості цільового ряду розглядається дані щодо споживання електроенергії за 24 години 25 грудня 2017 року. Задача дослідження – визначити з якими періодами споживання електроенергії у минулому схожа динаміка споживання 25 грудня 2017 року. Дана задача має практичний сенс, оскільки 25 грудня – свято Різдва Христова для християн західного обряду, яке стало вихідним днем в Україні вперше

саме у 2017 році. Тому аналітиків енергетичної компанії зацікавило питання, з якими святами або іншими днями у минулих періодах буде схожим споживання електроенергії у цей день.

Так як в рамках задачі розглядаються саме погодинні виміри упродовж доби, то аналіз було виконано із урахуванням періодичної сезонної складової – 24 годинний період, що починається з 00 годин та закінчується о 23 годині кожної доби.

Відрізок вхідних даних з 26112 погодинних значень було перетворено у 1088 вхідних інтервалів для аналізу, кожен з яких містить 24 погодинних значення. В таблиці 5.9 наведені перші п’ять та останні два значення показника ступеня подібності рядів.

Таблиця 5.9 – Фрагмент відсортованої таблиці значень ступені подібності рядів енергоспоживання

Номер показника	Дата	День тижня	Ступінь подібності рядів, %
1	02.12.2017	субота	97,99
2	21.11.2015	субота	97,91
3	24.12.2016	субота	97,88
4	19.11.2016	субота	97,71
5	18.11.2017	субота	97,32
…	…	…	…
1087	01.05.2016	неділя	53,23
1088	12.04.2015	неділя	52,99

У таблиці 5.9 наведено лише фрагмент таблиці результатів, повна таблиця містить 1088 рядків. Як можна побачити з табл. 5.1, енергоспоживання 25 грудня 2017 року більше всього схоже на суботні дні у грудні та листопаді 2015-2017 років, але найближчий аналог - 2 грудня 2017 року.

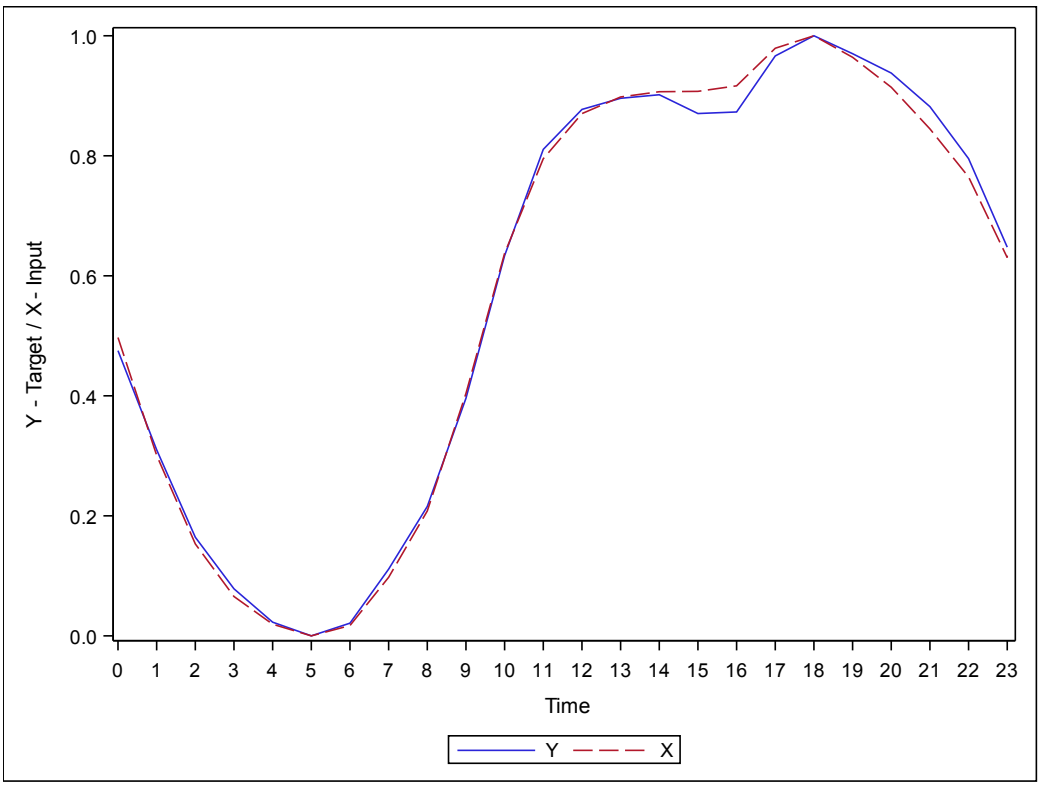


Рисунок 5.13 – Результати порівняння нормалізованих даних погодинного споживання електроенергії між 25 грудня 2017 року (Y) та 2 грудня 2017 року (X). Ступінь подібності рядів 97,99%.

Використовуючи пропонований підхід, енергетичні компанії можуть більш точно прогнозувати енергоспоживання населенням у відповідні передсвяткові, святкові та післясвяткові дні, а також незаплановані явища, такі як введення нових державних свят, яких не було раніше в країні, проведення міжнародних конкурсів та спортивних змагань, що призводять до аномального збільшення чи навпаки, зменшення споживання електроенергії, наприклад, Євробачення та Чемпіонат Європи з футболу.

Однак, слід зазначити, що даний підхід може бути застосований і у вирішенні задач прогнозування нелінійних нестаціонарних процесів, коли немає повної інформації про досліджуваний процес або інформація спотворена.

5.5 Інформаційна технологія, призначена для кластеризації даних в умовах, коли інформація є спотвореною або неповною

Не рідкість, коли на практиці потрібно вирішити задачу класифікації або сегментації об’єктів, осіб, процесів тощо за їхньою поведінкою у часі. Одними із підходів до вирішення цієї проблеми є підхід, оснований на виявленні подібності часових рядів.

У загальному вигляді задачу можна сформулювати наступним чином. Нехай, є *K* часових рядів, $Y = \{Y^{(1)}, \dots, Y^{(K)}\}$, кожен з яких описує поведінку деякого об’єкта в часі. Тоді, можна побудувати матрицю подібностей поведінки об’єктів, розмірності *K* x *K*:

$$Sim = \begin{pmatrix} Sim_{1,1} & \cdots & Sim_{1,K} \\ \vdots & \ddots & \vdots \\ Sim_{K,1} & \cdots & Sim_{K,K} \end{pmatrix} = \begin{pmatrix} Sim(Y^{(1)}, Y^{(1)}) & \cdots & Sim(Y^{(1)}, Y^{(K)}) \\ \vdots & \ddots & \vdots \\ Sim(Y^{(K)}, Y^{(1)}) & \cdots & Sim(Y^{(K)}, Y^{(K)}) \end{pmatrix}$$

де $Sim_{i,j} = Sim(Y^{(i)}, Y^{(j)})$ – це значення міри подібності між i – м та j – м часовими рядами, що описують поведінки відповідних об’єктів.

Зауважимо, що матриця має симетричний вигляд, тобто $Sim_{i,j} = Sim_{j,i}$, а діагональні елементи в загальному випадку не обчислюються, тому що ступінь подібності часового ряду з самим собою буде стовідсотковою.

З урахуванням наведених зауважень стає зрозумілим, що потрібно в загальному випадку зробити $\frac{K \cdot (K-1)}{2}$ попарних порівнянь між часовими рядами.

Після того як усі необхідні обчислення виконані та отримана матриця подібності, об’єкти об’єднуються в групи із використанням алгоритму кластеризації. Для визначення кількості кластерів у задачах прогнозування нелінійних нестаціонарних процесів, за допомогою методу мінімальної дисперсії Уорда, k-середніх та інших методів, що базуються на використанні

мінімізації суми квадратів усередині кластера для задач, як показали чисельні експерименти доцільно застосовувати критерій кубічної кластеризації (CCC), який розраховується із використанням методів Монте-Карло [188] (5.1):

$$CCC = ln\left(\frac{1 - E(R^2)}{1 - R^2}\right)\frac{\sqrt{\frac{np^*}{2}}}{\left(0.001 + E(R^2)\right)^{1.2}} \tag{5.1}$$

Формулу (5.1) було отримано емпірично, намагаючись стабілізувати дисперсію між різною кількістю спостережень, змінних та кластерів [188].

$$E(R^2) = 1 - \left|\frac{\sum_{j=1}^{p^*}\frac{1}{n+u_j} + \sum_{j=p^*+1}^{p}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p}u_j^2}\right| \cdot \left|\frac{(n-q)^2}{n}\right| \cdot \left|1 + \frac{4}{n}\right|$$

$$R^2 = 1 - \frac{p^* + \sum_{j=p^*+1}^{p}u_j^2}{\sum_{j=1}^{p}u_j^2}$$

$$u^* = \prod_{j=1}^{p^*} s_j$$

$$c = \left(\frac{v^*}{q}\right)^{\frac{1}{p^*}}$$

$$u_j = \frac{s_j}{c}$$

де q – кількість кластерів, з якими порівнюють p^*, v – розмір гіперкуба, c – довжина ребра гіперкуба, який отримують при діленні гіперкуба v на q частин, u_j – кількість гіперкубів вздовж -го виміру основного гіперкубу, s_j – корінь квадратний з -го власного числа, що описує множину даних X, n – кількість спостережень.

CCC-критерій використовується для визначені кількості кластерів за методо Уорда, наступним чином:

– послідовно перевіряються гіпотези, про кількість кластерів;
– для кожної гіпотези робиться розбиття на відповідну кількість кластерів та обчислюється значення CCC-критерію;
– фахівець аналізує отримані значення CCC-критерія та приймає рішення щодо відповідної кількості кластерів, які описують набір даних.

Емпірично експерти [188] з'ясували та рекомендують використовувати одне з двох правил.

Правило 1. Обирають ту кількість кластерів, для якої значення CCC-критерія перевищує 3.

Правило 2. Обирають ту кількість кластерів, для якої значення CCC-критерія стрімко зростає, після падіння.

Окрім наведеного вище підходу групування об'єктів на основі методів кластерного аналізу, можуть бути використані й інші методи кластеризації. Наприклад, на основі експертного досвіду аналітика, а саме, після того, як усі необхідні обчислення виконані та отримана матриця подібності, об'єкти об'єднуються в групи, за правилом: якщо міра подібності $Sim_{i,j} = Sim(Y^{(i)}, Y^{(j)})$ більша за деяке порогове значення, то i-й та j-й об'єкти, що описуються відповідно часовими рядами $Y^{(i)}$ та $Y^{(j)}$ належать одній групі.

Застосування пропонованого методу кластеризації рамках багатомодельного підходу доцільно продемонструвати на прикладі вирішення практичної задачі групування регіонів України обсягом капітальних інвестицій, спрямовуваних на охорону навколишнього середовища.

Вхідними даними є статистичні показники обсягу капітальних інвестицій в охорону навколишнього природного середовища в регіонах Укрархни за 2011-2018 рр. [173]. Всього досліджується 26 часових рядів. Для обчисленні значень схожості рядів виконувалася попередня нормалізація даних, а саме, нормування за діапазоном (3.5):

$$z_i = \frac{x_i - \min\limits_{i=1,..,m}\{x_i\}}{\max\limits_{i=1,..,m}\{x_i\} - \min\limits_{i=1,..,m}\{x_i\}}$$

Результати розрахунків представлені у Додатку І. В якості порогового значення схожості рядів за метрикою *ScoreSim* обрано значення 70%, тобто на графі зав'язків відображаються тільки ті зв'язки між областями що дорівнюють або перевищують 70% (найбільш суттєві).

На основі значень схожості рядів, отриманих за метрикою ScoreSim, (додаток І), використовуючи будь-який з методів кластеризації, можна згрупувати регіони, за ступенем подібності (близькості за метрикою ScoreSim).

Таблиця 5.10 – Результати кластерного аналізу, виконаного за методом Уорда (Ward) регіонів

Регіон	Номер кластеру	Кількість регіонів
Дніпропетровська, Запорізька, Київська, Миколаївська, Сумська, Тернопільська, Хмельницька, Черкаська області, місто Київ	1	9
Волинська, Івано-Франківська, Херсонська, Чернігівська, Чернівецька області	2	5
Кіровоградська, Луганська, Одеська, Рівненська, Запорізька області	3	5
Вінницька, Донецька, Закарпатська, Львівська, Полтавська, Харківська області	4	6

Код програми для імпортування даних та побудови дендограми за методом Уорда (Ward), в системі SAS Studio.

```
proc import datafile=….\Similarities.xlsx'
```

```
out=Similarities
dbms=XLSX
replace;
sheet='Similarity';
run;

ods graphics on;

proc cluster data=work.Similarities outtree=tree
method=ward plots=dendrogram;
   id _input_;
run;
```

Графічно результати кластеризації представлені на рис. 5.14.

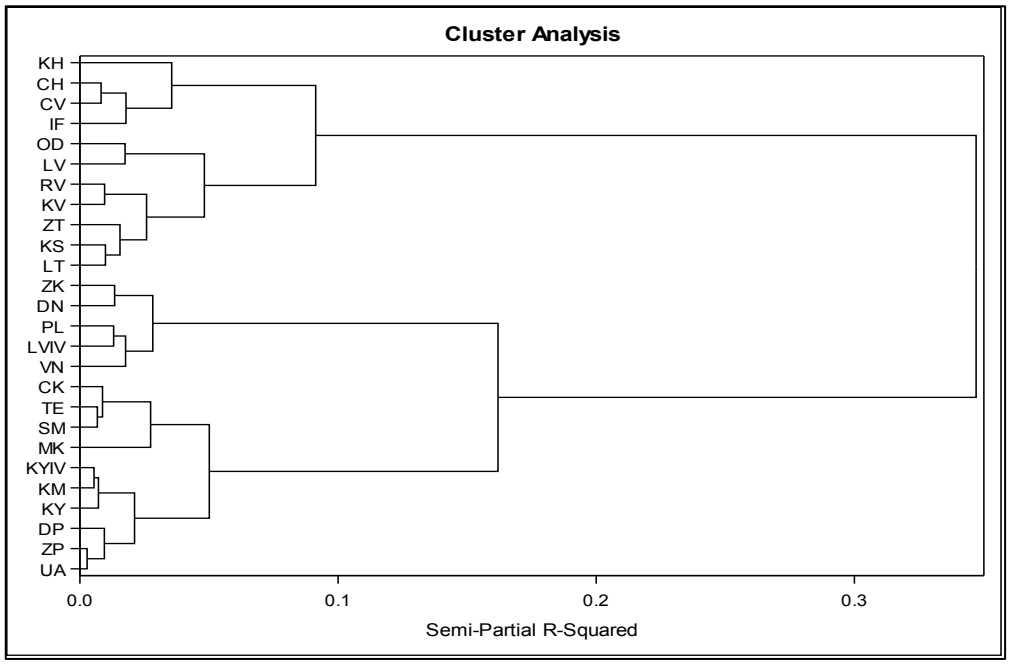


Рисунок 5.14 – Дендограма групування регіонів за обсягом капітальних інвестицій в охорону навколишнього середовища у чотири кластери методом Уорда (Ward), виконана в системі SAS Studio

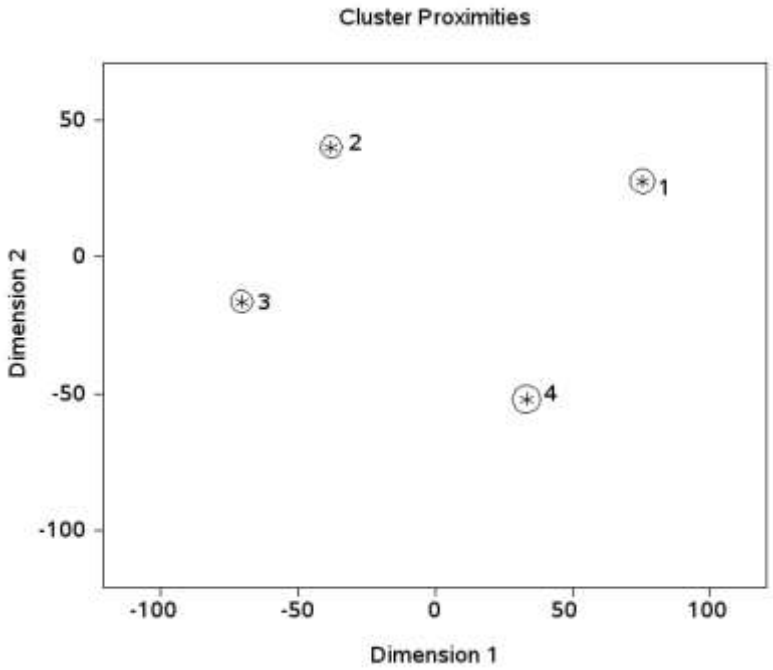


Рисунок 5.15 - Відстань між кластерами, результати роботи системи SAS Enterprise Miner

Для візуалізації результатів кластеризації було використано програму SAS Graph Network Visualization WorkShop (рис. 5.16).

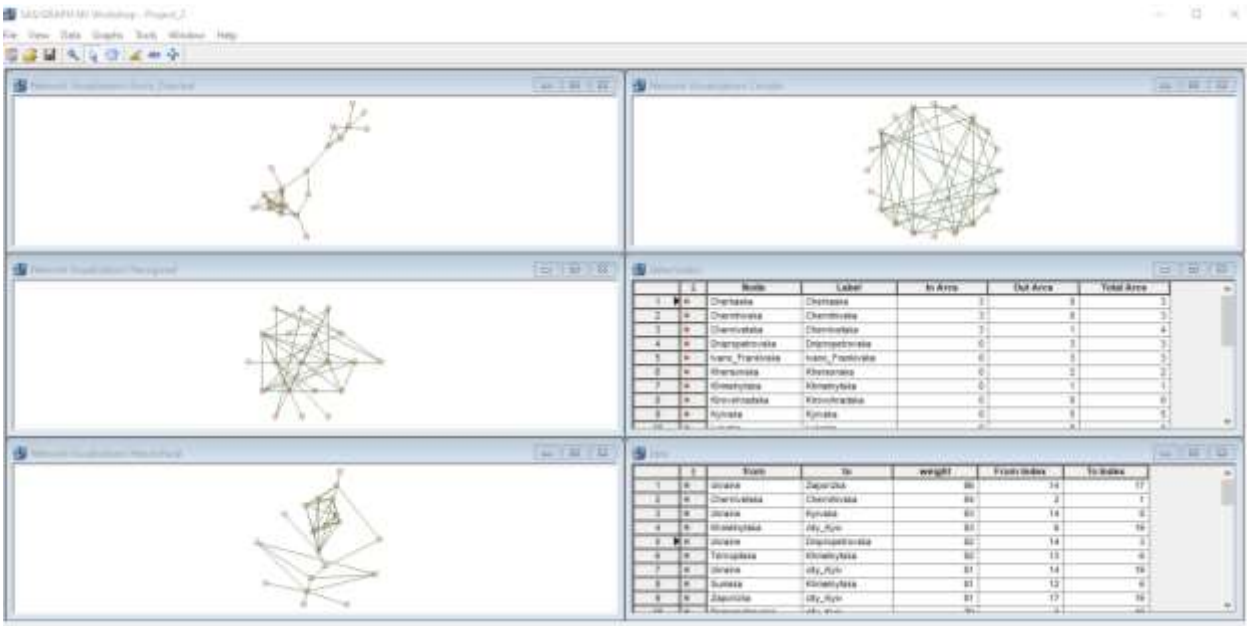


Рисунок 5.16 – Скріншот робочого вікна системи SAS Graph Network Visualization WorkShop

Отримано представлення у вигляді кругового графу; ієрархічного графу; гексагонального графу та багаторівневого графу.

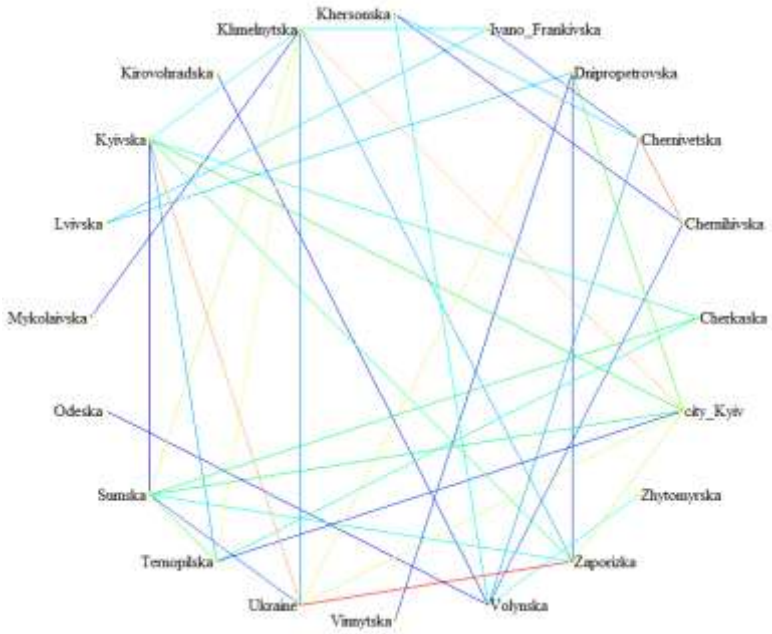


Рисунок 5.17- Круговий граф зв’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

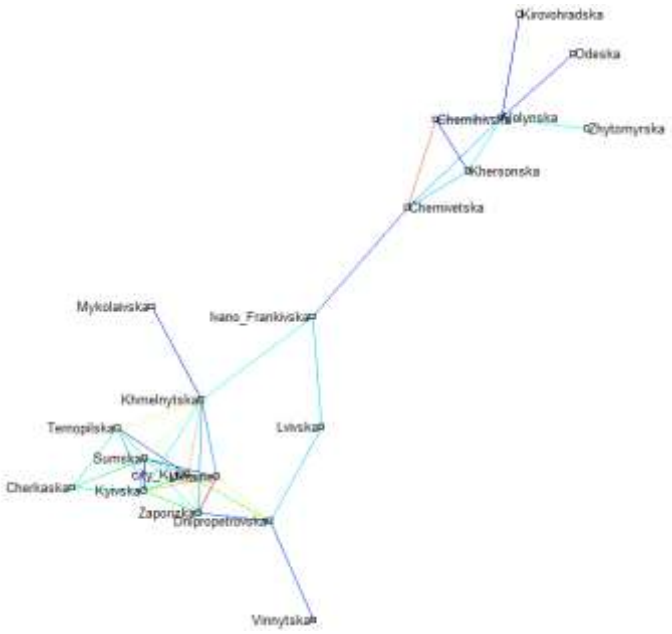


Рисунок 5.18 – Граф зав’язків за координатами користувача, відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

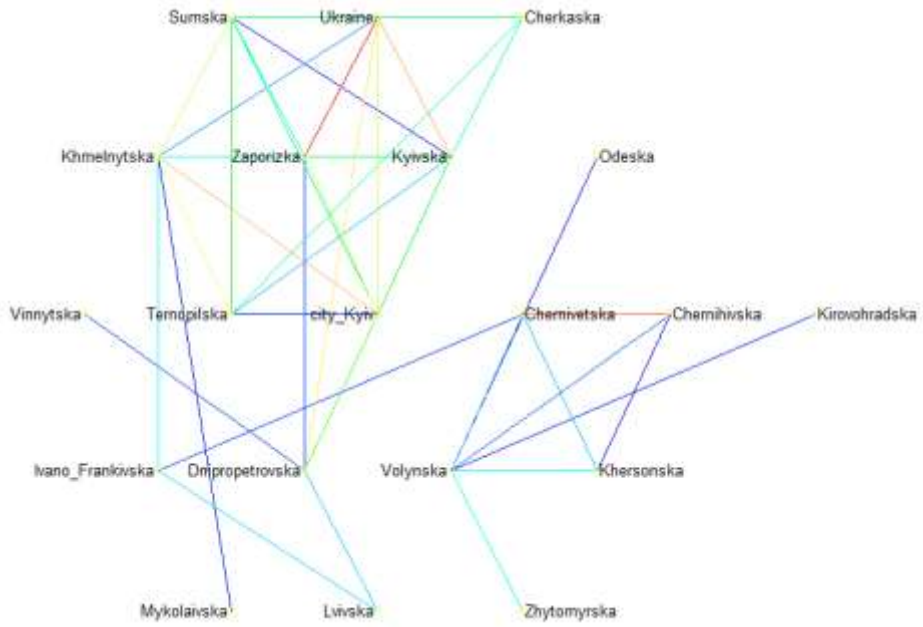


Рисунок 5.19 – Гексагональний граф зав’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

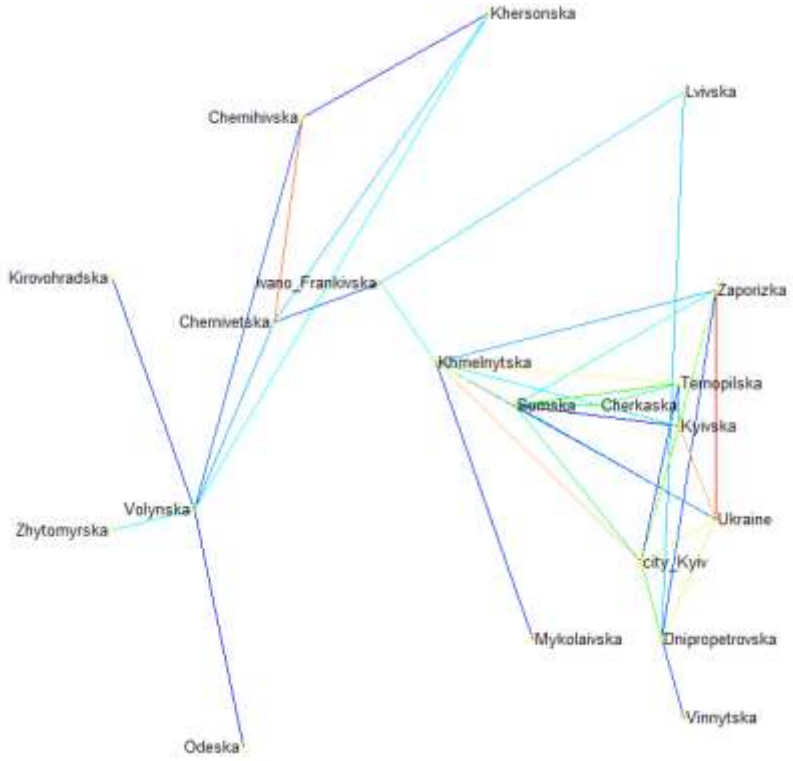


Рисунок 5.20 – Ієрархічний граф зав’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

Рис. 5.17 – 20 відображають одну і ту саме мережу, але з різним розташуванням вершин, при чому колір залежить від значення метрики ScoreSim.

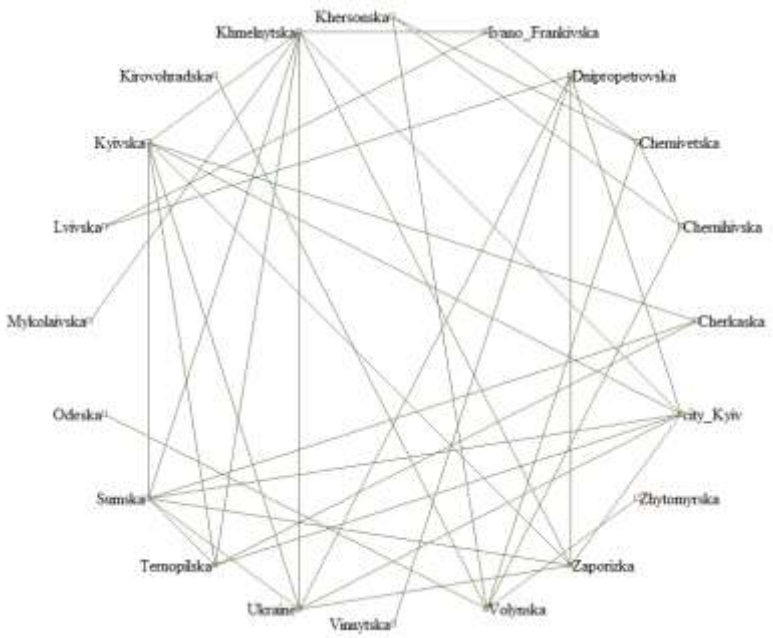


Рисунок 5.21 – Круговий граф зв’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%

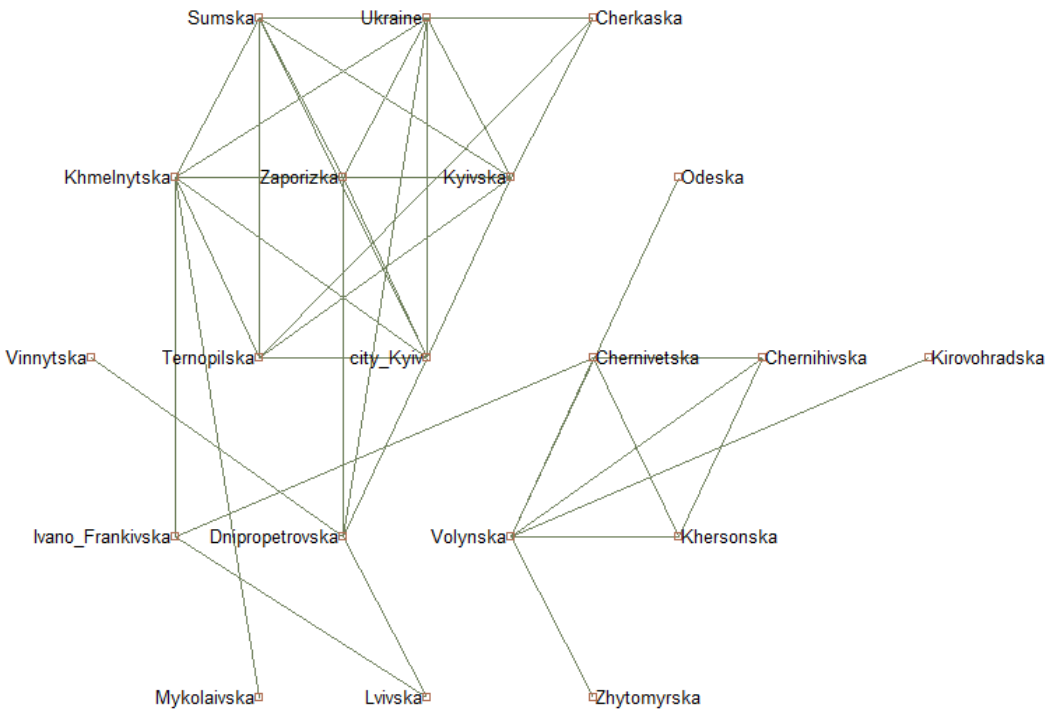


Рисунок 5.22.- Граф зв’язків, за координатами користувача, відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%

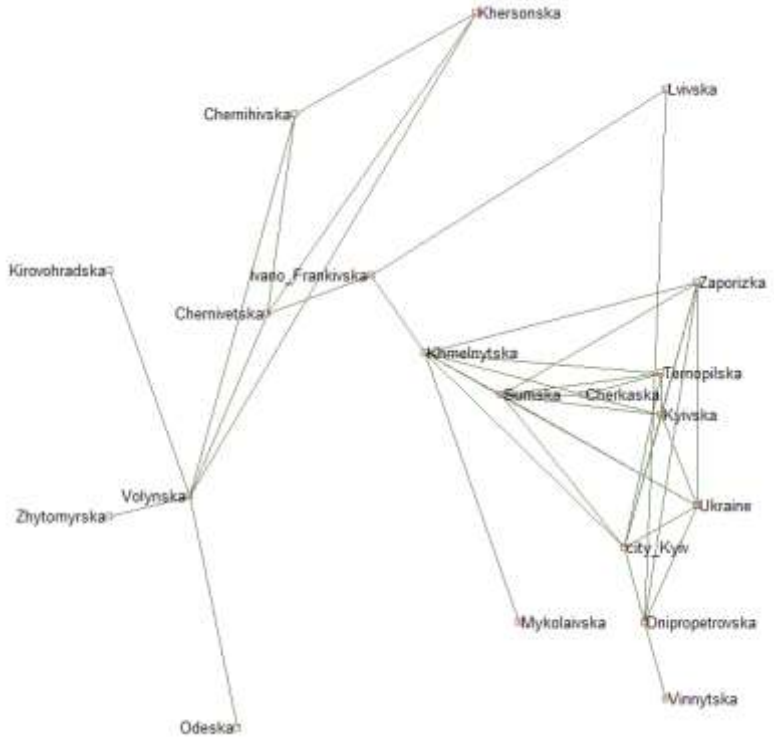


Рисунок 5.23 – Гексагональний граф зв’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%

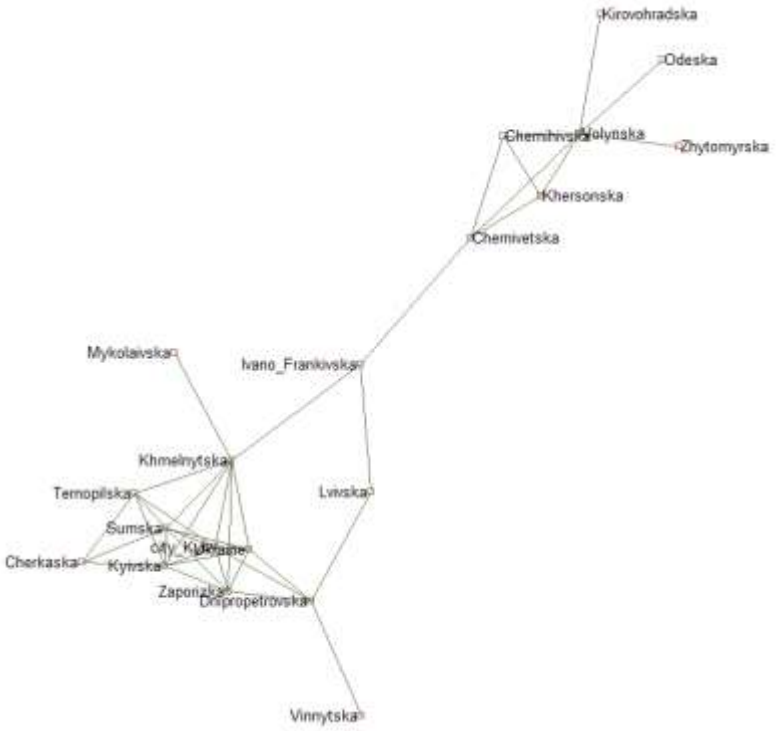


Рисунок 5.24 – Ієрархічний граф зв’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%

Як можна побачити з отриманих результатів (Додаток Н), між значеннями ступеня зв'язку для рядів за кореляцією Пірсона та метрикою подібності рядів ScoreSim є відмінності, насамперед для випадків, коли значення кореляції хоча і великі, але від'ємні, метрика ScoreSim прямує до нуля.

На основі результатів кластеризації наведених в Додатку Н щодо приналежності до кластеру регіону, за результатами аналізу методом Уорда (Ward) в системі SAS Enterprise Miner, можна зробити висновок, що:

– Дніпропетровська, Запорізька, Київська, Миколаївська, Сумська, Тернопільська, Хмельницька, Черкаська області та місто Київ належать до першого кластеру;

– Волинська, Івано-Франківська, Херсонська, Чернівецька та Чернігівська належать до другого кластеру;

– Житомирська, Кіровоградська, Луганська, Одеська та Рівненська належать до третього кластеру;

– Вінницька, Донецька, Закарпатська, Львівська, Полтавська та Харківська до четвертого кластеру.

Отже, використовуючи методику кластеризації даних, основану на теорії подібних процесів, отримано результати, які можна використовувати в ансамблях моделей, в тому числі і в якості змінної, що включена в модель як змінна, запропонована користувачем.

5.6 Інформаційна технологія використання неструктурованих даних за багатомодельного підходу

Значні обсяги різноманітної за змістом і за структурою інформації, стрімко накопичуються у відкритих Інтернет-джерелах. Текстових ресурси, такі як блоги, портали новин, статті в електронних виданнях містять актуальну інформацію, яка дозволяє розширити уяву про обєкт досліження за рахунок цілеспрямованого видобування знань. Традиційно, для цих цілей

застосовується розвідувальний пошук, коли аналізуються документи з різних джерел, в тому числі і з Інтернет-ресурсів, в результаті чого, дослідник одержує набір відомостей по темі пошукового запиту та виявляє підтеми досліджуваної проблеми, що в подальшому потребує обробки отриманих результатів експертами [187, 190].

Важливішим є виявлення потенційних проблем, які існують і можуть виникнути за поточного розвитку подій. Значна кількість науковців при вирішенні слабкоструктурованих проблем віддають перевагу використанню когнітивних [257], причинно-наслідкових моделей в поєднанні з математичними та економетричними моделями [244]. Інформаційною основою для побудови таких моделей, враховуючи постійне зростання обсягів інформації, що накопичується в мережі Інтернет є слабкоструктуровані Big Data. Тому, у процесі відбору чинників при формуванні сценаріїв, найбільш значимих змінних та змінних, які пропонує включити користувач, постає задача використання інформації із слабкоструктурованих масивів текстових Інтернет. З цією метою запропонована методика формування набору альтернатив та цільової установки дослідження, передбачає послідовну реалізацію набору кроків, які охоплюють як попередню обробку текстової інформації, формування набору правил для аналізу, відбір найбільш значимих критерії та цілей а також візуалізацію одержаних результатів:

1. Визначити коло питань, які цікавлять дослідників
2. Розробити таксономію проблемної сфери
3. Проаналізувати веб-ресурси з обраної проблематики, обравши оптимальні інструменти дослідження
4. Розробити правила для текстового аналізу
5. За потреби, визначити емоційне забарвлення інформації, що обробляється
6. Окреслити коло найбільш затребуваних питань
7. Сформувати матриці PEST та SWOT аналізу (за потреби)

8. Розробити сценарії розвитку подій враховуючи результати PEST та SWOT аналізу

9. Описати найбільш значимі фактори для кожного із варіантів сценаріїв

10. Розробити ймовірнісну модель для кожного із варіантів сценарію

11. Зробити висновки про можливі варіанти зниження (усунення) загроз та ризиків.

Для визначення емоційного забарвлення текстової інформації, запропоновано побудувати множини правил для відбору позитивних та негативних вподобань для кожної частини мови.

На етапах 1-3 та 5 запропоновано забезпечити автоматизоване опрацювання значних обсягів переважно тестової інформації за допомогою таких потужних сучасних інструментів – сімейство програмних засобів SAS Textual Analytics. Дані програмні засоби мають широкий спектр можливостей для аналізу структурованих та неструктурованих даних: це і SAS Web Crawler, SAS Search and Indexing, SAS Enterprise Content Categorization, SAS Ontology Management, SAS Text Miner та SAS Sentiment Analysis Studio (рис.5.25).

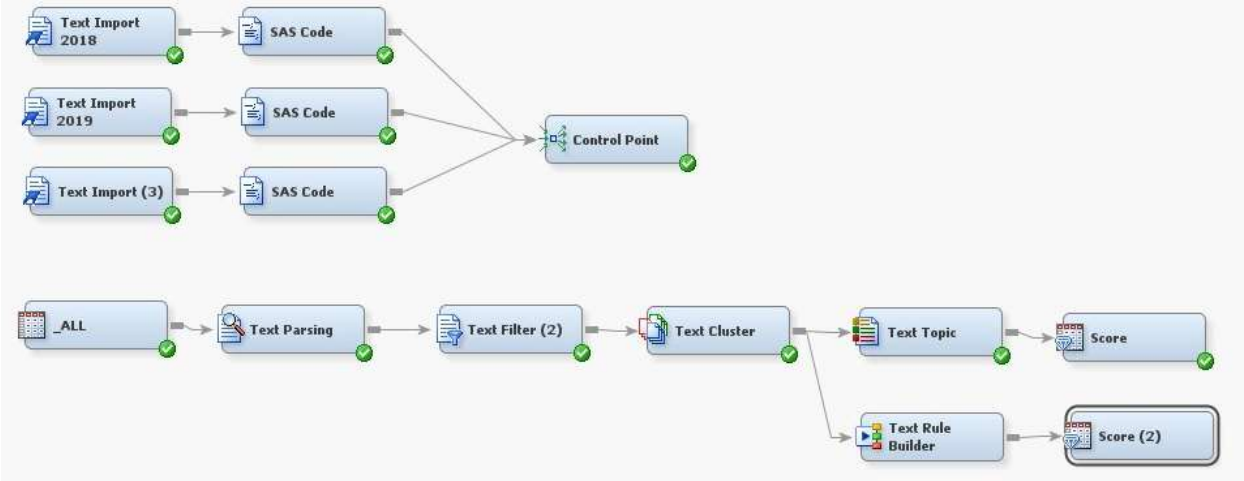


Рисунок 5.25 – Схема технологічного процесу з аналізу текстових повідомлень в системі SAS Text Miner

В ході дослідження (етапи 1-3) дані для відбору альтернатив першого та другого рівнів пропонується видобувати з веб-простору, використовуючи SAS Web Crawler – програмне забезпечення, що сканує сторінки мережі Інтернет та використовує вбудовану структуру посилань, за допомогою якої веб-сторінки пов’язані одна з одною. Веб-краулеру необхідно вказати посилання на сторінку, з якої необхідно розпочати сканування. Цю веб-сторінку ще називають точкою входу. У даному дослідженні точкою входу переважно є посилання безпосередньо на розділ новин певного веб-сайту. Звичайний веб-краулер починає свою роботу з дослідження точки входу і сканує розміщені на ній посилання, котрі, у свою чергу, посилаються на інші веб-сторінки. Віднайдені таким чином сторінки також скануються на наявність посилань на інші сторінки. Глибину такого пошуку можна обмежити шляхом встановлення значень певних параметрів. У ході парсингу веб-сторінок веб-краулер вилучає з них необхідний контент і зберігає його у вигляді файлів-документів.

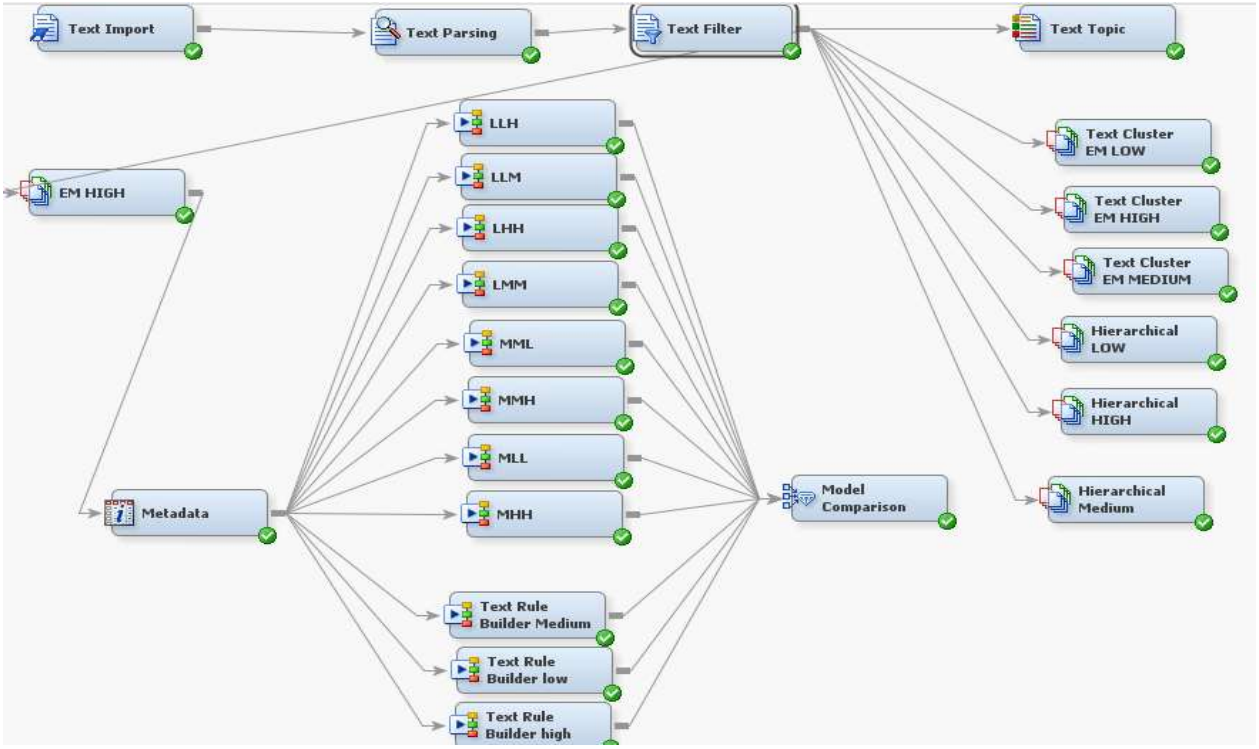


Рисунок 5.26 - Схема процесу аналізу контенту

Для проведення дослідження, передбаченого на етапі 5 запропоновано використовувати інструменти SAS Sentiment Analysis Studio, які дозволяють аналізувати тональність тексту на основі побудованих правил, будувати моделі, в основу яких покладені правила та гібридні моделі (на основі статистичних моделях та моделях на основі правил).

Перевірка працездатності методики на прикладі дослідження аналізу сільськогосподарських даних. Перевірку адекватності розробленої методики було виконано в ході аналізу Інтернет-джерел текстів щодо тематики виробництва в Україні окремих сільськогосподарських культур та пріоритетних напрямків їх реалізації на внутрішньому та світовому ринках для формування альтернатив першого та другого рівня задачі дворівневого багатокритеріального аналізу альтернатив у системі підтримки прийняття рішень розвитку агропромислового виробництва України. Дослідження виконано на прикладі основної експортної культури України – кукурудзи.

Для аналізу корпусу текстів було використано SAS Textual Analytics Suit, що складається з SAS Content Categorization та SAS Sentimental Analysis Studio за допомогою яких була розроблена таксономія із 776 правил [252].

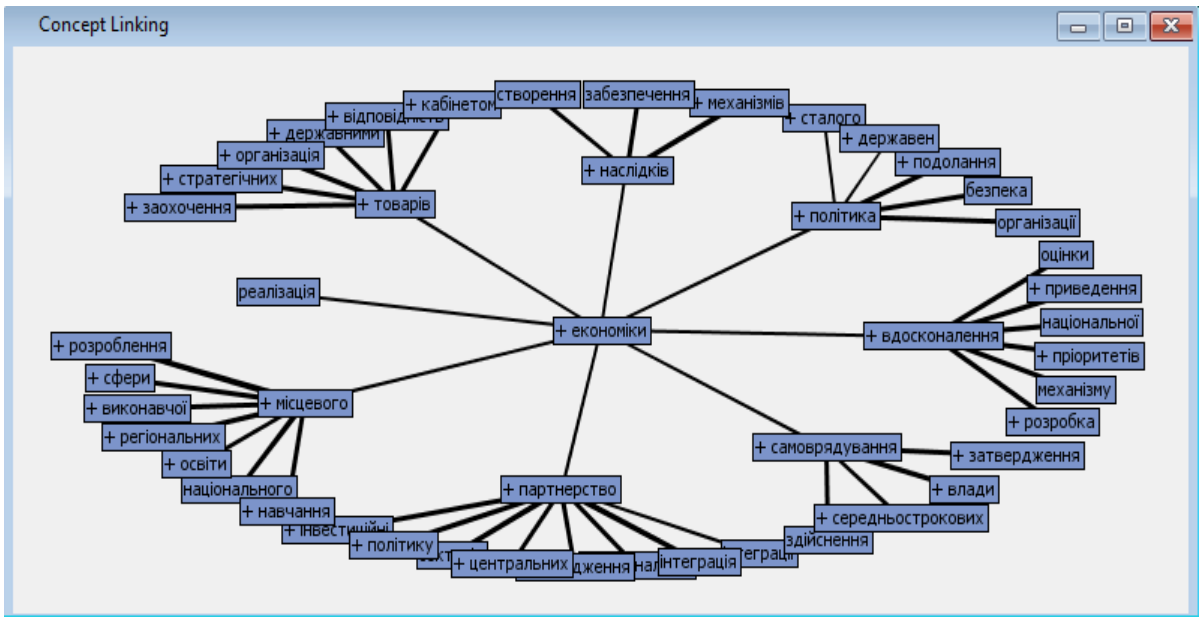


Рисунок 5.27 – Формування концептуальних-зв’язків

Використання інструментів тестової аналітики компанії SAS значно знизило трудомісткість одержання експертного висновку: було витрачено 40 людино-годин проти 900 людино-годин без використання вказаних інструментів.

Таблиця 5.10 – Приклад ієрархії кластерів у вигляді таблиці з описом кожного з термів в режимі High SVD resolution

Номер кластеру	Опис кожного з термів	Кількість документів	Частка в колекції документів
1.0	території +господарства її +або їх +видів вони чи +вирощування +України +є +Україні +зокрема +довкілля	1146	0,24
2.0	реалізація державної політики с. Основними м. національної р. Показники результати державного період проведення вирощування	936	0,20
3.0	+із проблеми +відходи +ніж вже для року ще +продажу	633	0,14
4.0	+млн т р +вчора землі+кукурудза т. +тис +ринку листопада +компанії	1673	0,36
5.0	потенціал результати період умови проведення системи рівень якщо після при +його +може їх +розвитку лише	919	0,20
6.0	м. національної рівень системи т. Проблеми при цьому + станом +даними +ринку	516	0,11

Для зручності побудови звітів отриманий набір даних, який по суті являє собою багатомірну OLAP-базу даних, було розкладено за чотирма рівнями.

