Міністерство освіти і науки
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова
праця на правах рукопису

БАБІЙ АНДРІЙ СТЕПАНОВИЧ

УДК 004.942

ДИСЕРТАЦІЯ

МОДЕЛІ, МЕТОДИ ТА ІНТЕЛЕКТУАЛЬНА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

05.13.06 - інформаційні технології

технічні науки

Подається на здобуття наукового ступеня кандидата наук

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

Науковий керівник
Єрохін Андрій Леонідович, доктор технічних наук, професор

Харків – 2017

АНОТАЦІЯ

Бабій А.С. Моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 «Інформаційні технології». – Харківський національний університет радіоелектроніки, Міністерство освіти і науки України, Харків, 2017.

У дисертаційній роботі запропоновано рішення актуальної науково-практичної задачі розробки моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для підвищення ефективності оцінювання поточного стану предметних областей в інформаційно-аналітичних системах.

Об'єктом дослідження є процес аналізу неоднорідних послідовностей при оцінюванні поточного стану предметної області. Предметом дослідження є моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей для оцінювання поточного стану предметної області на основі коротких вибірок даних.

Методи дослідження: при розробці та дослідженні методу визначення значущих чинників нечіткої регресійної моделі були використані методи теорії нечітких множин та регресійного аналізу. При розробці моделей і методів фільтрації компонент неоднорідних часових послідовностей були використані методи нечіткої апроксимації даних та аналізу динамічних рядів. При проведенні та аналізі результатів експериментальних досліджень були використані елементи математичної статистики.

Практична значимість одержаних результатів. На основі запропонованих моделей та методів удосконалено інформаційну технологію аналізу даних неоднорідних послідовностей, яка, на відміну від існуючих технологій, надає можливість при побудові моделі предметної області додатково враховувати

відомості подані у вигляді нечітких даних та здійснювати відбір значущих чинників, що надає можливість проведення аналізу в умовах коротких вибірок даних.

В роботі запропоновано метод визначення значущих чинників нечіткої регресійної моделі неоднорідних послідовностей даних, який, на відміну від існуючих, містить етапи підбору коефіцієнтів за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних та відбору підмножини значущих чинників з коефіцієнтами, що перевищують порогове значення та дозволяє запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій.

Етап 1. Для визначення функції належності скористаємось підходом на основі непрямого методу побудови функцій належності.

Етап 2. Подамо чіткі дані про фактори у вигляді матриці значень p пояснюючих змінних у *n* спостереженнях, а дані про центри нечіткої величини $\tilde{Y}$ у вигляді вектору значень.

Етап 3. Застосуємо підхід на основі методу підбору коефіцієнтів за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних. Задамо початкову оцінку $\hat{\mu}_A = 0$ вектора значень залежної змінної y.

Етап 4. Обчислимо вектор кореляцій: $\hat{c} = X^T (y - \hat{\mu}_A)$.

Етап 5. Знайдемо поточний набір індексів A, що відповідає ознакам із найбільшими абсолютними значеннями кореляцій: $A = \{ j : \left|\hat{c}_j\right| = \hat{C} \}$, де $\hat{C} = \max\limits_{j=1,\dots,n} \{\left|\hat{c}_j\right|\}$.

Етап 6. Знайдемо $s_j = sign(\hat{c}_j)$ для $j \in A$. Розрахуємо матриці X_A, ψ_A таким чином: $X_A = \left[s_{j_1} x_{j_1}, \dots, s_{j_{|A|}} x_{j_{|A|}} \right]$, $j = (j_1, \dots, j_{|A|}) \in A$, $\psi_A = (1_A^T \varsigma_A^{-1} 1_A)^{-\frac{1}{2}}$, де $s_j \in \{+1, -1\}$ і $\left|A\right|$- потужність множини A (кількість значень множини A), $\varsigma = X_A^T X_A$, 1_A – одинична матриця розміру $1 \times \left|A\right|$.

Етап 7. Розрахуємо вектор $a = X^T u_A$, де $u_A = X_A w_{A'}$, $w_A = \psi_A \varsigma_A^{-1} 1_A$.

Етап 8. Розрахуємо значення $\hat{\gamma} = \min_{j \in A}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\psi_A - a_j}, \frac{\hat{C} + \hat{c}_j}{\psi_A + a_j} \right\}$, де мінімум береться по всім додатнім значенням для кожного j.

Етап 9. Знаходимо значення $\hat{\mu}_A$ для наступної ітерації: $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$.

Етап 10. Процес повторюється *n* раз (де *n* – кількість факторів), починаючи з етапу 4. Для кожної ітерації обчислюється коефіцієнт *Ср* Маллоуза.

Етап 11. Для побудови нечіткої регресійної моделі обираємо набір коефіцієнтів, який буде відповідати мінімальному значенню коефіцієнта *Ср*

Після цього застосовується метод розв'язку задачі побудови лінійної регресійної моделі з лише обраним набором факторних змінних.