Синтез правил щодо аналізу досліджуваної системи на прикладі домена, обраного користувачем через виявлення станів «високий, низький,

зростаючий або спадаючий рівень потенційно негативного або позитивного показника» із додатковим вилученням фактуальної словникової інформації (словник-категоризатор країн, ієрархія понять).

В результаті проведеного дослідження одержано наступні результати. Виявлено, що найбільш обговорюваними темами щодо тенденцій розвитку світового ринку аграрної продукції є наявність генно-модифікованих організмів (7,42%), виробництво біодизелю (34,68%), проблеми вирощування кукурудзи (55,87%).

Щодо альтернатив другого рівня, пов'язаних із учасниками ринку кукурудзи, то загалом на блогах, форумах та сайтах новин, присвячених сільському господарству, найчастіше згадують про Україну – 21%, Росію – 17%, Європейський Союз – 17%, США – 10%, Польщу – 8,5% та інші країни – 26,5%. Найбільш важливими є альтернативи нарощування обсягів виробництва кукурудзи – 56%, виробництво біопалива – 34%, генетично модифіковані організми – 7% та інші– 4%.

В ході аналізу альтернатив значення цін та розвитку ринку сільськогосподарських культур, було виявлено, що розподіл пріоритетів змінювався протягом 2014-2016 маркетингових років. Якщо у 2015-му році структура альтернатив була наступною: ринкові розрахунки - 30%, інвестиційна діяльність – 15%, ринковий попит – 10%, ринкова пропозиція – 10%, розподіл ринку – 10%, ефективність вирощування культур – 5%. У 2016 році структура дещо відрізняється експертів більше цікавили питання про ринкові розрахунки (25%) та інвестиційну діяльність (20%).

Щодо емоційного забарвлення досліджуваної інформації, то слід відзначити зростання негативної тенденції щодо «квоти на продаж продукції» та «генно-модифіковані організми», позитивної – «кукурудза» та «Китай». Тобто, враховуючи негативний фактор наявності обмежених квот на продаж кукурудзи в Європейський Союз та нарощування обсягів виробництва кукурудзи в Україні, в якості пріоритетного ринку збуту зерна кукурудзи слід розглядати Китай.

Іншим прикладом використання неструктурованої інформації є задача кредитного скорингу.

Вхідною інормацією є персональні дані 3000 клієнтів банку на основі яких необхідно спрогнозувати ймовірність повернення кредиту.

У модель біли включені такі чинники як вік клієнта; коефіцієнт відношення рівня доходу до позик, що сплачуються щомісяця; наявність власного авто; кількість дітей; дохід клієнта; кількість непогашених кредитів; наявність власної квартира або будинку; трудовий клієнта (років).

В результаті кластеризації, виконаної за EM-методом, клієнти банку були згруповані у три кластери. В таблиці 5.11 представлені описи відповідних кластери та терми, виявлені з корпусу документів, що їх описують.

Таблиця 5.11 – Терми, що описують кластери клієнтів

Кластер	Терми, що описують кластер.
1	економічна криза, влада, пандемія, стурбованість, швидка позика, дитина
2	комунальні послуги, ліки, пенсія, субсидія, соціальний працівник
3	перспективи, власна справа, освіта, IT, подорож, автомобіль, відпочинок, цікаві місця

На основі виявлених залежностей були побудовані моделі логістичної регресії.

Модель 1 – класична модель логістичної регресії на стандартних показниках, сформованих на аплікаційних формах клієнтів, що заповнювалися в місцях видачі споживчого кредиту.

Модель 2 – розширена модель логістичної регресії, в яку включено показник щодо належності клієнта до відповідного кластеру, який було визначено на основі результатів SVD аналізу.

Рівняння моделі 1 для обчислення значення прогнозу-рейтингу (логіт балу) має вигляд:

$$logit(\hat{p}) = \exp(-4{,}6301 + 0{,}0524 \cdot \text{AGE} + 0{,}3368 \cdot \text{Ratio} + 0{,}1361 \cdot \text{CAR} + 0{,}129 \cdot \text{CHILDREN} + 0{,}00015 \cdot \text{INCOME} - 0{,}1056 \cdot \text{NUMLOAN} + 0{,}0232 \cdot \text{RESIDENCE} + 0{,}00153 \cdot \text{TIMEJOB})$$

Рівняння моделі 2 для обчислення значення прогнозу-рейтингу (логіт балу) має вигляд:

$$logit(\hat{p}) = \exp(-4{,}3826 + 0{,}0456 \cdot \text{AGE} + 0{,}3017 \cdot \text{Ratio} + 0{,}1381 \cdot \text{CAR} + 0{,}00828 \cdot \text{CHILDREN} - 0{,}1312 \cdot \text{CLUSTER}_1 - 0{,}2222 \cdot \text{CLUSTER\_2} + 0{,}00016 \cdot \text{INCOME} - 0{,}123 \cdot \text{NUMLOAN} + ,00198 \cdot \text{RESIDENCE} + 0{,}00149 \cdot \text{TIMEJOB})$$

Для переходу к значенню прогнозу-оцінки (ймовірності повернення кредиту) використовується рівняння:

$$\hat{p} = \frac{1}{1 + \exp(-logit(\hat{p}))}$$

Таблиця 5.12 – Значення оцінок параметрів та статистик моделі 1

Параметр	Оцінка параметра	Стандартна похибка	p-значення
Intercept	-4,6301	15,317	0,7624
AGE	0,0524	0,0046	<,0001
Ratio	0,3368	1,3145	0,8275
CAR=YES	0,1361	0,2047	0,5061
CHILDREN	0,129	0,0422	0,0022
INCOME	0,00015	0,00003	<,0001
NUMLOANS	-0,1056	0,0509	0,0381
RESIDENCE=YES	0,0232	0,1029	0,822
TIMEJOB	0,00153	0,000487	0,0016

Значення оцінок параметрів та статистик моделі 2 наведені в таблиці 5.13.

Таблиця 5.12 – Значення оцінок параметрів та статистик моделі 2.

Параметр та його опис	Оцінка параметра	Стандартна похибка	p-значення
Intercept (вільний член моделі)	-4,3826	15,3948	0,08
AGE (вік клієнта)	0,0456	0,00515	78,41
Ratio (коефіцієнт відношення рівня доходу до позик що сплачуються щомісячно)	0,3017	1,3919	0,05
CAR=YES (наявність власного авто)	0,1381	0,2056	0,45
CHILDREN (Кількість дітей)	0,00828	0,0437	0,04
Cluster = 1 (приналежність до кластеру)	-0,1312	0,0977	4,67
Cluster = 2 (приналежність до кластеру)	-0,2222	0,0824	7,72
INCOME дохід клієнта	0,00016	0,00003	28,3
NUMLOANS (кількість непогашених кредитів)	-0,123	0,0512	5,76
RESIDENCE= YES (наявність власної квартири або будинку)	0,00198	0,1033	0
TIMEJOB (трудовий стаж, років)	0,00149	0,00048	9,66

Як можна побачити з результатів, представлених у табл. 5.12 після включення в модель логістичної регресії нового регресору – «приналежність до кластеру», в модель було додано дві проектні змінні (приналежність к 1 та 2 кластерам). Це стандартна ситуація, коли номінальна змінна, з n рівнями, представляються в моделі у вигляді (n – 1) регресора індикаторного типу.

Аналіз значень коефіцієнтів підвищення шансів повернення кредиту (odds ratio) показав, що:

- збільшення віку на 1 рік збільшує ймовірність повернення кредиту на 4,7%;

- збільшення значення Ration на 1% збільшує ймовірність повернення кредиту на 0,2%;

- наявність власного автомобіля збільшує ймовірність повернення кредиту на 12%;

- збільшення кількості дітей збільшує ймовірність повернення кредиту на 0,08%;

- збільшення доходу на 1000 грн збільшуєймовірність повернення кредиту на 0,4%;

- наявність кожного додаткового непогашеного кредиту зменшує ймовірність повернення кредиту на 16%;

- наявність власного житла збільшує ймовірність повернення кредиту на 4%;

- наявність кожного додаткового року трудового стажу збільшує ймовірність повернення кредиту на 1%;

- за приналежності до 1 кластеру, на відміну від випадку принаалежності до 3 кластеру, зменшується ймовірність повернення кредиту на 12%, а приналежність до 2 кластеру, у порівняння з приналежністю до 3 кластеру, зменшує ймовірність повернення кредиту на 18%.

Порівняння результатів, отриманих за побудованими моделями представлене в табо.5.13.

Таблиця 5.13 – Статистичні характеристики отриманих моделей

Назва моделі	Похибка класифікації	Коефіцієнт Gini
Модель 1	32%	0,69
Модель 2	28%	0,72

Як можна побачити з табл. 5.13, підмішування до математичної моделі додаткового показника, отриманого на основі результатів кластеризації

клієнтів за текстовою інформацією, що їх описує, дозволяє зменшити похибку класифікації на 4%.

Отже, використання SAS Textual Analytics дозволило знизити вплив людського фактору на вибір альтернатив та значно пришвидшити опрацювання значних обсягів неструктурованої інформації з Інтернет-джерел та сформувати найбільш повне уявлення щодо розвитку ситуації, наявності ризиків та ознак кризових явищ, створити передумови для дослідження багатофакторних причинно-наслідкових залежностей, що можуть виникати серед різнорідної та неструктурованої інформації, джерелом якої є відомості, одержувані з офіційних та неофіційних джерел.

5.7 Статистичні методи та підходи, застосовані до оцінювання моделі

Існує багато різних статистичних критеріїв, що можна використовувати для вибору моделі серед декількох конкуруючих.

Залежно від рівня вимірювання цільової змінної можуть використовуватися різні підходи. Для цільової змінної інтервального типу використовуються усереднення або максимізація, для категоріального (номінального) типу усереднення, максимізація або голосування. Зазвичай в якості стандартного підходу при багатомодельному моделюванні використовується усереднення.

Усереднення, для випадку категорійних цільових змінних обчислює середнє значення апостеріорних ймовірностей окремих моделей. У випадку інтервальних цільових змінних обчислюється середнє значення для значень прогнозів-оцінок. Максимізація, для випадку категорійних цільових змінних обирає максимальне значення апостеріорної ймовірності серед усіх індивідуальних значень моделей.

Для випадку інтервально цільової змінної обирається в якості результату максимальне значення серед усіх прогнозів-оцінок.

Голосування – використовується тільки для категоріальних цільових змінних. При чому розрізняють два різновиди – усереднення та пропорція.

Усереднення – цей метод усереднює прогнози-оцінки з тих моделей, які вказують на первинний результат, і ігнорують будь-яку модель, яка вказує на вторинний результат. Наприклад, є три моделі M1, M2 і M3, які прогнозують первинний результат, і є модель M4, яка прогнозує вторинний результат. У цьому випадку результуюче значення апостеріорної ймовірності буде розраховано, як середнє за значеннями апостеріорного ймовірностей моделей M1, M2, M3.

Пропорція – цей метод ігнорує прогнози-оцінки і повертає замість них пропорцію моделей, що вказують на первинний результат. Наприклад, є три моделі M1, M2 і M3, які прогнозують первинний результат, і є модель M4, яка прогнозує вторинний результат. В цьому випадку пропорція моделей вказують на первинний результат дорівнює трьом четвертим.

Зазвичай моделі настроюються за допомогою оцінки статистики підгонки, обчисленої для перевірочних даних. У загальному випадку можна виконати розділення на два класи статистик:

- зведені статистики;
- статистичні графіки.

Зведені статистики перетворюють ефективність моделі в числові величини.В той час як статистичні графіки являють собою глобальний огляд ефективності моделі. Вони допомагають при порівнянні моделей для різноманітних сценаріїв застосування.

Більшість цих критеріїв ґрунтується на припущенні про те, що модель повинна мінімізувати варіабельність (середньоквадратичну похибку) при найменшій кількості можливих змінних в моделі (принцип економії).

До таких статистик належать наступні:

- коефіцієнт детермінації (R-квадрат).

- скоригований коефіцієнт детермінації (скоригований R-квадрат).
- статистика Маллоуза (Mallows) або Ср статистика.
- інформаційний критерій Акайке (AIC статистика).
- критерій Шварца-Байєса (SBC).

коефіцієнт детермінації:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2_{Adj} = 1 - \frac{SSE/df_E}{SST/df_T} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{(n-1)}{(n-p)}(1-R^2)$$

де $SSE -$ сума квадратів залишків, $SST -$ сума квадратів скоригованих за математичним сподіванням, $df_E -$ кількість ступенів свободи, пов'язаних із сумою квадратів залишків. $df_T -$ кількість ступенів свободи, пов'язаних із сумою квадратів взагалі. $n -$ кількість спостережень, $p -$ загальна кількість параметрів в моделі, враховуючи вільний член моделі.

Коефіцієнт детермінації (R-квадрат) є мірою частки варіабельності даних відгука, що пояснюється змінними-предикторами. R-квадрат завжи збільшується по мірі включення в модель все більшої кількості регресорів, як наслідок ця статистика не обов'язково є гарною мірою для використання при порівнянні моделей з різною кількістю предикторних змінних.

Скоригований R-квадрат – це аналог статистики R-квадрат, але додатково враховується кількість регресорів, що включаються до моделі. Отже, при порівнянні моделей з різною кількістю предикторних змінних доцільніше використовувати саме скоригований R-квадрат.

Статистика Маллоуза (Mallows) або Ср статистика, обчислюється за формулою [262]:

$$C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$$

де $MSE_p -$ середньоквадратична похибка, для моделі з p параметрами, $\mathrm{MSE}_{\mathrm{full}} -$ середньоквадратична похибка, для повної моделі з p параметрами. Використовується для оцінки істинної залишкової дисперсії.

n – кількість спостережень,

p – кількість параметрів моделі (кількість регресорів та вільний член моделі).

Інформаційний критерій Акайке [261]:

$$AIC = n \cdot ln\left(\frac{SSE}{n}\right) + 2p$$

де $n -$ кількість спостережень, $p -$ кількість параметрів в моделі, включаючи вільний член, SSE – сума квадратів залишків моделі.

Обидва ці критерії призначені для оцінки точності підгонки моделі в залежності від кількості параметрів у моделі. Чим менше значення критерія Акайке для моделі, тим краще модель. При цьому критерій Акайке прагне вибирати модель з більшою кількістю параметрів. Щоб подолати цю тенденцію, використовується критерій Шварца-Байєса, який фактично вводить в формулу більший штраф за додаткові параметри, що включаються в модель.

Критерій Шварца-Байєса:

$$SBC = n \cdot ln\left(\frac{SSE}{n}\right) + p \cdot ln(n)$$

Чим менше значення критерія Шварца-Байєса для моделі, тим краще модель.

Ср статистика є простим індикатором неточності моделі. Якщо статистика Ср набагато більше, ніж значення p, то це зазвичай вказує на недостатню специфікацію моделі [83].

Інший підхід до представлення формули Ср статистики:

$$C_p = \frac{SS(Residual)}{MS(Residual)} + 2p - n$$

де $SS(Residual) -$ сума квадратів залишків для моделі з $p-1$ змінною, $MS(Residual) -$ сума квадратів залишків для моделі з включенням усіх регресорів.

Коли модель вірно задана, то сума квадратів залишків є незміщеною оцінкою $(n-p)\sigma^2$, а Ср – незміщена оцінка $\frac{(n-p)\sigma^2}{\sigma^2} + 2p - n = p$. Таким чином Ср статистика апроксимаційно дорівнює p, коли модель вірно задана. Коли ж важлива змінна виключаються з моделі, сума квадратів залишків збільшується на число, що дорівнює варіабельності, яке пояснюється відповідним регресором, якщо цей регресор включити до моделі. Тобто Ср статистика збільшується і $C_p > p$ [69, 70, 94].

Існує багато автоматичних методів побудови моделей та вибору найкращої. Наприклад, побудова усіх можливих комбінацій регресорів з послідовним ранжуванням моделей за критерієм R-квадрат, скоригований R-квадрат або Ср критерій.

Окрім цього, є набір методів послідовного вибору регресорів, що включає підходи Forward, Backward, Stepwise, максимізації R-квадрат або мінімізації R-квадрат. Методи послідовного вибору регресорів мають рівні включення (SLENTRY=) та збереження (SLSTAY=) регресорів, що вказуються в операторі MODEL. Значення відповідних рівнів по замовчанню:

Forward SLENTRY=.50

Backward SLSTAY=.10

Stepwise SLENTRY=.15 nf SLSTAY=.15

Коли використовується R-квадрат, скоригований R-квадрат або Ср критерії у вигляді графіків то будуються лінії реферування що відповідають рівнянням $C_p = p$ та $C_p = 2p - p_{full}$, де $p_{full} -$ кількість параметрів в повній

моделі (за виключенням вільного члена моделі), а p –кількість параметрів в моделі, що аналізується (за виключенням вільного члена моделі). При цьому будується графік для статистики C_p в залежності від p. Д. Хоккінг в роботі [9] запропонував використовувати порогове значення $C_p \leq 2p - p_{full}$ для вибору моделі, що використовується для оцінки параметрів моделі. Для вирішення задач прогнозування Маллоуз в роботі [83] запропонував в якості порогового значення використовувати $C_p \leq p$, моделі для яких критерій C_p менше за p, заслуговують на подальше вивчення.

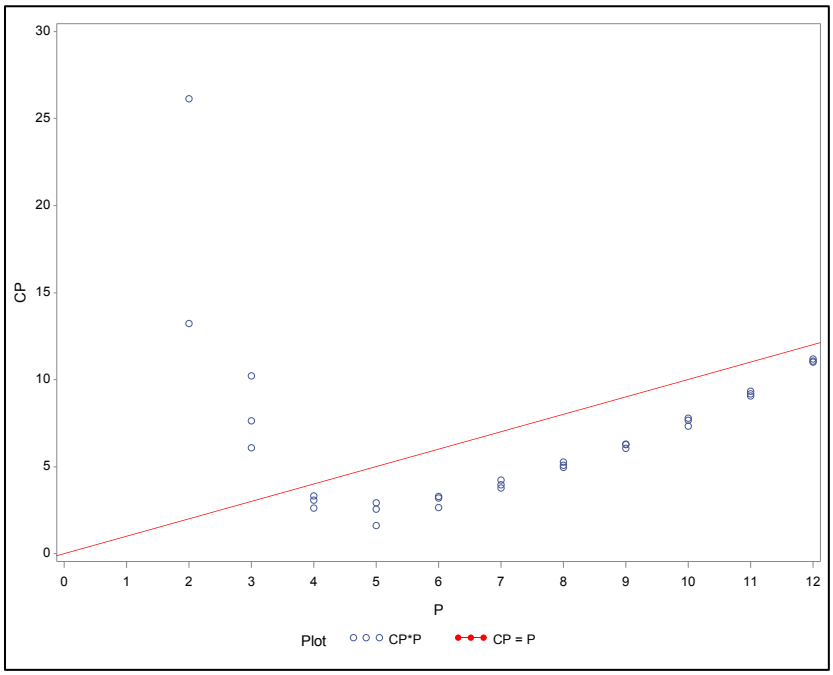


Рисунок 5.25 – Приклад графіка статистики Ср, з лінією реферування для рівняння $C_p = p$.

З наведеного на рис. 5.25 графіка, видно, що мінімальне значення Ср статистики досягається для моделі з п’ятьма параметрами. Для моделей з більше ніж п’ятьма параметрами Ср критерій збільшується. Однак нагадаємо, що Маллоуз запропонував використовувати модель з найменшою кількістю змінних, що відповідає пороговому значенню $C_p \leq p$, причому p –кількісь параметрів включая вільний член. Принаймні одна модель з чотирма параметрами (три регресори та вільний член) відповідає цьому критерію.

Зауважимо, що деякі моделі також були близькі до наведених вище, і майже так само хороші з точки зору Ср статистики. Бувають випадки, коли фахівцю з предметної області необхідно включити в модель певні змінні. Окрім того, при виборі моделі можуть бути використані міркування, щодо вартості включення того чи іншого регресора, виходячи із витрат на збір конкретної інформації, наявність систематичних пропусків в даних та різноманітних організаційно-управлінських рішень.

Для оцінювання якості прогнозів, обчислених за побудованими моделями, запропоновано процедуру автоматизованого вибору кращої прогнозної моделі, у якій для вибору було запропоновано інтегральний критерій якості, який включає 2-5 окремих статистичних критеріїв якості. Процедура автоматизації забезпечує можливість побудови, аналізу та вибору кращої із множини можливих моделей, кількість яких може сягати кількох сотень. Витрати часу залишаються при цьому цілком прийнятними для практичного використання пропонованої методики у інформаційній технології, призначеній для прогнозного моделювання.

Висновки до розділу 5

У даному розділі запропоновано використання різних типів інформаційних технологій прогнозування, в основу яких покладено поетапне розкриття невизначеностей різної природи, уточнення результатів моделювання на кожному етапі дослідження. Особливістю такого підходу є те, що особа, яка приймає рішення, має можливість отримати не лише «готове» рішення, а й брати участь у формуванні моделей, пропонувати власні сценарії та горизонти прогнозування з використанням фактичних даних.

Інформаційна технологія в даному випадку є зручним і зрозумілим інструментом користувача, робить процес прогнозування більш прозорим та

обгрунтованим, приховуючи при цьому всю складність застосовуваного математичного та технічного інструментарію.

В основу запропонованої технології покладено багатомодельний підхід, який ґрунтується на використанні множини різнотипних моделей: регресійний аналіз, байєсівське моделювання і прогнозування, технологія на основі застосування методу аналізу подібності, аналізу неструктурованих даних. При цьому кожний тип моделей призначений для виконання поставленої конкретної задачі. Регресійне моделювання забезпечує формування моделей на основі множини допустимих регресорів, їх оптимального вибору для конкретної моделі з наступним оцінюванням одно- або багатокрокового прогнозу.

Байєсівська технологія аналізу нелінійних нестаціонарних процесів забезпечує коректне виявлення та опис причинно-наслідкових зв'язків між змінними, обробку невизначеностей ймовірнісно-статистичного характеру, побудову адекватних моделей у випадку наявності прихованих змінних, пропусків вимірів і т. ін. Технологія на основі теорії подібних процесів забезпечує виконання аналізу даних у формі великих вибірок та наявності пропусків вимірів.

Запропонована методика моделювання, що має в основі пошук історичних аналогій поведінки досліджуваних процесів. В цьому випадку використовується історична вибірка даних та відрізок-шаблон для якого виконується пошук історичного аналогу, оскільки зазвичай для таких задач виміри прив'язані до часу.

Застосування запропонованого методу кластеризації в рамках багатомодельного підходу успішно продемонстровано на прикладі розв'язання практичної задачі групування регіонів України за обсягом капітальних інвестицій, спрямовуваних на охорону навколишнього середовища. Також успішно продемонстровано застосування багатомодельного підходу до аналізу неструктурованих даних.

Для розв’язання задачі автоматизації побудови прогнозуючих моделей в рамках створеної інформаційної технології аналізу нелінійних нестаціонарних процесів запропоновано множину статистичних критеріїв адекватності математичних моделей та якості оцінок прогнозів. Ефективність застосування запропонованої множини критеріїв проілюстровано відповідними прикладами аналізу даних.

РОЗДІЛ 6
ПОБУДОВА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПРОГНОЗУВАННЯ

6.1 Архітектурні рішення щодо створення інформаційної технології прогнозування нелінійних нестаціонарних процесів

При проектуванні інформаційних технологій, призначених для вирішення завдань математичного моделювання та прогнозування процесів різної природи, в тому числі й таких, що характеризуються нелінійністю та є нестаціонарними, перш за все необхідно передбачити, що отримані результати будуть використані у подальшій аналітичній роботі. Тому, доцільно об'єднати в єдину систему всі функціональні модулі які забезпечують збір, обробку, накопичення первинної інформації, сукупність процесів, що реалізують багатомодельний підхід до прогнозування. Концепція такої інформаційної технології [263] повинна ґрунтуватися на інтеграції чотирьох основних підсистем, що реалізують перераховані процедури і передбачає модульно-блочну побудову. Обов'язковою вимогою до такої інформаційної технології є забезпечення зручного інтерфейсу між користувачами, внутрішніми елементами системи і системою опрацювання результатів. Користувач повинен бути активним учасником всіх процесів, пов'язаних із обробкою даних та прогнозуванням, оскільки в більшості випадків вирішення практичних задач прогнозування, виникає проблема необхідності отримання експертних рішень на різних етапах прогнозування.

Інтерфейс з користувачем має бути реалізованим за допомогою відповідної підсистеми, яка повинна бути безпосередньо пов'язана з підсистемою збору та зберігання даних [264-269]. Така організація зберігання і одержання даних забезпечує можливість вибрати оптимальну систему управління даними для певної практичної задачі, якнайповніше врахувати особливості досліджуваної предметної області, що в майбутньому створить передумови для підвищення якості отримуваних прогнозів.

Головною є підсистема, що забезпечує реалізацію процесу аналізу і

розв’язання задач відповідно із загальною структурою вирішення задач прогнозування нелінійних нестаціонарних процесів. При цьому для здійснення певних процедур під’єднуються івикористовуються відповідні модулі, підключати які можна залежно від потреб користувача. Такі модулі призначені для імплементації розроблених методів і підходів, що застосовуються в процесі реалізації багатомодельного підходу та передбачають можливість подальшого вдосконалення і розвитку без необхідності коригування інших елементів інформаційної технології. Такий підхід забезпечує створення гнучкої архітектури, яка легко модифікується до розв’язання інших задач прогнозного моделювання, а також можливостей застосування інших програмних засобів, реалізованих на різних платформах.

Для розробки інформаційної технології аналізу й обробки інформації були досліджені реальні процеси, що пов’язані із функціонування складних систем різної природи. Аналіз та обробка інформації спрямовані на розробку методів аналізу й оцінювання кількісної та якісної інформації, визначення кількості та характеру зовнішніх впливів, збурень та їх типу (детерміноване чи стохастичне). Окрім того, необхідно було аналізувати попереднє встановлення можливості їх статистичного опису за допомогою конкретних типів розподілів випадкових величин, визначати та встановлювати основних типів ризиків та їхніх показників, ідентифікувати ключові невизначеності та розробляти методи їх подолання. Згідно з принципами багатомодельного підходу, розроблений метод синтезу інформаційних технологій застосований для розв’язування задач прогнозування розвитку нелінійних нестаціонарних процесів різної природи, ґрунтується на інтеграції різнотипної інформації й заснований на системному використанні методів аналізу даних, моделювання, прогнозування**.** Метод синтезу інформаційних технологій поєднує в собі задачі, які вирішуються на кожному етапі побудови прогнозу за допомогою інформаційних технологій: інформаційна технологія аналізу її оцінювання інформації, інформаційна технологія моделювання для побудови прогнозів, інформаційна технологія прогнозування, інформаційна технологія

аналітичних звітів, які будуть використані у процесі підтримки прийняття рішень. Кожна інформаційна технологія заснована на системному використанні різних інструментальних методів, які вирішують окремі завдання прогнозного моделювання. Групи інструментальних методів використовуються залежно від типу й механізму вибору та конкретної задачі. Методи можуть використовуватись як окремо – для розв'язування окремих (локальних) задач, так і системно для вирішення загальної проблеми. Особливістю архітектури є те, що вона має блок підготовки звітів. Загалом, розроблена інформаційна технологія орієнтована на подальшу інтеграцію до систем підтримки прийняття рішень різного типу. Розроблена інформаційна технологія може слугувати «платформою» для побудови різних додатки, за допомогою яких вирішуються задачі прогнозування ННП. Інформаційна технологія має ієрархічну структуру для аналізу даних, в тому числі і представлених у вигляді часових рядів. Особлива увага приділена обробці часових рядів статистичних даних, адже вони становлять основу для побудови моделей, що найбільш точно відображають характер досліджуваного процесу. Обробка статистичних даних, представлених часовими рядами, як правило, супроводжується невизначеностями різного типу, зокрема, такими як, структурні, статистичні й параметричні. Процес обробки даних та врахування невизначеностей різного типу, ієрархічність організації системи обробки даних, а також необхідність функціональної повноти, вимагають застосування системного підходу, що забезпечує можливість вирішення багатьох задач, які виникають у статистичному аналізі даних під час побудови моделі, оцінювання прогнозів та генерування результатів. Системною властивістю пропонованої інформаційної технології й те, що вона використовує окремі множини статистичних критеріїв для аналізу якості даних, адекватності побудованої моделі та якості оцінок прогнозів, згенерованих за моделлю. Перевагою розробленої інформаційної технології є й те, що для прогнозів, що побудовані за багатомодельного підходу є набір критеріїв для перевірки їх якості.

6.2 Конструювання прототипу інформаційної технології прийняття рішень із використанням засобів SAS

В рамках дослідження реалізовано інформаційну технологію, що представляє собою сукупність методів та моделей, призначених для прогнозування нелінійних нестаціонарних процесів. Передбачено, що дана інформаційна технологія має засоби вирішення задач інтелектуального аналізу даних, а саме: аналіз причинно-наслідкових зв'язків, класифікація, кластеризація даних, прогнозування та візуалізація результатів. Однією з головних переваг системи є можливість функціонування як без залучення експертів, на основі навчальних даних, так і з залученням експертів. В подальшому, запропонована інформаційна технологія може бути інтегрована у відповідну систему підтримки прийняття рішень

Методика інтелектуального аналізу даних, що має назву DMTwoStage реалізована автором із використанням спеціалізованого програмного забезпечення компанії SAS Institute. За технологічним рівнем розроблена система належить до настільних програм, але також може бути розвернута й у вигляді клієнт-серверної архітектури на технологічній платформі SAS.

Структурна схема інтеграції розробленої інформаційної технології до платформи SAS представлена на рис. 6.1. Структура має три основних підсистеми, архітектура – блочно-модульна. Головна підсистема аналізу призначена для аналізу даних на основі запропонованої методики DMTwoStage. Тобто на першому етапі відбувається побудова прогнозу-рішення на основі теорії мереж Байєса, а на другому прогнозу-оцінки на основі використання регресійного аналізу. Блок побудови структури БМ складається з модулів побудови топології мережі із використанням MDL або MRE методів.

Цей модуль дає можливість будувати БМ й «вручну», використовуючи графічний інтерфейс.

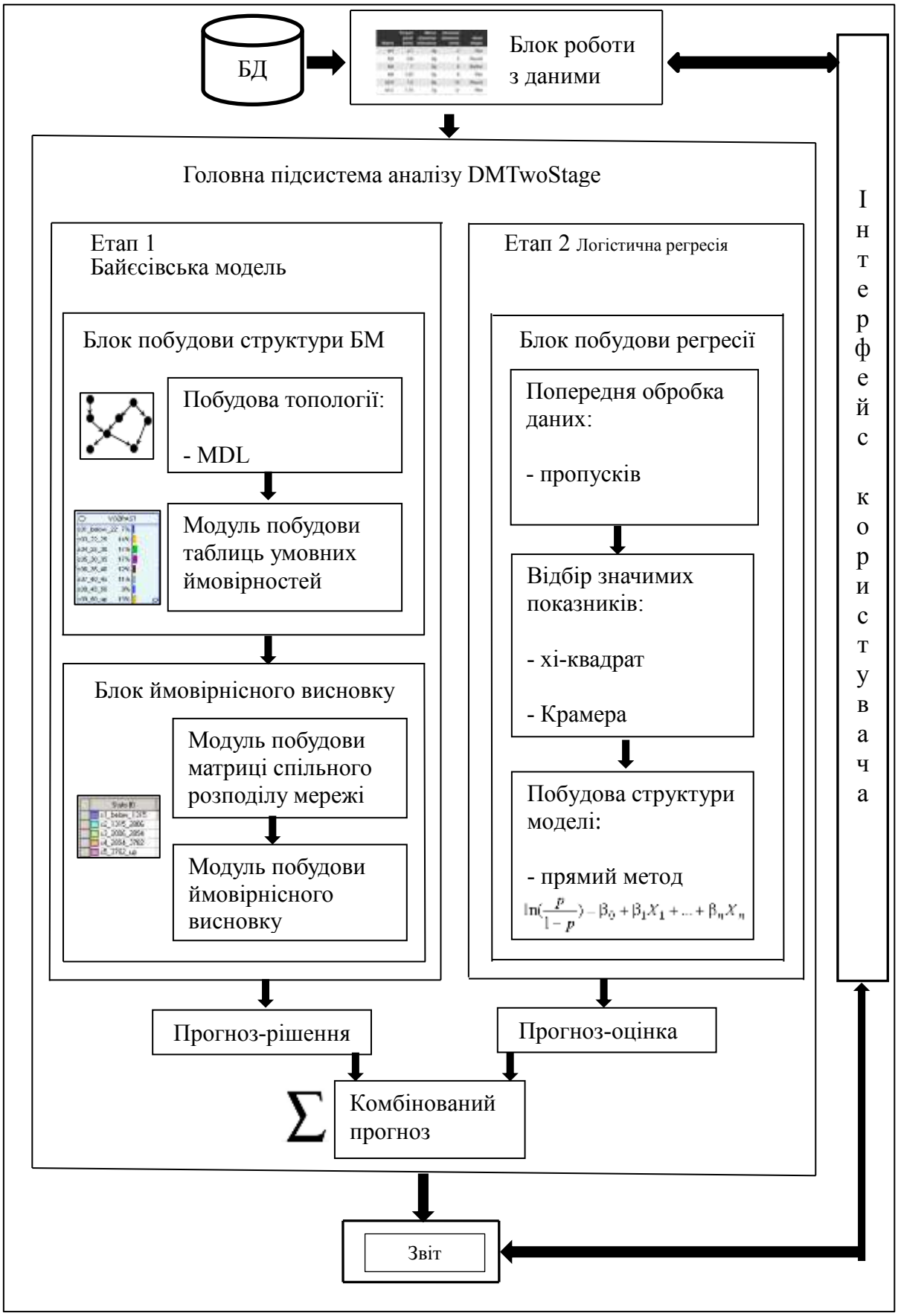


Рис. 6.1 Структура інтеграції розробленої інформаційної технології з платформою SAS на основі DMTwoStage

Блок складається з таких модулів: модуль побудови структури БМ; модуль побудови таблиць умовних ймовірностей; модуль побудови таблиць умовних ймовірностей (цей модуль дозволяє редагувати ім’я вершини, змінювати стани вершин, додавати, видаляти, перейменовувати їх та заповнювати таблиці умовних ймовірностей).

Блок ймовірнісного висновку БМ складається з модулів:

1) побудови матриці спільного розподілу мережі, призначений для побудови матриці емпіричних значень сумісного розподілу ймовірностей всієї БМ.

2) побудови ймовірнісного висновку, здійснює ймовірнісний висновок в мережі на основі матриці емпіричних значень сумісного розподілу ймовірностей.

Пристрої вводу-виводу дозволяють користувачеві завантажувати дані для побудови БМ та для навчання параметрів прихованих вершин, а також зберігати результати роботи в зовнішні файли.

Підсистема інтерфейсу користувача пов’язує користувача, внутрішні елементи системи, зовнішні запам’ятовуючі пристрої та пристрої вводу-виводу. Інтерфейс надає можливість користувачеві завантажувати дані, вводити команди і запити в систему, отримувати графічне представлення даних і результатів, зберігати отримані результати.

Підсистема зберігання інформації складається із загальної бази даних, бази моделей та бази знань (за потреби), які призначені для накопичення навчальних даних, готових моделей на основі БМ.

Блок роботи з даними призначений для пошуку оптимальної структури мережі Байєса, завантаження даних для побудови БМ, а також зберігання даних, отриманих як результат генерування вибірки. Даний блок складається з наступних модулів: модуль завантаження та зберігання даних; модуль пошуку оптимальної структури БМ. Модуль завантаження та зберігання даних призначений для завантаження даних з файлу та вивантаження згенерованих даних.

Модуль пошуку оптимальної структури БМ використовує евристичний алгоритм побудови БМ для знаходження топології мережі з можливістю вибору способу обчислення взаємозв'язку між вершинами серед наступних: значення взаємної інформації, коефіцієнти Пірсона, Крамера, Чупрова і лямбда Гудмана.

Для нормального функціонування програми необхідна ЕОМ з наступними характеристиками: чотириядерний процесор; тактова частота кожного ядра не менше 2,1 МГц; вільний дисковий простір: 30 Гігабайт і більше для програмних модулів системи, навчальних даних і формування баз даних (останнє значення не фіксоване і може змінюватися в процесі експлуатації); оперативна пам'ять 8 Гігабайт і більше; 32 або 64 розрядна операційна система Windows; пристрій безперебійного живлення ЕОМ для можливості автономної роботи програми; клавіатура та комп'ютерна «миша»; пристрій для запису даних і результатів на оптичні носії. У конфігурації робочої станції (Personal workstation SAS EM) є компоненти:

- клієнт SAS EM;
- SAS Foundation;
- Metadata Server;

що знаходяться на одному (головному) комп'ютері.

Компоненти зв'язані між собою за допомогою запатентованої компанією SAS технології IOM

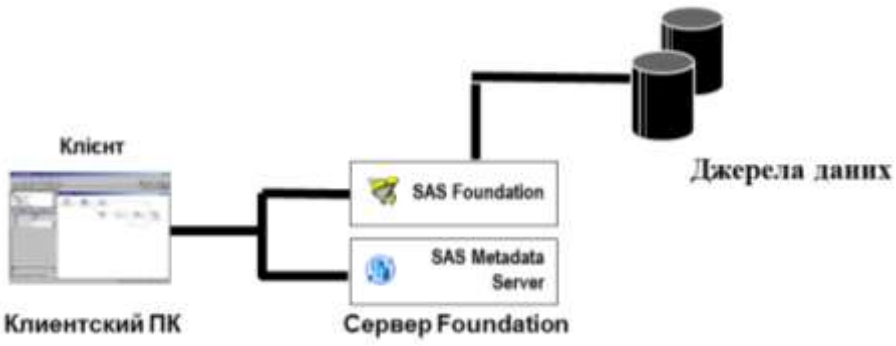


Рис. 6.2 Архітектура "Клієнтський комп'ютер" на базі SAS Foundation (пбудовано на основі матеріалів компанії SAS Institute)

Клієнт SAS EM – це тільки частина більш великого набору програм. За своєю суттю клієнт SAS EM – це просто вікно з Java інтерфейсом. Фактично робота з аналізу даних виконується програмним продуктом відомим як SAS Foundation.

SAS Foundation – це ще одна назва мови і набору процедур SAS. У свою чергу SAS Foundation підтримується ще одним програмним продуктом – SAS Metadata Server.

Metadata Server відстежує інформацію про доступ до даних і системнії архітектури.

У конфігурації «клієнт-сервер» до загального набору конфігурації «Клієнтський комп'ютер» додається компонент «Аналітична платформа» (Analytics Platform).

Аналітична платформа забезпечує зв'язок між:

- клієнтом SAS EM
- SAS Foundation
- серверами Metadata Server

Все, з чим має справу аналітик – це інтерфейс клієнта SAS EM. Аналітичний процес – це послідовність кроків, необхідних для виконання прикладної аналітичної задачі.

Основні кроки аналітичного процесу:

1. Завдання цілей аналізу
2. Вибір спостережень
3. Вилучення вхідних даних
4. Перевірка вхідних даних
5. Відновлення вхідних даних
6. Перетворення вхідних даних
7. Застосування аналізу
8. Генерування методів застосування
9. Інтеграція методу
10. Збір результатів

11. Оцінювання зібраних результатів
12. Уточнення цілей аналізу

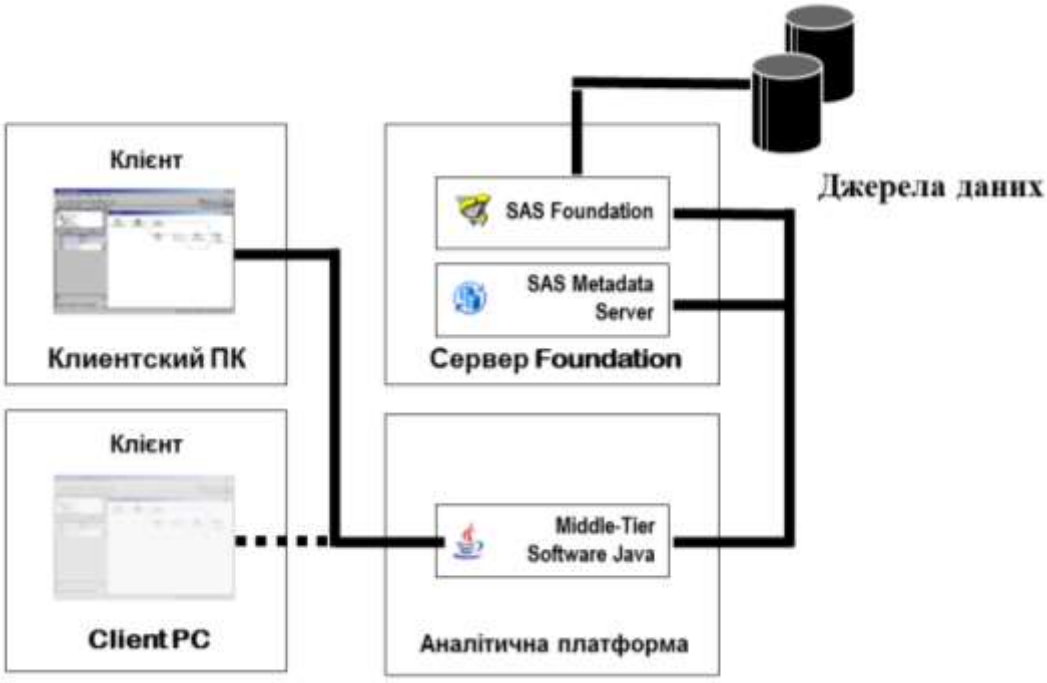


Рис. 6.3 Клієнт-серверна архітектура СППР на базі SAS Foundation

Така конфігурація (рис. 6.3) дозволяє багатьом клієнтам з'єднуватися з іншими серверами SAS Foundation. Це краща конфігурація для розподілених клієнт- серверних програм. Системний адміністратор повинен встановити і конфігурувати ці компоненти, зазвичай на декількох незалежних головних комп'ютерах. Після установки на платформі, архітектура системи, задача налагодження зв’язків між компонентами системи вже не важлива кінцевому користувачеві-аналітику. За винятком необхідності пам’ятати, що всі дані зчитуються на SAS Foundation Server, а не на клієнтському ПК.

6.3 Створення інформвційної технології для вирішення завдань моделювання і прогнозування

В сучасному світі галузь інформаційних технологій розвивається швидкими темпами, при цьому кожний рік вони зростають. Тому у цій роботі

ставиться задача розробки інформаційної технології, яка б задовольняла наступним вимогам:

− реалізація описаних у попередніх розділах алгоритмів оцінювання та прогнозування нелінійних нестаціонарних процесів;

− забезпечення можливості доступу користувачів до системи через мережу Інтернет;

− створення можливості використання іншими розробниками сервісу прогнозування використанню веб–сервісів;

− розробка «простого» та інтуїтивно зрозумілого інтерфейсу користувача, створеного за рахунок автоматизованих та прихованих від користувача опцій налаштувань;

− забезпечення можливості подальшого розширення та модифікації функцій системи.

Рисунок 6.4 – Структурна схема системного підходу до організації процесу прогнозування при вирішенні завдань дослідження

Зокрема, формулювання та розв’язання задач прогнозування нелінійних нестаціонарних процесів основане на застосуванні методології

системного аналізу, яка ґрунтується на комплексному використанні методів попередньої обробки і статистичного аналізу даних, математичного і статистичного моделювання, прогнозування та оптимального оцінювання станів процесів довільної природи. Використання запропонованої концепції забезпечує отримання високоякісних (за точністю) коротко- та середньострокових прогнозів за умови наявності інформативних даних, а також формування на їх основі альтернативних оптимальних і раціональних рішень, що також передбачає універсальність застосування запропонованої концепції до аналізу широкого класу нелінійних нестаціонарних процесів. На рис. 6.5 подано структурну схему системного підходу до організації процесу прогнозування.

Структура інформаційної технології підтримки прийняття рішень подана на рис. 6.5.

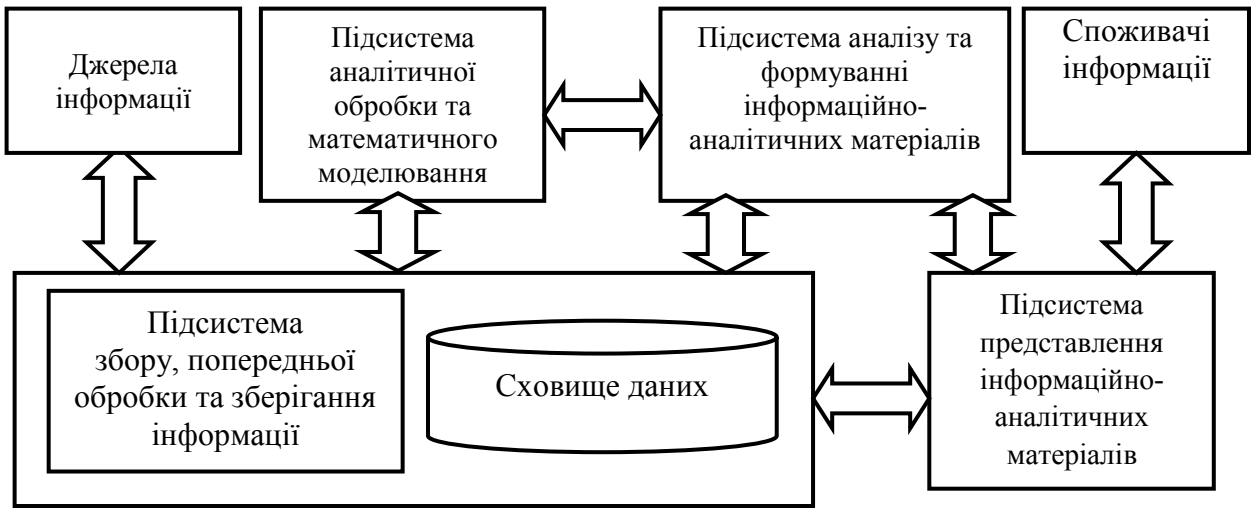


Рисунок 6.5 – Структура інформаційної технології підтримки прийняття рішень

Пропонований підхід ґрунтується на аналізі досліджуваного процесу, встановленні типів наявних характерних невизначеностей, оцінюванні структури і параметрів моделі та прогнозів.

Для розв’язання цього комплексу задач спроектовано реалізацію пропонованої інформаційної технології у СППР з відповідною архітектурою (рис.6.6).

Рисунок 6.6 – Спрощена концептуальна схема інтерації розробленої інформаційної технології до СППР

Пропонована архітектура дозволяє будувати додатки, за допомогою яких реалізується весь комплекс задач прогнозування нелінійних нестаціонарних процесів. Пропонована інтеграція до СППР забезпечується гнучкою модульною структуру, наявністю засобів аналізу даних у вигляді часових рядів сформованих на основі структурованої та неструктурованої інформації.

Блок збору та підготовки даних орієнтований, в тому числі, на використання API-інтерфейсу, що дає можливість опрацьовувати великі обсяги даних, отримуваних з мережі Інтернет, виконувати весь комплекс

робіт попередньої обробки даних, в тому числі і з використанням інструментів інтелектуального аналізу даних.

У блоці зберігання даних можуть бути використані системи управління базами даних такі як, MySql, PostgreSQL та аналогічні. У блоці прогнозного моделювання передбачено використання таких аналітичних систем як SAS, R, Python, є можливість інтеграції до інших аналітичних платформ.

Блок прогнозного моделювання має модуль тестової перевірки ідей та гіпотез, що дає можливість виконувати значну кількість чисельних експериментів для апробації пропонованих методик прогнозного моделювання. Також у блоці прогнозного моделювання передбачено можливість вибору горизонту прогнозування та побудови сценаріїв. Особливістю модуля представлення результатів є можливість для користувачів самостійно генерувати звіти про якість прогнозів. Пропонована архітектура СППР пропонована до інтеграції з іншими технологічними рішеннями (аналітичними платформами) (рис. 6.7).

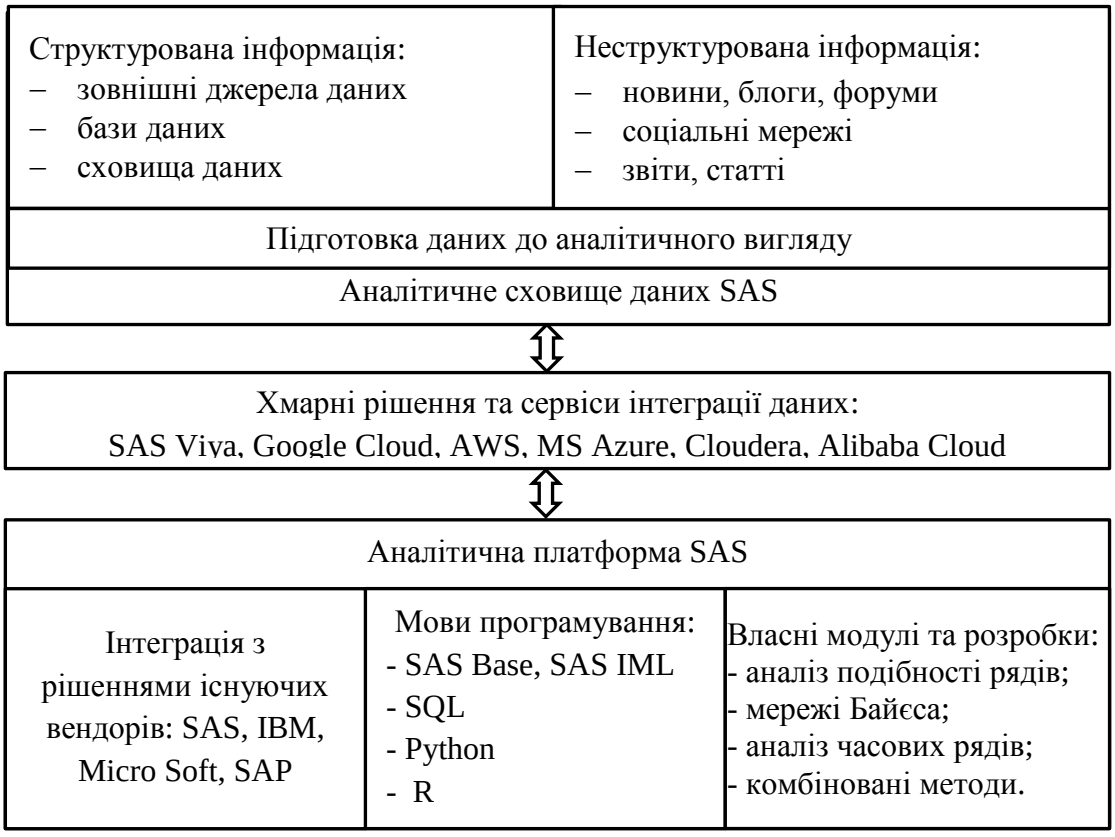


Рисунок 6.7 – Архітектура СППР на базі SAS Foundation

Така конфігурація дозволяє багатьом клієнтам з'єднуватися з серверами SAS Foundation. У конфігурації клієнт-сервер до загального набору конфігурації «Клієнтський комп’ютер» додається компонент «Аналітична платформа» (Analytics Platform). Аналітична платформа забезпечує зв'язок між клієнтом SAS Enterprise Miner, SAS Foundation, серверами Metadata Server.

Аналітичне сховище даних призначене для зберігання та інтеграції як структурованої, так і неструктурованої інформації у вигляді, придатному для аналізу та моделювання. Система формує специфічні набори даних (метадані), призначені для використання у аналітичній системі. Для певних типів аналізу можуть існувати інші специфічні метадані. Схема визначення метаданих в аналізі для вибірки дозволяє повідомляти SAS Enterprise Miner передати таку інформацію:

- налаштування ролі кожної змінної в аналізі (цільова або регресорна);
- рівень вимірювань кожної змінної – розрізняє безперервні чисельні і категоріальні змінні;
- роль набору даних в аналізі – повідомляє системі, як використовувати вибраний набір даних в аналізі (навчальний, валідаційний, тестовий, скорінговий, транзакційний, корпус текстів).

Аналітична платформа дозволяє використовувати як готові аналітичні рішення від існуючих вендорів, так і власні методи та підходи реалізовані у вигляді модулів програм на загальних та спеціалізованих мовах програмування. Хмарні рішення та сервіси інтеграції даних дозволяють виконувати задачі по розгортанню аналітичних рішень, баз та сховищ даних, налаштування програмних інтерфейсів для запуску побудованих аналітичних моделей та обміну з ними даними від різноманітних сервісів та рішень.

Аналітичне сховище даних призначене для зберігання та інтеграції як структурованої так і не структурованої інформації у вигляді, придатному для аналізу та моделювання. Аналітична платформа дозволяє використовувати як готові аналітичні рішення від існуючих вендорів, так і власні методи та

підходи реалізовані у вигляді модулів програм на загальних та спеціалізованих мовах програмування. Хмарні рішення та сервіси інтеграції даних дозволяють виконувати задачі по розгортанню аналітичних рішень, баз та сховищ даних, налаштування програмних інтерфейсів для запуску побудованих аналітичних моделей та обміну з ними даними від різноманітних сервісів та рішень.

6.4. Застосування DMTwoStage на базі технології SEMMA для SAS Enterprise Miner

SAS Enterprise Miner – це інтегрований компонент системи SAS, створений для виявлення в масивах даних інформації. Розроблений спеціально для пошуку та аналізу прихованих закономірностей у даних [272]. Enterprise Miner включає в себе ефективні методи статистичного аналізу, відповідну методологію виконання проектів дослідження даних (SEMMA) і зручний графічний інтерфейс користувача.

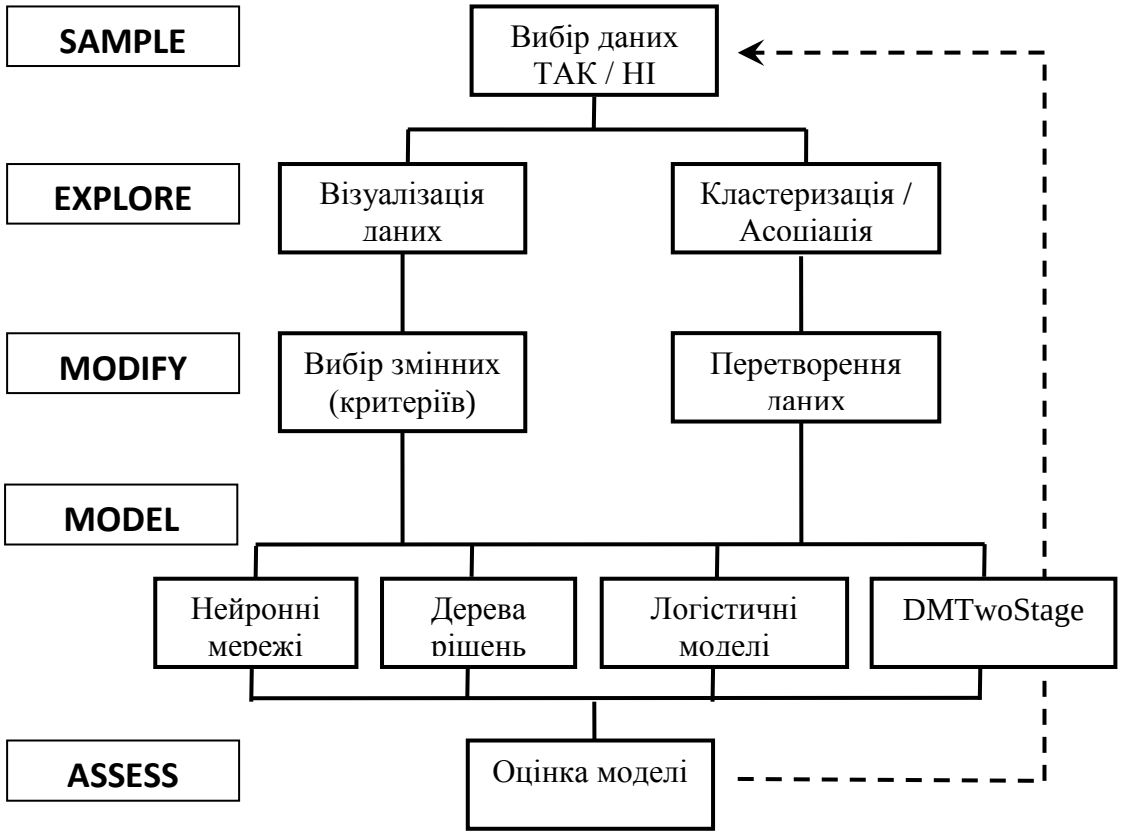


Рисунок 6.8 – Використання DMTwoStage

Для перетворення вхідних даних до формату, з яким здатний працювати модуль побудови математичної моделі у вигляді мережі Байєса запропоновано використання наступних методів об’єднання:

- на основі значення порогу
- на основі зваженої сукупності спостережень
- на основі згладжування значення зваженої сукупності спостережень

Існує багато методів для перекодування нечислових вхідних змінних, але варто відзначити, що проста нумерація рівнів категоріальної змінної є одним з найгірших підходів (для прогнозного моделювання). Багато моделей роблять допущення про наявність безперервних асоціацій між вхідний і цільовий змінними – невеликі зміни вхідних змінних повинні призводити до невеликих змін цільової змінної. При використанні простої нумерації, два суміжних нумерованих рівня можуть мати істотну відмінність [37].

Таблиця 6.1 - Приклад категоріальної змінної з десятьма станами

Рівень	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I	D_J
A	1	0	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0	0
F	0	0	0	0	0	1	0	0	0	0
G	0	0	0	0	0	0	1	0	0	0
H	0	0	0	0	0	0	0	1	0	0
I	0	0	0	0	0	0	0	0	1	0
J	0	0	0	0	0	0	0	0	0	1

Найбільш коректним підходом є створення проектних змінних (dummy-coding) або нумерація на основі використання цільової змінної (target-based enymeration).

Стандартним підходом є використання створення проектних змінних на основі рівнів категоріальних змінних. Але проблема виникає у випадку опрацювання змінних з великою кількістю рівнів – так звана проблема перевиробництва (overgeneration). Модель, побудована для такого випадку, якщо і буде точною, то лише для навчального набору даних, а от для валідаційних і тестового наборів виникнуть проблеми у вигляді погіршення точності моделі.

Простим, але разом з тим ефективним рішенням для випадку надвиробництва, є використання значення порогу. За цим підходом для створення проектних змінних обираються тільки ті, що мають рівні які перевищують задане граничне значення. На у табл. 6.2 показаний приклад, з якого видно, що G H I J мають значення менше порогу. Фіктивний зв'язок, між рівнем змінної та цільової змінної, як правило характеризує рівень, що рідко зустрічається.

Таблиця 6.2 – Об’єднання на основі значення порогу

Рівень	N(i) частота	Коментар
A	1562	Без змін
B	970	Без змін
C	223	Без змін
D	111	Без змін
E	85	Без змін
F	50	Без змін
G	23	Об’єднати в один спільний рівень
H	17	
I	12	
J	5	

Іншим підходом є нумерування рівнів, що відрізняється від звичайного підходу завдяки врахуванню значення цільової змінної. На рисунку нижче наведений приклад обчислення кількість первинних відгуків цільової змінної

по відношенню до появи заданого рівня регрессору. Після чого виконується сортування за значенням ймовірності – P(i). Фактично за своєю суттю – це варіант перетворення номінальною змінною в ординарну.

Таблиця 6.3 - Об'єднання значень цільової змінної (сортування по *P(i))*

Рівень	N(i)(частота)	$\sum Y_i$(кількість первинних відгуків)	P(i) (ймовірність первинного відгуку)
J	5	5	1
I	12	6	0,5
B	970	432	0,45
F	50	20	0,4
G	23	8	0,35
D	111	36	0,32
H	17	5	0,29
A	1562	430	0,28
E	85	23	0,27
C	223	45	0,2

Таблиця 6.4 - Об'єднання на значень цільової змінної – створення ординарної змінної

Рівень	N(i) (частота)	$\sum Y_i$(кількість первинних відгуків)	P(i)(ймовірність первинного відгуку)
1	5	5	1
2	12	6	0,5
3	970	432	0,45
4	50	20	0,4
5	23	8	0,35
6	111	36	0,32
7	17	5	0,29
8	1562	430	0,28
9	85	23	0,27
10	223	45	0,2

На практиці аналітики можуть використовувати різні варіації вказаних методів. Замість того, щоб виконувати нумерування числами, нерідко застосовується більш складний підхід – зважена сукупність (weight of

evidence). Метод зважених сукупностей добре використовувати для змінних з великою кількістю рівнів.

Таблиця 6.5 – Об’єднання на основі значення зважених сукупностей

Рівень	N(i) (частота)	$\sum Y_i$(кількість первинних відгуків)	Log(P(i) / (1-P(i)))
CL1	5	5	.
	12	6	0.00
	970	432	-0.10
	50	20	-0.18
CL2	23	8	-0.27
	111	36	-0.32
CL3	17	5	-0.38
	1562	430	-0.42
	85	23	-0.43
CL4	223	45	-0.60

Таблиця 6.6 - Об’єднання на основі значення зважених сукупностей в кластери

Рівень	N(i) (частота)	$\sum Y_i$(кількість первинних відгуків)	Log(P(i) / (1-P(i)))
CL1	1037	463	-0.09
CL2	134	44	-0.31
CL3	1664	458	-0.42
CL4	223	45	-0.60

Ідея полягає в тому, що обчислюється значення логарифмічної функції корисності (ЛФК). Такий підхід використовується при побудові дерев рішень, як наслідок, рівні змінних з близькими значеннями ЛФК групуються

в один кластер, як показано у табл. 6.6. Кількість кластерів, як правило невелика, що дозволяє в подальшому використовувати малу кількість проектних змінних.

Метод згладженого значення зваженої сукупності (smoothed weight of evidence) – це модифікація попереднього, що увібрала в себе ідею ідеї байєсівського статистичного підходу. Вважається, що наявна апріорна інформація про розподіли цільової змінної щодо середнього значення за рівнями. Це так званий апріорний розподіл, відображається інформацією щодо очікуваних значень цільової змінної для кожного рівня.

Приклад обчислення. Нехай очікуване значення (математичне сподівання) для регресора = 24, а для цільової змінної = 8. Тоді перерахунок значень, для рівня J, ЛФК буде виглядати наступним чином:

$$\frac{5+8}{5+24} = 0{,}4482 \approx 0{,}45$$

і так для кожного рівня (у чисельнику додається математичне очікування цільової змінної, а в знаменнику мат очікування регрсору).

Даний підхід забезпечує суттєве зменшення зміщення прогнозу у порівнянні з простим WOE, особливо для випадків коли у змінній десятки або сотні рівнів [37].

Таблиця 6.7 - Об’єднання на основі згладженого значення зважених сукупностей

Рівень	N(i)	MY	$\sum Y_i$	MX	P(i)	Log(P(i) / (1-P(i)))
J	5	+24	5	+5	0,45	-0,09
I	12	+24	6	+5	0,39	-0,19
B	970	+24	432	+5	0,44	-0,1
F	50	+24	20	+5	0,38	-0,22
G	23	+24	8	+5	0,34	-0,29
D	111	+24	36	+5	0,33	-0,32
H	17	+24	5	+5	0,32	-0,33
A	1562	+24	430	+5	0,28	-0,42
E	85	+24	23	+5	0,28	-0,4
C	223	+24	45	+5	0,21	-0,56

Для визначення найкращого методу перекодування категоріальних змінних було проведено серію обчислювальних експериментів.

Порівняння методів виконувалося на основі значення середньоквадратичного зміщення (MSBs – mean squared biases) моделі логістичної регресії, що побудована на штучно згенерованих даних:

- згенеровано 100 вибірок даних
- кожна згенерована вибірка складається з n=10 000 записів
- співвідношення первинних та вторинних відгуків n0: n1 = 3:1
- кардинальне число рівнів вхідних змінних змінюється від 20 до 200.
- співвідношення ступня розрідженості до білого шуму – середнє.

На рисунку 6.9 наведені результати порівняння для 5 методів:

– Threshold – методи об’єднання рівнів категоріальних змінних на основі значення порогу;

– WOE – зважений метод статистичної значумості рівнів категоріальної змінної;

– SWOE – метод згладженого значення зваженої сукупності, для категоріальної змінної. SWOE8 та SWOE24 – методи SWOE із вказанням вікна зглагодження відповідно у 8 та 24 елементи.

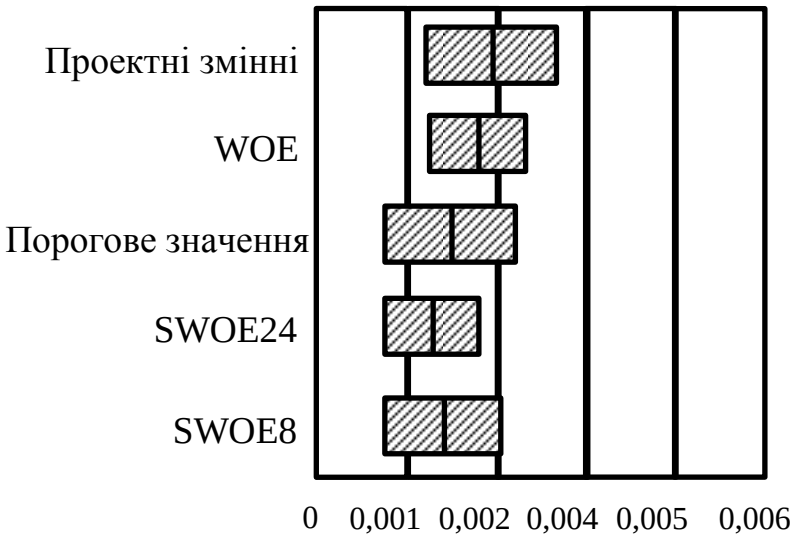


Рисунок 6.9 – Результати порівняння методів перекодування категоріальних змінних

Кожен метод випробувався на 100 різних тестових згенерованих наборах даних. Можна побачити, що методи об’єднання на основі значення порогу та на основі згладжування значення зваженої сукупності спостережень дали приблизно однакові результати. Було з’ясовано, що:

1. Методи об’єднання на основі значення порогу (Threshold) та на основі згладжування значення зваженої сукупності спостережень (SWOE) дали приблизно однакові результати.
2. Звичайний метод WOE та створення проектних змінних непридатний для випадку великої кардинальності вхідних категоріальних змінних навіть при середньому рівні ступеня розрідженості і білого шуму.
3. Найкращі результати для випадку великої кардинальності показав SWOE

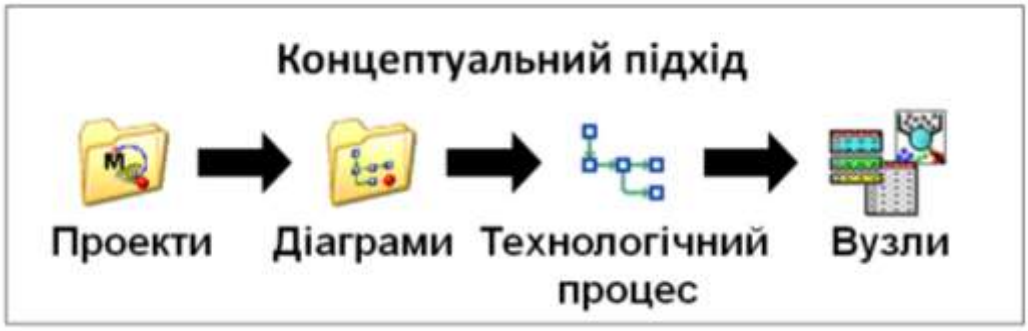


Рисунок 6.10 – Концептуальний підхід щодо організації елементів аналізу

За кадром (на сервері SAS Foundation, де знаходиться проект) організація набагато складніша.

У самому верху ієрархії розташований проект SAS EM, відповідний директорії на SAS Foundation (або іншій теці до якої має доступ користувач).

При описі проекту автоматично створюються директорії проекту:

1) Datasource – джерела даних

2) Meta – метадані технологічного процесу

3) Reports – звіти

4) System – системна інформація щодо технологічного процесу та sas коди проекту.

5) Workspace – містить діаграми. Діаграми зазвичай відносяться до однієї області аналізу або одному проекту. Кожна діаграма незалежна і ніяка інформація не може бути передана від однієї діаграми до іншого.

Технологічний процес – це послідовність завдань або вузлів, пов'язаних стрілками в інтерфейсі, і він задає порядок аналізу. Організація ходу процесу знаходиться у файлі em_dgraph.sas7bdat.

Кожен вузол діаграми відповідає окремому підкаталогу в директорії діаграми.

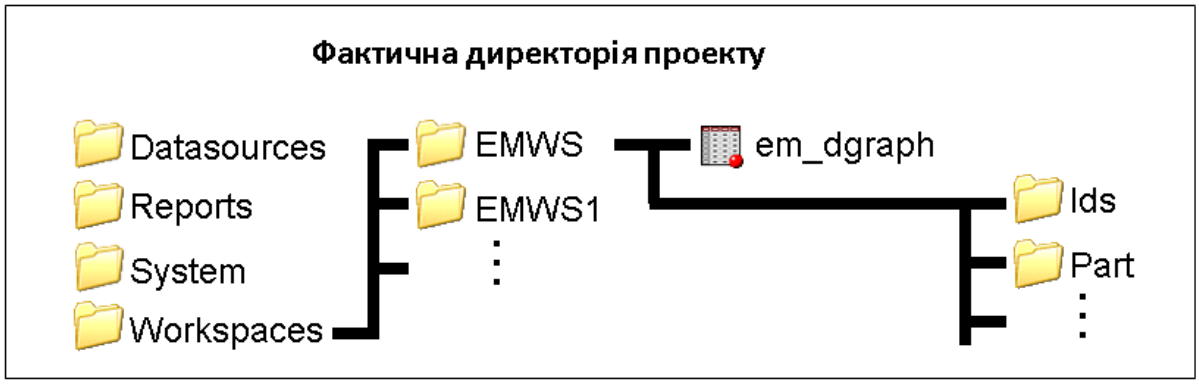


Рисунок 6.11 – Фізична організація файлів

Після того як створені новий проект і діаграми в системі аналітик приступає до створення джерела даних.

Джерело даних – це зв'язок між існуючою SAS таблицею і SAS EM. Щоб задати джерело даних, потрібно вибрати SAS таблицю для аналізу і задати метадані, що відповідають цілям аналізу.

Визначення метаданих в аналізі для набору даних дозволяє повідомляють SAS EM передати таку інформацію:

1. Роль кожної змінної в аналізі – повідомляє SAS EM призначення змінної в поточному аналізі.

2. Рівень вимірювань кожної змінної – розрізняє безперервні чисельні і категоріальні змінні.

3. Роль набору даних в аналізі – повідомляє SAS EM, як використовувати вибраний набір даних в аналізі.

Крім перерахованих метаданих можуть існувати інші специфічні метадані для певних типів аналізу.

Концептуальна схема роботи інформаційно-аналітичної системи, представлена на рис. 6.12 складається з блоків:

- зовнішніх джерела даних у вигляді баз та сховищ даних, від різноманітних вендорів відповідних IT рішень.

- аналітичне сховище SAS, що виконує функції універсального агрегатора даних та приведення їх до прийнятного, для подальшого використання в інформаційно-аналітичній системі вигляді. На цьому етапі використовуються як стандартні мови програмування для написання запитів до баз даних, наприклад, SQL, Ruby, SAS Base, DS2, або SCL так і спеціалізовані галузеві рішення, наприклад, DIS (Data Integration Studio) та SAS Inforation Map Studio.

- сервер метаданих дозволяє виконувати передачу інформаційних потоків в рамках всії IT інфраструктури між користувачами та програмно-апаратними комплексами, що в неї входять.

- аналітичне програмне забезпечення, в загальному випадку, може включати не тільки SAS рішення, а також рішення інших вендорів, наприклад, SPSS, R, Eviews, Netica, BayesiaLab та інші.

Налаштування та побудова технологічного процесу з аналізу даних дозволяє дані перетворити на інформацію у вигляді аналітичних моделей, скорингових кодів програм, звітів та API інтерфейсів.

Обрана SAS таблиця повинна бути доступна для сервера SAS Foundation через відповідну бібліотеку. Може бути використана будь-яка SAS таблиця, що знаходиться в бібліотеці SAS, включаючи ті, що зберігаються у форматах зовнішніх по відношенню до SAS. Наприклад, таблиці реляційних баз даних

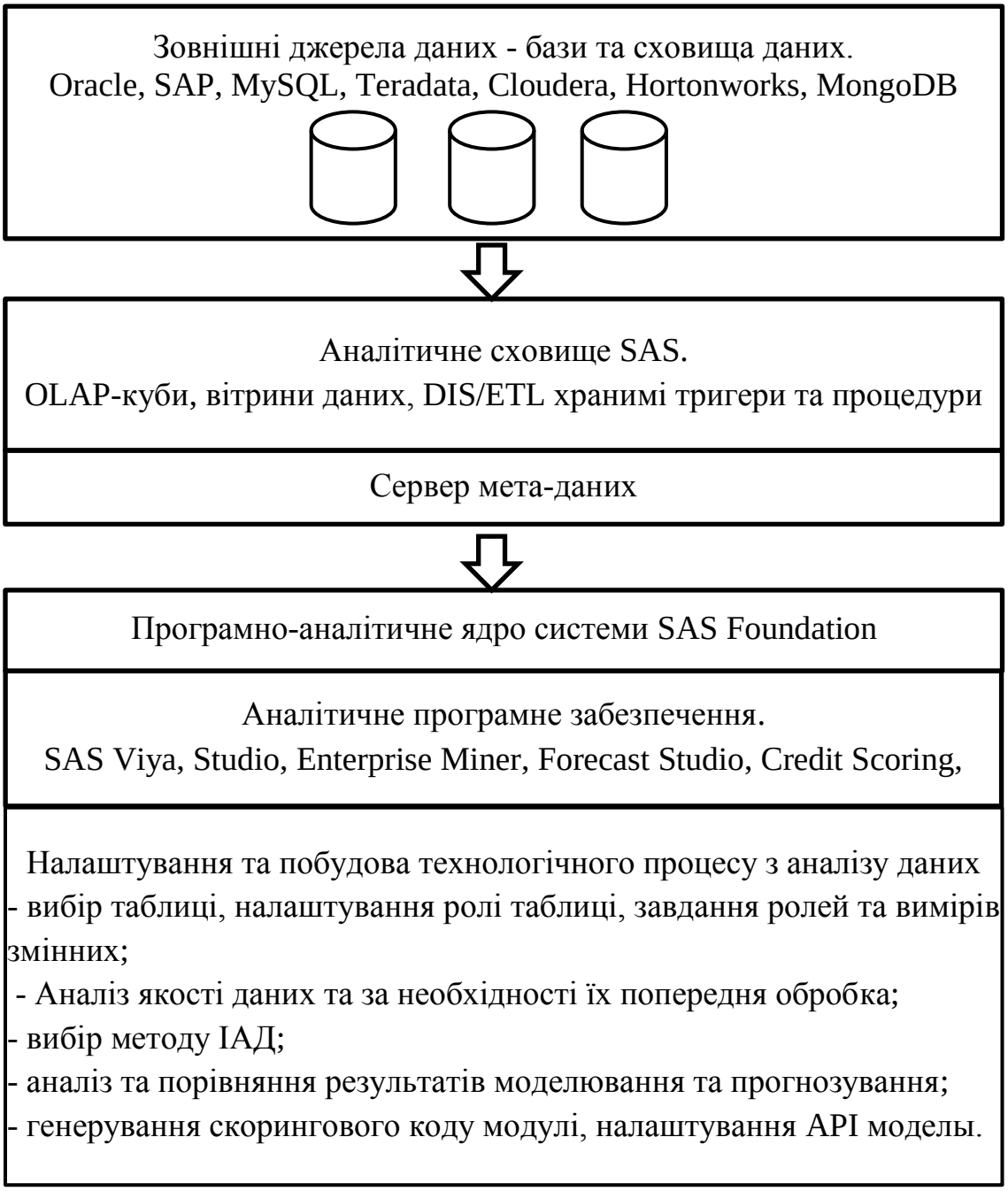


Рисунок 6.12 – Концептуальна схема функціонування інформаційної системи.

Будь-яке джерело даних, описаний для використання в SAS EM, повинен бути майже повністю готовий для аналізу. Як правило заснована робота з підготовки даних виконується до опису джерела даних.

Висновки до розділу 6

В роботі створені інформаційні технології для розв’язування задач прийняття рішень щодо управління розвитком соціально-економічних систем, які ґрунтуються на інтеграції різнотипної інформації й засновані на системному використанні методів аналізу даних, моделювання, методів прогнозування й методів прийняття рішень, розглянутих у попередніх розділах**.**

Запропоновано метод синтезу інформаційних технологій, який поєднує в собі задачі, які вирішуються на кожному етапі прийняття рішень за допомогою таких сучасних інформаційних технологій: інформаційна технологія аналізу й оцінювання інформації, інформаційна технологія моделювання для побудови математичних моделей, інформаційна технологія прогнозування, інформаційна технологія підтримки прийняття рішень. Кожна інформаційна технологія заснована на системному використанні інструментальних методів і методик, які вирішують окремі завдання прогнозного моделювання та прийняття рішень.

При цьому групи інструментальних методів використовуються залежно від типу й механізму вибору та конкретної задачі прийняття рішень. Методи можуть використовуватись як окремо – для розв’язування окремих задач прийняття рішень, так і системно для вирішення загальної проблеми.

Запропонована інформаційна СППР відрізняється застосуванням спеціальних мов програмування SAS/Baseта SAS/IML і має гнучку архітектуру, що є важливим фактором для модифікації розроблених методів і методик аналізу даних, а також продовження виконання досліджень в даній області. Методи, що були розроблені та реалізовані призначені насамперед для автоматизації процесу інтелектуального аналізу даних що описують досліджувані процеси.

РОЗДІЛ 7
ЗАСТОСУВАННЯ СТВОРЕНИХ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У ЗАДАЧАХ ПРОГНОЗУВАННЯ

7.1 Інформаційна технологія прогнозування на основі застосування комбінацій регресійних моделей та МБ

Ефективним є застосування багатомодельного підходу, в основу якого покладено використання комбінацій регресійних моделями та байєсівських мереж для прогнозування нелінійних нестаціонарних процесів у еколого-економічних системах. Для застосування даного підходу запропоновано методику, реалізовано за допомогою наступних кроків:

Крок 1. Завантаження даних.

Крок 2. Розділення даних на 2 підвибірки: навчальну та перевірочну.

Крок 3. Побудова моделей.

Крок 4. Порівняння моделей за статистичними критеріями якості.

Крок 5. Побудова загального звіту за виконаним обчислювальним експериментом.

Прогнозування нелінійних нестаціонарних процесів виконано на матеріалах урожайності зернових. Нестаціонарність досліджуваного процесу підтверджено розрахунками значення математичного сподівання показника урожайності, яке відрізняється у аналізовані роки у 1,62 рази, а дисперсія – у 6,13 рази. Вхідними даними для побудови моделей є часові ряди значень NDVI та погодних факторів. Побудована модель лінійної регресії має вигляд (7.1):

$$\begin{aligned} \text{Productivity} = {} & 25{,}352 \cdot \text{Intercept} + 89{,}646 \cdot \text{ndvi\_17\_18} - 29{,}044 \cdot \\ & \text{ndvi\_19\_20} + 74.230 \cdot \text{ndvi\_23\_24} - 89{,}578 \cdot \text{ndvi\_25\_26} + 52{,}028 \cdot \\ & \text{ndvi\_27\_28} + 0{,}330 \quad \text{temp\_17\_18} + 0{,}711 \cdot \text{temp\_19\_20} - 4{,}118 \cdot \\ & \text{temp\_21\_22} \end{aligned} \qquad (7.1)$$

Регресори типу ndvi_17_18 позначають різницю між значенням нормалізованого відносного індексу вегетації (NDVI) на 18-у та 17-у тижнях; відповідно, temp_17_18 позначає різницю між температурою навколишнього середовища на 18-у та 17-у тижнях вегетації.

В ході дослідженні було побудовано низку регресійних моделей з різними комбінаціями регресорів – факторів, що впливають на урожайність зернових (табл. 7.1).

Таблиця 7.1 – Статистичні характеристики регресійних моделей, побудованих для прогнозування урожайності озимої пшениці

Фактори моделі множинної регресії	RMSE	MAPE	DW
Тільки добрива	1,19	24,78	2,25
Тільки культура попередник	1,24	26,39	2,07
Добрива та культура попередник	1,17	24,59	2,31
Добрива, культура-попередник, добрива-культура - попередник	1,17	24,59	2,32
Тільки температура	0,93	20,58	1,41
Тільки опади	1,05	19,43	1,51
Тільки температура та опади	0,69	13,62	1,17
Добрива, культура попередник, температура та опади	0,52	10,01	1,95

Також були побудовані моделі за методом LARS з вибором моделі за значенням ASE та SBC на перевірочній вибірці, методом PLS, моделювання залежності урожайності озимої пшениці від факторів навколишнього середовища за лінійною регресією з відбором значущих змінних за допомогою дерева рішень, за методом градієнтного бустингу, дві моделі у формі мережі Байєса. Для побудови другої мережі Байєса, робились такі перетворення:

1. Обчислені перші різниці від показників NDVI, температури та кількості опадів.
2. Дані перетворені в дискретну форму за методом рівних проміжків

Таблиця 7.2 – Значення показників якості регресійних моделей залежності врожайності озимої пшениці від NDVI і погодних факторів

Модель	Вибірка					
	Навчальна			Перевірочна		
	RMSE	SSE	MAX E	RMSE	SSE	MAX E
Регресійна	5,21	977,74	13,64	5,54	459,81	11,73
LARS	5,21	977,73	13,64	5,57	465,68	11,74
LARS за критерієм мінімального SBC	5,54	1104,23	15,74	6,3	595,91	12,65
PLS	5,21	977,73	13,64	5,57	465,68	11,74
Градієнтного бустингу	3,63	475,25	10,58	4,94	366,78	13,04

Як видно з табл. 7.2 найкращою виявилась модель, побудована за методом градієнтного бустингу. Досліджуючи отримані результати моделювання залежності врожайності озимої пшениці від NDVI та погодних факторів на основі регресійних моделей, можна зробити висновок, що найкраща модель отримана за методом градієнтного бустингу. Показники якості на навчальній вибірці такі: RMSE = 3,63; SSE = 475,25; MAX E = 10,58; на перевірочній вибірці: RMSE = 4,94; SSE = 366,78; MAX E = 13,04). Інші моделі показали гірші і схожі результати: RMSE від 5,21 до 5,54; SSE від 977,73 до 1105,32; MAX E від 13,64 до 15,74. На перевірочній вибірці отримано такі значення: RMSE від 5,54 до 6,42; SSE від 459,81 до 595,91; MAX E від 11,73 до 13,16.

Також побудовані мережі Байєса за двома методами: за вихідними даними, перетвореними в інтервальні значення за методом оптимального розбиття на групи з використанням алгоритму дерева рішень. Для побудови

мереж використано спеціальний компонент Optimal Binning в системі SAS Enterprise Miner за вихідними даними, переведеними в різницеву форму і в інтервали за методом рівних інтервалів.

Для порівняння якості побудованих регресійних моделей з мережами Байєса, розраховано такі показники:

– для мереж Байєса розраховано відношення правильно класифікованих значень врожайності і вибиралась та мережа Байєса, у якої це показник вищий;

– для регресійних моделей вхідні дані показника врожайності і показника врожайності, розрахованого за відповідною моделлю, перетворено в інтервали відповідно до мережі Байєса, з якою виконувалось порівняння.

Отримані мережі Байєса порівнювались за загальною точністю класифікації, показники якої склали 55,36% та 70,98% для кукурудзи, і 78,43% та 47,06% для пшениці.

Порівнюючи результати прогнозування врожайності для обох досліджуваних культур, отримані за регресійними моделями та байєсівськими мережами, кращими є отримані методом градієнтного бустингу (відсоток класифікації на тестовому наборі даних склав: мережі Байеса – 78,43%, метод градієнтного бустингу – 86,11%, регресійні моделі, PLS, – 69,44, PLS, LARS – 69,44%).

Серед регресійних моделей урожайності пшениці, кращою виявилась модель, отримана за методом градієнтного бустингу. Показники якості на навчальній вибірці такі: RMSE = 3,63, SSE = 475,25, MAX E = 10,58; на перевірочній вибірці: RMSE = 4,94, SSE = 366,78, MAX E = 13,04. Серед байєсівських мереж, кращими за показником точності класифікації виявилися такі: для кукурудзи – модель, побудована за вихідними даними, переведеними в інтервальний вимір за методом рівних інтервалів (точність класифікації – 70,98%); для озимої пшениці – мережа, побудована за вихідними даними, переведеними в інтервальний вимір за методом градієнтного бустингу (точність класифікації – 78,43%).

7.2 Моделі, методи та інформаційна технологія прогнозування ННП за різних часових горизонтів та різних сценаріїв

Математичні методи та підходи та методи довго-, середньо-, короткострокового та дуже короткострокового прогнозування навантаження енергосистеми були застосовані для вирішення задачі прогнозування нелінійних нестаціонарних процесів в енергетиці. Для проведення обчислювальних експериментів та оцінювання коефіцієнтів рівнянь було використано реальні статистичні дані однієї з енергетичних компаній України. Схема застосованої методики представлена на рис. 7.

Для побудови довгострокової математичної моделі на річних інтервалах було використано такі макроекономічні чинники, як значення ВВП, загально чисельності населення та чисельність працюючих на виробництві, а також індекс, що відображає стан розвитку економіки та кількість населення. Отримана модель має характеристики R^2 = 0,98, MAPE=0,2%.

При середньостроковому прогнозуванні, на місячних відрізках даних, було побудовано модель авторегресії дванадцятого порядку та Вінтерса з мультиплікативною сезонністю, для яких значення статистики MAPE становить 2,64% та 2,6% відповідно.

Менш точні результати було отримано при короткостроковому прогнозуванні на квартальних даних, похибка авторегресії четвертого порядку сягнула значення статистики MAPE в 6,9%.

Для випадку короткострокового прогнозування на годинних часових проміжках, навчання моделей відбувалося на даних з 1 січня 2015 року по 31 вересня 2017 року, а тестування на відрізку від 1 жовтня до 31 грудня. Перед початком моделювання було досить ретельно виконано графічний та статистичний аналіз даних на наявність типових патернів та відмінностей поведінки за різних умов. Було виявлено періоди різкого зростання та спаду

споживання електроенергії, що використано при побудові моделей в якості додаткових чинників.

Аналіз даних в добовому та місячному розрізах температури та навантаження виявив залежність у вигляді оберненої фази в залежності від доби року. Як наслідок, температуру необхідно враховувати в моделі у вигляді поліному третього порядку разом із спеціалізованими комбінованими змінними на основі температури та місяця, температури та години доби, температури та дня неділі.

Було побудовано чотири моделі для короткострокового прогнозування у вигляді нейронної мережі, експоненційної моделі з мультиплікативною сезонністю, GLM моделі ARIMAX та GLM моделі із урахуванням свят, статистична характеристика MAPE дорівнює 3.19%, 2.55%, 2.43% та 2.39% відповідно.

На наступному етапі будують двохступеневі моделі. Це друга група методів, що будується із урахуванням результатів аналізу залишків одноступеневих моделей, включає наступні моделі:

1. GLM UCM (UCM – Unobserved Components Model) – модель компонентів, що не спостерігаються.

2. Нейронна мережа.

3. Моделі класу експоненційного згладжування (експоненційного згладжування, подвійного експоненційного згладжування, експоненційного згладжування з сезонністю, експоненційного згладжування з лінійним трендом (модель Хольта), Вінтерса з адитивною або мультиплікативною сезонністю, модель з урахуванням дрейфуючого тренду).

4. Модель GLM у вигляді ARIMAX.

5. Комбінована модель. Як правило це деяка комбінація прогнозів декількох моделей.

Більш детальний аналіз результатів показав, що загалом 93% попереднього значення навантаження враховується в наступному, при цьому присутній тренд в часових даних, а підмішування в модель регресорів що

описують свята підвищує точність прогнозування за критерієм MAPE на 0,04%.

Узагальнення результатів обчислювального експерименту представлене в таблиці 7.3, критерієм для порівняння якості побудованих моделей використано показник середньої похибки в процентах.

Таблиця 7.3 – Значення середньої похибки в процентах (MAPE) для різних класів моделей за реалістичного сценарію на різних горизонтах прогнозу, %

Назва моделі	Горизонт прогнозування			
	Дуже коротко-строкова модель, на 24 години вперед	Коротко-строкова модель, на 2 тижні вперед	Середньо-строкова модель, на 3 місяці вперед	Довгострокова модель
GLM-B	3,62	2,97	1,46	1,61
GLM-BR	3,32	2,74	2,15	1,67
GLM-BRW	3,32	2,74	2,15	1,67
GLM-BRWH	3,27	2,79	2,09	1,83
GLM UCM	2,68	2,05	1,70	3,13
Нейронна мережа	3,19	2,59	2,08	3,88
Експоненційне згладжування	2,55	2,01	1,86	3,15
GLM модель ARIMAX	2,39	1,97	1,57	3,09
Комбінована модель	2,44	2,03	1,75	1,82

Застосування мультимодельного підходу, попередньої підготовки даних дозволяє одержати значно кращі результати прогнозування, а використання програмного забезпечення SAS Energy Forecasting [270-271] – урахувати галузеві особливості досліджуваної компанії чи галузі національної економіки.

7.3 Інформаційна технологія прогнозного моделювання в умовах невизначеності за необхідності швидкого прийняття рішень

Використання запропонованого багатомодельного підходу, у системі підтримки прийняття інвестиційних рішень на ринку криптовалют забезпечило отримання прогнозів високої якості та вибір оптимальної торгівельної стратегії навіть в умовах невизначеності за необхідності швидкого прийняття рішень. Схема пропонованої системи підтримки прийняття рішень має структуру, представлену на рис. 6.4. Слід зазначити, що модуль управління збором даних з бірж орієнтований на використання API інтерфейсу та роботу переважно з біржами Bitfinex, Bitstamp, Huobi і т. ін.;. у модулі зберігання даних з використанням баз даних переважно застосовуються системи управління базами даних MySql, PostgreSQL та аналогічні; у модулі прогнозного моделювання передбачено використання таких аналітичних систем як SAS, R, Python, особливістю блоку прогнозного моделювання є наявність модуля тестової перевірки ідей та гіпотез, що викликано необхідністю проведення значної кількості чисельних експериментів з метою апробації пропонованих методик прогнозного моделювання у блоці прогнозного моделювання також реалізовано модуль вибору стратегії торгівлі криптовалютою; модуль представлення результатів надає можливість користувачам додаткового генерувати звіти про якість прогнозів на ретроспективних даних, а також про прибутковість поточних угод.

Використаний у дослідженні набір даних про угоди щодо купівлі-продажу криптовалюти складається з 957510 рядків, кожний з яких містить агреговану інформацію за кожні 15 секунд роботи біржі за період с 24.06.2018 по 18.12.2018 г. Агрегована інформація складається з таких елементів, як часова мітка в форматі день-місяць-рік-година-хвилина-секунда (DateTime), планова ціна пропозиції 1 біткоіна (Bid), загальний оборот на

біржі за 15 секунд (Aggregated_tick_volume), кількість угод на біржі за кожні 15 секунд її роботи (Trades_count).

На етапі попередньої обробки даних було досліджено фактори, що впливають на кількість та періодичність утворення пропусків у часових рядах біржових даних щодо торгів біткоін. Чисельні експерименти показали, що наявності на годинному фреймі 7,23% часових інтервалів містять від 80 до 100% пропусків, і 4,99% годинних інтервалів – від 50 до 80% пропусків. Однак, на добових інтервалах відсоток пропусків складає 5,74% і 4,92%, відповідно.

Для прогнозування обсягів торгів криптовалюти в рамках відповідного часового інтервалу на основі часових рядів, що містять пропуски даних, пропонується наступний підхід:

Крок 1. обчислити кількість хвилин за той проміжок часу для якого є дані щодо прогнозованого фінансового інструменту за формулою:

$$Indicator60 = \sum_{t=1}^{60} I(t),$$

де $I(t) = 1$, якщо для відповідного t −го хвилинного фрейму значення ряду заповнено, в іншому випадку $I(t) = 0$.

Крок 2. обчислити сумарне значення досліджуваного показника для годинного часового фрейму на основі наявних хвилинних фреймів:

$$Volume60 = \sum_{t=1}^{60} Volume(t),$$

де $Volume(t)$ − значення показника хвилинного фрейму t.

Крок 3. обчислити уточнене значення досліджуваного показника для відповідного часового фрейму:.

$$Volume\_bias\_60 = Vol\ ume60 \cdot \frac{60}{Indicator60}$$

Для інших часових фреймів використовується аналогічний підхід. Лише базис для розрахунків використовується рівний не 60, а фактичному значенню. Наприклад, для чотирьохгодинних фреймів, він дорівнює 240.

В таблиці 7.4 наведене порівняння результатів прогнозування обсягів торгів валютної пари BTC-USD після застосування різних методів

експоненційного згладжування для заповнення пропусків у часових рядах. В якості критерію для порівняння результатів прогнозного моделювання, використано показник середньоквадратичного відхилення (RMSE) на годинному та добовому фреймах.

Таблиця 7.4 – Порівняння результатів прогнозування обсягів торгів валютної пари BTC-USD за застосування різних методів експоненційного згладжування

Модель	Значення показника середньоквадратичного відхилення (RMSE)	
	на 15-ти хвилинному фреймі	за добового фрейму
Експонеційного згладжування з сезонністю	1,16	24,32
Вінтерса з адитивною складовою	1,17	17,01
Демпфуючого тренду	1,09	15,65
Подвійного експоненційного згладжування (модель Брауна)	1,35	18,57
Експоненційного згладжування з лінійним трендом (модель Хольта)	1,11	16,45
Експонеційна з мультиплікативною сезонністю	1,18	19,75
Просте експоненційне згладжування	1,09	16,71
Вінтерса з мультиплікативною складовою	1,21	21,66

Як видно з таблиці 7.4, найкращий результат одержано за застосування для заповнення пропусків у часових рядах даних моделей експоненційного згладжування с демпфуючим трендом.

Поряд із математичним моделюванням були застосовані інструменти технічного аналізу, використання яких також дозволяє аналізувати та передбачати загальні тенденції на ринку біткоїнів зокрема: осциляторами та індикаторами, такими як MACD, RSI, SMA та ін. У чисельному експерименті було використано індикатор Order Indicator. Ще однією частиною чисельного експерименту є дослідження ефективності застосування математичних моделей за запропонованого варіанту адаптивного методу прогнозування. Дослідження виконане на чотирьохгодинному часовому

фреймі, за показниками якості прогнозного моделювання. Кращою виявилась математична модель нейронної мережі (двошарова мережа прямого розповсюдження). Перший прихований шар формується на основі вхідних змінних $Amount_k$ з двох елементів H_{11} и H_{12}, а на основі змінних $Volume_k$ - другий прихований шар – також з двох елементів H_{21} та H_{22}, в той час як змінні $Price(t\text{-}1)$, …, $Price\ (t\text{-}12)$ напряму впливають на цільову змінну. Особливістю цієї мережі є те, що в кожному із прихованих шарів використовуються по два прихованих елементи. Для визначення оптимальної кількості прихованих елементів у прихованому шарі у загальному випадку можуть використовуватися різні підходи: простий емпіричний перебір; кластерний аналіз, виконуваний для всього набору прецедентів з метою виявлення оптимальної кількості кластерів, у які можуть бути об'єднані прецеденти; метод головних компонентів.

У нейронній мережі для визначення оптимальної кількості прихованих елементів у прихованому шарі використано метод головних компонентів та експертні оцінки фахівців. В якості функції активації використано гіперболічний тангенс. Тобто, математична модель нейронної мережі має вигляд:

$$g_0^{-1}(E(\Pr ice(t))) = \omega_0 + \omega_1 \cdot H_{11} + \omega_2 \cdot H_{12} + \omega_3 \cdot H_{21} + \omega_4 H_{22} + \omega_5 \cdot \Pr ice(t-1) + \ldots + \omega_{16} \cdot \Pr ice(t-12)$$

де $H_{11} = g_{11}(\omega_{01} + \omega_{11} \cdot Amount_{16} + \ldots + \omega_{71} \cdot Amount_{96})$,

$H_{12} = g_{12}(\omega_{02} + \omega_{12} \cdot Amount_{16} + \ldots + \omega_{72} \cdot Amount_{96})$, $H_{21} = g_{21}(\omega_{03} + \omega_{13} \cdot Volume_{16} + \ldots + \omega_{73} \cdot Volume_{96})$,

$H_{22} = g_{22}(\omega_{04} + \omega_{14} \cdot Volume_{16} + \ldots + \omega_{74} \cdot Volume_{96})$

За результатами розрахунку показників якості моделей, значення похибки RMSE рівне 31, а правильне визначення напрямку тренду ціни відбулось у 63% випадків.

За дотримання даної стратегії доходність угод з «купівлі-продажу» склала 4900 дол США. У розрахунку доходності враховувався чистий прибуток після вирахування комісії біржі за угоди. Всього, під час експерименту було виконано 139 торгівельних операцій, з яких 57 – на

підвищення курсу біткоїну, а 82 - на його зниження, середня доходність по кожній угоді склала 35 дол США.

Отже, на підставі проведеного експерименту запропоновано універсальну методику для формування інвестиційної політики на ринку криптовалют, яка реалізується у чотири етапи.

Етап 1. На основі наявних статистичних історичних даних щодо реалізації обраних стратегій протягом місяця розрахувати:

- вектор дохідності кожної зі стратегій протягом місяця;
- коваріаційну матрицю зміни дохідності всіх стратегій.

Етап 2. Для подальшого аналізу обирати стратегії з позитивну доходністю.

Етап 3. Використовуючи методи оптимізації на основі вектору доходностей стратегій та коваріаційної матриці сформувати відповідну інвестиційну політику. При цьому застосувати лінійні обмеження експертного характеру, зокрема, частка кожної стратегії не має перевищувати порогове значення (наприклад, 30%) та при цьому бути більшою за нуль, щоб забезпечити диверсифікацію інвестиційного портфелю.

Етап 4. Після формування інвестиційної політики, наявні активи розділяються у визначеній пропорції для кожної з рекомендованих стратегій та протягом місяця використовуються для біржових операцій. По закінченню місяця усі наведені на кроках 1 – 3 операції повторюються та формується рекомендація щодо оптимальної інвестиційної політики на наступний місяць.

Ефективність запропонованої методики була перевірена експериментальним шляхом. Вона була застосована під час торгівлі криптовалютою біткоїн в парі з доларами США (BTC-USD) на біржі Bitfinex у період з 01.07.2018 р. по 01.04.2019 р. за 12 стратегіями, які вирізнялись використовуваними індикаторами, їх комбінаціями, індексами, сигналами з бірж. Використання запропонованої методики дозволило після восьми місяців застосування отримати майже 43 центи прибутку на 1 доллар, вкладений у біткоіни (табл. 7.5).

Таблиця 7.5 – Аналіз доходності портфеля криптовалют

Рік	Місяць	Дохідність за інвестиційним портфелем, %	Приріст капіталу на 1 долар, вкладений у біткоїн, долл
2018	Серпень	-7,47	-0,07
	Вересень	17,56	0,09
	Жовтень	-0,56	0,08
	Листопад	-3,06	0,05
	Грудень	11,18	0,17
2019	Січень	14,01	0,40
	Лютий	5,67	0,41
	Березень	1,50	0,43

У таблиці 7.6 представлені статистичні характеристики ціни біткоїн за період з липня 2018 по квітень 2019 р.

Таблиця 7.6 - Статистичні характеристики ціни біткоїн у липні 2018 - березні 2019 р.

Рік, місяць		Вартість біткоіна, долл			Стандартне відхилення
		Середньомісячна	Мінімальна	Максимальна	
2018	Липень	7335,16	6113,2	8437,4	713,04
	серпень	6608,63	5960,2	7617	372,28
	Вересень	6565,78	6119,3	7410,2	314,27
	Жовтень	6539,58	6226,9	7017,6	148,68
	Листопад	4800,98	3691,5	6578,4	930,48
	Грудень	3793,42	3239	4384	278,44
2019	Січень	3754,63	3431	4157,5	187,20
	Лютий	3778,44	3434,5	4219,6	214,75
	Березень	3991,56	3776,8	4104,2	78,18

Отже, як показало дослідження, пропонована інформаційна технологія

є універсальною, гнучкою та адаптивною. Застосування мультимодельного підходу, в тому числі і для попередньої підготовки даних дозволяє одержати значно кращі результати прогнозування.

Програмна реалізація запропонованої системи підтримки прийняття рішень виконана в середовищі SAS Enterprise Guide 7.1 із використанням мови програмування SAS Base та спеціалізованої мови оптимізаційного моделювання OPTMODEL, призначеної для побудови моделей та використанню відповідних алгоритмів рішення оптимізаційних задач.

7.4 Інформаційна технологія короткострокового прогнозування ННП в умовах неповної чи недостовірної інформації

Інформаційна технологія, побудована на основі запропонованого багатомодельного підходу, використана у системі підтримки прийняття рішень для короткострокового прогнозування дозволила отримати прогнози високої якості за використання неповних даних щодо розвитку ННП. Для побудови прогнозу використано методи виявлення подібності процесів та ймовірнісно-статистичні моделі у формі мереж Байєса. Дана інформаційна технологія може бути використана як доповнення до існуючих систем прогнозування поширення інфекційних захворювань, оскільки запропонована інформаційна технологія більше орієнтована на вирішення задач короткострокового прогнозування (зокрема поширення коронавірусної хвороби). У зв’язку з тим, що на даний час немає значної кількості накопичених даних, що описують динаміку поширення цієї хвороби, особливостей перебігу її серед різних категорій населення в регіонах, тощо, запропоновано використання такого технологічного ланцюга, який би дозволяв виявляти країни із схожим перебігом епідемії, процеси, подібні за динамікою, широтою охоплення, структурою захворівших, тощо. У розробці використано програмне забезпечення SAS Enterprise Miner 14.1 та SAS Enterprise Guide.

На момент розробки системи в медичних базах даних містилась інформація щодо захворюваності на коронавірусну хворобу у 210 країнах. Для побудови моделей з них було обрано 50 (за принципом найбільшої кількості зареєстрованих випадків захворювання. Для виявлення схожих за ознаками (рівень смертності та індекс охорони здоров'я) країн було виконано кластерний аналіз статистичних даних. В якості вхідних даних використовувалися індекс смертності та індекс охорони здоров'я. Кластерний аналіз виконано в два етапи:

Етап 1. Із використанням методу Уорда було виконано перевірку гіпотез, щодо розбиття країн від 2 до 10 кластерів. На основі значення індексу CCC було з'ясовано, що оптимальна кількість кластерів дорівнює 6.

Етап 2. На основі методу кластерного аналізу *k*-середніх, із задаванням попередньої стандартизації вхідних змінних аналізу (з нульовим математичним сподіванням та одиничною дисперсією).

На основі кластерного аналізу було визначено країни-аналоги за рівнем смертності та станом медичної галузі.

Для подальшого дослідження були опрацьовані статистичні дані з Інтернет щодо захворюваності на коронавірусну хворобу у країнах, які входять до аналізованого кластеру та обчислено кумулятивні статистики із використанням розробленого автором програмного забезпечення. Підготовка даних для аналізу відбувалась в три етапи, а саме:

Етап 1. Введено змінну week (номер тижня). Відлік тижнів починається не від першого випадку, а від того дня коли було офіційно зареєстровано 100 випадків зараження.

Етап 2. Обчислено кумулятивне щотижневе значення підтверджених випадків зараження вірусом для кожної з країн, що увійшли до аналізованого кластеру.

Етап 3. Обчислено кумулятивне щотижневе значення підтверджених випадків смерті від вірусу для кожної з країн, що входять до аналізованого кластеру.

Порівняння часових рядів кількості захворівших на коронавірусну хворобу у досліджуваній країні і країнах-аналогах здійснюється автоматизовано за допомогою SAS STAT у наступні кроки (результати представлені у таблиці 7.7.

Крок 1. На основі вхідного цільового та часового ряду країни-аналога будується матриця відстаней.

Крок 2. Від елементу (1,1) до елементу (*m*,*n*) матриці відстаней будується оптимальний шлях з урахуванням обмежень на стиснення та розширення.

Крок 3. Обчислюються статистичні характеристики побудованого оптимального шляху (табл. 7.7).

Таблиця 7.7 – Статистичні характеристики результатів аналізу на ступінь подібності часових рядів

Показник	Країна				
	Білорусь	Угорщина	Нігерія	Румунія	Сербія
Процент пропусків в даних, %	2	0	5	0	0
Процент прямих співставлень шляху, %	72.00	83.33	86.96	87.50	83.33
Процент стиснень, %	12.00	8.333	4.348	8.333	8.333
Процент розширень, %	16.00	8.333	8.696	4.167	8.333
Загальний процент перетворень даних, %	28.00	16.67	13.04	12.50	16.67
Значення функціоналу вартості Cost	0.168	4,08	0.397	0.148	0.64
Міра подібності	59,88	74,21	89,64	96,79	88,16

Як можна побачити з таблиці 7.7, застосування міри подібності до аналізу нелінійних нестаціонарних процесів дозволяє відображати ступень подібності (схожості) часових рядів, навіть у випадках, коли їх довжини відрізняються. На відміну від стандартної кореляції Пірсона, яка відображає

наявність лінійного зв’язку, міра подібності відображає ще й нелінійний. Міра подібності обчислюється на основі розробленої методики, що описана у третьому розділі дисертації.

Очевидним є факт, що відсутність тестів впливає на об’єктивну статистичну картину урахування кількості хворих та померлих. Тому, в роботі пропонується використовувати підхід уточнення даних показників, розрахованих на основі даних країн-аналогів із використанням вагових коефіцієнтів, розрахованих для різних груп кластерів. Скориговані значення основних показників захворюваності в подальшому будуть використані для прогнозування поширення захворюваності серед населення в країні, що досліджується. Для прогнозування побудовано більше 700 моделей, зокрема, із застосуванням методів експоненційного згладжування (просте, подвійне, з лінійним трендом, з демпфуючим трендом) для прогнозування кількості хворих (померлих) в країні (табл. 7.8).

Таблиця 7.8 – Порівняння значення статистичної характеристики MAPE результатів прогнозування, отримане за різних методів експоненційного згладжування

Назва країни	Кількість етапів однокрокового прогнозування (для порівняння)	Експоненційне згладжування			
		просте	подвійне	з лінійним трендом	з демпфуючим трендом
Україна	17	11,78	1,95	1,87	1,9
Румунія	18	9,96	1,45	1,45	1,47
Білорусь	16	3,37	0,65	0,65	0,64
Угорщина	18	8,83	3,64	3,64	3,68
Нігерія	18	9,79	1,49	1,61	1,59
Сербія	18	5,91	1,48	1,48	1,48
Казахстан	17	13,72	7,09	7,25	7,04
Італія	21	1,92	0,55	0,55	0,54
США	20	8,38	0,93	0,93	0,93

Як можна побачити з табл. 7.8, в більшості випадків кращі результати прогнозування отримано за використання методу подвійного експоненційного згладжування. Застосування теорії подібності процесів для вирішення задачі прогнозування поширення коронавірусної хвороби дало можливість не лише отримати результати високої якості, а й опрацювати інформаційну невизначеність. Оскільки, як було підтверджено в ході дослідження, недостатнє охоплення населення тестуванням впливає на об'єктивну статистичну картину урахування кількості хворих та померлих внаслідок коронавірусної хвороби, тому, для уникнення спотворення вхідних даних, в роботі використано підхід уточнення значень показників із урахуванням значень аналогів-країн.

Висновки розділу 7

У сьомому розділі подано низку практичних прикладів успішного використання запропонованої інформаційної технології в СППР для моделювання нелінійних нестаціонарних процесів на конкретних прикладах з використанням фактичних статистичних даних.

Математичні методи та підходи до довго-, середньо-, короткострокового та дуже короткострокового прогнозування навантаження енергосистеми були застосовані для вирішення задачі прогнозування нелінійних нестаціонарних процесів в енергетиці. Для проведення обчислювальних експериментів та оцінювання коефіцієнтів рівнянь було використано фактичні статистичні дані однієї з енергетичних компаній України. Аналіз даних в добовому та місячному розрізах температури та навантаження виявив залежність у вигляді оберненої фази в залежності від доби року. Як наслідок, температуру необхідно враховувати в моделі у вигляді поліному третього порядку разом із спеціалізованими комбінованими змінними на основі температури та місяця, температури та години доби, температури та дня неділі. Було побудовано чотири моделі для

короткострокового прогнозування у вигляді нейронної мережі, експоненційної моделі з мультиплікативною сезонністю, GLM моделі ARIMAX та GLM моделі із урахуванням свят, статистична характеристика MAPE дорівнює 3.19%, 2.55%, 2.43% та 2.39% відповідно.

Ефективність запропонованої методики аналізу нелінійних нестаціонарних процесів перевірена експериментальним шляхом. Вона була застосована під час торгівлі криптовалютою біткоїн в парі з доларами США. Використання запропонованої методики дозволило після восьми місяців застосування отримати майже 43 центи прибутку на 1 доллар, вкладений у біткоіни).

Створена система може бути використана як доповнення до існуючих систем прогнозування поширення інфекційних захворювань, оскільки запропонована інформаційна технологія більше орієнтована на вирішення задач короткострокового прогнозування (зокрема поширення коронавірусної хвороби). У зв’язку з тим, що на даний час немає значної кількості накопичених даних, що описують динаміку поширення цієї хвороби, особливостей перебігу її серед різних категорій населення в регіонах, запропоновано використання такого технологічного ланцюга, який би дозволяв виявляти країни із схожим перебігом епідемії, процеси, подібні за динамікою, широтою охоплення, структурою захворівших тощо.

Застосування теорії подібності процесів для вирішення задачі прогнозування поширення коронавірусної хвороби дало можливість не лише отримати результати високої якості, а й опрацювати інформаційну невизначеність. Оскільки, як було підтверджено в ході дослідження, недостатнє охоплення населення тестуванням впливає на об’єктивну статистичну картину урахування кількості хворих та померлих внаслідок коронавірусної хвороби, тому для уникнення спотворення вхідних даних, в роботі використано підхід уточнення значень показників із урахуванням значень аналогів-країн.

ВИСНОВКИ ПЕРСПЕКТИВИ ВИКОНАННЯ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

На основі виконаних теоретичних та експериментальних досліджень вирішено важливу науково-прикладну проблему в галузі інформаційних технологій – підвищення ефективності прийняття рішень в управлінні розвитком нелінійних нестаціонарних процесів різної природи методами та засобами сучасних інформаційних технологій, інтелектуального аналізу даних, математичного моделювання і створення нових методів обробки даних і прогнозування. При цьому отримано такі нові результати:

1. Виконано аналіз сучасних методів моделювання та прогнозування розвитку процесів різної природи в умовах наявності нелінійності, нестаціонарності та невизначеності.

2. Розроблено системну методологію побудови математичних моделей для опису та прогнозування процесів, що характеризують розвиток нелінійних нестаціонарних процесів в умовах невизначеностей структурного, статистичного і параметричного характеру.

3. Удосконалено методи прогнозування нелінійних нестаціонарних процесів різної природи на основі використання ймовірнісно-статистичних та комбінованих методів прогнозування, що дозволяє зменшити похибки оцінок прогнозів на 5-30%.

4. Запропоновано метод оцінювання параметрів математичних моделей та їх ансамблів, що дозволило подолати зміщеність оцінок прогнозів та підвищити на 5-20% статистичні показники адекватності моделей.

5. Розроблено метод розкриття невизначеностей різних типів за допомогою ймовірнісно-статистичних методів з метою коректного розв’язання задач математичного моделювання і прогнозування розвитку обраних процесів.

6. Побудовано математичні моделі та ансамблі моделей обраних нелінійних нестаціонарних процесів для прогнозування їх розвитку з метою

підтримки прийняття управлінських рішень в умовах інформаційної невизначеності і з використанням створеної системи підтримки прийняття рішень.

7. Розроблено та виконано апробацію запропонованих методів і підходів до моделювання та прогнозування, запропоновано підходи, що забезпечують адекватний опис причинно-наслідкових зав’язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних нелінійних нестаціонарних процесів, що дозволило підвищити точність прогнозування їх подальшого розвитку на 10-20%.

8. Виконано аналіз адекватності побудованих моделей та формування висновків на їх основі щодо можливості їх застосування для розв'язання задач оцінювання і прогнозування розвитку нелінійних нестаціонарних процесів, зокрема, соціально-економічних та еколого-економічних, у межах використання створеної інформаційної системи підтримки прийняття рішень.

9. На основі системного підходу побудовано та реалізовано інформаційну технологію для розв'язання задач моделювання та прогнозування нелінійних нестаціонарних процесів з метою її подальшого використання у системах підтримки прийняття рішень.

У процесі виконання подальших досліджень за темою дисертаційної роботи доцільно розв’язати такі задачі:

– розширити застосування запропонованих у дисертаційному дослідженні методів і методик аналізу нелінійних нестаціонарних процесів (ННП) на виконання аналізу даних у частотній області, що сприятиме розширенню номенклатури математичних моделей і отриманню альтернативних методів прогнозування;

– застосувати для аналізу ННП сучасні нейронні мережі та нейронечіткі моделі, що сприятиме подальшому розширенню інструментарію для боротьби з невизначеностями амплітудного та іншого типів;

– активно використати для аналізу ННП, з відповідними модифікаціями, сучасні методи оптимальної та байєсівської фільтрації даних з метою подальшого підвищення їх якості та підвищення адекватності математичних прогнозуючих моделей;

– розробити розширену номенклатуру нелінійних елементів для підвищення адекватності математичного опису нелінійних нестаціонарних процесів;

– удосконалити існуючі методи оцінювання параметрів математичних моделей з метою їх поширення на розподіли ймовірностей даних різних типів з одночасним підвищенням якості оцінок параметрів моделей;

– на основі зазначеного вище створити узагальнену комплексну інформаційну технологію моделювання і прогнозування розвитку нелінійних нестаціонарних процесів, що характеризуються довільними типами розподілів ймовірностей та нелінійностями стосовно змінних і параметрів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Bates D. M., Watts D. G. Nonlinear Regression Analysis and Its Applications. New York: John Wiley & Sons, 1988. 392 p.

2. Arora S., Little M. A., McSharry P. E. Nonlinear and nonparametric modeling approaches for probabilistic forecasting of the US gross national product. *Studies in Nonlinear Dynamics and Econometrics*. 2013. Vol. 17, No. 4. P. 395–420.

3. Pearson R. K., Oggunaike B. A. Nonlinear process identification. Chapter 2. *Nonlinear process control*. Prentice-Hall, Inc., NJ, USA, 1997. P. 11–110. URL: https://dl.acm.org/doi/book/10.5555/248020 *(Last accessed: 18.07.2020).*

4. Márcio G., Garcia P., Marcelo C., Medeiros M. C., Gabriel F. R. Vasconcelos. Real-time inflation forecasting with high-dimensional models: the case of Brazil. *International Journal of Forecasting*. 2017. Vol. 33, issue 3. P. 679–693.

5. Szeles M. R., Muñoz R. M. Analyzing the regional economic convergence in Ecuador. Insights from parametric and nonparametric models. *Romanian Journal of Economic Forecasting.* 2016, Vol. XIX (2). P. 43–65.

6. Zar J. H. Biostatistical Analysis. Third Edition. New York : Prentice-Hall, Inc., 1996. 662 p.

7. Demirtas H., Sally A., Recai M. Plausibility of multivariate normality assumption when multiply imputing non-gaussian continuous outcomes: a simulation assessment. Technical report 2006-001. Chicago, C: Division of Epidemiology and Biostatistics University of Illinois at Chicago., 2006. 23 p. URL: http://dx.doi.org/10.1080/10629360600903866.

8. Hoaglin D.C., Mosteller F., Tukey J.W. Understanding Robust and Exploratory Data Analysis. New York : John Wiley & Sons, 2000. 472 p.

9. John P. W. M. Statistical Design and Analysis of Experiments. New York : The Macmillan Company, 1971. 756 p.

10. Myers R. H. Response Surface Methodology. Virginia : Polytechnic and State University, 1988. 705 p.

11. Wilcox R.R. Keselman H.J. Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*. 2003. Vol. 8 (3). P. 254-274.

12. Zgurovsky M. Z., Zaychenko Y. P. Big Data: Conceptual Analysis and Applications. International Publishing Switzerland, 2019. 277 p.

13. Zgurovsky M. Z., Zaychenko Y. P. The Fundamentals of Computational Intelligence: System Approach. Switzerland : Springer International Publishing, 2016. 375 p.

14. Zhang Zhengyou. Parameter Estimation Techniques: A Tutorial with Application to Conic. *INRIA Research Report.* 2004. №.2676, 34 p. URL: http://research.microsoft.com/en-us/um/people/zhang/Papers/ZhangIVC-97-01.pdf *(Last accessed: 19.07.2020).*

15. Азарсков В. Н., Житецкий Л. С., Николаенко С. А. О сходимости стандартных алгоритмов последовательного обучения нейросетевых моделей в нестохастической среде. *Нейроинформатика-2011* : сб. науч. трудов XIII Всерос. науч.-тех. конф. Москва : НИЯУ МИФИ, 2011. Ч. 2. С. 258–260.

16. Berry M. J. A., Linoff G. S. Data Mining Techniques. 2nd edition. New York : Wiley Publishing, Inc., 2004. 670 p.

17. Tangirala A. K. Non-stationary processes : Course of lectures “Introducing to random process”. IIT, Madras, India, 2021. 18 p.

18. Márcio G., Garcia P., Marcelo C., Medeiros M. C., Gabriel F. R. Vasconcelos. Real-time inflation forecasting with high-dimensional models: the case of Brazil. *International Journal of Forecasting*

19. Bidyuk P., Prosyankina-Zharova T., Terentiev O. Modelling Nonlinear Nonstationary Processes in Macroeconomy and Finances. *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing* / Ed. Hu Z., Petoukhov S., Dychka I., He M. Cham : Springer, 2019. Vol. 754. P. 735–745. URL: http://doi.org/10.1007/978-3-319-91008-6_72

20. Коршевнюк Л. О., Терентьєв О. М., Бідюк П. І. Методика побудови математичних моделей динамічних процесів. *Системний аналіз та інформаційні технології(SAIT 2013) :* матеріали 15-ї міжнар. наук.-техн. конф. (Київ, 27–31 трав. 2013 р.). Київ : ННК "ІПСА" НТУУ "КПІ", 2013. С. 288–289.

21. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. Киев : Радио и связь, 1987. 120 с.

22. Кашкарьов А. О., Терентьєв О. М. Аналіз випадкових процесів за допомогою швидкого перетворення Фур'є. *Праці Таврійського державного агротехнологічного університету*, 2010. Вип. 10, Т. 10. С. 158–162.

23. Бард Й. Нелинейное оценивание параметров : пер. с англ. Москва : Статистика, 1979. 349 с.

24. *Sel'kov E. E. (1968). Self-Oscillations in Glycolysis 1. A Simple Kinetic Model. European Journal of Biochemistry. 1968. 4 (1). S. 79–86.* URL: https://febs.onlinelibrary.wiley.com/doi/full/10.1111/j.1432-1033.1968.tb00175.x *(Last accessed: 14.12.2020).*

25. Polozhaenko S. A., Abdullach O. M. Methods of mathematical design and machine authentication o fnon – statics anomalous diffusive processes. *Informatics and mathematical method sinmodelling*. 2015. Vol. 5, N 3. P. 249–258.

26. Krak I., Barmak O., Manziuk E. Using visual analytics то develop human and machine-centric models: A review of approaches and proposed information technology, *Computition alIntelligence.* 2020. Vol. 1. P. 1–26.

27. Roberts F. S., Discrete mathematical models with applications to social, biological and environmental objectives. *New York* : Englewood Cliffs : Prentice-Hall. 1986. XVI, 559 p.

28. Jordanova T. An Introduction to Stationary and Non-Stationary Processes. 2020. URL: https://www.investopedia.com/articles/trading/07/stationary.asp#:~:text=A%20non%2Dstationary%20process%20with,constant%20and%20independent%20of%20time.&text=It%20specifies%20the%20value%20at,trend%2C%20and%20a%20stochastic%20component *(Last accessed: 21.06.2020).*

29. Cheng C., Sa-Ngasoongsong A, Faruk O., Le T. Time Series Forecasting for Nonlinear and Nonstationary Processes: A Review and Comparative Study. *IIE Transactions,* 2015. DOI: 10.1080/0740817X.2014.999180.

30. Explained: Linear and nonlinear systems. *MIT News*. URL: https://news.mit.edu/2010/explained-linear-0226 *(Last accessed: 10.01.2021).*

31. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. Москва : Финансы и статистика, 2003. 416 с.

32. Данилов В. Я, Бідюк П. І, Жиров О. Л., Бідюк О. П. Прогнозування кредитоспроможності фізичних осіб за допомогою байєсівських мереж. *Економіка: теорія та практика*. 2016. №2 (8). С. 56–65.

33. Литвиненко В. І., Кожухівська О. А., Фефелов А. О. Метод прогнозування гетероскедастичних процесів з використанням синтезованих поліноміальних нейронних мереж. *Вісник Національного університету "Львівська політехніка". Серія: Інформаційні системи та мережі*. 2015. №829. С. 201–214. URL: http://nbuv.gov.ua/UJRN/VNULPICM_2015_829_16 (дата звернення: 22.12.2020)

34. Терентьєв О. М., Кириченко В. Е., Связінська Н. О., Просянкіна-Жарова Т. І. Прогнозування фінансових ризиків з використанням наївного і доповненого деревом класифікаторів на основі байєсівських мереж. *Наукові вісті НТУУ КПІ.* 2016. №2. С. 60–68. URL: https://doi.org/10.20535/1810-0546.2016.2.63882

35. Снитюк В. Є. Прогнозування. Моделі. Методи. Алгоритми : навч. посіб. Київ : Маклаут, 2008. 364 с.

36. Светуньков И.С., Светуньков С.Г. Методы социально-экономического прогнозирования. Теория и методология. Москва: Юрайт, 2015. 351 с.

37. Brocklebank J. C., Dickey D. A. *SAS System for Forecasting Time Series* / Second Edition Cary N C. SAS Institute Inc., 2003. 418 p.

38. Bidyuk, P. Terentiev O., Prosyankina-Zharova T. Dynamic processes forecasting and risk estimation under uncertainty using decision support systems. IEEE First Ukraine Conference on Electrical and Computer Engineering

(UKRCON) : Proceedings of theProceedings of theconference (Kyiv, 29 May-2 June 2017). Kyiv : Igor Sikorsky Kyiv Polytechnic Institute. 2017. P. 795–800. (Scopus). URL: https://doi. 8100355

39. Badach A. Adaptive Models in Econometric Forecasting. IFAC Proceedings Volumes. 1980. Vol. 13, issue 5. P. 271–277.

40. Littell R. C., Milliken G. A., Stroup W. W., Wolfinger R. D. SAS® System for Mixed Models, Cary, NC : SAS Institute Inc.,1996. 828 p.

41. Dobrescu E. Modelling an Emergent Economy and Parameter Instability Problem. *Journal for Economic Forecasting of Institute for Economic Forecasting*. 2017. Vol. 0(2). P. 5–28.

42. Bidyuk P., Huskova V., Terentiev O. Client solvency estimation using intellectual data analysis approach, *Advanced Information Systems and Technologies (AIST 2018)* : Proceedings of the 6th International Conference (Sumy, 16–18 May 2018), Sumy : Sumy State University, 2018 . С. 12–16.

43. Bidyuk P., Prosyankina-Zharova T., Terentiev O., Medvedeva M. Adaptive modelling for forecasting economic and financial risks under uncertainty in terms of the economic crisis and social threats. *Технологічний аудит та резерви виробництва.* 2018. №4. С. 24–36. URL: https://doi.org/ 10.15587/2312-8372.2018.135483.

44. Bauer D. J. A Semiparametric Approach to Modeling Nonlinear Relations Among Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*. 2005. Vol. 12, issue 4. URL: https://www.tandfonline.com/ doi/abs/10.1207/ 15328007sem1204_1 *(Last accessed: 11.08.2020).*

45. Andreica, M., Popescu, M.ME., Micu D., Albu E. Adaptive management procedural model for support of economic organizations. *Proc of 10th International Management Conference "Challenges of Modern Management*. 2016. P. 295–301.

46. Bidiuk P. I., Prosiankina-Zharova T. I., Terentieev O. M., Lakhno V. A., Zhmud O. V. Intellectual technologies and decision support systems for the control

of the economic and financial processes. *Journal of Theoretical and Applied Information Technology*. 2019. Vol. 96. No. 1. P. 71–87. (Scopus Q3)

47. Bidyuk P. I, Menyailenko O. S., Polovtcev O. V. Methods of Forecasting., Lugansk : Alma Mater, 2008. 608 p.

48. Bidyuk P., Overmyer S., Prosyankina-Zharova T., Terentiev O. Methodology of Modelingand Forecasting Nonlinear Processesin Finances. *Наукові вісті НТУУ КПІ*. 2018. №1. С. 15–25. URL: https://doi.org/10.20535/1810-0546.2018.1.120361.

49. Trofymchuk O. M, Bidiuk P. I., Prosiankina-Zharova T. I., Terentiev O. M. Decision support systems for modelling, forecasting and risk estimation: monography. Riga : LAP LAMBERT Academic Publishing. 2019. 176 p.

50. Trofymchuk O. M., Bidyuk P. I., Prosyankina-Zharova T. I., Terentiev O. M. Decisionsupportsystemformodelingandforecastingecological processes. *Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях*: колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2020. С. 23–44.

51. Trofymchuk O. M., Bidyuk P. I., Prosyankina-Zharova T. I., Terentiev O. M. Forecasting nonstationary processes in demography, ecology, economy and finances. *Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях* : колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2018. С. 113–123.

52. Box, G. E. P., Cox D. R. An Analysis of Transformations. *Journal of the Royal Statistics Society.* 1964. P. 211–252.

53. Ding F., Chenb T. Identification of Hammerstein nonlinear ARMAX systems. *Automatica.* 2005. №41. P. 1479–1489. URL: http://www.paper.edu.cn/scholar/showpdf/NUT2EN2IMTD0ExeQh *(Last accessed: 18.09.2020).*

54. Fung E. H., Wong Y. K., Ho H. F., Mignolet M. P. Modelling and prediction of machining errors using ARMAX and NARMAX structures. *Applied Mathematical Modelling*. 2003. Vol. 27, issue 8. P. 611–627. URL:

https://www.sciencedirect.com/science/article/pii/S0307904X03000714 *(Last accessed: 19.03.2021).*

55. Stoviček K. Forecasting with ARMA models. The case of Slovenian inflation. *Prikazi in analize*. 2007. Vol. XIV/1. P. 23–55.

56. Бокс Дж., Дженкинс Г. М. Анализ временных рядов. Москва : Мир, 1974. 288 с.

57. Griepentrog G. L., Ryan J. M., Smith, L. D. Linear Transformation of Polynomial Regression Models. *The American Statistician*. 1982. Vol. 36, issue 3a. P. 171–174.

58. White H. A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrics*. 1980. Vol. 48. P. 817–838.

59. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society*. 1996. Vol. 58, №1. P. 267–288.

60. De Jong P., Heller G. Z. Generalized Linear Models for Insurance Data. New York : Cambridge University Press, 2008. 197 p.

61. Hastie and Tibshirani *Generalized Additive Models*. New York: Chapman and Hall, 1990. 175 p.

62. McCullagh P., Nelder J. A. Generalized Linear Models. Second Edition. London : Chapman and Hall, 1989. 532 p.

63. Nelder J. A., Wedderburn R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society A.* 1972. Vol. 135. P. 370–384.

64. Abdi H. Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences*. 2003. P. 792–795. URL: https://www.utdallas.edu/~herve/Abdi-PLS-pretty.pdf *(Last accessed: 11.08.2020).*

65. De Jong S. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*. 1993. Vol. 18. P. 251–263.

66. Tobias R. D. An Introduction to Partial Least Squares Regression." In Proceedings of the Twentieth Annual SAS Users Group International Conference, P. 1250–1257. Cary, NC: SAS Institute, Inc.,1995.

URL: http://support.sas.com/rnd/app/stat/papers/pls.pdf. *(Last accessed: 18.07.2020).*

67. Bradley E., Hastie T., Tibshirani R., Johnstone Least Angle Regression. *Annals of Statistics*. 2004. Vol. 32 (2). P. 407–499.

68. Jim G., Jeff T., Chip W. Applied Analytics Using SAS Enterprise Miner: Course Notes. SAS Institute Inc. Cary, NC, USA, 2010. 698 p.

69. Amrhein J., Dickey D., Kiernan K., Marovich P. Statistics 2: ANOVA and Regression Course Notes. SAS Institute Inc. Cary, NC, USA. 2012. 712 p.

70. Freund R. J., Littell R. C. SAS System for Regression. 3 Edition. Cary, NC : SAS Institute, Inc., 2000. 256 p.

71. Iman R. The Analysis of Complete Blocks Using Methods Based on Ranks. *Proceedings of the SAS Users Group International Conference.* Cary, NC : SAS Institute, Inc., 1988. Vol. 13. P. 970–978.

72. Breiman L., Friedman J., Olshen R., Stone C. Classification and Regression Trees. Florida : CRC Press, 2000. 360 p.

73. Chow C. K., Liu C. N. Approximating discrete probability distributions with dependence trees. *IEE Transactions on information theory*. 1968. Vol. IT-14, №3. P. 462–467.

74. Pashynska N., Snytyuk, V., Putrenko V., Musienko A. A decision tree in a classification of firehazard factors. *Восточно-Европейский журнал передовых технологий*. 2016. №5(10). С. 32–37. URL: http://nbuv.gov.ua/UJRN/Vejpte_2016_5%2810%29__6 (дата обращения: 12.12.2020).

75. Quinlan J. R. Induction of decision trees. *Machine Learning*. Boston : Kluwer Academic Publishers.1986. Vol. 1, № 1. P. 81–106.

76. Quinlan J. R. C4.5: programs for machine learning. San Francisco : Morgan Kaufmann, 1993. 302 p.

77. Galante R. Improving the Performance of Data Mining Models with Data Preparation Using SAS Enterprise Miner. *Proceeding SAS Global Forum*, April 26–29, 2015. Dallas, TX, USA, 2015. 21 p. URL:

https://support.sas.com/resources/papers/proceedings15/SAS1965-2015.pdf *(Last accessed: 11.08.2020).*

78. Adaptive fuzzy clustering of multivariates hort time series with unevenly distributed observations based on matrix neuro-fuzzy self-organizing network / Setlak G. et al. *Advancesin Intelligent Systemsand Computing*. 2018. Vol. 643. P. 308–315.

79. Schubert S., Lee T. Time Series Data Mining with SAS Enterprise Miner. *The SAS Global Forum 2011 Conference.* 2011. 12 p. URL: https://support.sas.com/resources/papers/proceedings11/160-2011.pdf. *(Last accessed: 12.09.2019).*

80. Gillies D. A. Lecture 11: Exact Inference / Course of lectures on «Intelligent data analysis and probabilistic inference». Imperial college of London : Department of computing, 2006. 4 p.

81. Gilks W. R., Richardson S., Spiegelhalter D. J. Markov Chain Monte Carlo in Practice. New York : Chapman & Hall/CRC, 2000. 486 p.

82. Bolstad W. M. Understanding computational Bayesian statistics. Hoboken (New Jersey): John Wiley & Sons, Ltd, 2010. 334 p.

83. Top 10 algorithms in data mining / Wu X. et al. *Knowledge and Information Systems.* 2008. Vol. 14, №1. P. 1–37.

84. Tyshchenko O., Kopaliani D., Bodyanskiy Ye. The least squares support vector machines based on a neo-fuzzyneuron. *Computational Modelsfor Businessand Engineering Domains* / Eds. By G. Setlak, K. Markov. Rzeszow-Sofia : ITHEA. 2014. P. 44–51.

85. Bayes T. An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*.1763. Vol. 53. P. 370–418.

86. Cooper G., Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*. 1992. Vol. 9. P. 309–347.

87. Bernardo J. M., Smith A. F. M. Bayesian Theory. New York : John Wiley & Sons, Ltd, 2000. 586 p.

88. Hautaniemi S. K., Korpisaari P. T., Jukka P. P. Saarinen. Target identification with dynamic hybrid Bayesian networks. *Proceedings of the forth SPIE's international symposium on sensor fusion: architectures, algorithms and applications*. Orlando (Florida US). 2000. P. 55–66.

89. Heckerman D. Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery*. 1997. № 1. P. 79–119.

90. Heckerman D. E., Horvitz E. J., Nathwani B. N. Toward normative expert systems : Part I The Path Finder Project. *Methods of information in medicine*. 1992. Vol. 31, №2. P. 90–105.

91. Heckerman D., Geiger D., Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. *Technical report MSR-TR-94-09*. Microsoft Research. 1994. 53 p.

92. Бідюк П. І., Кузнецова Н. В. Основні етапи побудови і приклади застосування мереж Байєса. *Системні дослідження та інформаційні технології*. 2007. №4. С. 26–39.

93. Бідюк П. І., Литвиненко В. І., Кроптя А. В. Аналіз ефективності функціонування мережі Байєса. Автоматика. *Автоматизація. Електротехнічні комплекси та системи*. 2007. № 2. С. 6–15. : URL: http://nbuv.gov.ua/UJRN/aaeks_2007_2_3 (дата звернення: 31.01.2020).

94. Härdle W. K., Hautsch N., Mihoci A. Local adaptive multiplicative error models for high-frequency forecasts. *Journal of applied econometrics* 2015. Vol. 30. P. 529–550.

95. Лукашин Ю. П. Адаптивные методы краткосрочного прогнозирования временных рядов. Москва : Финансы и статистика, 2003. 416 с.

96. Tsay R. S. Analysis of Financial Time Series. Hoboken : Wiley & Sons, Inc., 2010. 715 p.

97. Stanfill C., Waltz D. L. Toward memory-based reasoning. Communications of the ACM, 1986. Vol. 29, №12. P. 1213–1223.

98. Stepashko V. S., Efimenko S. N. Sequential Estimation of the Parameters of Regression Model. *Cybernetics and Systems Analysis*, 2005. Vol. 41. P. 631–634.

99.Leonard M., Sloan J., Lee ., Elsheimer B. An Introduction to Similarity Analysis Using SAS. *Proceedings of the SAS Global Forum 2007* . 2007. 22 p. URL: https://support.sas.com/rnd/app/ets/papers/similarityanalysis.pdf. *(Last accessed: 10.03.2018).*

100. Sankoff D., Kruskal J. Time warps, string edits, and macromolecules: The theory and practice of sequence comparison. Stanford, CA : CSLI Publications Stanford University Cordura, 1999. 408 p.

101. Gibbs B. P. Advanced Kalman Filtering, Least-squares and Modeling : a practical handbook. John Wiley & Sons, Inc., Singapore, 2011. 627 p. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470890042 *(Last accessed: 10.01.2019).*

102. Petersen I. R., Savkin A.V. Robust Kalman filtering for signals and systems with large uncertainties. Boston : Birkhouser USA, 1999. 214 p.

103. Балакришнан А. В. Теория фильтрации Калмана : пер. с англ. Москва : Мир, 1988. 168 с.

104. Axak N., Rosinskiy M. D. MapReduce Hadoop Models for Distributed Neural Network Processing of Big Data Using Cloud Services. *Conference on Computer Science and Information Technologies*. 2019. P. 387–400.

105. Bidyuk P., Prosyankina-Zharova T., Terentiev O., Medvedeva M. Adaptive modelling for forecasting economic and financial risks under uncertainty in terms of the economic crisis and social threats. *Технологічний аудит та резерви виробництва.* 2018. №4. С. 24–36. URL: https://doi.org/ 10.15587/2312-8372.2018.135483.

106. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Москва : Финансы и статистика,1986. 366 с.

107. Бард Й. Нелинейное оценивание параметров . Москва : Статистика, 1979. 349 с.

108. Бідюк П. І., Романенко В. Д., Тимощук О. Л. Аналіз часових рядів : навч. посіб. Київ : НТУУ КПІ, 2013. 600 с.

109. Freund R. J., Littell R. C. SAS System for Regression. 3 Edition. Cary, NC : SAS Institute, Inc., 2000. 256 p.

110. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды : пер. с англ. Москва : Наука, 1976. 736 с.

111. Кендэл М. Временные ряды : пер. с англ. Москва : Финансы и статистика, 1981. 199 с.

112. Меняйленко О. С., Шевчук О. Б. Аналіз сучасних інформаційних експертних систем економічного напряму. *Вісник Луганського національного університету імені Тараса Шевченка. Серія: Педагогічні науки.* Луганськ, 2012. № 21. С. 95–101. URL: http://nbuv.gov.ua/UJRN/vlup_2012_21_15 (дата звернення: 12.01.2021).

113. Stokes M. E., Davis C. S., Koch G. G. Categorical Data Analysis Using the SAS® System. 2 Edition. Cary, NC : SAS Institute Inc., 2000. 647 p.

114. Загальноекономічний і економіко-математичний словник Лопатнікова.
URL: https://lopatnikov.pro/?s=%D0%BD%D0%B5%D0%BB%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F+%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0 *(дата звернення: 17.02.2021).*

115. Кембриджський словник. URL: https://dictionary.cambridge.org/dictionary/english *(дата звернення: 11.01.2020).*

116. Нелінійна система. URL: https://uk.wikipedia.org/wiki/Нелінійна_система#:~:text=Нелінійна%20система%20-%20динамічна%20система%2C%20в,займається%20дослідженням%20нелінійних%20динамічних%20систем.

117. Nonlinear system. URL: https://en.wikipedia.org/wiki/Nonlinear_system#: ~:text=In%20mathematics%20and %20science%2C%20a, are% 20inherently%20nonlinear%20in%20nature.

118. Новицький І. В., Ус С. А. Випадкові процеси : навч. посіб. Дніпропетровськ : Нац. гірничий ун-т, 2011. 125 с.

119. Cheng J., Bell D. A., Liu W. Learning belief networks from data: an information theory based approach. *Proceedings of the sixth international conference on information and knowledge management (CIKM 1997)*. Las Vegas (Nevada),1997. P. 325–331.

120. Mirollo R. E., Strogatz S. H. Amplitude death in an array of limit-cycle oscillators.URL: https://link.springer.com/article/ 10.1007/BF01013676?LI=true *(Last accessed: 01.04.2020)*

121. Енциклопедія математики. URL: https://encyclopediaofmath.org/index.php?title=Chaos *(дата звернення: 15.03.2021)*

122. Boussinesq J. Theorie de I'intumescence Liquid, Appleteonde Solitaire au de Translation, se Propageantdansun Canal Rectangulaire. *Les Comptes Rendus de l'Académie des Sciences*, 1871. Vol. 72. P. 755–759.

123. Jenkins A. Self-oscillation. *Physics Reports*. 2013. 525 (2). P. 167–222. DOI:10.1016/j.physrep.2012.10.007.

124. Box G. E. Non-normality and Tests on Variance. *Biometrika*. 1953. №40. P. 318–335.

125. Dickey D. A., Fuller W. A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association. 1979.* №74 (366). P. 427 – 431.

126. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Прикладна статистика : навч. посіб. Вінниця : Едельвейс і К, 2013. 288 с.

127. Siddiqi N. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley and SAS Business, USA, 2015. 208 p.

128. Thode H. C. Jr. Testing for Normality. New York : Marcel Dekker, Inc., 2002. 495 p.

129. Демиденко Е. З. Линейная и нелинейная регрессия. Москва : Финансы и статистика, 1981. 302 с.

130. Гожий О. П. Підхід до оцінювання невизначеностей в задачах прогнозування. *Электротехнические и компьютерные системы*. 2015. № 19(95). С. 243–247.

131. Згуровський М. З., Панкратова Н. Д. Основи системного аналізу. Київ : Видавнича група BHV, 2007. 544 с.

132. Payzan-LeNestour E., Bossaerts P. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian Learning in Unstable Settings. *PLoS Computational Biology*. 2011. Vol. 7, issue 1. P. 1–14.

133. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society.* 1995. № 158. P. 419–466.

134. Rykov A. C. System analysis methods: multi-criteria and fuzzy optimization, modeling and expert estimates. Moscow : Economics, 1999. 1587 p.

135. Das B. Representing uncertainties using Bayesian Networks / Technical report DSTO-TR-0918, Defence Science and Technology Organization, Electronics and Surveillance Research Laboratory. Australian Department of Defense, 1999. 66 p.

136. Penman S. H. Financial forecasting, risk, and valuation: accounting for the future. *Abacus*. 2010. Vol. 46, issue 2. P. 211–228

137. McNeil A. J., Frey R., Embrechts P. Quantitative Risk Management. – Princeton. New Jersey : Princeton University Press, 2005. 538 p.

138. Козьменко О. В., Кузьменко О. В. Економіко-математичні методи та моделі (економетрика) : навч. посіб. Суми : Університетська книга, 2014. 406 с.

139. Ларичев О. И., Мошкович Е. М. Качественные методы принятия решений. Вербальный анализ решений. Москва : Наука. Физматлит, 1996. 208 с.

140. Shahar Y. Computing with Influence Diagrams and the Path Finder Project : Course of lectures on “Judgment and Decision Making in Information Systems”/ Ben-Gurion University of the Negev, department of Information System Engineering. 2008. Lecture 6. 26 p.

141. Trofymchuk O., Bidyuk P. Probabilistic and statistical uncertainty Decision Support Systems. *Visnyk of Lviv Polytechnic National University*. 2015. No. 826. P. 237–248.

142. Трофимчук О., Бідюк П., Кожухівська О., Кожухівський А. Ймовірнісно-статистичні невизначеності в системах підтримки прийняття рішень. *Вісник Національного університету «Львівська політехніка».* 2015. № 826. С. 237–248.

143. Galante R. Improving the Performance of Data Mining Models with Data Preparation Using SAS Enterprise Miner. *Proceeding SAS Global Forum*, April 26–29, 2015. Dallas, TX, USA, 2015. 21 p. URL: https://support.sas.com/resources/papers/proceedings15/SAS1965-2015.pdf *(Last accessed: 11.08.2020).*

144. Matignon R. Data-mining using SAS Enterprise Miner. 2007. 584 p. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470149019.html

145. Enders C.K. Applied Missing Data Analysis. NY : The Guilford Press, 2010. 401 p.

146. Multiple imputation in SAS / Institute for digital research and education. 2016 URL: http://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1 *(Last accessed: 11.08.2020).*

147. Zhang Zhongheng. Missing data imputation: focusing on single imputation. *Annals of Translational Medicine*. 2016. №4(1), 9. P. 11–18. DOI: 10.3978/j.issn.2305-5839.2015.12.38.

148. Olson D.L., Delen D. Advanced Data Mining Techniques. 2008. 181p.

149. Айвазян С. А. Теория вероятности и прикладная статистика. Москва : Юнити, 2001. 656 с.

150. Андерсен Т. Введение в многомерный статистический анализ : пер. с англ. Москва : Физматгиз, 1963. 500 с.

151. Андерсон Т. Статистический анализ временных рядов : пер. с англ. Москва : Мир, 1976. 755 с.

152. Гмурман В. Е. Теория вероятностей и прикладная статистика. Москва : Высшая школа, 2002. 479 с.

153. Xekalaki E., Degiannakis S. ARCH models for financial applications. Padstow, Cornwall, Greate Britain : TJ International, 2010. 558 p.

154. Seber G. A. F., Wild C. J. Nonlinear Regression. New York : John Wiley & Sons, Inc, 1989. 800 p.

155. Neter J., Kutner M. H., Wasserman W., Nachtsheim C. J. Applied Linear Statistical Models. Fourth Edition. New York : WCB McGraw Hill, 1996. 1424 p.

156. Hurvich C. M., Simonoff J S., Tsai C. L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society.* 1998. P. 271–293.

157. Anderson W. N., Kleindorfer G. B., Kleindorfer P. R., Woodroofe M. B. Consistent estimates of the parameters of a linear system. *The Annals of Mathematical Statistics*. 1969. Vol. 40, №6. P. 2064–2075.

158. Bowerman B. L., O'Connell R. T., Dickey D. A. Linear Statistical Models, an Applied Approach. Boston, MA: Duxbury Press, 1986. 1024 p.

159. Chartrand, R., Yin, W. Iteratively reweighted algorithms for compressive sensing. *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2008. P. 3869-3872.

160. Iman R. Some Aspects of the Rank Transform in Analysis of Variance Problems. *Proceedings of the SAS Users Group International Conference.* Cary, NC: SAS Institute, Inc., 1982. Vol. 7. P. 676–680

161. Peixoto J. L. Hierarchical Variable Selection in Polynomial Regression Models. *The American Statistician.* 1987. Vol. 41, No. 4. P. 311–313.

162. Miller D. M. Reducing Transformation Bias in Curve. Fitting. *The American Statistician.* 1984. Vol. 38, 2. P. 124–126.

163. Olejnik S. F., Algina J. Type I Error Rates and Power Estimates of Selected Parametric and Non-parametric Tests of Scale. *Journal of Educational Statistics*. 1987. Vol. 12. P. 45–61.

164. Searle S. R. Linear Models for Unbalanced Data. New York : John Wiley & Sons, Inc, 1987. 551 p.

165. SAS Institute Inc. Create a gradient boosting model of the data. Getting started with SAS® Enterprise Miner™ 7.1. URL: //support.sas.com/documentation/cdl/en/emgsj/64144/HTML/default/viewer.htm#p03iy98sk0 c9bvn1r6x7ppx8uj08.htm *(Last accessed: 11.08.2020).*

166. Huber P. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*. 1964. Vol. 35(1). P. 73-101

167. Goodall C. M-estimators of location: An outline of the theory. *Understanding Robust and Exploratory Data Analysis*. New York, 1983. P. 339-403.

168. Andrews' Wave. StatWiki internet resource. URL: http://141.20.100.229/mediawiki/statwiki/index.php/Andrews%E2%80%99_Wave . *(Last accessed: 21.02.2020).*

169. Martin A. T. The Data Augmentation Algorithm. *Tools for Statistical Inference* . New York, Inc : Springer-Verlag, 2011. P. 90–136. URL: https://link.springer.com/chapter/10.1007%2F978-1-4612-4024-2_5 *(Last accessed: 13.05.2020).*

170. Бідюк П. І. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Застосування інструментів SAS BASE для дослідженняефективностіметодів обробки пропусків у вибірках даних з метою підвищенняякості прогнозування показників продовольчої безпеки країни. *International conferenceon System Analysis and Information Technology (SAIT 2017*) : Proceedings of the 19-th. Conference (Kyiv, May 22–25 2017). Kyiv : National Technical University of Ukraine "KPI", 2017. С. 253–255.

171. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Методи заповнення пропусків даних у задачах прогнозного моделювання соціально-економічних процесів. *Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMCI 2016) :* матеріали міжнар. конф. (м. Залізний порт, 22–26 травня 2017 р.). Херсон : Вишемирський В. С., 2017. С. 185–187.

172. SAS Institute Inc. SAS/STAT User's Guide. Version 9.1. Cary, NC : SAS Institute, Inc., 2004. 5136 p.

173. Державна служба статистики України : офіц. сайт. URL: http://www.ukrstat.gov.ua/ (дата звернення: 12.01.2020).

174. Resource potential of agrarian sphere: problems and tasks of effective use / Sobkevich. V. et al. Kyiv : NISS, 2013. …p.

175. Kleinbaum D. G., Kupper L. L., Muller K. E. Applied Regression Analysis and Other Multivariable Methods. Boston, MA : PWS-Kent Publishing Company,1988. 718 c.

176. Kotz S., Johnson N. L., Read C. B. Encyclopedia of Statistical Sciences. New York : John Wiley & Sons, 1988. 336 p.

177. Ratkowsky D. Handbook of Nonlinear Regression Models. New York; Basel : Marcel Dekker, 1990. 241 p.

178. Peixoto J. L. A Property of Well-Formulated Polynomial regression Models. *The American Statisticia*. 1990, Vol. 44, No. 1. P. 26–30.

179. Jordan M. I., Ghahramani Z. G., Jaakkola T. S., Saul L. K. An introduction to variational methods for graphical models. *Machine Learning*. Cambridge : MIT Press, 1999. Vol. 37, №2. P. 183–233.

180. Anscombe F. Graphs in Statistical Analysis. *The American Statistician*. 1973. №27. P. 17–21.

181. Miller R. G., Jr. Beyond ANOVA: Basics of Applied Statistics. Boca Raton, FL : Chapman & Hall/CRC. 1997. 336 p.

182. O'Brien R. G. A General ANOVA Method for Robust Tests of Additive Models for Variances. *Journal of the American Statistical Association. 1979*. Vol. 74. P. 877–880.

183. Belsey D. A., Kuh E., Welsch R. E. Regression Diagnostics: Identifying Influential data and sources of collinearity. New York : John Wiley & Sons, *1980*. XV, 292 p.

184. SherginV. L., Chala L. E., Udovenko S. G. Fractal dimension of infinitely growing discrete sets. *14th International Conference on Advanced*

Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). 2018. P. 259–263.

185. Milliken G. A., Johnson D. E. Analysis of Messy Data, Volume 1: Designed Experiments. New York : Chapman & Hall, 1992. 690 p.

186. Agresti A. An Introduction to Categorical Data Analysis. New York: John Wiley & Sons, 1996. 718 p.

187. Chakraborty G., Pagolu M. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. Cary, NC: SAS Institute Inc. 2013. 564 p.

188. Art D., Gnanadesikan R., Kettenring, R. Data-based Metrics for Cluster Analysis. *Utilitas Mathematica*. *1983*. Vol. 21. P. 75–99.

189. Gorodetsky V. I., Drozdgin V. V., Jusupov R. M. Application of attributed grammar and algorithmic sensitivity model for knowledge representation and estimation. *Artificial Intelligence and Information, Control Systems of ROBOTSA.* North Holland : Elsevier Science, 1984. P. 232–237.

190. Терентьєв О. М., Просянкіна-Жарова Т. І., Бідюк П. І., Связинська Н. О., Кириченко В. Е. Text mining analysis of agriculture internet sources using SAS software. *Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту (ISDMCI 2016)* : матеріали міжнар. конф. (м. Залізний порт, 24–28 трав. 2016 р.). Херсон : ХНТУ, 2016. С. 244–246.

191. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України. *Реєстрація, зберігання і обробка даних.* 2016. Т. 18. № 3. С. 75–86.

192. Шолохов О. В., Терентьєв О. М. Просянкіна-Жарова Т. І. Підвищення ефективності соціальних комунікацій на основі аналізу інтернет-джерел засобами textmining. *Прикладні системи та технології в інформаційному суспільстві :* матеріали міжнар. наук.-практ. конф. (Київ, 30

верес. 2020 р.). Київ : Київський нац. ун-т імені Тараса Шевченка, 2020. С. 23–44.

193. Бідюк П. І., Коршевнюк Л. О., Терентьєв О. М. Підтримка розв'язання слабкоструктурованих задач в органах державної влади. *Системный анализ и информационные технологии (САІТ-2012) :* матеріали 14 міжнар. наук.-техн. конф. (Київ, 24 квіт. 2012 р.). Київ : ННК "ІПСА" НТУУ "КПІ". 2012. С. 169–170.

194. Wilbur W. J., Lipman D. J. Rapid similarity searches of nucleic acid and protein data banks. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/6572363/ (Last accessed: 18.07.2020).

195. Клейнберг Дж., Тардос Е. Алгоритмы: разработка и применение. Классика Computers Science. СПб.: Питер, 2016. 800 с.:

196. Andreassen S., Woldbye M., Falck B., Andersen S. K. MUNIN – a causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*. Los Altos, CA : Morgan Kaufmann Publishers, 1987. P. 366–373.

197. Dempster A. P., Laird N. M., Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological).* 1977. Vol. 39, No. 1, P.1-38.

198. Diez F. J., Mira J., Iturralde E., Zubillaga S. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*. 1997. Vol. 10. P. 59–73.

199. Falzon L. Using Bayesian network analysis to support centre of gravity analysis in military planning. *European Journal of Operational Research*. 2006. Vol. 170, № 2. P. 629–643.

200. Friedman N., Linia M., Nachman I. Using Bayesian networks to analyze gene expression data. *Journal of Computational Biology*. 2000. Vol. 7. P. 601–620.

201. Friedman N., Ninio M., Pupko T. A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology*. 2002. Vol. 9. P. 331–353.

202. Fung R., Favero B. Applying Bayesian Networks to information retrieval. *Communications of the ACM (CACM).* 1995. Vol. 38, № 3. P. 24–30.

203. Galan S. F., Aguado F., Diez F. J., Mira J. NasoNet, Joining Bayesian Networks and Time to Model Nasopharyngeal Cancer Spread. *AIME.*, Heidel Berg : Springer-Verlag, 2001. P. 207–216.

204. Gelman A., Carlin J. B., Stern H. S., Rubin D. B. Bayesian Data Analysis. New York : Chapman and Hall, CRC Press, 2000. 670 p.

205. Guo H. Probabilistic inference in Bayes networks / Report on research activities, Kansas State University, department of computing and information sciences, November 2000. 27 p.

206. Guo H. Real time Bayesian networks inference / Report on research activities. Kansas State University, department of computing and information sciences, March November 2001. 42 p.

207. Jensen F.V., Nielsen Th.D. Bayesian Networks and Decision Graphs. New York : Springer, 2007. 457 p.

208. Jilani T., Najamuddin M. Adaptive Bayesian modeling for time series forecasting. *Journal of Applied Environmental and Biological Sciences*. 2014. Vol. 4(7S). P. 99–106.

209. Jouffe L., Munteanu P. New search strategies for learning Bayesian networks. *Proceedings of tenth international symposium on applied stochastic models and data analysis (ASMDA 2001).* Compiegne (France). 2001. Vol. 2. P. 591–596.

210. Kim J. H., Pearl J. CONVINCE: A conversational inference consolidation engine. *IEEE Transactions on systems, Man and Cybernetics.* 1983. Vol. 17(2). P. 120–132.

211. Kozlov A.V., Singh J. P. A parallel Lauritzen-Spiegelhalter algorithm for probabilistic inference. *Proceedings of the Supercomputing'94 conference.* Washington, DC, November, 1994. P. 320–329.

212. Lauritzen S. L., Spiegelhalter D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal Royal Statistics Society, series B – statistical methodology*. 1988. Vol. 50, №2.
P. 157–194.

213. Lo B., Thiemjarus S., Yang G. Adaptive Bayesian networks for video processing. *IEEE International conference on image processing (ICIP*), Barcelona (Spain). 2003. 14–17 September. №1. P. 889–892.

214. Mani S., McDermott S., Valtorta M. MENTOR: a Bayesian Model for prediction of mental retardation in newborns. *Research in Developmental Disabilities*. 1997. Vol. 18, №5. P. 303–318.

215. Neil M., Fenton N. E., Tailor M. Using Bayesian networks to model expected and unexpected operational losses. *Risk Analysis*. 2005. Vol. 25, No. 4. P. 34–57.

216. Onisko A., Druzdzel M. J., Wasyluk H. A probabilistic causal model for diagnosis of liver disorders. *Proceedings of the Seventh Symposium on Intelligent Information Systems (IIS-98).* Malbork, Poland, 1998. P. 379–387.

217. Pearl J. Causality: models, reasoning, and inference. Cambridge University Press, 2000. 400 p.

218. Pourret O., Naim P., Marcot B. Bayesian Networks: A Practical Guide to Applications. Chichester, UK : Wiley, 2008. 448 p.

219. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I. The probabilistic model and inference algorithms / Shwe M. et. al. *Methods of Information in Medicine*. 1991. Vol. 30(4). P. 241–255.

220. Rossi P., Allenby G. M. Bayesian Statistics and Marketing. Hoboken. New Jersey : Wiley & Sons, Ltd., 2005. 350 p.

221. Santos E. J., Shimony S. E., Solomon E., Williams E. On a distributed anytime architecture for probabilistic reasoning : technical report TR-95-02, department of electrical and computer engineering / Air Force Institute of Technology. Ohio, USA : Wright-Patterson AFB, 1995. 18 p.

222. Shenoy P., Shenoy P. Bayesian Network models of portfolio risk and return. *Proceedings of the sixth annual international conference Computational Finance 1999* / New York University's Stern School of Business. Cambridge : MIT Press, 2000. P. 87–106.

223. Sierra B, Larranaga P. Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches. *Artificial Intelligence in Medicine.* 1998. Vol. 14. P. 215–230.

224. Spirtes P., Glymour C., Scheines R. Causation, prediction and search. *Adaptive computation and machine learning*. MIT press, 2001. 565 p.

225. Learning Bayesian networks from data: an information-theory based approach / Cheng J. et al. *The artificial intelligence journal (AIJ).* 2002. Vol. 137. P. 43–90. Littell R. C., Stroup W. W., Freund R. J. SAS® System for Linear Models. Cary, NC : SAS Institute Inc, 2002. 487 p.

226. Golmard J. L., Mallet A. Learning probabilities in causal trees from incomplete databases. Revued' Intelligence Articielle. 1991. Vol. 5. P.93–106.

227. Rebane G., Pearl J. The recovery of causal poly-trees from statistical data. // *International journal of approximate reasoning*. 1988. Vol. 2, № 3. P. 175–182.

228. Гасанова Л. Т., Терентьев О. М., Мельник І. В. Застосування мережі Байеса в медицині. *Вісник Університету "Україна"*. Київ : Університет "Україна", 2008. № 6. С. 93–96.

229. Бидюк П. И., Терентьев А. Н. Метод вероятностного вывода в байесовских сетях по обучающим данным. *Кибернетика и системный анализ*. 2007. № 3. С. 93–99.

230. Згуровський М. З., Бідюк П. І., Терентьєв О. М. Системна методика побудови байєсових мереж. *Наукові вісті НТУУ "КПІ"*. 2007. №4. С. 47–61.

231. Згуровський М. З., Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Байєсівські мережі в системах підтримки прийняття рішень : навч. посіб. Київ : Едельвейс, 2015. 300 с.

232. Бідюк П. І., Терентьєв О. М., Коновалюк М. М. Байєсівські мережі в технологіях інтелектуального аналізу даних. *Наукові праці [Чорноморського державного університету імені Петра Могили комплексу "Києво-Могилянська академія"]. Серія:Комп'ютернітехнології.* 2010. Т. 134. Вип. 121. С. 6–16.

233. Терентьев А. Н., Бидюк П. И., Коршевнюк Л. А. Алгоритм вероятностного вывода в байесовских сетях. *Системні дослідження та інформаційні технології*. 2009. № 2. С. 107–111

234. Bidyuk P. I., Davidenko V. I.,Trofimenko D. V.,Terentyev A. N. Comparative analysis of estimation methods of vertices correlation while Bayesian networks construction. *Journal of Automation and Information Sciences*, 2010. №42(11). P. 36–45. (Scopus)

235. Prosyankina-Zharova T. I., Terentiev O. M., Bidyuk P. I., Gasanov A. Features of SAS Enterprise Guide for probabilistic modeling system, macroeconomic analysis and forecasting. *On control and optimization with industrial applications (COIA-2015*) : Proceedings of the fifth international conference (Baku, 27–29 August 2015), Baku : Institute of applied mathematics BSU, 2015. P. 365–368.

236. Prosyankina-Zharova T., Terentiev O., Bidyuk P., Makukha M. Features of SAS Enterprise Guide for probabilistic modeling system, macroeconomic analysis and forecasting. *Journal of Mathematics and System Science*. 2016. №6. P. 112–122. (Index of Copurnicus)

237. Бідюк П. І., Коршевнюк Л. О., Терентьєв О. М., Просянкіна-Жарова Т. І. Аналіз інвестиційних і соціально-економічних процесів

методами моделювання обмежених множин багатовимірних даних. *Наукові вісті КПІ*. 2012. № 2. С. 87–93.

238. Собкевич О. В., Русан В. М., Юрченко А. Д., Ковальова О. В. Ресурсний потенціал аграрної сфери: проблеми та завдання ефективного використання. Київ : НІСД, 2013. 76 с.

239. Обидєннова Т. С. Використання моделей та методів для формування та прийняття ефективних управлінських рішень. *Вісник Харківського національного технічного університету сільського господарства імені Петра Василенка*. Харків, 2016. Вип. 172. С. 147–153. URL: http://nbuv.gov.ua/UJRN/Vkhdtusg_2016_172_18 (дата звернення: 21.12.2020)

240. Бідюк П. І., Кириченко В. Е., Связінська Н. О. Комп'ютерна програма "IMLBayesNet" : авторське свідоцтво № 64117 України ; заявл. 17.12.2015 № 64562 ; опубл. 16.02.2016.

241. Трофименко Д. В., Давиденко В. І. Комп'ютерна програма "Bayesian Network Master BNetMaster" : авторське свідоцтво № 34443 України ; заявл. 09.06.2010; № 34657 ; опубл. 09.08.2010.

242. Малахов Е. В., Востров Г. Н. Интеллектуальные системы информационной поддержки процедур принятия решений. *Материалы междунар. науч. конф. «Искусственный интеллект – 2000».* Таганрог : ТРТУ, 2000. С. 40–41.

243. Littell R. C., Milliken G. A., Stroup W. W., Wolfinger R. D. SAS® System for Mixed Models, Cary, NC : SAS Institute Inc.,1996. 828 p.

244. Довгий, С. О., Бідюк П. І., Трофимчук О. М. Системи підтримки прийняття рішень на основі статистично-ймовірнісних методів. Київ : Логос, 2014. 418 с. : рис., табл.

245. Modeling and forecasting financial and economic processes with decision support system / Bidyuk P. I. et al. *Наукові вісті КПІ.* 2019. № 5–6. С. 7–17. URL: https://doi.org/10.20535/kpi-sn.2019.5-6.176835

246. Свиридюк Я. В., Дегтярьов А. О. Терентьєв О. М. Застосування нечітко-множинного методу для оцінювання ризику банкрутства корпорації. *Вісник Університету «Україна». Серія: Інформатика, обчислювальна техніка та кібернетика*. 2010. № 8. С. 105–108.

247. Korbicz J., Bidyuk P., Kuznietsova N., Terentiev O., Prosiankina-Zharova T. Multivariate distribution model for financial risks management. *The 9th International Conference "InformationControlSystems& Technologies"(ICST2020)* (Odessa, Sept. 24–26, 2020). CEUR Workshop Proceedings, 2020. P. 416–429. (Scopus)

248. Terentyev A. N., Bidyuk P. I., Mironova A. V., Medin N. Y. Comparison of data mining methods while credit rating of natural persons. *Journal of Automation and Information Sciences*, 2009. № 41(10). P. 71–80. (Scopus)

249. Терентьєв О. М., Бідюк П. І., Кузнєцова Н. В. Система підтримки прийняття рішень для аналізу фінансових даних. *Наукові вісті КПІ*. 2011. №1. С. 48–61.

250. Космічний простір на службі аграрія URL: http://www.agroprofi.com.ua/statti/1115-kosmichnij-prostir-na-sluzhbi-agrarija.html. (дата звернення: 23.01.2018).

251. Моніторинг навколишнього середовища з використанням космічних знімків супутника NOAA : монографія / Довгий С. О. та ін. ; Ін-т телекомунікацій і глобал. інформ. простору НАН України, Нац. аерокосм. ун-т ім. М. Є. Жуковського "ХАІ". Київ : Пономаренко Є. В., 2013. 316 с. : іл.

252. Регрессионные модели оценки урожайности сельскохозяйственных культур по данным MODIS / Н. Н. Куссуль и др. *Современные проблемы дистанционного зондирования Земли из космоса*. 2012. Т. 9, № 1. С. 95–107.

253. Малахов Е. В., Лисенко М. І., Полін Є. Л. Організація систем інформаційного обігу між картографічними системами та банками даних. *Матеріали 35-ї наукової конференції молодих дослідників ОПУ –*

магістрантів «Сучасні інформаційні технології та телекомунікаційні мережі». Одеса : ОДПУ, 2000. С. 33.

254. Трофимчук О. М., Бідюк П. І., Просянкіна-Жарова Т. І., Терентьєв О. М. Адаптивне моделювання і прогнозування у системах підтримки прийняття рішень сільськогосподарського призначення. *Сучасні інформаційні технології управління екологічною безпекою, природокористуванням, заходами в надзвичайних ситуаціях* : колективна монографія / ред. С. О. Довгий. Київ : Юстон, 2017. С. 21–28.

255. Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І., Ефендієв В. В. Прогнозне моделювання нелінійних нестаціонарних процесів у рослинництві з використанням інструментів SAS Enterprise Miner. *Наукові вісті НТУУ КПІ*. 2017. №1. С. 24–36. URL: https://doi.org/10.20535/1810-0546.2017.1.87423.

256. Терентьєв О. М., Просянкіна-Жарова Т. І., Савастьянов В. В. Застосування когнітивного та ймовірнісного моделювання у задачах формування сценаріїв розвитку соціально-економічних систем. *Наукові вісті НТУУ КПІ*. 2016. №5. С. 37–47. URL: https://doi.org/10.20535/1810-0546.2016.5.79876.

257. Research semistructured problems of socio-economic systems: cognitive approach / Gorelova G.V. et al. Rostov na Donu : Publishing House of Rostov. University Press, 2006. 332 p.

258. Terentiev O. M., Prosiankina-Zharova T. I., Lahno V. A., Usatiuk Y. V. The features of the predictive computing modeling power system load in terms of reforming energy market. *Journal of Theoretical and Applied Information Technology*. 2020. Vol. 98. No. 2. P. 163–182. (Scopus Q3)

259. Терентьєв О. М., Связінська Н. О., Просянкіна-Жарова Т. І. Візуалізація типології регіонів України із використанням система SAS Enterprise Guide 7.1. *Геоинформационные системы и компьютерны етехнологии эколого-экономического мониторинга* : матеріали міжнар. наук.-

техн. конф. (Дніпропетровськ, 13–15 квітня 2016 р.). Днепропетровск : ГВУЗ «НГУ» МОН Украины, 2016. С. 21–25.

260. Щука Р. В., Іванов С. С., Терентьєв О. М., Бідюк П. І., Макуха М. П Комп'ютерна програма "SAS Similar Trajectories" № 72358 ; заявл.07.03.2017 ; опубл. 05.05.2017.

261. Akaike H. Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory* / eds: Petrov, Csaki.1973. P. 267–281.

262. Mallows C. L. Some Comments on Cp. *Technometrics.* 1973. Vol. 15. P. 661–675 p.

263. Power D. J. What is a DSS? *The On-Line Executive Journal for Data-Intensive Decision Support*. 1997. Vol. 1., N3. P. 46–54.

264. Системи баз даних та знань. Кн. 2 : Системи управління базами даних та знань : навч. посіб. Львів : Магнолія-2006, 2008. 584 с.

265. Kondratenko Y.P., Kondratenko G. V., Sidenko I. V. (2018) Knowledge-Based Decision Support System with Reconfiguration of Fuzzy Rule Base for Model-Oriented Academic-Industry Interaction. Applied Mathematics and Computational Intelligence. FIM 2015 / eds: Gil-Lafuente A., Merigó J., Dass B., Verma R. *Advances in Intelligent Systems and Computing*. Cham : Springer, 2018. Vol. 730. P. 101–112.

266. Моделі оптимізації багаторівневого зберігання даних / Теленик С. Ф. та ін. *Вісник НТУУ «КПІ». Серія: Інформатика, управління та обчислювальна техніка* : зб. наук. праць. 2015. Вип. 63. С. 48–53.

267. Берко А. Ю., Верес О. М., Пасічник В. В. Системи баз даних та знань. Кн. 1. Львів : Магнолія-2006, 2008. 440 с.

268. Бідюк П. І., Демківський Є. О. Проектування систем підтримки прийняття рішень : навч. посіб. Київ : КНУТД, 2005. 156 с.

269. Малахов Е. В. Вопросы организации иерархических информационных хранилищ. *Перспективы.* 1997. № 1. С. 122–123.

270. SAS Energy Forecasting 3.2: User's Guide, Second Edition. Cary, NC : SAS Institute Inc., 2017. 302 p. URL: https://support.sas.com/documentation/onlinedoc/ef/3.2/efug.pdf *(Last accessed: 10.05.2020).*

271. SAS Energy Forecasting: Fact Sheet. 2018. 4 p. URL: sas.com/energy-forecasting *(Last accessed: 16.08.2020).*

272. Домрачев В. Н., Костецкий Р. И. Терентьев А. Н. SAS BASE: Основы программирования. Киев : Эдельвейс, 2014. 304 с.

273. Колбасюк Д.А., Дегтярьов А. О., Терентьєв О. М. Комп'ютерна програма "Alpha" : авторське свідоцтво № 34191 України ; заявл. 21.05.2010; № 34419 ; опубл. 21.07.2010.

274. Стефанішин Д. В., Власюк Ю. С. До питання порівняльного аналізу водноенергетичних характеристик малих і великих гідроелектростанцій України у складі гідровузлів з водосховищами. *Математичне* моделювання в економіці. 2018. №2. URL: https://www.mmejournal.in.ua/index.php/mmejournal/article/view/35/27 (дата звернення: 12.11.2020).

ДОДАТКИ

ДОДАТОК А

АКТИ ВПРОВАДЖЕННЯ

Додаток містить акти впровадження результатів дисертації в:

- у навчальному процесі Інституту прикладного системного аналізу Національного технічного університету «Київський політехнічний інститут імені Ігоря Сікорського»

- державної служби України з лікарських засобів та контролю за наркотиками;

- ТОВ «Картезіан-Європа»;

- Smart Arbitrage Technologies Limited;

УКРАЇНА
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
03056, м. Київ, пр-т Перемоги, 37; тел. (+38 044) 204-94-55 тел./факс (+38 044) 236-10-43
http://www.mmsa.kpi.ua e-mail: ipsa@kpi.ua ЄДРПОУ 02070921

Довідка

про використання результатів дисертаційного дослідження Терентьєва Олександра Миколайовича на тему: "Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності" у навчальному процесі

Окремі теоретико-методологічні та практичні положення дисертаційного дослідження Терентьєва О. М. на тему: "Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності" використані в навчальному процесі Інституту прикладного системного аналізу при викладанні дисциплін "Вступ до інтелектуального аналізу даних", "Інструментальні засоби SAS обробки та аналізу сховищ даних" та "Аналіз часових рядів" на кафедрі математичних методів системного аналізу, спеціальності 122 Комп'ютерні науки освітньо-професійної програми "Системи і методи штучного інтелекту" та спеціалізації "Інтелектуальний аналіз даних в управлінні проектами", а також спеціальності 124 Системний аналіз освітньо-професійної програми "Системний аналіз і управління".

Довідка видана для подання в спеціалізовану вчену раду за місцем захисту дисертації на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.06 - Інформаційні технології.

Завідувач кафедри математичних методів системного аналізу, к.т.н. доц. О.Л. Тимощук

ДЕРЖАВНА СЛУЖБА УКРАЇНИ З ЛІКАРСЬКИХ ЗАСОБІВ ТА КОНТРОЛЮ ЗА НАРКОТИКАМИ
(Держлікслужба)

проспект Перемоги,120-А, м. Київ, 03115, тел/факс: (044) 422-55-77, e-mail: dls@dls.gov.ua, https://www.dls.gov.ua Код ЄДРПОУ 40517815

_____ № _____ На № _____ від _____

Акт впровадження результатів дисертаційної роботи Терентьєва О.М.

Аналіз, моделювання та прогнозування розповсюдження епідеміологічних захворювань, на сьогодення є однією з актуальних задач, що потребує обробку та аналіз великих обсягів даних та побудови спеціалізованих математичних моделей.

В рамках дисертаційного досліджування, за темою "Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності", Терентьєв О.М. запропонував методику обчислення міри подібності між часовими рядами, яка була використана для виявлення країн-аналогів України за характером захворюваності та смертності на вірус Covid-19, з подальшою можливістю каутеризації країн за ступенем схожості та обчислення скорегованих прогнозів щодо подальшого розвитку епідеміологічної ситуації.

Для проведення обчислень було використано відкриті дані з різних відкритих джерел використовувалися дані World Bank (www.worldbank.org) та JHU CSSE (coronavirus.jhu.edu).

Вхідний набір даних постійно оновлюється та включає щоденну кількість виявлених хворих та померлих, починаючи з 22 січня 2020 року для 205 країн, загалом більше 30 тисяч записів станом на березень 2021 року.

Використання відповідних моделей, разом із існуючими експертними підходами, а також наявними в Державній службі України з лікарських засобів та контролю за наркотиками даними, дозволяє прогнозувати необхідну кількість ліків та медичних засобів для уникнення дефіциту та нестачі в аптеках та медичних закладах охорони здоров'я, при розповсюджені епідеміологічних захворювань.

Цей акт не є документом для фінансових розрахунків.

Голова **Роман ІСАЄНКО**

Clinical Data Management and Outsourcing Provider

Акт впровадження результатів
дисертаційної роботи Терентьєва О.М.

Фармацевтична індустрія розробки нових ліків та проведення випробувань в клінічних умовах характеризується наявністю великої кількості інформації, що зберігається у вигляді масивів даних в базах даних, які потребують менеджменту та статистичної обробки на предмет виявлення значимих зав'язків впливу між показниками процесу, що досліджується, боротьби із пропусками даних, та виявлення аномальних значень.

Запропонована Терентьєвим О.М., в рамках дисертаційної роботи "Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності", методика побудови мережі Байєса з прихованими вершинами дозволяє вирішувати проблему неповноти даних досліджень. В основу обчислювальних процедур покладений структурний EM-алгоритм, який навчає мережі на основі штрафних ймовірнісних значень, які включають значення, отримані за допомогою байєсівського інформаційного критерію або принципу мінімальної довжини.

Це надає можливість зменшити інформаційну невизначеність неповних даних та робити статистично обґрунтовані рішення щодо результатів аналізу медичних досліджень.

Цей акт не є документом для фінансових розрахунків.

Голова ТОВ "Картезіан-Європа", к.ф.-м.н. Ю.М. Карташов

Тел.: 0505401046
03150, м. Київ, вул. Антоновича, 172. ТОЦ «Палладіум Сіті»

Smart Arbitrage TECHNOLOGIES

Smart Arbitrage Technologies Limited
Kemp House, 152-160 City Road, London, EC1V 2NX, GB
www.smartarbitrage.io

Акт впровадження результатів

дисертаційної роботи Терентьєва О.М.

Торгівля ф'ючерсами на криптовалютних торгівельних біржах – складний та неоднозначний процес з точки зору системного аналізу у зв'язку із наявністю великої кількості визначальних факторів. Зміна економічного стану світових країн-лідерів, США, Китаю, країн Європи, суттєво впливають на поведінку інституціональних та приватних інвесторів при формуванні портфелів цінних бумах та цифрових активів. Ф'ючерси на криптовалютні активи (біткоїн, єфір та інші) – це контракт на купівлю або продаж активу за певною ціною у майбутньому.

За допомогою розроблених Терентьєвом О.М. алгоритмів для комп'ютерної програми, в рамках дисертаційного дослідження за темою "Моделі, методи та інформаційні технології прогнозування нелінійних нестаціонарних процесів в умовах невизначеності", що використовують запропоновану методику побудови нелінійних нестаціонарних математичних моделей, на основі постійно оновлюваних даних з бірж (Binance Futures, Bitfinex, Bitstamp, Huobi та інші), прогнозуються напрямки тренду зміни ціни криптовалютних активів, Використання відповідних моделей, разом з існуючими експертними підходами, дає можливість зменшити втрати фінансової установи при роботі з ф'ючерсами, що досягається за рахунок уточнення прогнозів відповідних процесів на криптовалютних біржах.

Цей акт не є документом для фінансових розрахунків.

Директор **/Аптілон М.О./**

SWIFT
TRWIBEB1XXX
IBAN
BE63 9670 6008 6008

ДОДАТОК Б

ОГЛЯД ПРОГРАМНИХ ПАКЕТІВ ДЛЯ РОБОТИ З БАЙЄСОВИМИ МЕРЕЖАМИ

http://www.ai.mit.edu/~ murphyk/Software/BNT/bnsoft.html

Огляд пакетів програмного забезпечення для роботи з мережами Байєса, що був створений Кевіном Мерфі, наведено в таблицях Б.1 та Б.2. Перша таблиця містить назву програми, виробника, та адресу, де програма доступна он-лайн. Якщо програмне забезпечення є комерційним, але компанія має веб-адреси окремих інститутів або центрів дослідження мереж Байєса, це також вказується. Друга таблиця охоплює інформацію про основні можливості такі, як, наприклад, технічна інформація про доступність вихідного коду, платформи, графічний інтерфейс користувача та інтерфейс прикладного програмування; інформацію більш високого рівня функціональності, наприклад, підтримка різного типа вершин (дискретні і/або неперервні), мереж рішень, ненаправлених графів тощо. Також таблиця містить дані про можливості навчання мереж Байєса (параметричне та/або структурне), описує основні алгоритми ймовірнісного виводу (якщо є доступна інформація), показує чи є програмний продукт безкоштовним або комерційним. Повна назва значень кожного стовпця дана в таблиці Б.3.

Список Google програмного забезпечення для роботи з БМ можна знайти за адресою:

http://directory.google.com/Top/Computers/ArtificialIntelligence/Belief Networks/Software/ .

Bayesian Network Repository – сховище БМ – містить приклади мереж Байєса та набори даних для їх навчання:

http:www.cs.huji.ac.uil/labs/combio/Repository

Таблиця Б.1

Програмне забезпечення: назва, доступ в мережі, розробники

Назва	Доступ в мережі	Розробники
Analytica	http://www.lumina.com	Lumina (Henrion)
Bassist	http://www.cs.Helsinki.FI/research/fdk/bassist	U. Helsinki
Bayda	http://www.cs.Helsinki.FI/research/cosco/Projects/NONE/SW	U. Helsinki
BayesBuilder	http://www.mbfys.kun.nl/snn/Research/bayesbuilder/	Nijman (U. Nijmegen)
BayesiaLab	http://www.bayesia.com	Bayesia Ltd
Bayesware Discoverer	http://www.bayesware.com	Bayesware (Open Univ., UK)
B-course	http://b-course.cs.helsinki.fi	U. Helsinki
BN power constructor	http://www.cs.ualberta.ca/˜jcheng/bnpc.htm	Cheng (U.Alberta)
BNetMaster	mail to: vova.davidenko@gmail.com	Trofymenko, Davydenko, Terentev, (NTUU KPI, Kyiv, Ukraine)
BNT	http://www.ai.mit.edu/˜murphyk/Software/BNT/bnt.html	Murphy (prev U.C.Berkeley, now MIT)
BNJ	http://bndev.sourceforge.net/	Hsu (Kansas)
BucketElim	http://www.ics.uci.edu/˜irinar	Rish (U.C.Irvine)
BUGS	http://www.mrc-bsu.cam.ac.uk/bugs	MRC/Imperial College
Business Navigator 5	http://www.data-digest.com	Data Digest Corp
CABeN	http://www-pcd.stanford.edu/cousins/caben-1.1.tar.gz	Cousins et al. (Wash. U.)
CaMML	http://www.datamining.monash.edu.au/software/camml	Wallace, Korb (Monash U.)
CoCo+Xlisp	http://www.math.auc.dk/˜jhb/CoCo/information.html	Badsberg (U. Aalborg)
CIspace	http://www.cs.ubc.ca/labs/lci/CIspace/	Poole et al. (UBC)
Deal	http://www.math.auc.dk/novo/deal	Bottcher et al.
Ergo	http://www.noeticsystems.com	Noetic systems
GDAGsim	http://www.staff.ncl.ac.uk/d.j.wilkinson/software/gdagsim/	Wilkinson (U. Newcastle)
GMRFsim	http://www.math.ntnu.no/˜hrue/GMRFsim/	Rue (U. Trondheim)
GeNIe/SMILE	http://www.sis.pitt.edu/˜genie	U. Pittsburgh (Druzdzel)
GMTk	http://ssli.ee.washington.edu/˜bilmes/gmtk/	Bilmes (UW), Zweig (IBM)
gR	http://www.r-project.org/gR	Lauritzen et al.
Grappa	http://www.stats.bris.ac.uk/˜peter/Grappa/	Green (Bristol)
Hugin	http://www.hugin.com	Hugin Expert (U. Aalborg, Lauritzen/Jensen)
Hydra	http://software.biostat.washington.edu/statsoft/MCMC/Hydra	Warnes (U.Wash.)
Ideal	http://yoda.cis.temple.edu:8080/ideal/	Rockwell (Srinivas)

Назва	Доступ в мережі	Розробники
Java Bayes	http://www.cs.cmu.edu/~javabayes/Home/	Cozman (CMU)
MIM	http://www.hypergraph.dk/	HyperGraph Software
MSBNx	http://research.microsoft.com/adapt/MSBNx/	Microsoft
Netica	http://www.norsys.com	Norsys (Boerlage)
PMT	http://people.bu.edu/vladimir/pmt/index.html	Pavlovic (BU)
PNL	http://www.ai.mit.edu/ murphyk/Software/PNL/pnl.html	Eruhimov (Intel)
Pulcinella	http://iridia.ulb.ac.be/pulcinella/Welcome.html	IRIDIA
RISO	http://sourceforge.net/projects/riso	Dodier (U.Colorado)
TETRAD	http://www.phil.cmu.edu/tetrad/	CMU
UnBBayes	https://sourceforge.net/projects/unbbayes/	?
Vibes	http://www.inference.phy.cam.ac.uk/jmw39/	Winn & Bishop (U. Cambridge)
Web Weaver	http://snowhite.cis.uoguelph.ca/faculty info/yxiang/ww3/	Xiang (U.Regina)
WinMine	http://research.microsoft.com/~dmax/WinMine/tooldoc.htm	Microsoft
XBAIES 2.0	http://www.staff.city.ac.uk/~rgc/webpages/xbpage.html	Cowell (City U.)

Таблиця Б.2

Порівняння функціональних можливостей програм

Назва	Src	API	Exec	GUI	D/C	DN	Param	Struct	D/U	Infer	Free
Analytica	N	Y	WM	Y	G	Y	N	N	D	S	$
Bassist	C++	Y	U	G	N	N	Y	N	D	MH	O
Bayda	J	Y	WUM	Y	G	N	Y	N	D	?	O
BayesBuilder	N	N	W	Y	D	N	N	N	D	?	O
BayesiaLab	N	N	-	Y	Cd	N	Y	Y	CG	JT,G	$
Bayesware	N	N	WUM	Y	Cd	N	Y	Y	D	?	$
B-course	N	N	WUM	Y	Cd	N	Y	Y	D	?	O
BNPC	N	Y	W	Y	D	N	Y	CI	D	?	O
BNetMaster	D	N	W	Y	D	N	Y	Y	D	E(++)	O
BNT	M/C	Y	WUM	N	G	Y	Y	Y	DU	S,E(++)	O
BNJ	J	Y	-	Y	D	N	N	Y	D	JT,IS	O
BucketElim	C++	Y	WU	N	D	N	N	N	D	VE	O
BUGS	N	N	WU	Y	Cs	N	Y	N	Ds	GS	O
BusNav	N	N	W	Y	Cd	N	Y	Y	D	JT	$
CABeN	C	Y	WU	N	D	N	N	N	D	S(++)	O
CaMML	C	N	U	N	Cx	N	Y	Y	D	N	O
CoCo+Xlisp	C/L	Y	U	Y	D	N	Y	CI	U	JT	O
CIspace	J	N	WU	Y	D	N	N	N	D	VE	O
Deal	R	-	-	Y	G	N	M	Y	D	N	O
Ergo	N	Y	WM	Y	D	N	N	N	D	JT(+S)	$
GDAGsim	C	Y	WUM	N	G	N	N	N	D	E	O
GeNIe/SMILE	N	Y	WU	Y	D	Y	N	N	D	JT(+S)	O
GMRFsim	C	Y	WUM	N	G	N	N	N	U	MC	O
GMTk	N	Y	U	N	D	N	Y	Y	D	JT	O
gR	R	-	-	-	-	-	-	-	-	-	O
Grappa	R	Y	-	N	D	N	N	N	D	JT	O
Hugin	N	Y	WU	Y	G	Y	Y	CI	CG	JT	$
Hydra	J	Y	-	Y	Cs	N	Y	N	DU	MC	O
Ideal	L	Y	WUM	Y	D	Y	N	N	N	JT	$
Java Bayes	J	Y	WUM	Y	D	Y	N	N	N	JT,VE	O
MIM	N	N	W	Y	G	N	Y	Y	CG	JT	$
MSBNx	N	Y	W	Y	D	Y	N	N	D	JT	O
Netica	N	Y	WUM	Y	G	Y	Y	N	D	JT	$
PMT	M/C	Y	-	N	D	N	Y	N	D	O	O
PNL	C++	-	-	N	D	N	Y	Y	UD	JT	O
Pulcinella	L	Y	WUM	Y	D	N	N	N	D	?	O
RISO	J	Y	WUM	Y	G	N	N	N	D	PT	O
TETRAD	N	N	WU	N	G	N	Y	CI	UD	N	O
UnBBayes	J	Y	-	Y	D	N	N	Y	D	JT	O
Vibes	J	Y	WU	Y	Cx	N	Y	N	D	V?	O
Web Weaver	J	Y	WUM	Y	D	Y	N	N	D	?	O
WinMine	N	N	W	Y	Cx	N	Y	Y	UD	N	O
XBAIES 2.0	N	N	W	Y	G	Y	Y	Y	CG	JT	O

Таблиця Б.3

Опис функціональних можливостей програмного забезпечення для БМ

Назва	Функціональні можливості	Характеристики
Src	Вихідний код програми	N = немає J = Java; M = Mathlab; D = Delphi; L = Lisp C, C++, R – мови програмування
API	Інтерфейс прикладних програм	N = програма не може бути об‘єднана з вашим кодом, тобто має виконуватись як автономна програма Y = програма може бути об‘єднана
Exec	Операційна система	W = Windows (95/98/200/NT) U = Unix M = MacIntosh - = будь яка з компілятором
GUI	Графічний інтерфейс користувача	N = ні Y = так
D/C	Неперервні та/або дискретні типи вершин	D = лише дискретні вершини G = Гаусівські вершини (аналітично) Cs = неперервні вершини (вибірка) Cd = неперервні вершини (дискретизація) Cx = неперервні вершини (невідомий метод)
DN	Підтримка мереж рішень/діаграм впливу	N = ні Y = так
Param	Параметричне навчання	N = ні Y = так
Struct	Структурне навчання	N = Ні Y = Так Cl = тести на умовну незалежність K2 = алгоритм K2 Купера і Гершковіца
D/U	Типи графів	U = лише неорієнтовані графи D = лише орієнтовані графи UD = орієнтовані та неорієнтовані графи CG = ланцюгові графи (змішані орієнтовані та неорієнтовані)
Infer	Алгоритм ймовірнісного виводу	JT = зв’язне дерево (Junction Tree) VE = виключення змінних PT = Pearl’s poly-tree E = точний MH = Metropols Hastings MC = Markov chain Monte Carlo (MCMC) GS = Gibbs sampling IS = вибірка по значимості S = формування вибірки O = інший ++ = підтримка багатьох методів ? = не визначений N = немає (+S) = точний з формуванням вибірки
Free	Безкоштовна версія програми	O = Так (навіть якщо лише з навч. метою) $ = Ні, комерційна (+обмежені версії)

ДОДАТОК В

КОД КОМП'ЮТЕРНОЇ ПРОГРАМИ «BNETMASTER»

В.1 Процедура побудови структури мережі за начальними даними

```
function heurestic_alg(var x: TStrArray; method:Integer):TDIntArray;
var i2,i,j,j2,k,n_var,m:longint;
    mi_sort:TDRealArray;
    AFullC_count:TIntArray;
    V,List_OM:TDIntArray;
    Node_states,AFullC:TStrArray;
    VFullC,CPT:TDIntArray;
begin
  values_count(x,AFullC, AFullC_count);
  states_matrix(AFullC,Node_states);
  //0: mutual information
  //1: pearson coef
  //2: chuprov coef
  //3: cramer coef
  //4: lamdba coef
  case method of
    0: mi(AFullC,Node_states,AFullC_count, mi_sort);
    4: lambda(AFullC,Node_states,AFullC_count, mi_sort);
  else
    fi_matrix(AFullC,Node_states,AFullC_count,method,mi_sort);
  end;
  mi_sort:=sort_MI(mi_sort);
  n_var:=Length(AFullC[0]);
  //VFullC:=form_VFullC(mi_sort, n_var,0,List_OM);
  V:=var_matrix(n_var);

  for i:=0 to trunc(n_var*(n_var-1)/4) do
 // for i:=0 to length(mi_sort[0])-3 do
  begin
  //формуємо нову матрицю VFullC
  try
    Finalize(VFullC);
  except
  end;
  VFullC:=form_VFullC(mi_sort, n_var,i,List_OM);

  //видаляємо з VFullC циклічні моделі
  VFullC:=delete_cycle2(VFullC,V,n_var);
```

```
          //MDL();
          try
            Finalize(List_OM);
          except
          end;

          List_OM:=MDL(AFullC, Node_states, AFullC_count, VFullC);

          end; //for i:=1 to Length(mi_sort[0])-1 do

          //перетворення в матрицю з 0 і 1. беремо першу модель
          m:=1;//first model
          SetLength(result,n_var,n_var);
          for j:=0 to Length(List_OM[m-1]) - 1 do
            begin
              if List_OM[m-1,j]=-1 then result[V[2,j],V[1,j]]:=1;
              if List_OM[m-1,j]=1 then result[V[1,j],V[2,j]]:=1;
            end;
          end;


        //перетворення в матрицю з 0 і 1
        function matr_from_model(var model_list:TDIntArray;
      m,n_var:Integer):TDIntArray;
        var V:TDIntArray;
           j:Integer;
        begin
          V:=var_matrix(n_var);
          SetLength(result,n_var,n_var);
          for j:=0 to Length(model_list[m-1]) do
            begin
              if model_list[m-1,j]=-1 then result[V[2,j],V[1,j]]:=1;
              if model_list[m-1,j]=1 then result[V[1,j],V[2,j]]:=1;
            end;
        end;

//-------------------------------------------------------------------------------------------------
-----------------------------
    function MDL(AFullC, Node_states:TStrArray; AFullC_count:TIntArray;
VFullC:TDIntArray):TDIntArray;
    var
alpha,i2,i,j,j2,num_par,count,count_parent,count_model,k,r,value,num_child,n_var
,kn2,kn3:longint;
       List_parent,Row_count,List_opt_model:TIntArray;
```

```
        CPT,V:TDIntArray;
        Full_parent:TStrArray;
        MLKG_opt_model:TRealArray;
        model_entropy,model_entropy_mlkg,model_kg,
kg,temp,sum,L,MLKG,L_optimal,denominator, numerator, pen_sum, sum_k3,
sum_k2,sum_mlkg:real;

      begin
      L_optimal:=1.7*power(10,308);
      count_model:=0;
      n_var:=Length(AFullC[0]);
      V:=var_matrix(n_var);

      for i2:=Low(VFullC) to High(VFullC) do
         begin

         //формуємо CPT - модель в матричному виді
         SetLength(CPT,n_var,n_var);
         for j2:=Low(VFullC[i2]) to High(VFullC[i2]) do
           begin
             if VFullC[i2,j2]=-1 then CPT[V[2,j2],V[1,j2]]:=1;
             if VFullC[i2,j2]=1 then CPT[V[1,j2],V[2,j2]]:=1;
           end;

         model_entropy:=0;
         model_entropy_mlkg:=0;
         model_kg:=0;
         pen_sum:=0;

         //Цикл по вершинам моделі
         for j2:=Low(CPT) to High(CPT) do
         begin
          count_parent:=0;
          List_parent:=0;  //формуємо список предків для чергової вершини
          for i:=Low(CPT[j2]) to High(CPT[j2]) do
           begin
            if i<>j2 then
              begin
               If CPT[i,j2]=1 then
                 begin
                   count_parent:=count_parent+1;
                   SetLength(List_parent, count_parent);
                   List_parent[count_parent-1]:=i;
                 end;
              end;
```

```
  end;

// останньою в списку предків записуємо дорічню вершину
count_parent:=count_parent+1;
SetLength(List_parent,count_parent);
List_parent[count_parent-1]:=j2;

num_par:=1;
for i:=Low(List_parent) to High(List_parent) do
 begin
  j:=List_parent[i];
  num_par:=num_par*(High(Node_states[j])+1);
 end;
SetLength(Full_parent,num_par,High(List_parent)+1);

value:=num_par;
      //формуємо Full_parent - матрицю всіх можливих комбінацій
станів вершин в List_parent
for k:=Low(List_parent) to High(List_parent) do
 begin
  i:=List_parent[k];
  value:=floor(value/(High(Node_states[i])+1));
  r:=0;
  for j:=Low(Full_parent) to High(Full_parent) do
   begin
    if (j mod value)=0 then r:=r+1;
    if r>(High(Node_states[i])) then r:=0;
    Full_parent[j,k]:=(Node_states[i,r]);
   end;
 end;

//Підраховуємо кіл-ть співпадінь комбінацій з Full_parent та AFullC
SetLength(Row_count,num_par);
for i:=Low(Full_parent) to High(Full_parent) do
        begin
         value:=0;
         for j:=Low(AFullC) to High(AFullC) do
          begin
          count:=0;
    for k:=Low(List_parent) to High(List_parent) do
           begin
                  r:=List_parent[k];
                  if Full_parent[i,k]=AFullC[j,r] then count:=count+1;
           end;
          if count=Length(List_parent) then value:=value+AFullC_count[j];
```

```
            end;
          Row_count[i]:=value;
        end;

//Обчислення емпіричної ентропії j2-Ї вершини
i:=0;
sum:=0;
sum_mlkg:=0;

num_child:=List_parent[High(List_parent)];
while  i<High(Full_parent) do
       begin
         denominator:=0;
         for j:=0 to High(Node_states[num_child]) do
           begin
             i:=i+1;
             denominator:=denominator+Row_count[i-1];
           end;

  //for mlkg
  sum_k3:=0;
  alpha:=Length(Node_states[num_child]);
  for kn3:=alpha to Floor(denominator+alpha-1) do
    begin
      sum_k3:=sum_k3+logN(exponenta,kn3);
    end;
  pen_sum:=pen_sum+sum_k3;

  //for mlkg+mdl
  i:=i-Length(Node_states[num_child]);
         for j:=0 to High(Node_states[num_child]) do
           begin
     sum_k2:=0; //mlkg
             i:=i+1;
             numerator:=Row_count[i-1];
             if ((numerator<>0) and (denominator<>0)) then
     begin
       sum:=sum-numerator*logN(exponenta,(numerator/denominator));
     end;

     //mlkg
     if (numerator<>0) then
     begin
       for kn2:=1 to Floor(numerator) do
         begin
```

```
                sum_k2:=sum_k2+logN(exponenta, kn2);
              end;
            end;
            sum_mlkg:=sum_mlkg+sum_k2;
                end;
          end;

          model_entropy:=model_entropy+sum;
          model_entropy_mlkg:=model_entropy_mlkg+sum_mlkg;

          //Обчислення кількості умовних ймовірностей для j2-Ї вершини
          kg:=1;
          for i:=Low(List_parent) to (High(List_parent)-1) do
           begin
            j:=List_parent[i];
                kg:=kg*(High(Node_states[j])+1);
           end;
          j:=List_parent[High(List_parent)];
          kg:=kg*(High(Node_states[j]));

          model_kg:=model_kg+kg;
          Finalize(Full_parent);
          Finalize(List_parent);

        end; //кінець циклу по вершинах

        L:=model_entropy+((model_kg/2)*logN(exponenta,
sum_int(AFullC_count)));
        MLKG:=-model_entropy_mlkg+pen_sum;

        if abs(L_optimal-L)<0.0000000001 then
          begin
            count_model:=count_model+1;
            SetLength(List_opt_model,count_model);
            List_opt_model[count_model-1]:=i2;
            SetLength(MLKG_opt_model,count_model);
            MLKG_opt_model[count_model-1]:=MLKG;
          end
        else
          begin
            if L_optimal>L then
              begin
                L_optimal:=L;
                count_model:=1;
                SetLength(List_opt_model,count_model);
```

```
            List_opt_model[count_model-1]:=i2;
            SetLength(MLKG_opt_model,count_model);
            MLKG_opt_model[count_model-1]:=MLKG;
           end;
        end;

       Finalize(CPT);

     end; //кінець циклу по моделях
    i2:=0;
    if Length(MLKG_opt_model)<>1 then
     begin

       for i:=1 to Length(MLKG_opt_model)-1 do
        begin
         if (MLKG_opt_model[i2]>MLKG_opt_model[i]) then
          i2:=i;
        end;
     end;

    SetLength(Result,1,High(VFullC[0])+1);
//    for k:=Low(List_opt_model) to High(List_opt_model) do
//     begin
       for j:=0 to High(VFullC[0]) do
        begin
         Result[0,j]:=VFullC[List_opt_model[i2],j];
        end;
//     end;
     Finalize(List_opt_model);
     Finalize(MLKG_opt_model);
    end;
```

В.2 Модуль псевдовипадкового генерування вибірки

```
unit TSpreadGeneratorUnit;

interface

uses TServiceUnit, Contnrs, TDiscreteBayesianNodeUnit,
TDiscreteBayesianNodesUnit,
     TypesUnit, TBaseStatesUnit, TBaseSatesIteratorUnit, TBayesianNodeUnit,
TStateUnit;

type TSpreadGenerator = class(TService)
protected
   PropagationOrder:TObjectList;
   procedure calculateForNode(aNode:TDiscreteBayesianNode);
   function generatePropagationOrder(aNodes:TDiscreteBayesianNodes):
        TObjectList;
   function GetName():string;override;
public
   constructor CreateWithNodes(aNodes:TDiscreteBayesianNodes);overload;
   destructor Destroy();override;
   function Generate(aNumberOfCorteges:integer):TwoDimStringDynArray;
end;

implementation

//==============================Public=====================
==================
constructor
TSpreadGenerator.CreateWithNodes(aNodes:TDiscreteBayesianNodes);
begin
  inherited CreateWithNodes(aNodes);
end;

destructor TSpreadGenerator.destroy();
begin
  PropagationOrder.Free();
  inherited destroy();
end;

function
TSpreadGenerator.Generate(aNumberOfCorteges:integer):TwoDimStringDynArra
y;
```

```
var
  theData:TwoDimStringDynArray;
  theRandomValue:double;
  theValue:double;
  i, j, k, theIndex:integer;
  theNodes:TDiscreteBayesianNodes;
begin
  SetLength(theData, aNumberOfCorteges, self.Nodes.Count);
  Randomize();
  theNodes := TDiscreteBayesianNodes(self.Nodes);

  if (PropagationOrder <> nil) then
  begin
    PropagationOrder.Free();
  end;
  PropagationOrder :=
       self.generatePropagationOrder(TDiscreteBayesianNodes(self.Nodes));

  for i := 0 to aNumberOfCorteges - 1 do
  begin
    for j := 0 to theNodes.Count - 1 do
    begin
      if (theNodes[j].States.InstancedState <> nil) then
        theNodes[j].States.InstancedState.Instanced := false;
    end;

    for j := 0 to theNodes.Count - 1 do
    begin
      self.calculateForNode(TDiscreteBayesianNode(self.PropagationOrder[j]));
      theRandomValue := Random;
      theValue := 0;
      theIndex := -1;
      while theValue < theRandomValue do
      begin
        theIndex := theIndex + 1;
        theValue := theValue +

TDiscreteBayesianNode(self.PropagationOrder[j]).States[theIndex].Probability;
      end;

theData[i][self.Nodes.IndexOf(TDiscreteBayesianNode(self.PropagationOrder[j]))
] :=
          TDiscreteBayesianNode(self.PropagationOrder[j]).States[theIndex].Name;
```

```
    TDiscreteBayesianNode(self.PropagationOrder[j]).States[theIndex].Instanced
:= true;
   end;
  end;

  Result := theData;
end;

//==============================Private=====================
=================

function TSpreadGenerator.GetName():string;
begin
  Result :='SpreadDataGenerator';
end;

function
TSpreadGenerator.generatePropagationOrder(aNodes:TDiscreteBayesianNodes):T
ObjectList;
var
i,j:integer;
thePropagationOrder:TObjectList;
begin

thePropagationOrder := TObjectList.Create(false);

while thePropagationOrder.Count < aNodes.Count do //ïîêà âñå âåðøèíè íå ïðîéäåíû
  for j:=0 to Nodes.Count-1 do
   if (thePropagationOrder.IndexOf(Nodes[j]) = -1) then //åñëè âåðøèíà åùå íå
ïðîéäåíà, ïûòàåìñÿ åå ïðîéòè
   begin
     i:=0;
     while (i<=aNodes[j].NumberOfParents-1) and
(thePropagationOrder.IndexOf(aNodes[j].Parents[i]) <> -1) do
       i:=i+1;

     if i=aNodes[j].NumberOfParents then   //åñëè âñå ðîäèòåëüñêèå âåðøèíû
ïðîéäåíû òî ìîæíî ïðîéòè äî÷åðíþþ âåðøèíó
     begin
       thePropagationOrder.Add(aNodes[j]);
     end;
   end;

  Result := thePropagationOrder;
end;
```

```
procedure TSpreadGenerator.calculateForNode(aNode:TDiscreteBayesianNode);
var
i,j:integer;
theParentProbability, theProbability:double;
theParentStates:TBaseStates;
theIterator:TBaseSatesIterator;
aFlag:boolean;
theSum:double;
begin
  for i := 0 to aNode.States.Count - 1 do
    aNode.States[i].Probability := 0;

  theIterator := aNode.ProbabilityStructure.GetIterator();
  while (theIterator.Finished = false) do
  begin
    theParentStates := theIterator.GetStates();
    aFlag := true;
    for j := 0 to theParentStates.Count - 1 do
    begin
      if (TState(theParentStates[j]).InstanceState = isNotInstanced) then
      aFlag := false;
    end;

    if aFlag then
    begin
      theParentProbability := 1;
      for j := 0 to theParentStates.Count - 1 do
      begin
        theParentProbability := theParentProbability *
            theParentStates[j].Probability;
      end;

      for j := 0 to aNode.States.Count - 1 do
      begin
        theProbability := theParentProbability *
            aNode.ProbabilityStructure.Probability(theParentStates, aNode.States[j]);
        aNode.States[j].Probability := aNode.States[j].Probability +
            theProbability;
      end;
    end;
      theIterator.Next();
      theParentStates.Free;
  end;
```

```
  theSum := 0;
  for j := 0 to aNode.States.Count - 1 do
  begin
    theSum := theSum + aNode.States[j].Probability;
  end;

  for j := 0 to aNode.States.Count - 1 do
  begin
    aNode.States[j].Probability := aNode.States[j].Probability / theSum;
  end;

  theIterator.Free;
end;

end.
```

ДОДАТОК Г

СВІДОЦТВО ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА КОМП'ЮТЕРНУ ПРОГРАМУ «BNETMASTER»

Додаток містить:

– свідоцтво про реєстрацію авторського права на комп'ютерну програму «BNETMASTER».

УКРАЇНА
Міністерство освіти і науки України
Державний департамент інтелектуальної власності

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 34443

Комп'ютерна програма "Bayesian Network Master" ("BNetMaster")
(вид, назва твору)

Автор(и) **Трофименко Дмитро Вікторович, Давиденко Володимир Іванович, Терентьєв Олександр Миколайович**
(повне ім'я, псевдонім (за наявності))

Дата реєстрації 09.08.2010

Голова Державного департаменту інтелектуальної власності М.В.Паладій

М.П.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ ДЕПАРТАМЕНТ ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ

Україна, МСП 03680, Київ-35, вул. Урицького, 45
Тел.: (044) 494-06-06 Факс: (044) 494-06-67

РІШЕННЯ

ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА ТВІР

Державний департамент інтелектуальної власності розглянув заяву

Трофименко Дмитро Вікторович, пров. Ковальський, 5, кім. 414, м. Київ, 03057
(повне ім'я автора, адреса)

заявка від **09.06.2010** № **34657**

про реєстрацію авторського права на твір і прийняв рішення зареєструвати авторське право на твір **Комп'ютерна програма "Bayesian Network Master"("BNetMaster"); Трофименко Дмитро Вікторович, Давиденко Володимир Іванович, Терентьєв Олександр Миколайович**
(вид, повна, скорочена (за наявності) назва твору, повне ім'я, псевдонім (за наявності) автора (ів))

Внесення відомостей до Державного реєстру свідоцтв про реєстрацію авторського права на твір та видача свідоцтва будуть здійснені за умови сплати збору за оформлення і видачу свідоцтва про реєстрацію авторського права на твір відповідно до п.3 постанови Кабінету Міністрів України від 27 грудня 2001 року № 1756 "Про державну реєстрацію авторського права і договорів, які стосуються права автора на твір".

Якщо протягом трьох місяців від дати одержання заявником рішення про реєстрацію авторського права на твір Державний департамент не одержав документ про сплату збору за оформлення і видачу свідоцтва у розмірі та порядку, визначених законодавством, або копію документа, що підтверджує право на звільнення від сплати зазначеного збору, заявка вважається відхиленою і реєстрація авторського права та публікація відомостей про реєстрацію Державним департаментом не проводиться.

Голова Державного департаменту інтелектуальної власності **М.В.Паладій**

ДОДАТОК Д

КОД КОМП'ЮТЕРНОЇ ПРОГРАМИ «IMLBAYESNET»

Д.1 Процедура побудови структури мережі Байєса за навчальними даними на SAS/IML

```
/* Program for real Asia data*/

varn = {"F1" "F2" "F3" "F4" "F5" "F6" "F7" "F8"};
use sasuser.asia;
read all var varn into X;
close sasuser.asia;     /* Initializing data */

pairs = ncol(X)*(ncol(X)-1)/2;

PXi = j(ncol(X),2);
do i = 1 to ncol(X);
PXi[i,1] = nrow(X[,i]) - ncol(loc(X[,i])); /* i stands for node's number*/
PXi[i,2] = ncol(loc(X[,i]));
end;
PXi = PXi / nrow(X);   /* First responds to Xi=0, second to Xi=1*/

PXiXj = j(pairs,6,0); /* variable1 variable2 00 01 10 11 */
r=0;
do p1 = 1 to ncol(X)-1;
do p2 = p1+1 to ncol(X); /*Our 2 variables cycle*/
r = r+1;
do i = 1 to nrow(X);   /* We make cycle under ALL records*/
t = j(1,2);
t[1] = p1;
t[2] = p2;
PXiXj[r,{1 2}] = t; /* Our var1 var2*/
if X[i,t] = {0 0} then PXiXj[r,3] = PXiXj[r,3] + 1; /*00*/
if X[i,t] = {0 1} then PXiXj[r,4] = PXiXj[r,4] + 1; /*01*/
if X[i,t] = {1 0} then PXiXj[r,5] = PXiXj[r,5] + 1; /*10*/
if X[i,t] = {1 1} then PXiXj[r,6] = PXiXj[r,6] + 1; /*11*/
end;
end;
end;
PXiXj[,{3 4 5 6}] = PXiXj[,{3 4 5 6}] / nrow(X); /* Estimation of common probability by 2 variables */
```

```
        MI = j(pairs,3,0);
        do j = 1 to nrow(MI);
        MI[j,1] = PXiXj[j,1];
        MI[j,2] = PXiXj[j,2];
        do i = 3 to ncol(PXiXj);
        if PXiXj[j,i] ^=0 then
        MI[j,3] = MI[j,3] + PXiXj[j,i] * log( PXiXj[j,i] /
(PXi[PXiXj[j,1],1+mod(int((i-3)/2),2)]*PXi[PXiXj[j,2],1+mod(i-3,2)]) );
        end;
        end;

        call sortndx(ndx,MI,3,3); /* Show order of indexes of sorted MI (needs for
find them in PXiXj)*/
        call sort(MI,3,3);

        G = j(pairs,3,0); /*first node, second node, value: 1 means . -> . ,  2 means .
<- . , 0 means .  .*/
        P = j(ncol(X), ncol(X),0); /* Parents of variables*/
        L = j(ncol(X),1); /* Length of description for each variable*/
        H = j(ncol(X),1); /* Entropy of each variable without parents*/
        K = j(ncol(X),1); /* Weight of each variable*/
        LA2 = j(pairs,1);

        do i = 1 to ncol(X);
        H[i] = -nrow(X)*( PXi[i,1]*log(PXi[i,1]) + PXi[i,2]*log(PXi[i,2]) );
        K[i] = (2##sum(P[i,]))*log(nrow(X))/2;
        L[i] = H[i] + K[i];
        end;

        LALL = sum(L);

        *USED = j(1,ncol(X)); /* List of used variables (maybe this will need for
uncycles) */

         /*Let's investigate each 3 ways to join 2 nodes*/
        do i = 1 to pairs; /* all combinations by 2 nodes*/
        L1TEST = 0; H1TEST = 0; K1TEST = 0;
        L2TEST = 0; H2TEST = 0; K2TEST = 0;
        L3TEST = 0;
        i1 = MI[i,1];
        i2 = MI[i,2];
        P1 = P;
        P2 = P;
        P1[i2,i1] = 1;
        P2[i1,i2] = 1;
```

```
        /* Is there an arrow between current two nodes?*/
        do j = 3 to 6; /* summ for {0,1} states of granny nodes and of child nodes*/
        jP = 1+mod(int((j-3)/2),2);
        ij = 2*mod(j-3,2)+j-int((j-2)/2);
        if PXiXj[ndx[i],j] ^=0  then
        H1TEST = H1TEST - nrow(X)*PXiXj[ndx[i],j] * log( PXiXj[ndx[i],j] /
PXi[i1,jP] ); /* i1 -> i2*/
        if PXiXj[ndx[i],ij] ^=0  then
        H2TEST = H2TEST - nrow(X)*PXiXj[ndx[i],ij] * log( PXiXj[ndx[i],ij] /
PXi[i2,jP] ); /* i1 <- i2*/
        end;
        K1TEST = (2##sum(P1[i2,]))*log(nrow(X))/2;
        K2TEST = (2##sum(P2[i1,]))*log(nrow(X))/2;
        L1TEST = H1TEST + K1TEST + L[i1];
        L2TEST = H2TEST + K2TEST + L[i2];
        L3TEST = L[i1] + L[i2]; /* no arrows*/
        print i1 i2 K1TEST H1TEST L1TEST K2TEST H2TEST L2TEST L3TEST
LALL;
        minL = min(L1TEST, L2TEST, L3TEST);
        if minL = L1TEST then  do; LALL=LALL-L3TEST+L1TEST;
L[i2]=H1TEST + K1TEST; P=P1; G[i,1]=i1; G[i,2]=i2; G[i,3]=1; minL=-1; end;
        if minL = L2TEST then  do; LALL=LALL-L3TEST+L2TEST;
L[i1]=H2TEST + K2TEST; P=P2; G[i,1]=i1; G[i,2]=i2; G[i,3]=2; end;
        LA2[i] = LALL;
        end;

        print
        PXi[colname={"PXi=0", "PXi=1"} format=12.2]
        PXiXj[colname={"Xi", "Xj", "Xi Xj=0 0", "Xi Xj=0 1", "Xi Xj=1 0", "Xi
Xj=1 1"} format=12.4]
        pairs
        MI [colname={"Xi", "Xj", "MI"}]
        ndx H K L
        "Columns = parents of row's variable" P
        LALL G[colname={"Xi", "Xj", "type of an arrow"}];

        /* now graph building*/

        i_i = t(1:pairs);
        declare LinePlot lp;
        lp = LinePlot.Create("Line", i_i, LA2 );
        lp.ShowObs(false);
        lp.SetLineWidth(2);
        lp.SetAxisLabel(XAXIS, "Number of iteration");
        lp.SetAxisLabel(YAXIS, "Minimum Description Length value");
```

Д.2 Лістинг генератора та інструменту аналізу на SAS Base

```
libname prog2 'c:\prog2';

%let numlen=5;
%let namelen=5;
%let upcase='Y';
%let Clients_n=40;
%let Office_n=10;
%let Discount_n=5;

proc format;
	value gender
		1 = 'M'
		2 = 'F';
	value marital
		1 = 'Married'
		2 = 'Not married';
	value mon
		1 = 'January'
		2 = 'February'
		3 = 'March'
		4 = 'April'
		5 = 'May'
		6 = 'June'
		7 = 'July'
		8 = 'August'
		9 = 'September'
		10 = 'October'
		11 = 'November'
		12 = 'December';
	value reg
		1 = 'Kievskaya'
		2 = 'Chernigovskaya';
	value cit
		1 = 'Kiev'
		2 = 'Fastov'
		3 = 'Chernigov';
	value disc
		1 = '5% discount'
		2 = '10% discount'
		3 = '15% discount';
run;
```

```
data rannameadr(keep=ranadr ranname sk);

length ranname $&namelen. rannum $&numlen.;

alpha="abcdefghjklmnpqrstuvwxyz";
num="0123456789";

if &upcase.='Y' then data=upcase(alpha);
else data=alpha;

do iObs= 1 to &Clients_n.;
sk=iObs;
ranname= ' ';

do iLen=1 to &namelen.;
ranname= trim(ranname)||substr(data,ceil(ranuni(0)*(length(alpha))), 1);
end;

rannum= ' ';
do iLen=1 to &numlen.;
rannum= trim(rannum)||substr(num,ceil(ranuni(0)*(length(num))),1);
end;

ranadr=compress(ranname||'-'||rannum);
output;
end;
run;

data prog2.CUSTOMER_DIM;
        format CUSTOMER_SK CUSTOMER_ID 8. BIRTH_DT mmddyy8.
CUSTOMER_NM $40. CUSTOMER_ADR $50.
        CUSTOMER_GENDER $1. MARITAL_STATUS $11.;
        set rannameadr;
                CUSTOMER_ID=100000*ranuni(76);
                BIRTH_DT=290870+10000*ranuni(14);
           gend = round(ranuni(99)+1);
                marr = round(ranuni(11)+1);
                CUSTOMER_GENDER=put(gend,gender.);
                MARITAL_STATUS=put(marr,marital.);
                CUSTOMER_SK=sk;
                CUSTOMER_NM=ranname;
                CUSTOMER_ADR=ranadr;
                drop ranname ranadr gend marr sk;
        output;
run;
```

```
proc print data=prog2.CUSTOMER_DIM noobs;
title 'CUSTOMER_DIM';
run;

data prog2.TIME_DIM;
	format TIME_SK YEAR QUARTER 8. MONTH $10.;
	YEAR=2000;
	do ii=1 to 12;
		TIME_SK=ii;
		MONTH=put(ii,mon.);
		QUARTER=1+int((ii-1)/3);
		output;
	end;
	drop ii;
run;

proc print data=prog2.TIME_DIM noobs;
title 'TIME_DIM';
run;

data prog2.OFFICE_DIM;
	format OFFICE_SK OFFICE_ID 8. REGION CITY $20.;

	do ii=1 to &Office_n.;
		OFFICE_ID=100000*ranuni(0);
		OFFICE_SK=ii;
		cc = round(2*ranuni(0)+1);
		if (cc=1) or (cc=2) then rr=1;
		else rr=2;
		REGION=put(rr,reg.);
		CITY=put(cc,cit.);
		output;
	end;
	drop ii rr cc;

run;

proc print data=prog2.OFFICE_DIM noobs;
title 'OFFICE_DIM';
run;

data prog2.DISCOUNT_DIM;
	format DISCOUNT_SK DISCOUNT_ID 8. DISCOUNT_NM $30.;
	do ii=1 to &Discount_n.;
```

```
                DISCOUNT_ID=100000*ranuni(0);
                DISCOUNT_SK=ii;
                nm = round(2*ranuni(0)+1);
                DISCOUNT_NM=put(nm,disc.);
                output;
        end;
        drop ii nm;
run;

proc print data=prog2.DISCOUNT_DIM noobs;
title 'DISCOUNT_DIM';
run;

proc sql;
create table TEMP as
        select CUSTOMER_SK, TIME_SK
        from prog2.CUSTOMER_DIM, prog2.TIME_DIM
        order by CUSTOMER_SK, TIME_SK;
quit;

data prog2.CUSTOMER_FACT;
        format CUSTOMER_SK TIME_SK OFFICE_SK DISCOUNT_SK
CUSTOMER_CNT PRODUCT_SUM
        PRODUCT_CNT VALUE_AVG REVENUE_SUM 8.;
        set TEMP;
        OFFICE_SK=round(9*ranuni(0)+1);
        DISCOUNT_SK=round(4*ranuni(0)+1);
        CUSTOMER_CNT=1;
        PRODUCT_SUM=10000*ranuni(0);
        PRODUCT_CNT=round(100*ranuni(0));
        VALUE_AVG=PRODUCT_SUM/PRODUCT_CNT;
        REVENUE_SUM=PRODUCT_SUM*(10*ranuni(0)+10)/100;
run;

proc print data=prog2.CUSTOMER_FACT noobs;
title 'CUSTOMER_FACT';
run;
```

Д3 Лістинг процедури налаштування моделі на SAS Base

```
***Note: This code requires that SAS Enterprise Miner be licensed and
installed;
**declare a list of the four stay patterns to be used in macro loops;
%let pattern_list=WKDY WKND EXTD CMBO;
**declare the location where the score code files should be placed;
%let score_path=F:\Trip Purpose Model\Score Code - Preliminary Input
Models;
**********************************************************************
**********;
**** Step 1: Fit Predictive Models to derive factor variables ****;
**********************************************************************
**********;
%macro stay_pattern_loop;
**loop through each stay pattern and fit separate models for each;
%do i=1 %to 4;
%let pattern=%scan(&pattern_list.,&i.);
***build dtree and regression models for different nominal variables;
***start by creating dataset with nominal variables and building dmdb;
***select just the records for the current stay pattern;
***source datasets should be sorted and indexed on stay_pattern for
efficiency;
data mdl_tmp.pattern_data_train;
set mdl_tmp.model_data_train;
where stay_pattern="&pattern.";
**only keep target and inputs to be used for models;
keep target_purpose_business_ind rate_type_cd property_cd;
run;
data mdl_tmp.pattern_data_validate;
set mdl_tmp.model_data_validate;
where stay_pattern="&pattern.";
keep target_purpose_business_ind rate_type_cd property_cd;
run;
** PROC DMREG requires the creation of a data modeling database with
PROC DMDB;
proc dmdb data=mdl_tmp.pattern_data_train
out=mdl_tmp.pattern_data_train_dmdb
dmdbcat=mdl_tmp.dmdb_cat batch;
class target_purpose_business_ind (desc) rate_type_cd property_cd;
target target_purpose_business_ind;
run;
proc dmdb data=mdl_tmp.pattern_data_validate
out=mdl_tmp.pattern_data_validate_dmdb
```

```
dmdbcat=mdl_tmp.dmdb_cat batch;
class target_purpose_business_ind (desc) rate_type_cd property_cd;
target target_purpose_business_ind;
run;
***Model 1 - Property Specific Factor;
***example of decision tree to bin properties using PROC ARBOR;
***declare the fileref for the pattern-specific score code for model;
***PROC ARBOR also creates a “description” file that is useful to keep;
filename scr_prop "&score_path.\Score - Property Specific Factor -
&pattern..sas";
filename dsc "&score_path.\Description - Property Specific Factor -
&pattern..txt";
proc arbor data=mdl_tmp.pattern_data_train
leafsize=5000
criterion=entropy
maxbranch=10
maxdepth=1
missing=bigbranch;
target target_purpose_business_ind /level=nominal;
input property_cd /level=nominal;
assess validata=mdl_tmp.pattern_data_validate;
describe file=dsc;
code file=scr_prop leafid prediction noresidual;
run;quit;
***model 2 - Rate Type factor;
***example of regression for nominal inputs using PROC DMREG;
filename scr_rtyp "&score_path.\Score - Rate Type Factor - &pattern..sas";
proc dmreg data=mdl_tmp.pattern_data_train
dmdbcat=mdl_tmp.dmdb_cat
validata=mdl_tmp.pattern_data_validate
outest=mdl_parm.tpm_rate_type_code_&pattern.
namelen=32
noprint;
class target_purpose_business_ind rate_type_cd;
code file=scr_rtyp group=&pattern._rtyp;
model target_purpose_business_ind = rate_type_cd
/selection=none;
performance cpucount=12 threads;
run;
***delete data;
proc datasets library=mdl_tmp nolist;
delete pattern_data_train pattern_data_validate;
quit;
***score data against each model to transform nominal variables to interval;
***keep relevant variables;
```

```
	***dataset will then be segmented based on factor variables;
	data mdl_tmp.pattern_data_train;
	set mdl_tmp.model_data_train;
	where stay_pattern="&pattern.";
	%include scr_prop;
	property_specific_factor=p_target_purpose_business_ind1;
	%include scr_rtyp;
	rate_type_code_factor=p_target_purpose_business_ind1;
	keep target_purpose_business_ind property_specific_factor
rate_type_code_factor;
	run;
	data mdl_tmp.pattern_data_validate;
	set mdl_tmp.model_data_validate;
	where stay_pattern="&pattern.";
	%include scr_prop;
	property_specific_factor=p_target_purpose_business_ind1;
	%include scr_rtyp;
	rate_type_code_factor=p_target_purpose_business_ind1;
	keep target_purpose_business_ind property_specific_factor
rate_type_code_factor;
	run;
	**Example of using PROC ARBOR to create a decision tree for segmenting
data;
	**Inputs are the various factor variables created;
	**At most 4 segments (2 branches x 2 depth) will be created for each stay
pattern;
	filename scr_seg "&score_path.\Score - Model Segment - &pattern..sas";
	filename dsc "&score_path.\Description - Model Segment - &pattern..txt";
	proc arbor data=mdl_tmp.pattern_data_train
	leafsize=5000
	criterion=probchisq
	maxbranch=2
	maxdepth=2
	missing=bigbranch;
	target target_purpose_business_ind /level=nominal;
	input property_specific_factor rate_type_code_factor /level=interval;
	assess validata=mdl_tmp.pattern_data_validate;
	describe file=dsc;
	code file=scr_seg leafid prediction noresidual;
	run;quit;
	***delete datasets to keep directory clean;
	proc datasets library=mdl_tmp nolist;
	delete pattern_data_train pattern_data_validate
	pattern_data_train_dmdb pattern_data_validate_dmdb;
	quit;
```

```
        %end;
        %mend stay_pattern_loop;
        ***execute first macro;
        %stay_pattern_loop;
        ************************************************************
***********;
        **** Step 2: Create single score code file for each factor variable****;
        ************************************************************
***********;
        **List, in order, the 4-character alias used in fileref for each factor;
        **this should be the same as the fileref used below, such as scr_prop;
        %let file_list=prop rtyp seg;
        **List of the factor variables to be created, in same order as file_list;
        %let var_list=property_specific_factor rate_type_code_factor
model_segment;
        **Declare fileref for final score code;
        filename fin_prop "&score_path.\Score - Property Specific Factor - All
        Patterns.sas";
        filename fin_rtyp "&score_path.\Score - Rate Type Factor - All
Patterns.sas";
        filename fin_seg "&score_path.\Score - Model Segment - All Patterns.sas";
        ***This macro loops through each factor;
        ***First, it writes out a header for the file;
        ***Then is loops through the individual score code for each stay pattern and
        inserts it into the final file, wrapped in “if-then-else” code;
        ***Finally, it writes out a footer for the file with assignment and drop
        statements;
        %macro score_file_loop;
        ***first write headers for each score code file;
        %do v=1 %to 3;
        %let file=%scan(&file_list.,&v.);
        %let var=%scan(&var_list.,&v.);
        data _null_;
        file fin_&file.;
        put "**** Trip Purpose Model ****;";
        put "**** Score Code for &var. ****;";
        put "**** To be pasted into Transform Variable Node in Enterprise Miner
****;";
        run;
        %end;
        ***loop through four stay patterns, writing out score code for each file;
        %do p=1 %to 4;
        %let pattern=%scan(&pattern_list.,&p.);
        filename scr_prop "&score_path.\Score - Property Specific Factor -
&pattern..sas";
```

```
filename scr_rtyp "&score_path.\Score - Rate Type Factor - &pattern..sas";
filename scr_seg "&score_path.\Score - Model Segment - &pattern..sas";
%do v=1 %to 3;
%let file=%scan(&file_list.,&v.);
%let var=%scan(&var_list.,&v.);
data _null_;
file fin_&file. mod;
put "if stay_pattern='&pattern.' then do;";
run;
data _null_;
file fin_&file. mod;
infile scr_&file.;
input;
put _infile_;
run;
data _null_;
file fin_&file. mod;
put "end;";
put "else";
run;
%end;
%end;
***write out closing code for each file;
%do v=1 %to 3;
%let file=%scan(&file_list.,&v.);
%let var=%scan(&var_list.,&v.);
data _null_;
file fin_&file. mod;
put "do;";
put "p_target_purpose_business_ind1=.;";
put "end;";
put;
%if &var.=model_segment %then %do;
put "&var.=_node_;";
%end;
%else %do;
put "&var.=p_target_purpose_business_ind1;";
%end;
put;
put "drop i_target_purpose_business_ind _node_ _leaf_
p_target_purpose_business_ind1 p_target_purpose_business_ind0";
put " q_target_purpose_business_ind1 q_target_purpose_business_ind0
v_target_purpose_business_ind1 v_target_purpose_business_ind0";
put " u_target_purpose_business_ind _warn_;";
run;
```

```
%end;
%mend score_file_loop;
%score_file_loop;
```

ДОДАТОК Е

СВІДОЦТВО ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА КОМП'ЮТЕРНУ ПРОГРАМУ «IMLBAYESNET».

Додаток містить:

– свідоцтво про реєстрацію авторського права на комп'ютерну програму «IMLBayesNet».

УКРАЇНА

ДЕРЖАВНА СЛУЖБА ВЛАСНОСТІ УКРАЇНИ ІНТЕЛЕКТУАЛЬНОЇ

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 64117

Комп'ютерна програма "IMLBayesNet"

(вид, назва твору)

Автор(и) Терентьєв Олександр Миколайович, Бідюк Петро Іванович, Кириченко Вікторія Едуардівна, Связінська Наталія Олександрівна

(повне ім'я, псевдонім (за наявності))

Дата реєстрації 16.02.2016

Голова Державної служби інтелектуальної власності України

А.Г. Жарінова

М.П.

ДЕРЖАВНА СЛУЖБА ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ УКРАЇНИ
Україна, 03680, МСП, м. Київ-35,
вул. Урицького, 45
Тел. (044) 494-06-06
Факс (044) 494-06-67
E-mail: post@sips.gov.ua

STATE INTELLECTUAL PROPERTY SERVICE OF UKRAINE
Ukraine, 03680, MSP, Kyiv-35,
45, Urytskogo str.
Tel. (044) 494-06-06
Fax (044) 494-06-67
E-mail: post@sips.gov.ua

РІШЕННЯ

ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА ТВІР

Державна служба інтелектуальної власності розглянула заяву

Терентьєв Олександр Миколайович, вул. Виборзька, 31/37-а, кв. 73 , м. Київ, 03056
(повне ім'я автора, адреса)

заявка від **17.12.2015** № **64562**

про реєстрацію авторського права на твір і прийняла рішення зареєструвати авторське право на твір **Комп'ютерна програма "IMLBayesNet"; Терентьєв Олександр Миколайович, Бідюк Петро Іванович, Кириченко Вікторія Едуардівна, Связінська Наталія Олександрівна**
(вид, повна, скорочена (за наявності) назва твору, повне ім'я, псевдонім (за наявності) автора (ів))

Внесення відомостей до Державного реєстру свідоцтв про реєстрацію авторського права на твір та видача свідоцтва будуть здійснені за умови сплати збору за оформлення і видачу свідоцтва про реєстрацію авторського права на твір відповідно до п.3 постанови Кабінету Міністрів України від 27 грудня 2001 року № 1756 "Про державну реєстрацію авторського права і договорів, які стосуються права автора на твір".

Якщо протягом трьох місяців від дати одержання заявником рішення про реєстрацію авторського права на твір Державна служба не одержала документ про сплату збору за оформлення і видачу свідоцтва у розмірі та порядку, визначених законодавством, або копію документа, що підтверджує право на звільнення від сплати зазначеного збору, заявка вважається відхиленою і реєстрація авторського права та публікація відомостей про реєстрацію Державною службою не проводиться.

Голова Державної служби інтелектуальної власності **А.Г. Жарінова**

ДОДАТОК Ж
КОД КОМП'ЮТЕРНОЇ ПРОГРАМИ «SAS SIMILAR TRAJECTORIES» ТА НАСТАНОВА ІЗ ВИКОРИСТАННЯ

Ж.1 Мета розробки

Комп'ютерна програма "SASSimilarTrajectories" була розроблена для аналізу, моделювання та прогнозування статистичних даних. Програмний продукт реалізований на платформі SAS Foundation 9.4. В основі роботи програми лежить метод подібних траєкторій.

Метод подібних траєкторій являє собою потужний інструмент для аналізу даних. За його допомогою можна здійснювати прогнозування для числових рядів.

Для тестування програми були використані реальні дані, зі стандартної бібліотеки системи SAS Base, набору даних AIR (рисунок Ж.1), що описує історіюпродажівавіабілетів на основі якої можна порівняти результати роботи нових методів з уже існуючими, та даних МВС України по кількості зареєстрованих правопорушень пов'язаних з наданням публічних послуг у місті Дніпро (рисунок Ж.2).

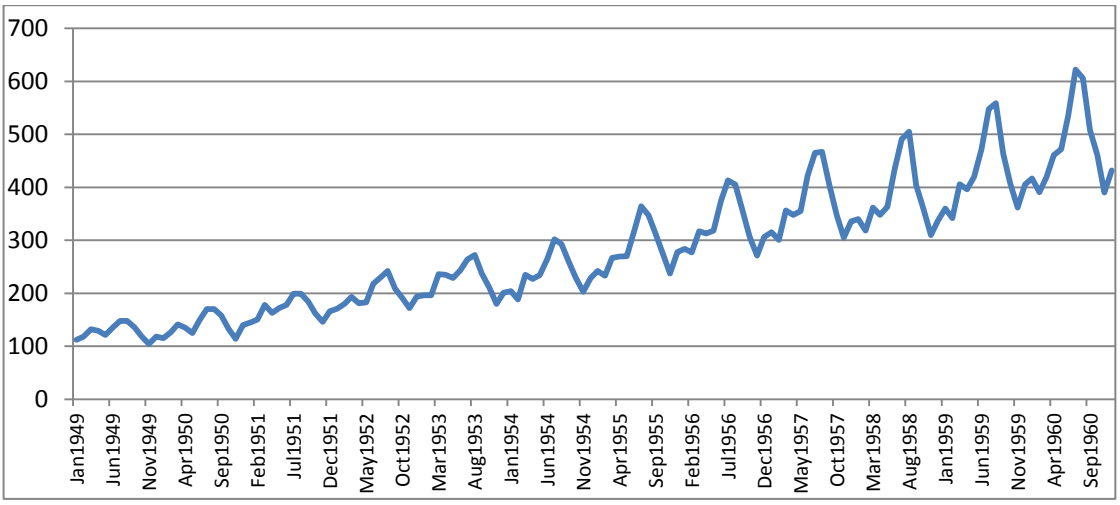


Рисунок Ж.1 – Тестовий графік даних з файлу SASAIR

За допомогою відомих даних можна перевірити правильність і швидкість роботи розробленої програми.

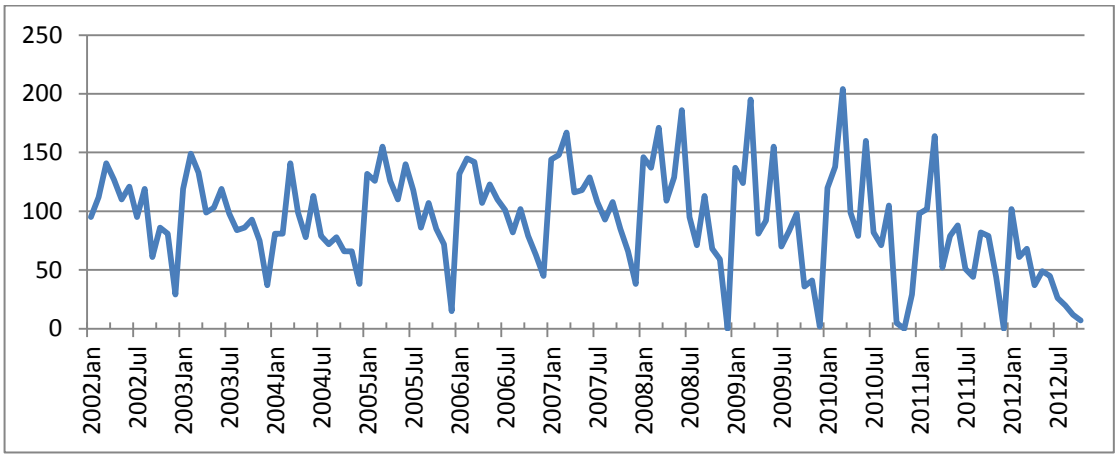


Рисунок Ж.2 – Графік правопорушень, пов’язаних з наданням публічних послуг у місті Дніпро

Ж.2 Мова програмування

Комп’ютерна програма **“SASSimilarTrajectories”** використовує методику прогнозування часових рядівна основіданих, які містять тисячі спостережень. Для реалізації алгоритмубув використаний SASBase 9.4, який являє собою потужну і гнучку мову процедурного програмування в динамічному середовищі, розраховану на програмістів, статистиків, дослідників та аналітиків високого рівня.

Ж.3 Запуск програми SASEnterprise Guide

Для того, щоб запустити програму “SASSimilarTrajectories”необхідно спершу запустити SAS EnterpriseGuide 7.1.

SAS Enterprise Guide– це інтерфейс для написання й виконання коду, а також графічний інтерфейс для створення запитів, звітів та графіків Після запуску SAS Enterprise Guide з’явиться вікно Welcome.

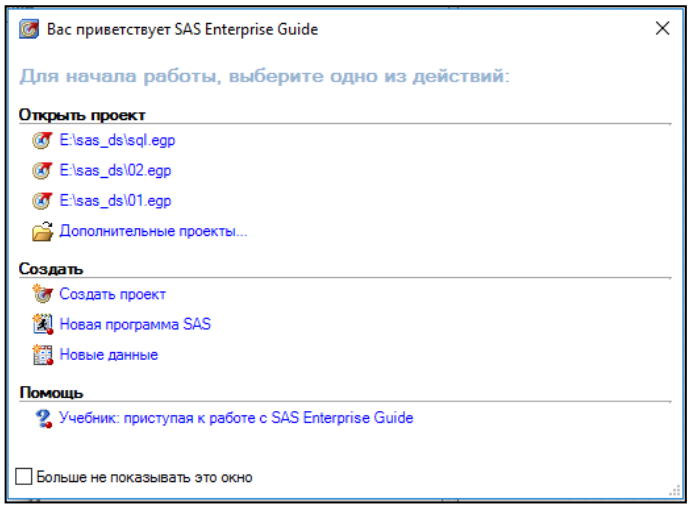


Рисунок Ж.3 – Вигляд вікнаWelcome

Вибираємо потрібний проект серед запропонованих у вкладці «Открыть проект» або натискаємо на «Дополнительные проекты»:

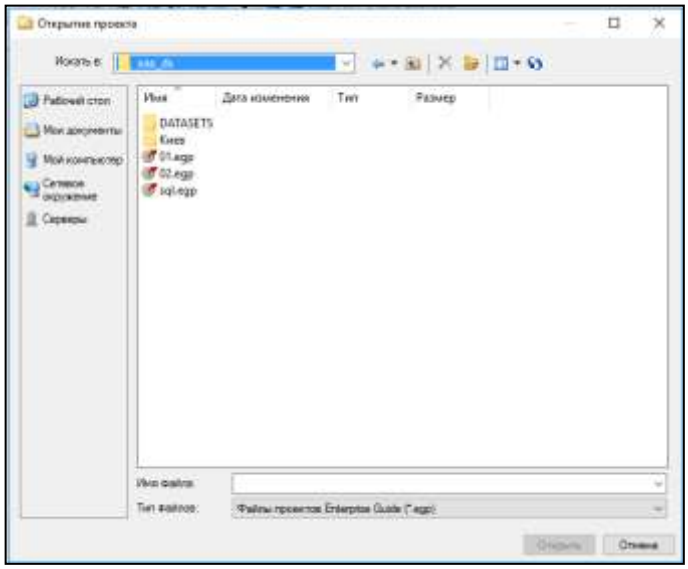


Рисунок Ж.4 – Вигляд вікна «Дополнительные проекты» (Open File)

У вікні Open File вказуємо місце знаходження папки sas_ds, у якій вибираємо файл з назвою 02.egp.

У вікні SAS Enterprise Guide схема проекту під назвою 02.egp.

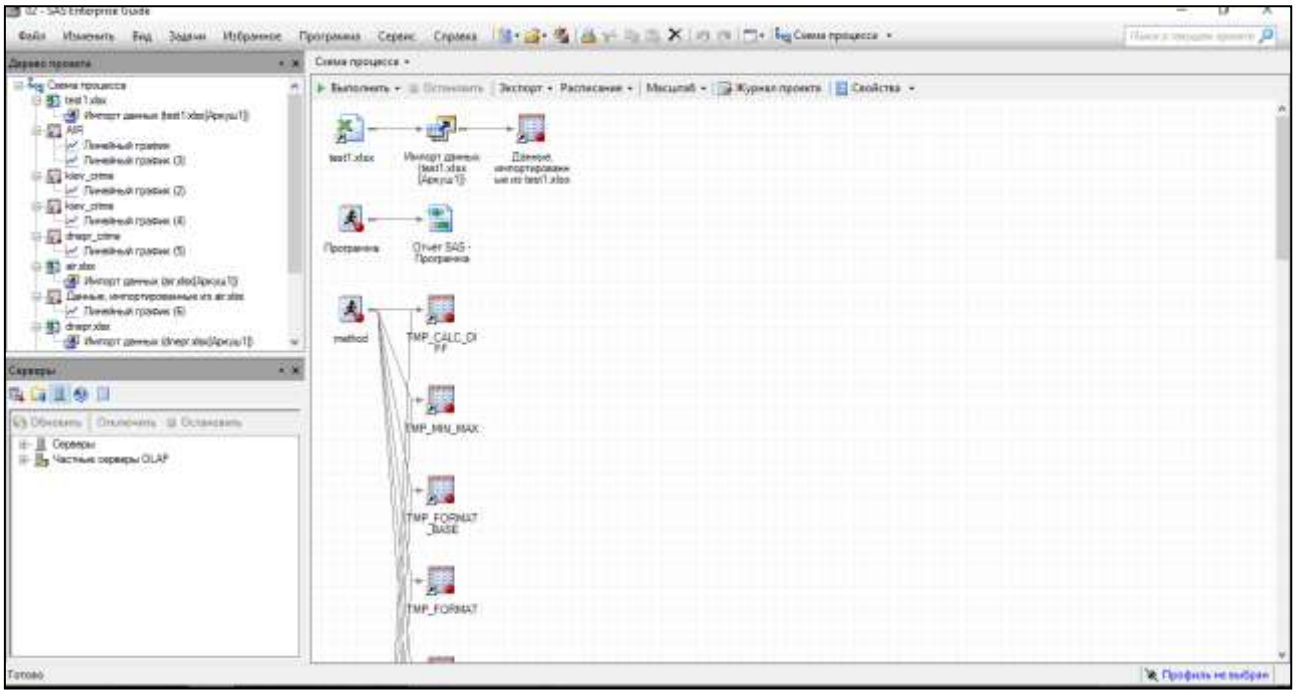
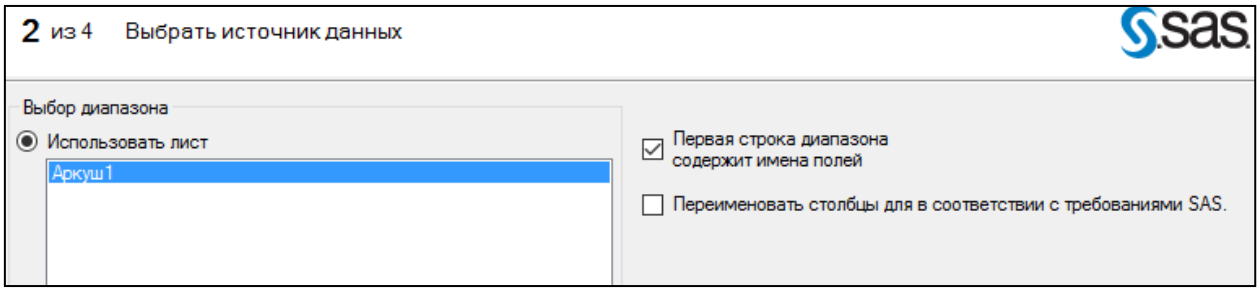


Рисунок Ж.5 – SAS EnterpriseGuide

Ж.4 Завантаження даних

Перед початком роботи програми дані слід представити у спеціальному форматі: пари значень (дата – число).

Після представлення дані імпортуються в SAS:



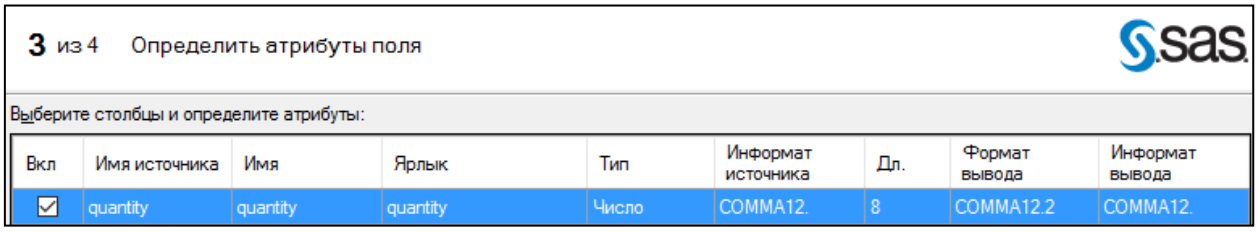


Рисунок Ж.6 – Імпорт даних в SAS, кроки вбудованого візарда 2 та 3.

У результаті імпорту дані представлені у приятному для платформи SAS вигляді. Відтепер можна приступати до їх обробки.

Налаштування параметрів для методу подібних траєкторій відбувається у файлі method.Код налаштування приведено нижче.

```
%letinterv_num = 16;
%letforcast = 2;
%let FIRST_OBS = %eval(&row_num - 4 + 1);
```

Ж.5 Запуск програми “SASSimilarTrajectories”

Для того, щоб запустити програму необхідно відкрити файл method та натиснути на головній панелі «Выполнить» або натиснути на клавіатурі клавішуF5:

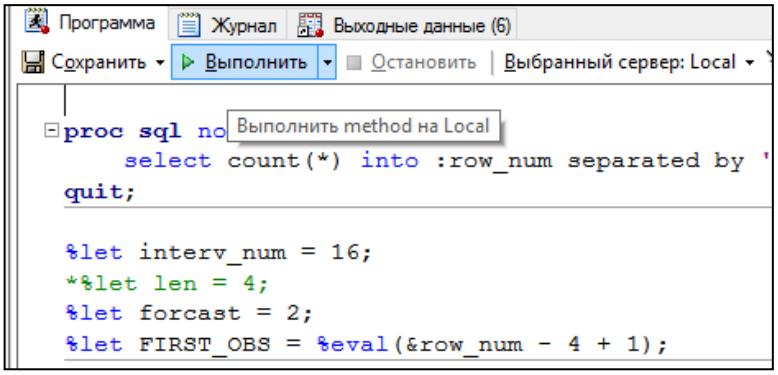


Рисунок Ж.7 – Запуск програми

Далі відбувається підрахунок перших різниць вхідного вектору та запис результатів у нову змінну diff. Суть методу полягає у визначені діапазонів, що відповідають конкретній літері алфавіту $B=\{b_k\}$ з наступним співставленням кожної ділянки з опорним вектором $<e_i>$. Для цього слід знайти максимальне та мінімальне значення серед даних вектора перших різниць. У фрагменті програми, наведеному нижче, відбувається пошук екстремальних значень вектора перших різниць.

```
procmeansdata=tmp_calc_diffminmaxnoprint;
vardiff;
outputout=tmp_min_maxmin=minmax=max;
run;
data_null_;
        settmp_min_max                                                  ;
        callsymput('max',max+1);
        callsymput('min',min-1);
run;
```

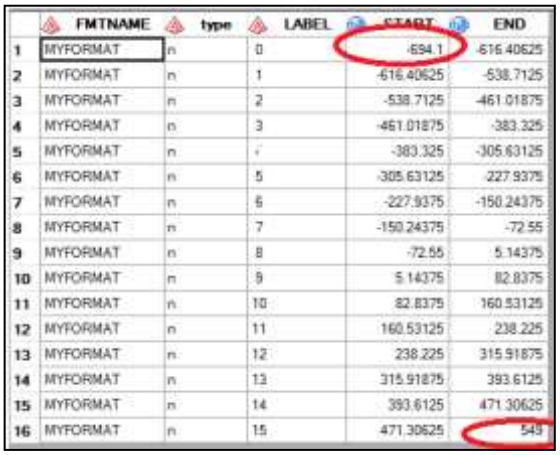
На основі вектора перших різниць, а також величини алфавіту визначаємо ширину інтервалів для кожної букви алфавіту з метою подальшого прогнозу. Для цього в SAS передбачено інструментарій для побудови власного (користувацького) формату даних. Для того, щоб застосувати новостворений формат до вектору перших різниць потрібно виконати наступний код:

```
datatmp_format_base;
        retain FMTNAME 'MYFORMAT'type'n';
        dropvarboundrank;
        var = (&max - &min) / &interv_num;
        bound = &min;
        rank = -1;
        dountil(bound>&max);
                rank + 1;
                LABEL = put(rank, 2.);
                START = bound;
                bound = bound + var;
                END = bound;
                output;
        end;
run;
procformatcntlin=tmp_format_basecntlout=tmp_format;

datatmp_res;
        settmp_calc_diff;
```

```
        formatdiffMYFORMAT10.;
run;
```

У результаті чого отримуємо повністю сформований алфавіт з конкретними значеннями діапазонів, що однозначно задають конкретну літеру алфавіту (рис. Ж.8).



	FMTNAME	type	LABEL	START	END
1	MYFORMAT	n	0	-694.1	-616.40625
2	MYFORMAT	n	1	-616.40625	-538.7125
3	MYFORMAT	n	2	-538.7125	-461.01875
4	MYFORMAT	n	3	-461.01875	-383.325
5	MYFORMAT	n	4	-383.325	-305.63125
6	MYFORMAT	n	5	-305.63125	-227.9375
7	MYFORMAT	n	6	-227.9375	-150.24375
8	MYFORMAT	n	7	-150.24375	-72.55
9	MYFORMAT	n	8	-72.55	5.14375
10	MYFORMAT	n	9	5.14375	82.8375
11	MYFORMAT	n	10	82.8375	160.53125
12	MYFORMAT	n	11	160.53125	238.225
13	MYFORMAT	n	12	238.225	315.91875
14	MYFORMAT	n	13	315.91875	393.6125
15	MYFORMAT	n	14	393.6125	471.30625
16	MYFORMAT	n	15	471.30625	549

Рисунок Ж.8 – Результат прив’язки

Після створення алфавіту відбувається процес побудови прогнозних значень згідно з методом подібних траєкторій. Код операції наведено нижче:

```
data _null_;
        vector_base = substr(symget('diff_str'),3);
        vector_search = symget('vector');
        res = 0;
        dountil(find(vector_base, vector_search, res+1) = 0);
                res = find(vector_base, vector_search, res+1);
                putres=;
        end;
        put'end';
run;
```

Ж.6 Результати роботи програми

В результаті роботи алгоритму були побудовані прогнозні моделі, гафіки яких показані на рисунку 8(для тестових даних AIR) та на рис. Ж 9 (для даних МВС України у місті Дніпро).

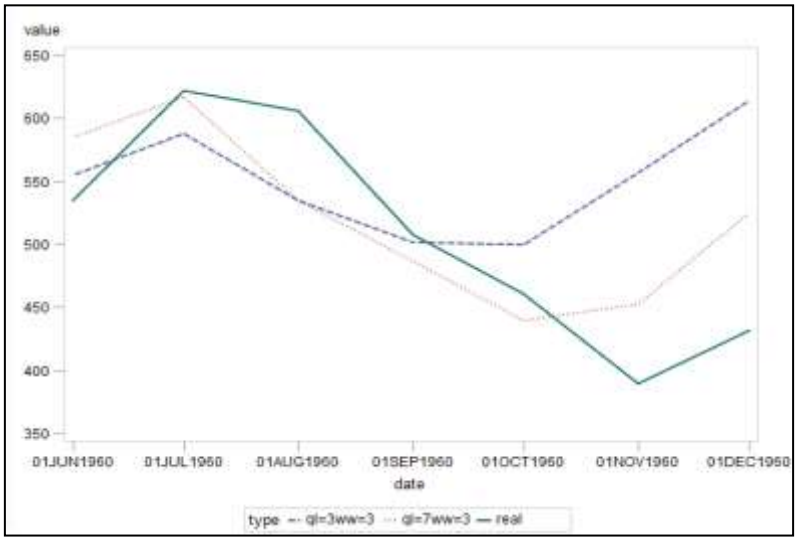


Рисунок Ж.9 – Графік прогнозу

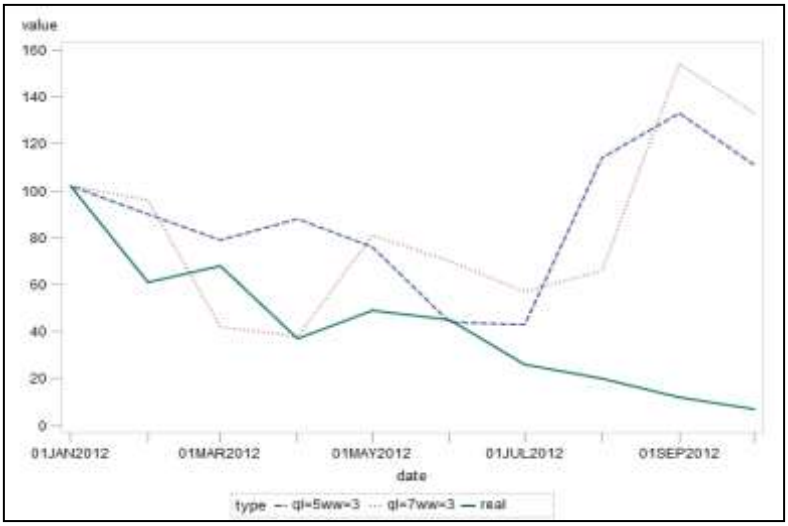


Рисунок Ж.10 – Графік прогнозу

Результати прогнозування наведені у таблиці Ж.1.

Таблиця Ж.1 – статистики прогнозних моделей

	RMSE	MAPE	DW	R^2
Тестовий графік AIR ql=3, ww=3	9881	16,50169	0,808534	0,345968
Тестовий графік AIR ql=3, ww=7	3009	9,748089	0,996792	0,673292
Графік кількості правопорушень у м.Дніпроql=5, ww=3	1935	114,1657	1,03231	0,72647
Графік кількості правопорушень у м.Дніпроql=7, ww=3	960,6833	82,12633	0,960026	0,864263

Ж.7 Вимоги до завантажувальних даних

Статистичні дані в текстовому форматі, наприклад, txt, dat, csv,що розділені стандартними символами тире, пробіл, знак табуляції, крапка, крапка з комою. У завантажувальному файлі повинно бути лише два стовпці даних.

Ж.8 Текст програми

```
proc sql noprint;
        select count(*) into :row_num separated by ' ' from sasuser.input;
quit;

%let interv_num = 16;
*%let len = 4;
%let forcast = 2;
%let FIRST_OBS = %eval(&row_num - 4 + 1);

* Calculating DELTA_A ;

data tmp_calc_diff;
        set sasuser.input;
        retain prev_quantity;
        diff = prev_quantity - quantity;
```

```
        prev_quantity = quantity;
run;

* Determine MIN and MAX ;

proc means data=tmp_calc_diff min max noprint;
var diff;
output out=tmp_min_max min=min max=max;
run;

* Assign MIN and MAX to macro variables ;

data _null_;
        set tmp_min_max ;
        call symput('max',max+1);
        call symput('min',min-1);
run;

* Createing user format to arrange DELTA_A ;

data tmp_format_base;
        retain FMTNAME 'MYFORMAT' type 'n';
        drop var bound rank;
        var = (&max - &min) / &interv_num;
        bound = &min;
        rank = -1;
        do until(bound > &max);
                rank + 1;
                LABEL = put(rank, 2.);
                START = bound;
                bound = bound + var;
                END = bound;
                output;
        end;

run;

proc format cntlin=tmp_format_base cntlout=tmp_format;

* Applying user format to DELTA_A;

data tmp_res;
        set tmp_calc_diff;
        format diff MYFORMAT10.;
run;
```

```
* Select vector with searching value ;

data tmp_vector;
        set tmp_res (firstobs=&FIRST_OBS);
run;

* Creating vector_base and vector_search;

proc sql noprint;
        select diff into :diff_str separated by ' ' from tmp_res;
        select diff into :vector separated by ' ' from tmp_vector;
quit;

* Main algorithm ;
*options linesize=50;
data _null_;
        vector_base = substr(symget('diff_str'),3);
        vector_search = symget('vector');
        res = 0;
        do until(find(vector_base, vector_search, res+1) = 0);
                res = find(vector_base, vector_search, res+1);
                put res=;
        end;
        put 'end';
run;
```

ДОДАТОК И

СВІДОЦТВО ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА КОМП'ЮТЕРНУ ПРОГРАМУ «SASSimilarTrajectories»

Додаток містить:

– свідоцтво про реєстрацію авторського права на комп'ютерну програму «SASSimilarTrajectories».

УКРАЇНА

ДЕРЖАВНА СЛУЖБА ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ УКРАЇНИ

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 71750

Комп'ютерна програма "SASSimilarTrajectories"

(вид, назва твору)

Автор(и) **Щука Роман Володимирович, Іванов Сергій Сергійович, Терентьєв Олександр Миколайович, Бідюк Петро Іванович, Макуха Михайло Павлович**

(повне ім'я, псевдонім (за наявності))

Дата реєстрації 05.05.2017

Державна служба інтелектуальної власності України. Ідентифікаційний код 37552556

Голова Державної служби інтелектуальної власності України

В.о. Голови А.А. Малиш

ДЕРЖАВНА СЛУЖБА ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ УКРАЇНИ
Україна, 03680, МСП, м. Київ-35,
вул. Урицького, 45
Тел. (044) 494-06-06
Факс (044) 494-06-67
E-mail: post@sips.gov.ua

STATE INTELLECTUAL PROPERTY SERVICE OF UKRAINE
Ukraine, 03680, MSP, Kyiv-35,
45, Urytskogo str.
Tel. (044) 494-06-06
Fax (044) 494-06-67
E-mail: post@sips.gov.ua

РІШЕННЯ

ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА ТВІР

Державна служба інтелектуальної власності розглянула заяву

Щука Роман Володимирович, пров. Ковальський, 5, кв. 401, м. Київ, 03057
(повне ім'я автора, адреса)

заявка від **07.03.2017** № **72358**

про реєстрацію авторського права на твір і прийняла рішення зареєструвати авторське право на твір **Комп'ютерна програма "SASSimilarTrajectories"; Щука Роман Володимирович, Іванов Сергій Сергійович, Терентьєв Олександр Миколайович, Бідюк Петро Іванович, Макуха Михайло Павлович**
(вид, повна, скорочена (за наявності) назва твору, повне ім'я, псевдонім (за наявності) автора (ів))

Внесення відомостей до Державного реєстру свідоцтв про реєстрацію авторського права на твір та видача свідоцтва будуть здійснені за умови сплати збору за оформлення і видачу свідоцтва про реєстрацію авторського права на твір відповідно до п.3 постанови Кабінету Міністрів України від 27 грудня 2001 року № 1756 "Про державну реєстрацію авторського права і договорів, які стосуються права автора на твір".

Якщо протягом трьох місяців від дати одержання заявником рішення про реєстрацію авторського права на твір Державна служба не одержала документ про сплату збору за оформлення і видачу свідоцтва у розмірі та порядку, визначених законодавством, або копію документа, що підтверджує право на звільнення від сплати зазначеного збору, заявка вважається відхиленою і реєстрація авторського права та публікація відомостей про реєстрацію Державною службою не проводиться.

В.о. Голови Державної служби інтелектуальної власності **А.А.Малиш**

ДОДАТОК К
КОД КОМП'ЮТЕРНОЇ ПРОГРАМИ «ALPHA» ТА НАСТАНОВА ІЗ ВИКОРИСТАННЯ

Аналіз та побудова моделей.

У випадку, якщо вже було завантажено файл (з даними откліку) або файли з регресорами (залежними від даних откліку) відбувається аналітична обробка даних. Автоматично розраховуються статистичні характеристики процесу, що досліджується, та будується графік. Аналіз даних візуально представляється у вкладці «График данных», як показано нижче:

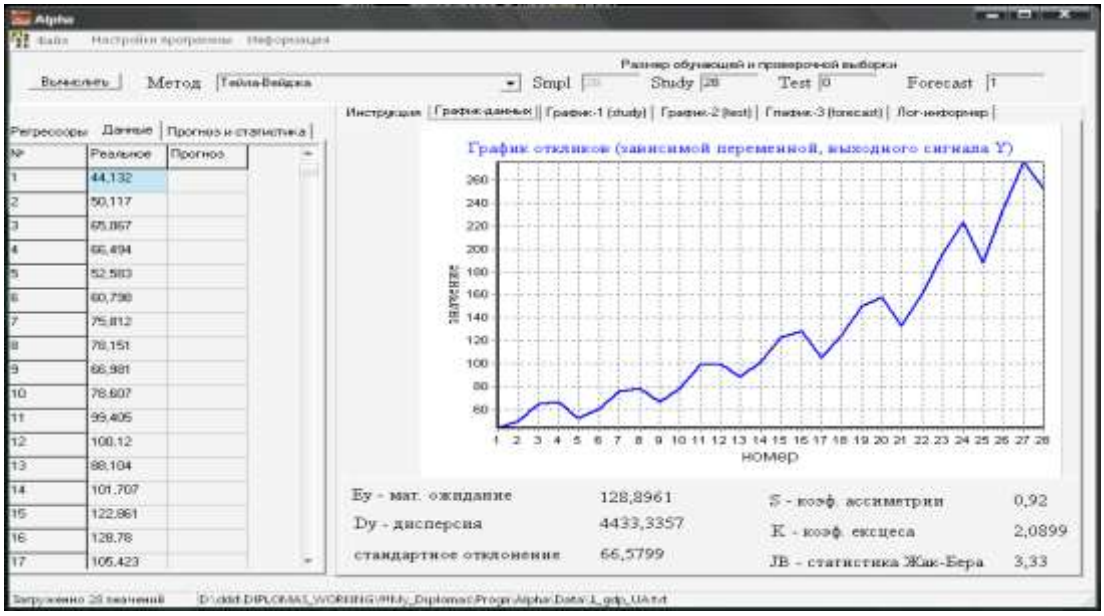


Рисунок К.1 – Графік аналізу даних ряду.

Для побудови моделі процесу виконується розбиття вибірки. Редагуючи відповідні поля, як показано нижче, можна редагувати розмір навчальної вибірки, кількість кроків прогнозування та розмір тестової вибірки.

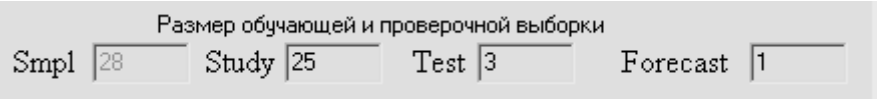


Рисунок К.2 – Окно параметрів моделі

Smpl – розмір всієї навчальної вибірки, що завантажена в програму. Цей параметр не можна змінити.

Study – розмір навчальної вибірки, тобто кількість значень з усієї множини значень, які були завантажені в програму та будуть використані для побудови математичної моделі.

Test – використовується для виявлення прогнозуючих властивостей моделі (дозволяє порівнювати реальні та прогнозні значення).

Необхідно зауважити, що Smpl= Study+ Test.

Forecast – задає число кроків прогнозування (горизонт прогнозування, тобто на яку кількість точок вперед будується прогноз).

Далі необхідно зробити вибір математичного методу для побудови моделі процесу. Для цього необхідно встановити один з восьми можливих методів у кейс-поле, як показано на рис. К.3:

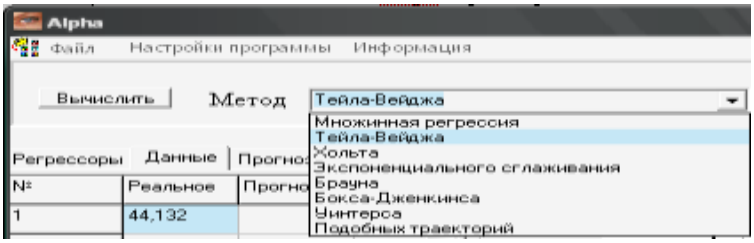


Рисунок К.3 – Окно вибору моделі

За допомогою кнопки Настройки программы для окремих методів встановлюються початкові налаштування, які можна корегувати:

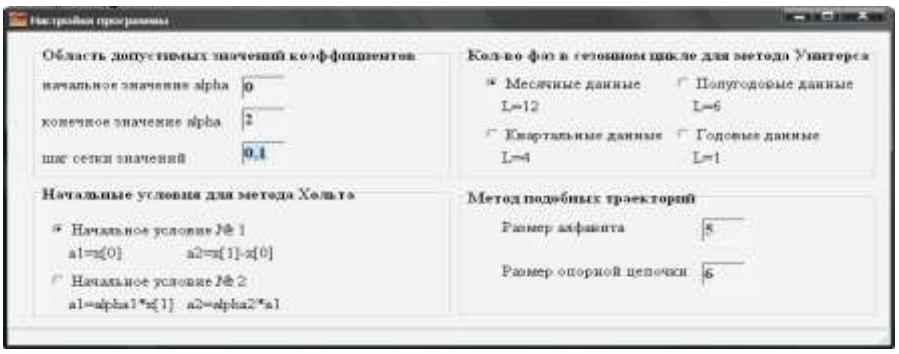


Рисунок К.4 – Окно налаштування моделі

Після чого необхідно натиснути на кнопку Вычислить.

Після розрахунків вибраного методу, відтворені дані по моделі заносяться до таблиці Данные у стовпчик Прогноз, як показано нижче.

Рисунок К.5 Окно результатів

Після цього з’являється вкладка График-1 (study), де відображаються реальні дані з даними, які побудовані по навчальній вибірці отриманої моделі:

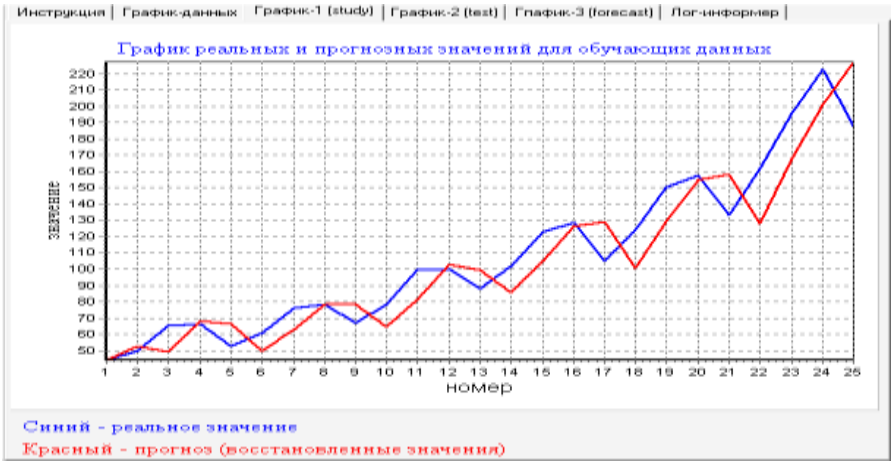


Рисунок К.6 – Графіки результатів моделювання.

Основні результати моделювання і ходу виконання обчислень заносяться до вкладки Лог-информер, як показано нижче:

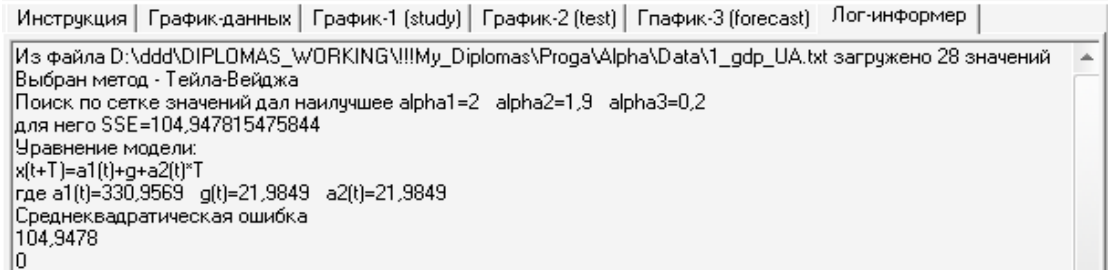


Рисунок К.7 – Лог-вікно системи.

Прогнозування даних.

У випадку коли поле Test 3 - приймає не нульове значення або коли значення поля Forecast 1 - не менше одиниці, тоді відбувається розрахунок прогнозних значень при натисненні кнопки Вычислить.

Виведення результатів прогнозування на задану кількість кроків вперед заноситься до таблиці Прогноз и статистика. Нижче представлені статистичні критерії якості прогнозу:

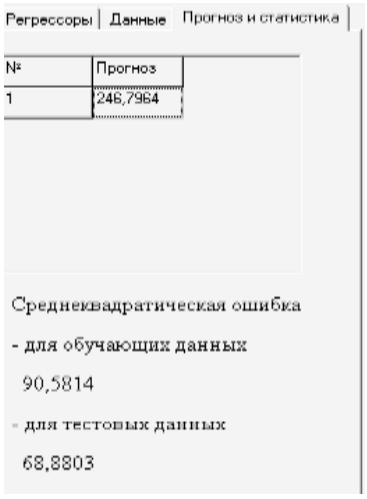


Рисунок К.8 – Вікно прогнозу.

У вкладці Гпафик-3 (forecast) відображаються прогнозні значення на задану кількість кроків вперед.

Якщо параметр Forecast=0, то прогноз не будується і графік також. Для випадку Forecast=1 обчислюється прогноз на один крок вперед, але графік також не будується.

Виведення результатів прогнозування для тестових даних відображається у таблиці Данные у стовпчик Прогноз в кінці ряду.

Після цього з’являється вкладка График-2 (test), де відображаються реальні дані з даними, які побудовані по тестовій вибірці

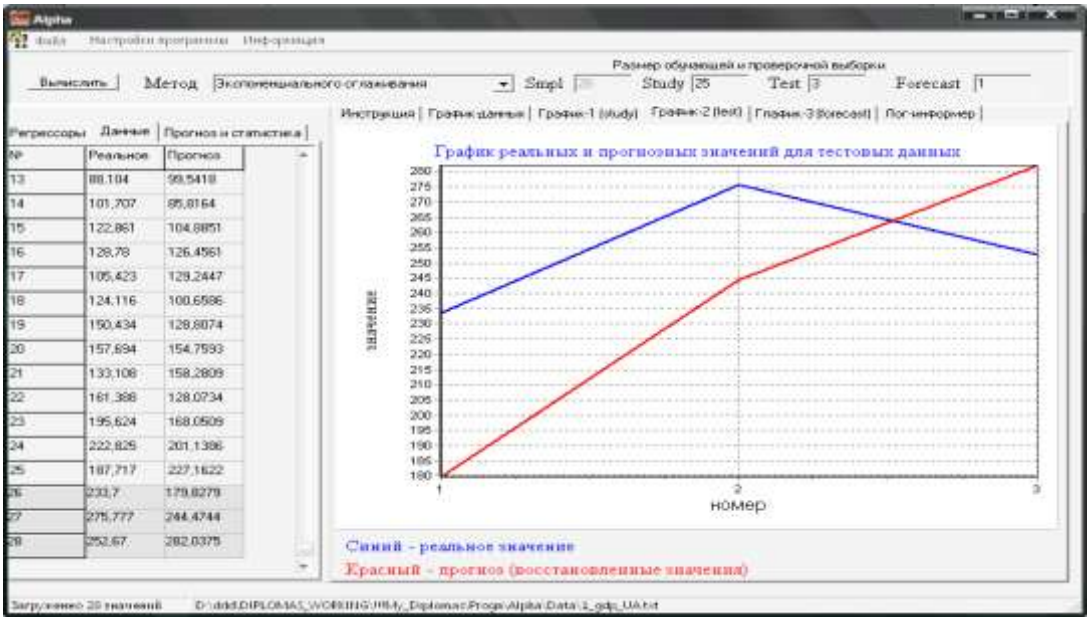


Рисунок К.9 – Результат прогнозу на декілька кроків

Інформація про результати прогнозування також заносить до вкладки Лог-информер, як показано нижче:

Рисунок К.10 – Вікно лог-інформеру системи

Інформація про програму

Для кращого розуміння роботи з розробленою програмою, користувачу пропонується ознайомитись з файл-інструкцією через довідкову. Для виклику інструкції необхідно натиснути на вкладку Информация, як показано нижче:

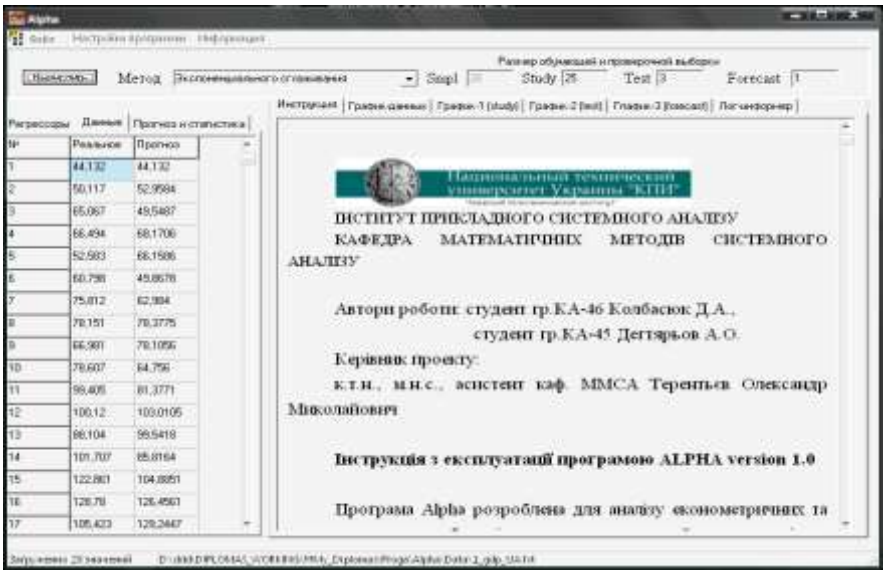


Рисунок К.11 – Справочна система програми

Вимоги до завантажуваних даних.

Прикладом хавантажуваних даних може бути текстовий файл, що має вигляд:

17

34

44

….

68

Рисунок К.12 – Вхідні дані для аналізу.

Будь-який інший варіант представлення даних буде сприйматись програмою як некоректний, з неможливістю його подальшої обробки.

Для завантаження даних необхідно натиснути кнопку 

, як показано нижче.

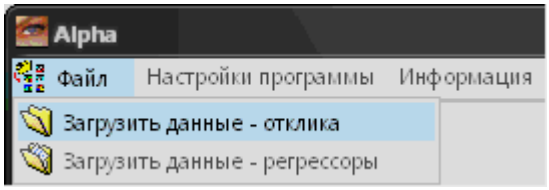


Рисунок К.13 – Вікно завантаження даних в систему

Для завантаження пояснюючих змінних у вигляді регресорів необхідно натиснути кнопку 

, як показано нижче.

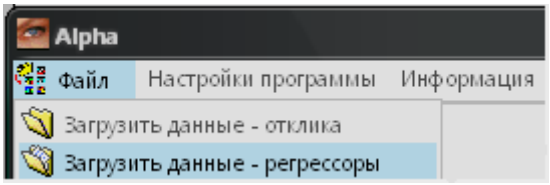


Рисунок К.13 – Вікно завантаження даних-регресорів в систему

У діалоговому вікні необхідно вибрати потрібний файл даних для аналізу:

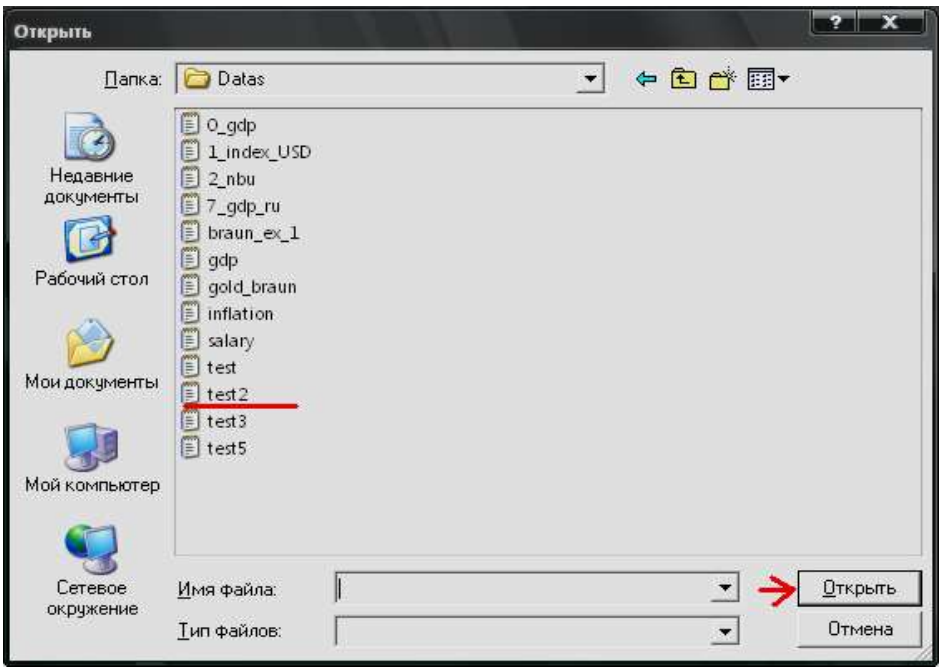


Рисунок К.14 – Вікно вибору файлів

Після того як файл з даними обраний, значення часового ряду відображаються в таблиці в стовпчику Реальное, як показано на рисунку:

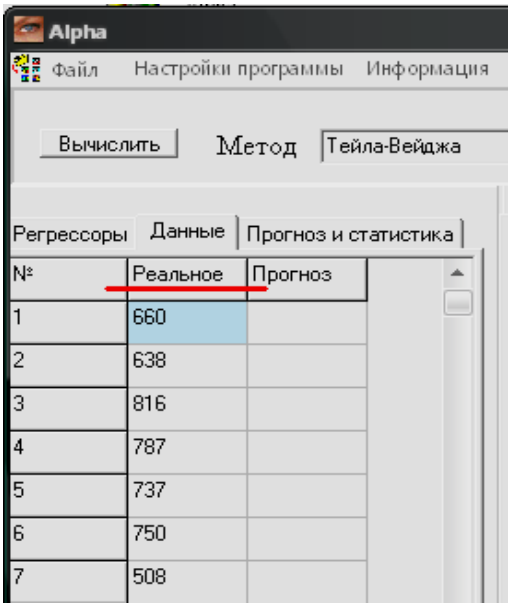


Рисунок К.15 – Вікно відображення даних

Системні вимоги.

- Операційна система Windows 2000/NT/XP.
- Тактова частота процесора не менше 400 Мгц.
- 10 Мбайт вільного місця на жорсткому диску.
- Клавіатура та комп’ютерна миша.

К.1 Частини коду комп’ютерної програми для аналізу, моделювання та прогнозування економетричних даних „Alpha”

Модуль AMath.

Модуль AMath представляє собою набір функцій та процедур побудови моделей і прогнозів часових наборів даних. В ньому реалізовані наступні методи: експоненційного згладжування, Хольта, Тейла-Вейджа, Брауна, Бокса-Дженкінса, Вінтерса, подібних траєкторій і множинної регресії.

Для використання цього модуля необхідне середовище програмування Delphi7 або її нова версія.

Вхідні параметри:

x_Sg – вектор прогнозних значень;

size_study – розмір навчальної вибірки;

size_test – розмір тестової(прогнозної) вибірки;

size_forecast –кількість кроків прогнозування;

alpha_min – мінімальне значення яке може приймати альфа;

alpha_max – максимальне значення яке може приймати альфа;

alpha_step –розмір сітки при підборі коефіцієнта згладжування альфа.

Вихідні параметри:

Res_ExpSmooth - змінна типу запис, яка складається з таких полів:

S : array_of_Real; // вектор прогнозних значень;

SSE_Study : real; // помилка СКП для навчальних даних;

SSE_Test : real; // помилка СКП для тестової вибірки;

Loginfo : TListBox; // вектор строкових повідомлень про процес обчислення.

К.2 Програмний код (деяких алгоритмів)

Функція expsmooth.

// Функція побудови прогнозу на основі метода експоненційного згладжування

// при цьому для підбору коефіцієнта alpha виконується пошук по сітці значень.

```
function expsmooth (x_Sg : array_of_Real; size_study, size_test, size_forecast : integer; alpha_min, alpha_max, alpha_step : real): Res_ExpSmooth;

Var
    x, S : array_of_real;
    i : integer;
    alpha, alpha_best : real;
    SSE, SSE_min : real;
    R : Res_ExpSmooth;
    bs : string;
```

```
Begin
  try
    finalize(R.LogInfo);
  except
  end;

  try
    finalize(R.S);
  except
  end;

  try
    finalize(R.F);
  except
  end;

  R.SSE_Study:=1.7e+36;
  R.SSE_Test:=1.7e+36;
  SetLength(x, size_study);
  SetLength(S, size_study);
  R.LogInfo:=TListBox.Create(Form_Main);
  R.LogInfo.Parent:=Form_Main;
  R.LogInfo.Visible:=False;
  R.S:=TListBox.Create(Form_Main);
  R.S.Parent:=Form_Main;
```

```
R.S.Visible:=False;
R.F:=TListBox.Create(Form_Main);
R.F.Parent:=Form_Main;
R.F.Visible:=False;

for i:=Low(x_Sg) to size_study-1 do
  Begin
    x[i]:=x_Sg[i];
  End;

// + пошук найкращого коефіцієнта згладжування alpha
alpha:=alpha_min;
SSE_min:=1.7e+36;

while alpha <= (alpha_max+0.000001) do
  Begin

    // Обчислення прогнозних значень для навчальної вибірки
    S := expsmooth1(x,alpha);

    // обчислення СКП для навчальних даних
    SSE:=0;
    for i:=low(S) to High(S)-1 do
      Begin
        SSE:=SSE+power((x[i+1]-S[i]),2);
```

```
        End;
      SSE:=power(SSE,(0.5));

      if SSE<SSE_min then
        Begin
          SSE_min:=SSE;
          alpha_best:=alpha;
        End;

        R.LogInfo.Items.Add('alpha='+FloatToStr(alpha)+'
SSE='+FloatToStr(SSE));

      alpha:=alpha+alpha_step;

    End; // while alpha

  alpha:=alpha_best;
  R.SSE_Study:=SSE_min;

  R.LogInfo.Items.Add('Поиск по сетке значений дал наилучшее
alpha='+FloatToStr(alpha));
  R.LogInfo.Items.Add('для него SSE='+FloatToStr(SSE_min));

  R.LogInfo.Items.Add('Уравнение модели:');
  bs:='S(k)=alpha*y(k)+(1-alpha)*S(k-1)';
  R.LogInfo.Items.Add(bs);
```

```
finalize(x);
finalize(S);
SetLength(S, (size_study + size_test));

// Обчислення прогнозних значень для тестової вибірки
S := expsmooth1(x_Sg,alpha);

//обчислення СКП для тестових даних
SSE:=0;
for i:=size_study to size_study+size_test-1 do
  Begin
    SSE:=SSE+power((x_Sg[i]-S[i-1]),2);
  End;
SSE:=power(SSE,(0.5));
R.SSE_Test:=SSE;

for i:=Low(S) to High(S) do
  Begin
    R.S.Items.Add(FloatToStr(S[i]));
  End;

   for i:=1 to size_forecast do
  Begin
    R.F.Items.Add(FloatToStr(S[High(S)]));
```

```
  End;

 finalize(S);

 Result:=R;
End;
```

Функція Holt

```
function Holt (x_Sg : array_of_Real; size_study, size_test, size_forecast : integer; alpha_min, alpha_max, alpha_step : real): Res_ExpSmooth;

 Var

  alpha1, alpha1_best : real;
  alpha2, alpha2_best : real;
  x, S : array_of_real;
  i : integer;
  SSE, SSE_min : real;
  R : Res_ExpSmooth;

Begin

  try
   finalize(R.LogInfo);
  except
```

```
end;

try
  finalize(R.S);
except
end;

try
  finalize(R.F);
except
end;

R.SSE_Study:=1.7e+36;
R.SSE_Test:=1.7e+36;
SetLength(x, size_study);
SetLength(S, size_study);
R.LogInfo:=TListBox.Create(Form_Main);
R.LogInfo.Parent:=Form_Main;
R.LogInfo.Visible:=False;

R.S:=TListBox.Create(Form_Main);
R.S.Parent:=Form_Main;
R.S.Visible:=False;
R.F:=TListBox.Create(Form_Main);
R.F.Parent:=Form_Main;
```

```
R.F.Visible:=False;

for i:=Low(x_Sg) to size_study-1 do
  Begin
    x[i]:=x_Sg[i];
  End;

// + пошук найкращих коефіцієнтів згладжування alpha1 та alpha2
alpha1:=alpha_min;
SSE_min:=1.7e+36;

while alpha1 <= (alpha_max+0.000001) do
  Begin
    alpha2:=alpha_min;

    while alpha2 <= (alpha_max+0.000001) do
      Begin

        // Обчислення прогнозних значень для навчальної вибірки
        S := holt1(x,alpha1,alpha2);

        // Обчислення СКП для навчальних даних
        SSE:=0;
        for i:=low(S) to High(S)-1 do
          Begin
```

```
          SSE:=SSE+power((x[i+1]-S[i]),2);
        End;
      SSE:=power(SSE,(0.5));

      if SSE<SSE_min then
        Begin
          SSE_min:=SSE;
          alpha1_best:=alpha1;
          alpha2_best:=alpha2;
        End;

    R.LogInfo.Items.Add('alpha='+FloatToStr(alpha)+'
SSE='+FloatToStr(SSE));
    R.LogInfo.Items.Add('alpha1='+FloatToStr(alpha1)+'
alpha2='+FloatToStr(alpha2)+'   SSE='+FloatToStr(SSE));

        alpha2:=alpha2+alpha_step;

      End; // while alpha2

    alpha1:=alpha1+alpha_step;

  End;  // while alpha1

  alpha1:=alpha1_best;
  alpha2:=alpha2_best;
```

```
R.SSE_Study:=SSE_min;

R.LogInfo.Items.Add('Поиск по сетке значений дал наилучшее alpha1='+FloatToStr(alpha1)+'  alpha2='+FloatToStr(alpha2));
R.LogInfo.Items.Add('для него SSE='+FloatToStr(SSE_min));

finalize(x);
finalize(S);

SetLength(S, (size_study + size_test));

// Обчислення прогнозних значень для тестової вибірки
S := holt1(x_Sg,alpha1, alpha2);

// Обчислення СКП для тестових даних
SSE:=0;
for i:=size_study to size_study+size_test-1 do
  Begin
    SSE:=SSE+power((x_Sg[i]-S[i-1]),2);
  End;
SSE:=power(SSE,(0.5));
R.SSE_Test:=SSE;

for i:=Low(S) to High(S) do
  Begin
    R.S.Items.Add(FloatToStr(S[i]));
```

```
        End;

      for i:=1 to size_forecast do
        Begin
          R.F.Items.Add(FloatToStr(a1t_holt+i*a2t_holt));
        End;
    R.LogInfo.Items.Add('a1t_holt='+FloatToStr(a1t_holt)+'
'+'a2t_holt='+FloatToStr(a2t_holt));

        R.LogInfo.Items.Add('Уравнение модели:');
        R.LogInfo.Items.Add('x(t+T)=a1(t)+a2(t)*T');
        R.LogInfo.Items.Add('где
a1(t)='+FloatToStr(floor(a1t_holt*10000)/10000)+'
a2(t)='+FloatToStr(floor(a2t_holt*10000)/10000));

        finalize(S);

      Result:=R;
    End;
```

Функція MRegress.

```
    function MRegress (x_Sg : array_of_Real; size_study, size_test : integer):
Res_ExpSmooth;

      Var
        i, j, size : integer;
```

```
    S : array_of_real;
    str : string;
    H,YY,T : Matrix;
    R : Res_ExpSmooth;
   SSE : real;
Begin
    try
      finalize(R.LogInfo);
    except
    end;
    try
      finalize(R.S);
    except
    end;
    try
      finalize(R.F);
    except
    end;

    R.SSE_Study:=1.7e+36;
    R.SSE_Test:=1.7e+36;
    R.LogInfo:=TListBox.Create(Form_Main);
    R.LogInfo.Parent:=Form_Main;
    R.LogInfo.Visible:=False;
    R.S:=TListBox.Create(Form_Main);
```

```
R.S.Parent:=Form_Main;
R.S.Visible:=False;
R.F:=TListBox.Create(Form_Main);
R.F.Parent:=Form_Main;
R.F.Visible:=False;

size:=size_study;
Setsize(H,size,Variable_Count+1);
Zero(H);
Setsize(YY,size,1);
Zero(YY);
// Заповнення матриць H та YY
for i:=0 to size-1 do
 Begin
  H.Data[i,0]:=1;
  YY.Data[i,0]:=x_Sg[i];
 End;
// Заповнення матриці H - додаються стовпчики (регресори моделі)
for i:=0 to size-1 do
for j:=1 to Variable_Count do
 Begin
  H.Data[i,j]:=StrToFloat(MatrixofData[i, j-1]);
 End;

Setsize(T,Variable_Count+1,1);
```

```
// Обчислення оцінок за допомогою МНК
T:=Mul(Mul(Inverse(Mul(Transpose(H),H)),Transpose(H)),YY);

// Обчислення прогнозних значень для навчальної та тестової вибірок
SetLength(S,size_study+size_test);
for i:=0 to size_study+size_test-1 do
 Begin
  S[i]:=T.Data[0,0];
  for j:=1 to Variable_Count do
   S[i]:=S[i]+T.Data[j,0]*StrToFloat(MatrixofData[i, j-1]);
 End;

    // Обчислення СКП для навчальних даних
     SSE:=0;
     for i:=0 to size_study-1 do
       Begin
         SSE:=SSE+power((x_Sg[i]-S[i]),2);
       End;
     SSE:=power(SSE,(0.5));
     R.SSE_Study:=SSE;

    // Обчислення СКП для тестових даних
    SSE:=0;
    for i:=size_study to size_study+size_test-1 do
```

```
                Begin
                  SSE:=SSE+power((x_Sg[i]-S[i]),2);
                End;
                SSE:=power(SSE,(0.5));
                R.SSE_Test:=SSE;

            for i:=Low(S) to High(S) do
              Begin
                R.S.Items.Add(FloatToStr(S[i]));
              End;

            R.LogInfo.Items.Add('Уравнение модели:');
            str:='';
            for i:=1 to Variable_Count do
              str:=str+'a'+IntToStr(i)+'*x'+IntToStr(i)+'(t)+';
            Delete(str,length(str),1);
            R.LogInfo.Items.Add('x(t)=a0+'+str);
            for i:=0 to Variable_Count do
             R.LogInfo.Items.Add('где
a'+IntToStr(i)+'='+FloatToStr(floor(T.Data[i,0]*10000)/10000));
            R.F.Items.Add('0'); // Немає прогнозних значень вперед
            finalize(S);
            finalize(x_Sg);
          Result:=R;

        End;
```

ДОДАТОК Л

СВІДОЦТВО ПРО РЕЄСТРАЦІЮ АВТОРСЬКОГО ПРАВА НА КОМП'ЮТЕРНУ ПРОГРАМУ «ALPHA»

Додаток містить:

– свідоцтво про реєстрацію авторського права на комп'ютерну програму «Alpha».

УКРАЇНА
Міністерство освіти і науки України
Державний департамент інтелектуальної власності

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 34191

Комп'ютерна програма "Alpha"
(вид, назва твору)

Автор(и) **Колбасюк Дмитро Анатолійович, Дегтярьов Андрій Олександрович, Терентьєв Олександр Миколайович**
(повне ім'я, псевдонім (за наявності))

Дата реєстрації 21.07.2010

Голова Державного департаменту інтелектуальної власності М.В.Паладій

М.П.

ДОДАТОК М

КОД КОМП'ЮТЕРНОЇ ПРОГРАМИ «SAS TIME SERIES SIMILARITY»

Опис програми.

За посиланням в Інтернеті:

https://sites.google.com/view/ot-codes-01

можна завантажити відповідні коди програм, наведені в цьому додатку дисертаційної роботи.

Для запуску програми та редагування коду, можна використовувати будь-який редактор sas-коду, наприклад, SAS Base, SAS Enterprise Guide, SAS Studio, тощо.

Опис фрагментів коду програми.

Усі наведені нижче в цьому підрозділі коди програм можна виконати, наприклад, в системі SAS Studio.

На першому етапі необхідно створити або завантажити набір даних для аналізу. Для цього можна використати, як показано нижче, крок даних та створити відповідний набір даних для подальшого аналізу.

```
DATA Example_01;
INFILE DATALINES DLM=' ' MISSOVER DSD ;
INPUT T Y X;
DATALINES;
1998 90.5 90.4
1999 94.6 93.1
2000 112.3 .
2001 127.7 .
```

```
2002 117.8 101.2
2003 93.3 89
2004 166.8 119.7
2005 114.6 100.1
2006 115.5 102.5
2007 105.6 93.5
;
run;
```

В результаті буде створено набір даних, що містить три змінні: T – рік, Y – цільову змінну та X – вхідну змінну.

Створення значень оберненої цільової змінної, що буде використовуватися для обчислення ScoreAvg0.

```
data &file_in._2;
set &file_in.;
X=-Y;
run;
```

Обчислення кількості непустих значень цільової змінної та запис у макро-змінну n.

```
proc sql NOPRINT;
select count(*) into :n from &file_in._2 where Y ne .; quit;
run;
%let n=&n.;
%let m=&n.;
```

Початок обчислювального блоку на мові програмування SAS IML.

```
proc iml;
```

Завантаження значень змінних T, Y та X з набору даних у матрицю A.

```
use work.&file_in._2;
read all var {T Y X} into A;
```

Завантаження непустих значень змінної Y у вектор Y0.

```
read all var {Y} into Y0;

   Y=j(&n.,1,.);
   i=0;
   do j=1 to nrow(Y0);
     IF (Y0[j] ^= .) then
       do;
         i=i+1;
              Y[i]=Y0[j];
       end;
   end;
```

Завантаження непустих значень змінної X у вектор X0.

```
   read all var {X} into X0;

   X=j(&m.,1,.);
   i=0;
   do j=1 to nrow(X0);
     IF (X0[j] ^= .) then
       do;
```

```
      i=i+1;
      X[i]=X0[j];
    end;
        end;
```

Алгоритм нормування цільового ряду даних за діапазоном.

```
Y_max=Max(Y0);
  Y_min=Min(Y0);

  Y1=j(nrow(Y0),1,.);
  do j=1 to nrow(Y0);
   Y1[j,1]=(Y0[j,1]-Y_min)/(Y_max-Y_min);
  end;

  Y=j(&n.,1,.);
  i=0;
  do j=1 to nrow(Y1);
   IF (Y1[j] ^= .) then
    do;
      i=i+1;
          Y[i]=Y1[j];
    end;
        end;
```

Алгоритм нормування вхідного ряду даних за діапазоном.

```
  X_max=Max(X0);
  X_min=Min(X0);
```

```
X1=j(nrow(X0),1,.);
do j=1 to nrow(X0);
  X1[j,1]=(X0[j,1]-X_min)/(X_max-X_min);
end;

X=j(&m.,1,.);
i=0;
do j=1 to nrow(X1);
  IF (X1[j] ^= .) then
    do;
      i=i+1;
      X[i]=X1[j];
    end;
        end;
```

Алгоритм стандартизованого нормування цільового ряду даних.

```
Y_mean=Mean(Y0);
      Y_std=STD(Y0);

Y1=j(nrow(Y0),1,.);
do j=1 to nrow(Y0);
  Y1[j,1]=(Y0[j,1]-Y_mean)/(Y_std);
end;
      IF "&test_mode."="1" then do; print Y1; end;

Y=j(&n.,1,.);
i=0;
do j=1 to nrow(Y1);
  IF (Y1[j] > .) then
```

```
      do;
       i=i+1;
            Y[i]=Y1[j];
      end;
         end;
```

Алгоритм стандартизованого нормування вхідного ряду даних.

```
X_mean=Mean(X0);
       X_std=STD(X0);

  X1=j(nrow(X0),1,.);
  do j=1 to nrow(X0);
   X1[j,1]=(X0[j,1]-X_mean)/(X_std);
  end;
  IF "&test_mode."="1" then do; print X1; end;

  X=j(&m.,1,.);
  i=0;
  do j=1 to nrow(X1);
  IF (X1[j] > .) then
   do;
    i=i+1;
      X[i]=X1[j];
   end;
end;
```

Алгоритм формування матриці відстаней. Якщо макрозміна metric дорівнює (1) abs_dif, то за методом нормування різниць значень, у випадку (2) sqr_dif, за методом квадрату різниць значень.

```
D=J(&n.,&m., .);
  do j=1 to &n.; /* row index */
    do i=1 to &m.; /* column index */
          if "&metric."="abs_dif" then D[j,i]=abs(X[i]-Y[j]);
          if "&metric."="sqr_dif" then D[j,i]=(X[i]-Y[j])*(X[i]-Y[j]);
    end;
  end;
```

Алгоритм формування діагонального шляху між рядами Y та Y0.

```
Length=0;
Path=j((&n.), 5, .);
     do j=1 to (min(&m., &n.)); /* 2-1 */
           i=j;
           Length=Length+1;
           Path[Length,1]=Length;
       Path[Length,2]=j;
                Path[Length,3]=i;
       Path[Length,4]=D[j,i];
          end; /* 2-1 */
```

Алгоритм обчислення значення вартості діагонального шляху між рядами Y та Y0. Фактично обчислюється значення ScoreAvg0.

```
  Path_opt_D=J(&n.,&m., .);
  Score_opt=0;
  do j=1 to Length;
    Vect=Path[j,];
        j1=Path[j,2];
```

```
        i1=Path[j,3];
   Path_opt_D[j1, i1]=Path[j,4];
        Score_opt=Score_opt+Path[j,4];
  end;
```

Кінець використання мови SAS IML.

```
run;
```

Вищенаведені основні коди програм реалізовано у вигляді макросу %SCORE_0 який обчислює значення ScoreAvg0 для формули , що міститься у макрозмінній Score_AVG_INV, яка в свою чергу використовується у макросі %SCORE_1, основні коди програм якого наведені нижче.

Алгоритм побудови оптимального шляху між цільовим та вхідним часовими рядами. RL та CL – це макро-змінні, що містять значення обмежень на стиснення та розширення відповідно.

```
j=1;    i=1;
  Length=Length+1;
  Path[Length,1]=Length;
  Path[Length,2]=j;
  Path[Length,3]=i;
  Path[Length,4]=D[j,i];

  do while ((j<=(&n.-1)) & (i<=(&m.-1)));
    D1=&inf.; D2=&inf.; D3=&inf.;
    RL=&RL.;
    CL=&CL.;
```

```
        i1=i;   j1=j+1;  if ((-&RL.<=(i1-j1)) & ((i1-j1)<=&CL.) & (i1<=&m.) &
(j1<=&n.)) then  D1=D[j1, i1];
        i2=i+1; j2=j;    if ((-&RL.<=(i2-j2)) & ((i2-j2)<=&CL.) & (i2<=&m.) &
(j2<=&n.)) then  D2=D[j2, i2];
        i3=i+1; j3=j+1;  if ((-&RL.<=(i3-j3)) & ((i3-j3)<=&CL.) & (i3<=&m.) &
(j3<=&n.)) then  D3=D[j3, i3];

        If &n.>=&m. then
          do;
           IF D2<=MIN(D1, D2, D3) THEN DO; i=i2; j=j2; END;
           IF D1<=MIN(D1, D2, D3) THEN DO; i=i1; j=j1; END;
           IF D3<=MIN(D1, D2, D3) THEN DO; i=i3; j=j3; END;
          end;

        If &n.<&m. then
          do;
           IF D1<=MIN(D1, D2, D3) THEN DO; i=i1; j=j1; END;
           IF D2<=MIN(D1, D2, D3) THEN DO; i=i2; j=j2; END;
           IF D3<=MIN(D1, D2, D3) THEN DO; i=i3; j=j3; END;
    end;

        Length=Length+1;
        Path[Length,1]=Length;
        Path[Length,2]=j;
        Path[Length,3]=i;
        Path[Length,4]=D[j,i];
 end;

 if ((j<&n.) & (i=&m.)) Then;
  do;
```

```
   i=&m.;
    do j1=j+1 to &n.;
      Length=Length+1;
      Path[Length,1]=Length;
      Path[Length,2]=j1;
      Path[Length,3]=i;
      Path[Length,4]=D[j1,i];
    end;
   end;

 if ((j=&n.) & (i<&m.)) Then;
  do;
   j=&n.;
   do i1=i+1 to &m.;
      Length=Length+1;
      Path[Length,1]=Length;
      Path[Length,2]=j;
      Path[Length,3]=i1;
      Path[Length,4]=D[j,i1];
   end;
         end;
```

Алгоритм для видалення зайвих шляхів та переходів, що можуть бути замінені на діагональні переходи.

```
 Length_opt=2;
 do raw=2 to Length-1;
  Length_opt=Length_opt+1;
  j0=Path[raw-1,2];
  i0=Path[raw-1,3];
```

```
  j1=Path[raw,2];
  i1=Path[raw,3];
  j2=Path[raw+1,2];
  i2=Path[raw+1,3];

  if ((j0+1)=j1) & (i0=i1) & ((j0+1)=j2) & ((i0+1)=i2) & (Path[(raw-1),5]=.) then
    do;
      Path[raw,5]=1; Length_opt=Length_opt-1;
    end;

  if (j0=j1) & ((i0+1)=i1) & ((j0+1)=j2) & ((i0+1)=i2) & (Path[(raw-1),5]=.) then
    do;
      Path[raw,5]=1; Length_opt=Length_opt-1;
    end;
end;

Path_opt_2=j(Length_opt,5,.);
j=0;
do row=1 to Length;
  flag=Path[row,5];
  if flag=. then
    do;
          j=j+1;
      Path_opt_2[j,1]=j;
              Path_opt_2[j,2]=Path[row,2];
              Path_opt_2[j,3]=Path[row,3];
              Path_opt_2[j,4]=Path[row,4];
         end;
        end;
```

Алгоритм обчислення значення фнкції вартості шляху Score_Avg1=ScoreAvg(p1). Отримане значення збурігається як макрозмінна Score_Avg.

```
Path_opt_2_D=J(&n.,&m., .);
 Score_opt_2=0;
 do j=1 to Length_opt;
  Vect=Path_opt_2[j,];
      j1=Path_opt_2[j,2];
      i1=Path_opt_2[j,3];
  Path_opt_2_D[j1, i1]=Path_opt_2[j,4];
      Score_opt_2=Score_opt_2+Path_opt_2[j,4];
 end;
      call symputx("Score", Score_opt_2);
```

Реалізовані макро-програми %SCORE_0 та %SCORE_1 використовуються в основній макро-програмі %Macro_Cycle, яка на основі заданих значень обмежень на стиснення та розширення цільового ряду даних робить перебір усіх можливих комбінацій оптимальних шляхів, а в результаті будується матриця що містить значення відповідних статистик якості для кожного варіанту. Нижче наведений основний код макро-програми %Macro_Cycle.

```
%DO j=&RL. %to &RL.;
%DO i=&CL. %to &CL.;

 %let RL=&j; %let CL=&i;
 %PUT &=RL &=CL;
```

```
 %let test_mode=0;
 %Score_1;

 proc append base=stats_model_all data=stats_model;
 run;
%END;
%END;

%IF &n. >= &m. %THEN
 %DO;
   proc sort data=stats_model_all out=stats_model_all_sort;
    BY  Score_AVG RL CL;
   RUN;
 %END;

%IF &n. < &m. %THEN
 %DO;
   proc sort data=stats_model_all
           out=stats_model_all_sort;
    BY  Score_AVG CL RL;
   RUN;
 %END;
```

На основі отриманого файлу з результатами роботи макро-програми %Macro_Cycle, обирається найкраща модель за критерієм мінімального значення Score_AVG та найменшою кількістю стиснень (RL) та розширень (CL).

Для обчислення статистики міри подібності рядів за формулою (21) використовується наступний код програми:

```
data Test1;
set stats_model_all_sort;
call symput("RL",RL);
call symput("CL",CL);
call symput("Score_Avg",Score_Avg);

Score_Sim=Round(100*(1-(Score_AVG / symgetn("Score_AVG_INV"))),.01);
call symput("Score_Sim",Score_Sim);

output;
STOP;
      run;
```

Для побудови графіків використовується процедура SGPLOT з оператором SERIES.

```
proc sgplot data=&file_out._2;
series x=T y=Y / break;
xaxis type=discrete label="Time";
yaxis label="Y - Target / X - Input";
series x=T y=X / break;
run;
```

Для побудови гістограм використовується процедура SGPLOT з оператором HISTOGRAM.

```
proc sgplot data=&file_out. (Rename=(Err=&metric.));
  histogram &metric. / nbins=5 showbins;
  density &metric. / type=kernel;
run;
```

ДОДАТОК Н

ПРИКЛАДИ ВИКОРИСТАННЯ КОМП'ЮТЕРНОЇ ПРОГРАМИ «SAS TIME SERIES SIMILARITY»

Н.1 Практичний приклад 1. Кластеризація регіонів України за інвестиціями в охорону навколишнього середовища

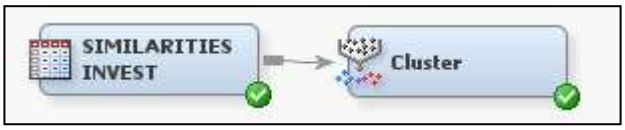


Рисунок. Н.1 – Схема технологічного процесу в системі SAS Enterprise Miner

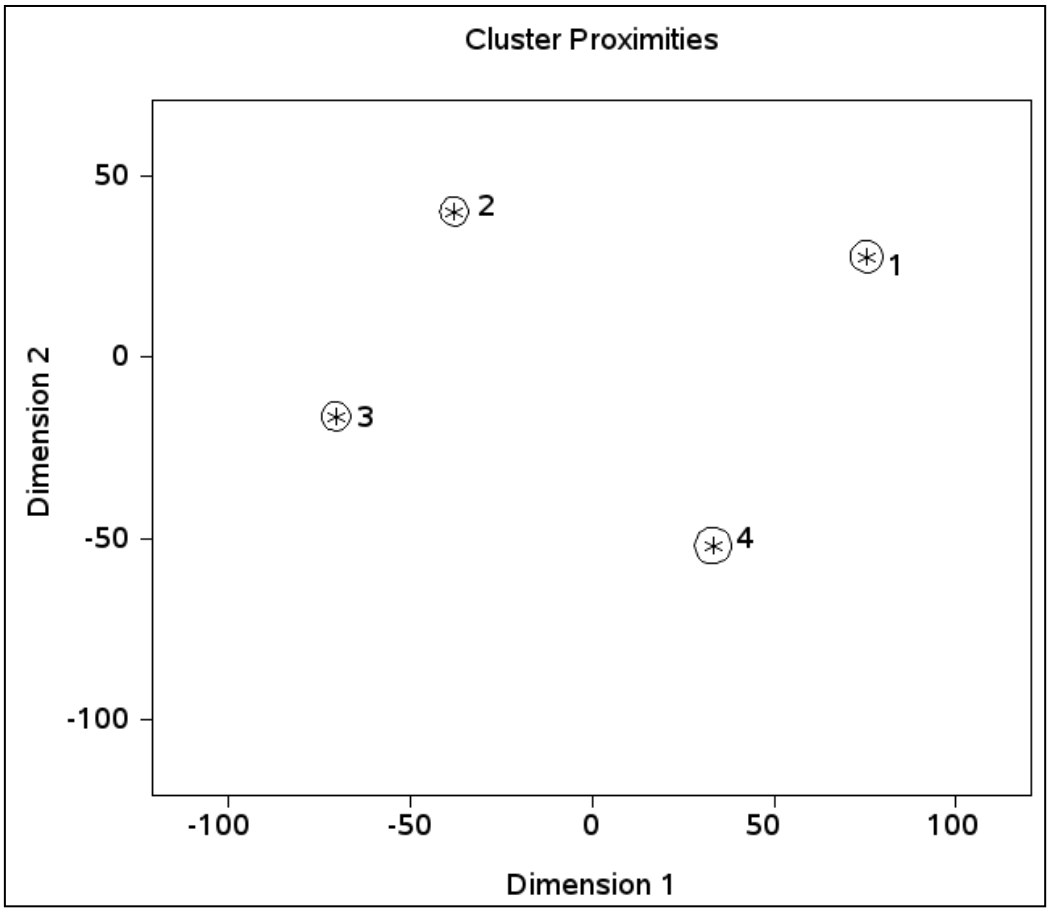


Рисунок Н.2 – Відстань між кластерами, результати роботи системи SAS Enterprise Miner

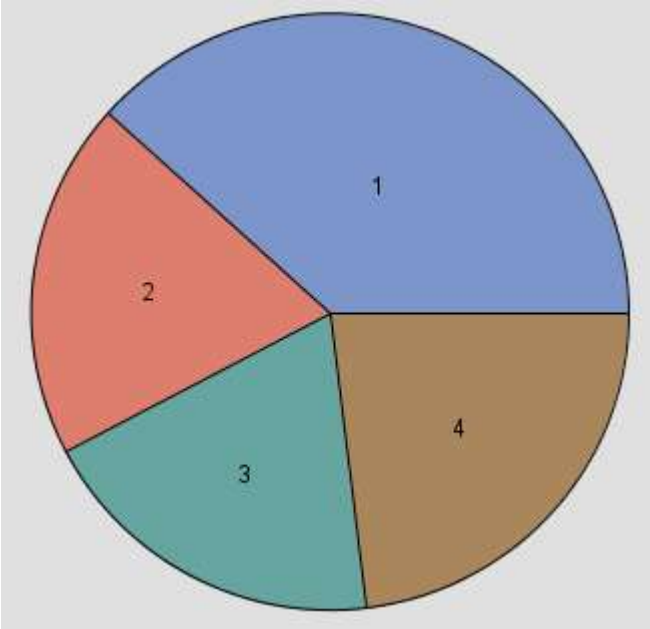


Рисунок Н.3 – Кругова діаграма розподілу кластерів за частотою, результати роботи системи SAS Enterprise Miner

Таблиця. Н.1 Кількість регіонів, що увійшли до відповідного кластеру.

Номер кластеру	Кількість регіонів
1	10
2	5
3	5
4	6

Для візуалізації отриманих результатів у вигляді мереж було використано програму SAS Graph Network Visualization WorkShop, у вигляді

- кругового графу;
- ієрархічного графу;
- гексагонального графу
- багаторівневого графу.

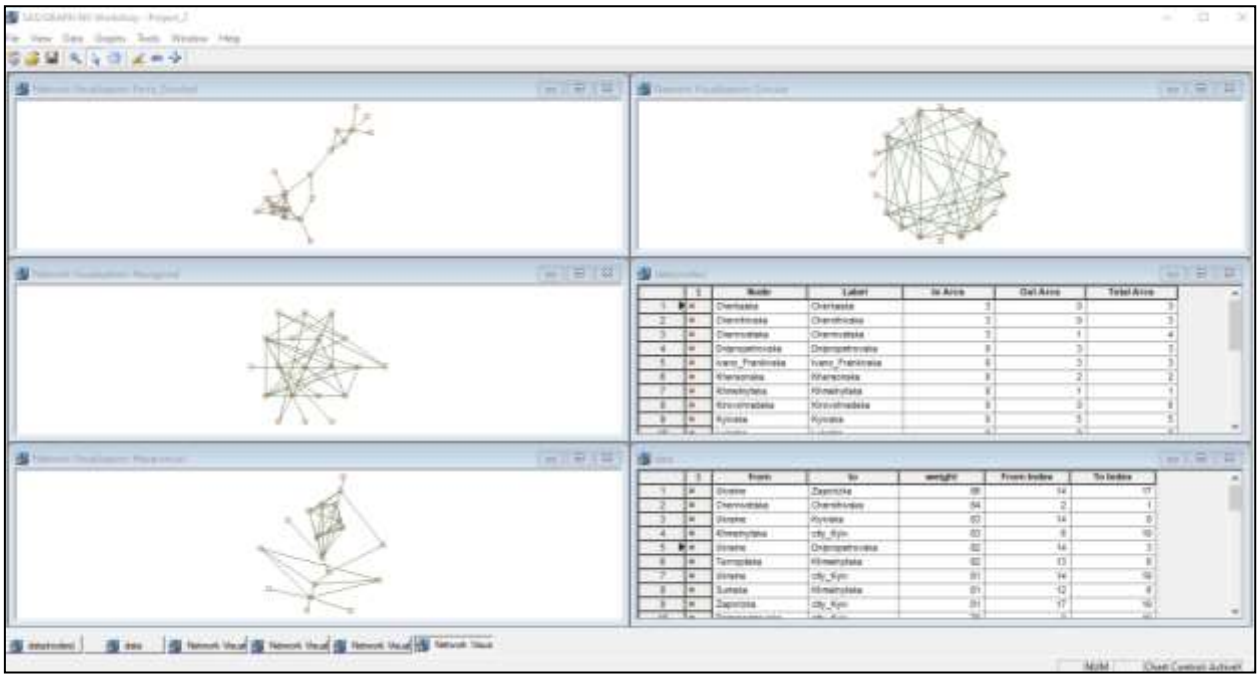


Рисунок Н.4 – Скріншот системи SAS Graph Network Visualization WorkShop для розглянутого прикладу

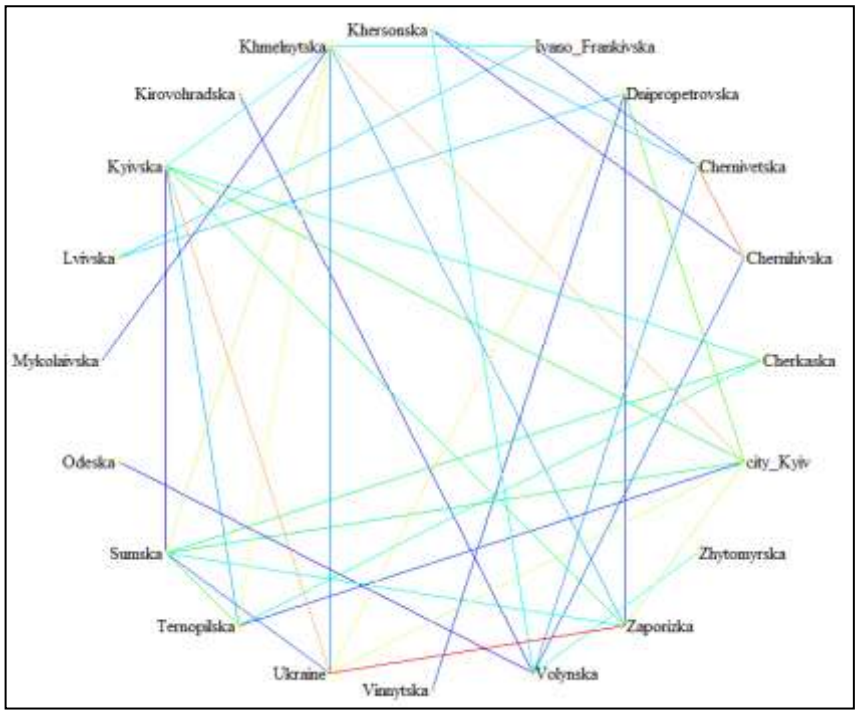


Рисунок Н.5 – Круговий граф зав’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

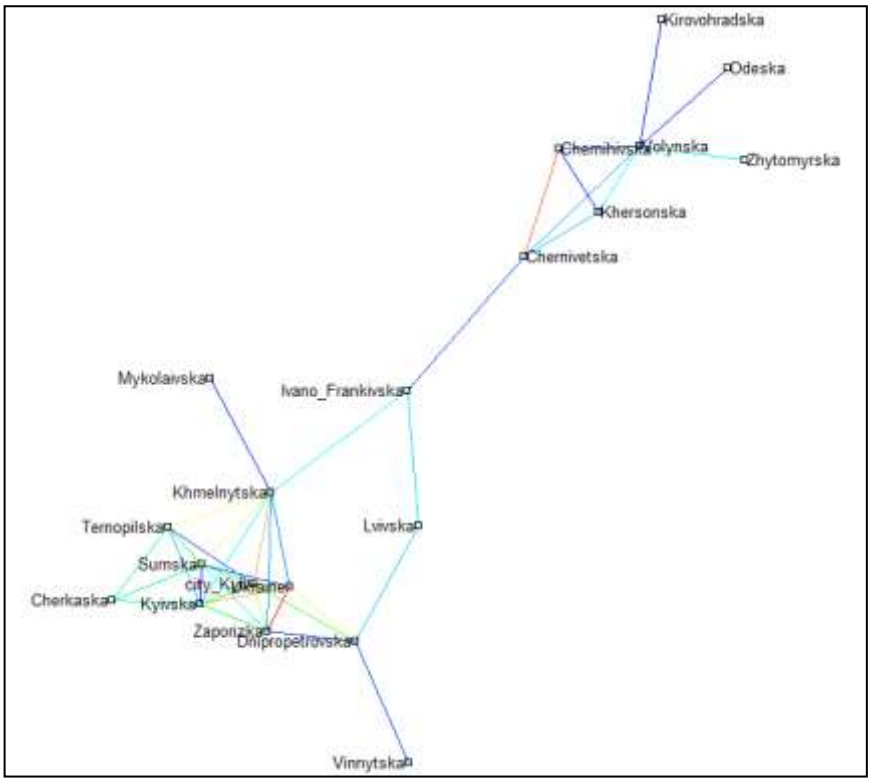


Рисунок Н.6 – Граф зав’язків за координатами користувача, відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

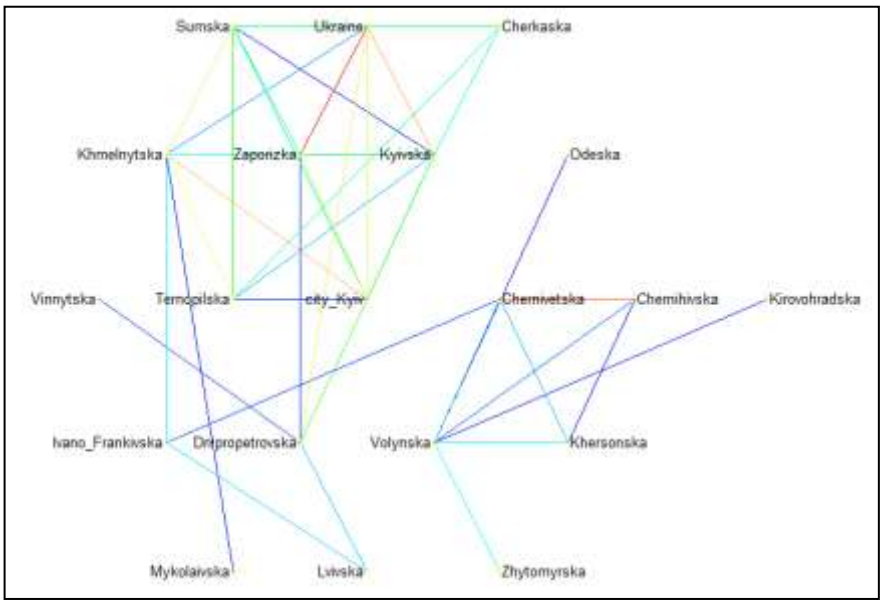


Рисунок Н.7 – Гексагональний граф зав’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

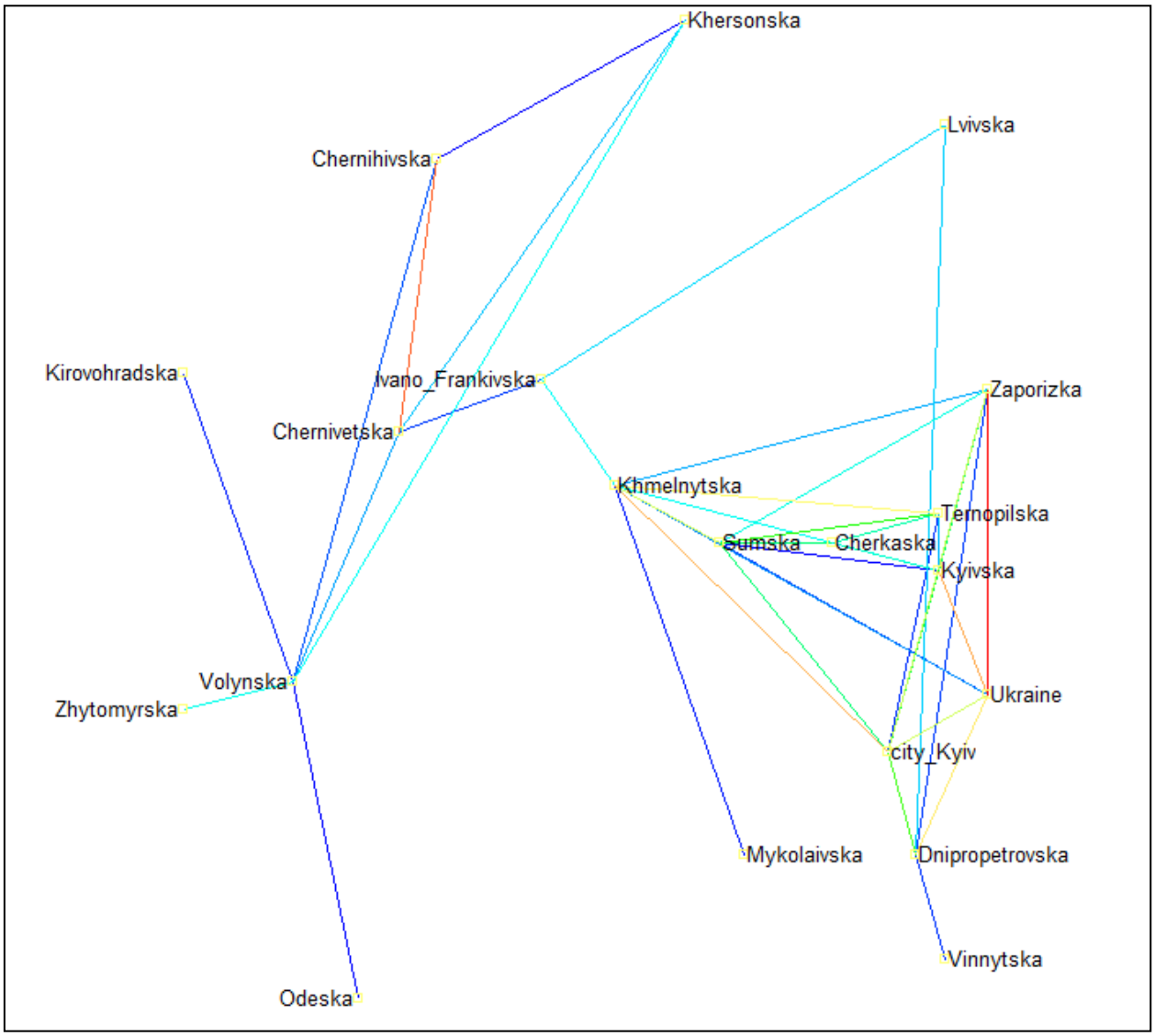


Рисунок Н.8 – Ієрархічний граф зав’язків відображає тільки ті зв’язки між областями що дорівнюють або перевищують 70%. Колір залежить від значення метрики ScoreSim

Таблиця Н.2 – Значення схожості рядів за метрикою ScoreSim

	UA	VN	LT	DP	DN	ZT	ZK	ZP	IF	KY	KV	LV	LVIV	MK	OD	PL	RV	SM	TE	KH	KS	KM	CK	CV	CH	KYIV
UA	100	60	20	82	45	39	55	86	42	83	28	20	65	56	17	55	44	71	63	48	29	72	60	35	44	81
VN	60	100	32	70	51	25	59	57	53	42	26	15	60	40	46	60	39	56	53	38	22	59	43	31	47	65
LT	20	32	100	42	50	74	37	38	59	36	70	58	43	56	70	43	58	41	55	49	74	47	42	72	71	42
DP	82	70	42	100	45	39	63	71	52	69	25	10	73	47	24	59	33	62	49	45	22	64	44	28	37	79
DN	45	51	50	45	100	27	66	43	42	36	33	38	59	26	26	53	46	36	45	67	24	33	33	37	59	39
ZT	39	25	74	39	27	100	29	39	47	40	53	42	43	62	39	30	59	32	51	19	66	40	34	46	51	47
ZK	55	59	37	63	66	29	100	53	43	41	21	21	66	22	27	56	30	35	29	47	15	41	25	15	41	61
ZP	86	57	38	71	43	39	53	100	46	76	37	20	68	55	21	57	54	74	66	39	31	72	63	37	44	81
IF	42	53	59	52	42	47	43	46	100	52	52	37	73	45	58	67	65	60	64	46	61	74	55	71	69	64
KY	83	42	36	69	36	40	41	76	52	100	40	26	61	63	21	48	48	70	73	56	42	74	75	51	48	77
KV	28	26	70	25	33	53	21	37	52	40	100	47	40	41	51	51	68	23	41	22	62	41	38	49	60	43
LV	20	15	58	10	38	42	21	20	37	26	47	100	39	24	49	46	48	11	21	41	47	22	16	40	55	27
LVIV	65	60	43	73	59	43	66	68	73	61	40	39	100	27	31	72	44	51	57	40	28	63	39	43	52	68
MK	56	40	56	47	26	62	22	55	45	63	41	24	27	100	35	32	54	60	65	34	48	70	50	37	40	66
OD	17	46	70	24	26	39	27	21	58	21	51	49	31	35	100	44	48	21	32	32	49	33	11	39	54	29
PL	55	60	43	59	53	30	56	57	67	48	51	46	72	32	44	100	54	35	27	26	30	36	17	35	52	52
RV	44	39	58	33	46	59	30	54	65	48	68	48	44	54	48	54	100	25	38	14	52	41	21	40	56	48
SM	71	56	41	62	36	32	35	74	60	70	23	11	51	60	21	35	25	100	78	55	46	81	76	58	55	76
TE	63	53	55	49	45	51	29	66	64	73	41	21	57	65	32	27	38	78	100	52	57	82	75	64	58	70
KH	48	38	49	45	67	19	47	39	46	56	22	41	40	34	32	26	14	55	52	100	46	53	54	59	64	56
KS	29	22	74	22	24	66	15	31	61	42	62	47	28	48	49	30	52	46	57	46	100	48	46	73	70	44
KM	72	59	47	64	33	40	41	72	74	74	41	22	63	70	33	36	41	81	82	53	48	100	64	44	47	83
CK	60	43	42	44	33	34	25	63	55	75	38	16	39	50	11	17	21	76	75	54	46	64	100	64	54	67
CV	35	31	72	28	37	46	15	37	71	51	49	40	43	37	39	35	40	58	64	59	73	44	64	100	84	59
CH	44	47	71	37	59	51	41	44	69	48	60	55	52	40	54	52	56	55	58	64	70	47	54	84	100	49
KYIV	81	65	42	79	39	47	61	81	64	77	43	27	68	66	29	52	48	76	70	56	44	83	67	59	49	100

Таблиця Н.3 – Значення кореляцій Пірсона між рядами.

	UA	VN	LT	DP	DN	ZT	ZK	ZP	IF	KY	KV	LV	LVIV	MK	OD	PL	RV	SM	TE	KH	KS	KM	CK	CV	CH	KYIV
UA	1	0,58	-0,35	0,83	0,15	-0,22	0,37	0,93	0,14	0,96	-0,18	-0,5	0,56	0,28	-0,38	0,41	-0,03	0,76	0,67	0,24	-0,48	0,75	0,55	-0,18	-0,26	0,89
VN	0,58	1	0,2	0,74	0,24	-0,27	0,58	0,42	0,55	0,42	-0,15	-0,58	0,59	0,21	0,42	0,45	0,15	0,7	0,61	0,38	-0,28	0,63	0,29	0,23	0,22	0,63
LT	-0,35	0,2	1	-0,22	-0,19	0,5	-0,22	-0,26	0,39	-0,42	0,35	0,08	-0,17	0,23	0,82	0,08	0,31	-0,16	0,12	-0,08	0,7	-0,02	-0,38	0,62	0,61	-0,1
DP	0,83	0,74	-0,22	1	0,13	-0,2	0,6	0,76	0,42	0,69	-0,18	-0,59	0,78	0,18	-0,01	0,49	-0,09	0,72	0,47	0,35	-0,49	0,62	0,24	-0,14	-0,16	0,87
DN	0,15	0,24	-0,19	0,13	1	-0,67	0,39	-0,09	-0,11	0	-0,39	0,18	0,41	-0,4	-0,01	0,29	-0,02	-0,17	-0,25	0,59	-0,33	-0,09	-0,14	-0,16	0,26	-0,07
ZT	-0,22	-0,27	0,5	-0,2	-0,67	1	-0,43	-0,06	-0,03	-0,13	0,45	0,11	-0,42	0,61	0,28	-0,35	-0,01	-0,1	0,22	-0,51	0,5	0,15	-0,12	-0,05	-0,23	0,01
ZK	0,37	0,58	-0,22	0,6	0,39	-0,43	1	0,18	0,24	0,19	-0,1	-0,25	0,53	-0,09	0,14	0,51	-0,04	0,29	0,01	0,68	-0,44	0,29	-0,03	-0,08	0,1	0,44
ZP	0,93	0,42	-0,26	0,76	-0,09	-0,06	0,18	1	0,22	0,91	0,04	-0,45	0,51	0,22	-0,39	0,45	-0,01	0,67	0,64	0,07	-0,31	0,66	0,44	-0,04	-0,23	0,87
IF	0,14	0,55	0,39	0,42	-0,11	-0,03	0,24	0,22	1	-0,02	0,31	-0,41	0,68	0,03	0,59	0,7	0,49	0,26	0,29	-0,12	0,28	0,33	-0,24	0,64	0,56	0,44
KY	0,96	0,42	-0,42	0,69	0	-0,13	0,19	0,91	-0,02	1	-0,16	-0,52	0,36	0,34	-0,54	0,23	-0,07	0,76	0,71	0,06	-0,45	0,72	0,71	-0,23	-0,41	0,8
KV	-0,18	-0,15	0,35	-0,18	-0,39	0,45	-0,1	0,04	0,31	-0,16	1	0	-0,03	-0,14	0,24	0,33	0,2	-0,29	0,13	-0,4	0,47	-0,03	-0,16	0,42	0,2	-0,07
LV	-0,5	-0,58	0,08	-0,59	0,18	0,11	-0,25	-0,45	-0,41	-0,52	0	1	-0,44	-0,2	0,02	-0,23	-0,02	-0,67	-0,48	0,18	0,16	-0,4	-0,55	-0,16	0,1	-0,53
LVIV	0,56	0,59	-0,17	0,78	0,41	-0,42	0,53	0,51	0,68	0,36	-0,03	-0,44	1	-0,17	0,13	0,78	0,2	0,36	0,19	0,22	-0,24	0,37	-0,08	0,13	0,24	0,62
MK	0,28	0,21	0,23	0,18	-0,4	0,61	-0,09	0,22	0,03	0,34	-0,14	-0,2	-0,17	1	0,11	-0,3	-0,05	0,42	0,58	-0,17	0,02	0,68	0,27	-0,24	-0,36	0,47
OD	-0,38	0,42	0,82	-0,01	-0,01	0,28	0,14	-0,39	0,59	-0,54	0,24	0,02	0,13	0,11	1	0,19	0,3	-0,11	0,03	0,07	0,45	0,01	-0,49	0,48	0,59	-0,09
PL	0,41	0,45	0,08	0,49	0,29	-0,35	0,51	0,45	0,7	0,23	0,33	-0,23	0,78	-0,3	0,19	1	0,53	0,21	0,18	0,2	0,16	0,3	-0,16	0,55	0,61	0,52
RV	-0,03	0,15	0,31	-0,09	-0,02	-0,01	-0,04	-0,01	0,49	-0,07	0,2	-0,02	0,2	-0,05	0,3	0,53	1	0,2	0,27	-0,38	0,68	0,26	0,05	0,69	0,69	0,11
SM	0,76	0,7	-0,16	0,72	-0,17	-0,1	0,29	0,67	0,26	0,76	-0,29	-0,67	0,36	0,42	-0,11	0,21	0,2	1	0,78	0,01	-0,28	0,77	0,71	0,03	-0,17	0,76
TE	0,67	0,61	0,12	0,47	-0,25	0,22	0,01	0,64	0,29	0,71	0,13	-0,48	0,19	0,58	0,03	0,18	0,27	0,78	1	-0,23	-0,05	0,89	0,68	0,13	-0,14	0,67
KH	0,24	0,38	-0,08	0,35	0,59	-0,51	0,68	0,07	-0,12	0,06	-0,4	0,18	0,22	-0,17	0,07	0,2	-0,38	0,01	-0,23	1	-0,55	-0,02	-0,25	-0,2	0,1	0,18
KS	-0,48	-0,28	0,7	-0,49	-0,33	0,5	-0,44	-0,31	0,28	-0,45	0,47	0,16	-0,24	0,02	0,45	0,16	0,68	-0,28	-0,05	-0,55	1	-0,17	-0,29	0,68	0,62	-0,26
KM	0,75	0,63	-0,02	0,62	-0,09	0,15	0,29	0,66	0,33	0,72	-0,03	-0,4	0,37	0,68	0,01	0,3	0,26	0,77	0,89	-0,02	-0,17	1	0,51	-0,04	-0,15	0,84
CK	0,55	0,29	-0,38	0,24	-0,14	-0,12	-0,03	0,44	-0,24	0,71	-0,16	-0,55	-0,08	0,27	-0,49	-0,16	0,05	0,71	0,68	-0,25	-0,29	0,51	1	-0,14	-0,41	0,32
CV	-0,18	0,23	0,62	-0,14	-0,16	-0,05	-0,08	-0,04	0,64	-0,23	0,42	-0,16	0,13	-0,24	0,48	0,55	0,69	0,03	0,13	-0,2	0,68	-0,04	-0,14	1	0,86	0
CH	-0,26	0,22	0,61	-0,16	0,26	-0,23	0,1	-0,23	0,56	-0,41	0,2	0,1	0,24	-0,36	0,59	0,61	0,69	-0,17	-0,14	0,1	0,62	-0,15	-0,41	0,86	1	-0,11
KYIV	0,89	0,63	-0,1	0,87	-0,07	0,01	0,44	0,87	0,44	0,8	-0,07	-0,53	0,62	0,47	-0,09	0,52	0,11	0,76	0,67	0,18	-0,26	0,84	0,32	0	-0,11	1

Н.3 Практичний приклад 3. Аналіз між потужністю та кількістю малих гідро електростанцій, за статистичними даними 2009-2017 років.

В якості практичного прикладу показано аналіз між потужністю та кількістю малих гідро електростанцій, за статистичними даними 2009-2017 років.

Мала гідроенергетика в Україні демонструє нарощування потужності та кількості малих гідроелектростанцій, в [226] наведені відповідні статистичні дані у вигляді таблиці.

Таблиця. Н.4 – Динаміка введення МГЕС в Україні в 2009-2017.

Рук	Встановлена потужність МГЕС, МВт	Кількість МГЕС
2009	49,2	46
2010	62,6	60
2011	70,8	72
2012	73,5	80
2013	75,3	90
2014	79	98
2015	82,2	111
2016	85,5	123
2017	94,6	136

На основі запропонованої методики обчислення подібності часових рядів було виконано аналіз між потужністю та кількістю введених гідроелектростанцій в Україні в 2009-2017 роках. В результаті, ступінь схожості рядів дорівнює 88,33% що не сильно дивує, так як візуально за графіком можна побачити, що обидва ряди мають позитивну кореляційну залежність, як показано на рисунку нижче, і збільшення кількості введених в експлуатацію станцій призводить до збільшення сумарної потужності електроенергії, що виробляється ними.

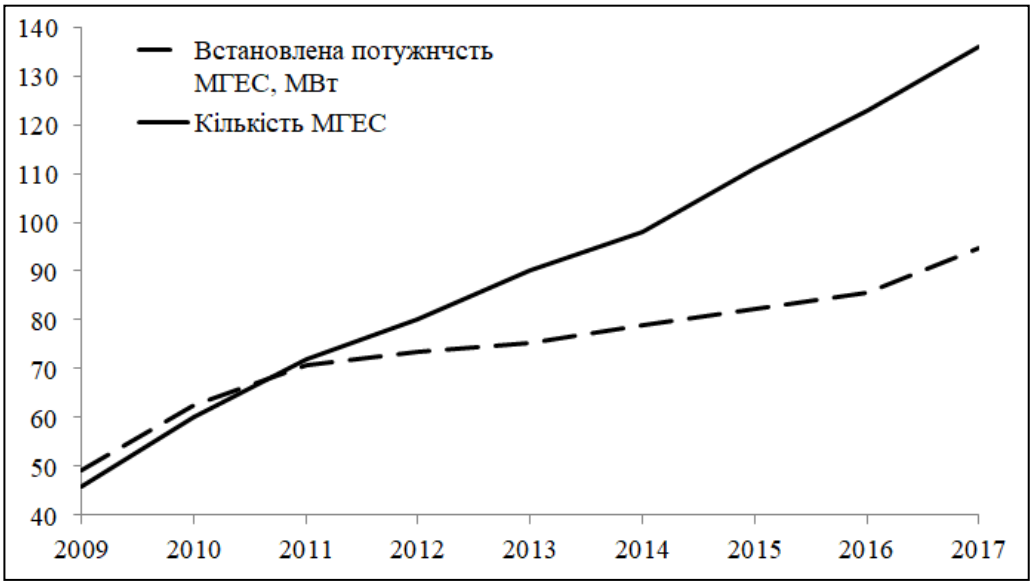


Рисунок. Н.9 – Графіки введення кількості малих гідроелектростанцій (суцільна лінія) та встановленної потужності в Мега Ваттах (штрихова лінія).

Н.3 Приклад побудови шляху для методики аналізу подібності часових рядів

По-перше треба задати обмеження на перетворення. Нехай в рамках даного прикладу, $RL = 2$ – обмеження на стиснення цільової змінної, а $CL = 1$ – обмеження на розширення цільової змінної. В табл. Н.5 наведені значення матриці відстаней D з урахуванням наведених обмежень на перетворення цільового ряду.

Таблиця Н.5 – Значення матриці відстаней D з урахуванням обмежень на перетворення цільового ряду $RL = 2$ та $CL = 1$.

Індекс часу	i =1	i =2	i =3	i =4	i =5	i =6	i =7	i =8
j =1	0,0456	0.1335						
j =2	0,0081	0,0799	0.3436					
j =3	0,2401	0,1522	0,1116	0.2857				
j =4		0,3539	0,0901	0,4875	0.5124			
j =5			0,0395	0,3578	0,6422	0.0037		
j =6				0,0367	0,9633	0,3248	0.4031	
j =7					0	0,6384	0,5603	0.8534
j =8						0,0457	0,1239	0,1692
j =9							0,1121	0,1811
j =10								0,0513

Ітерація 1.

Побудова оптимального шляху починається з елементу D[1,1] матриці відстаней.

Таблиця Н.6 – Ітерація 1.

Індекс часу	i =1
j =1	**0,0456**

Ітерація 2.

В якості наступного елемента може бути обраний $D[2,1] = 0{,}0081$, $D[1,2] = 0{,}1335$ або $D[2,2] = 0{,}0799$. Обирається $D[2,1]$ тому що $0{,}0081$ – найменше серед усіх значень.

Таблиця Н.7 – Ітерація 2.

Індекс часу	$i = 1$	$i = 2$
$j = 1$	**0,0456**	0.1335
$j = 2$	**0,0081**	0,0799

Ітерація 3.

Серед найближчих елементів до $D[2,1]$ є три елементи $D[3,1] = 0{,}2401$, $D[2,2] = 0{,}0799$ та $D[3,2] = 0{,}1522$, серед яких обирається із найменшим значенням різниці відстаней елемент $D[2,2]$.

Таблиця Н.8 – Ітерація 3.

Індекс часу	$i = 1$	$i = 2$
$j = 1$	**0,0456**	0.1335
$j = 2$	**0,0081**	**0,0799**
$j = 3$	0,2401	0,1522

Ітерація 4.

Серед найближчих елементів до $D[2,2]$ є три допустимих елементи $D[3,2] = 0{,}1522$, $D[2,3] = 0{,}3436$ та $D[3,3] = 0{,}1116$, серед яких обирається $D[3,3]$, тому що $0{,}1116$ – найменше серед всіх значення.

Таблиця Н.9 – Ітерація 4.

Індекс часу	$i = 1$	$i = 2$	$i = 3$
$j = 1$	**0,0456**	0.1335	
$j = 2$	**0,0081**	**0,0799**	0.3436
$j = 3$	0,2401	0,1522	**0,1116**

Ітерація 5.

Серед найближчих елементів до D[3,3] є три допустимих елементи $D[4,3] = 0{,}0901$, $D[343] = 0{,}2857$ та $D[4,4] = 0{,}4875$, серед яких обирається D[4,3], тому що $0{,}0901$ – найменше значення.

Таблиця Н.10 – Ітерація 5.

Індекс часу	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$j = 1$	**0,0456**	0.1335		
$j = 2$	**0,0081**	**0,0799**	0.3436	
$j = 3$	0,2401	0,1522	**0,1116**	0.2857
$j = 4$		0,3539	**0,0901**	0,4875

Ітерація 6.

Серед найближчих елементів до D[4,3] є три допустимих елементи $D[5,3] = 0{,}0395$, $[4,4] = 0{,}4875$ та $D[5,4] = 0{,}3578$, серед яких обирається D[5,3], як найменше.

Таблиця Н.11 – Ітерація 6.

Індекс часу	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$j = 1$	**0,0456**	0.1335		
$j = 2$	**0,0081**	**0,0799**	0.3436	
$j = 3$	0,2401	0,1522	**0,1116**	0.2857
$j = 4$		0,3539	**0,0901**	0,4875
$j = 5$			**0,0395**	0,3578

Ітерація 7.

Серед найближчих елементів до D[5,3] є два допустимих елементи $D[5,4] = 0{,}3578$ та $D[6,4] = 0{,}0367$, серед яких обирається D[6,4], тому що $0{,}0367 < 0{,}3578$.

Таблиця Н.12 – Ітерація 7.

Індекс часу	$i=1$	$i=2$	$i=3$	$i=4$
$j=1$	**0,0456**	0.1335		
$j=2$	**0,0081**	**0,0799**	0.3436	
$j=3$	0,2401	0,1522	**0,1116**	0.2857
$j=4$		0,3539	**0,0901**	0,4875
$j=5$			**0,0395**	0,3578
$j=6$				**0,0367**

Ітерація 8.

Серед найближчих елементів до $D[6,4]$ є два допустимих елементи $D[6,5] = 0,9633$ та $D[7,5] = 0$, серед яких обирається $D[7,5]$, тому що $0 < 0,9633$.

Таблиця Н.13 – Ітерація 8.

Індекс часу	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$j=1$	**0,0456**	0.1335			
$j=2$	**0,0081**	**0,0799**	0.3436		
$j=3$	0,2401	0,1522	**0,1116**	0.2857	
$j=4$		0,3539	**0,0901**	0,4875	0.5124
$j=5$			**0,0395**	0,3578	0,6422
$j=6$				**0,0367**	0,9633
$j=7$					**0**

Ітерація 9.

Серед найближчих елементів до $D[7,5]$ є два допустимих елементи $D[7,6] = 0,6384$ та $D[8,6] = 0,0457$, серед яких обирається $D[8,6]$, тому що $0,0457 < 0,6384$.

Таблиця Н.14 – Ітерація 9.

Індекс часу	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$
$j=1$	**0,0456**	0.1335				

$j=2$	**0,0081**	**0,0799**	0.3436			
$j=3$	0,2401	0,1522	**0,1116**	0.2857		
$j=4$		0,3539	**0,0901**	0,4875	0.5124	
$j=5$			**0,0395**	0,3578	0,6422	0.0037
$j=6$				**0,0367**	0,9633	0,3248
$j=7$					**0**	0,6384
$j=8$						**0,0457**

Ітерація 10.

Серед найближчих елементів до D[8,6] є два допустимих елементи $D[8,7] = 0{,}1239$ та $D[9,7] = 0{,}1121$, серед яких обирається D[9,7], тому що $0{,}1121 < 0{,}1239$.

Таблиця Н.15. Ітерація 10.

Індекс часу	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$
$j=1$	**0,0456**	0.1335					
$j=2$	**0,0081**	**0,0799**	0.3436				
$j=3$	0,2401	0,1522	**0,1116**	0.2857			
$j=4$		0,3539	**0,0901**	0,4875	0.5124		
$j=5$			**0,0395**	0,3578	0,6422	0.0037	
$j=6$				**0,0367**	0,9633	0,3248	0.4031
$j=7$					**0**	0,6384	0,5603
$j=8$						**0,0457**	0,1239
$j=9$							**0,1121**

Ітерація 11.

Серед найближчих елементів до D[9,7] є два допустимих елементи $D[9,8] = 0{,}1811$ та $D[10,8] = 0{,}0513$, серед яких обирається D[10,8], тому що $0{,}0513 < 0{,}1811$.

Таблиця Н.16 Ітерація 11.

Індекс часу	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
$j=1$	**0,0456**	0.1335						
$j=2$	**0,0081**	**0,0799**	0.3436					
$j=3$	0,2401	0,1522	**0,1116**	0.2857				
$j=4$		0,3539	**0,0901**	0,4875	0.5124			
$j=5$			**0,0395**	0,3578	0,6422	0.0037		
$j=6$				**0,0367**	0,9633	0,3248	0.4031	
$j=7$					**0**	0,6384	0,5603	0.8534
$j=8$						**0,0457**	0,1239	0,1692
$j=9$							**0,1121**	0,1811
$j=10$								**0,0513**