Отримала подальший розвиток тренд-сезонна модель неоднорідних послідовностей, в якій, на відміну від існуючих моделей, трендова складова подається у вигляді інтерпольованих усереднених значень із врахуванням функції належності, яка асоційована із кожним нечітким розділенням, що дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень і тим самим підвищити ефективність моделювання стану предметних областей в інформаційно-аналітичних системах:

Пропонується модель сезонного явища на основі декомпозиційного підходу, яку можна представити у вигляді, який докладно описаний в другому розділі: $X_i = U_i + V_i + \varepsilon_i$, де $i = 1, \ldots, N$.

Трендова компонента U_i відшукується з використанням ітеративного застосування F-перетворення, тобто створивши n нечітких розбиття Руспіні, для кожного з них буде визначений центр розбиття i_k, функція належності A_k і значення U_k, що представляє собою точки, які належать тренду динамічного ряду: $U_k = \sum X_{t_i} A_k(t_i) / \sum A_k(t_i), k = 1, \ldots, n$.

У випадку, якщо значення тренду U_i необхідно відшукати для значень які відмінні від центрів нечіткого розбиття, тоді використовуються значення F-

компоненти, що передує шуканому - U_{d0} та наступне за ним U_{d1}, після чого розраховують $U_i^{(1)}$ за допомогою інтерполяції :

Середню сезонну хвилю пропонується подати у вигляді $V_j^{(1)} = \sum_{i=1}^{m} \bar{l}_{ij} / m$, де $\bar{l}_{ij}$ середні значення місячних відхилень розраховуються за допомогою ділення $l_{ij}^{'}$ - окремі місячні відхилення на середнє квадратичне відхилення за рік.

Отримав подальший розвиток метод фільтрації компонент неоднорідних часових послідовностей, в якому, на відміну від існуючих, для виявлення тренду початкова послідовність ітеративно розбивається на скінчену кількість нечітких розділів, для кожного з яких розраховується усереднене значення із врахуванням функції належності, яка асоційована із нечітким розділенням, що дозволяє підвищити ефективність оцінювання зміни стану предметної області за рахунок фільтрації коливань різних періодів та виділення трендової складової:

Етап 1. Розділимо динамічний ряд на n нечітких розбиттів Руспіні розміром T_0. Для цього визначимо n рівновіддалених точок з індексом t_k, де $k = 1..n$, які належать до цих нечітких частин, причому $t = 1..N$, $t_k = 1 + h(k-1), k = 1,...n,$ де можна відзначити, що виконуються умови: $N > n, h = (N-1)/(n-1)$.

Етап 2. Визначимо n базисних функцій, $A_1 \ldots A_n$ які покривають всі частини динамічного ряду та відповідають таким умовам, що A_k - неперервна, A_k - монотонно зростає на $[t_{k-1}, t_k]$ і монотонно спадає на $[t_k, t_{k+1}]$, для кожної функції виконуються умови $A_k : [1..N] \to [0,1], A_k(t_k) = 1$.

Визначимо, що $A_k(t) = 0$, якщо $t \notin (t_{k-1}, t_{k+1})$, при цьому вважаємо що $t_0 = t_1 = 1, t_{n+1} = t_n = N$ і для всіх $t \in [1..N]$ виконується рівняння: $\sum A_k(t) = 1$

Виходячи з перерахованих вище умов, для даного випадку скористаємося трикутною базисною функцією.

Етап 3. Використовуючи базисні функції перетворимо даний динамічний ряд X в кортеж з n дійсних чисел $[U_1...U_n]$, які визначаються за допомогою F-перетворення.

Етап 4. Для кожного року i обчислюється σ_i - середнє квадратичне відхилення, на яке діляться окремі місячні (квартальні) відхилення відповідного року.

Етап 5. З «нормованих» таким чином відхилень обчислюється в першому наближенні середня сезонна хвиля $V_j^{(1)} = \sum_{i=1}^{m} \bar{l}_{ij} / m$.

Етап 6. Середня сезонна хвиля множиться на середнє квадратичне відхилення кожного року і віднімається від рівнів початкового емпіричного ряду: $\overline{=}{x}_{ij} = x_{ij} - V_j^{(1)} \cdot \sigma_i$.

Етап 7. Цей ряд знову піддається нечіткому згладжуванню, для місячних даних розмір розбиття Руспіні повинен буде складати чотири або шість точок, в залежності від того, на скільки інтенсивні дрібні коливання). В результаті одержується нова оцінка тренда $U_i^{(2)}$.

Етап 8. Розрахуємо відхилення початкового емпіричного ряду $\{x_i\}$ від ряду $\{U^{(2)}\}$, одержаного на етапі 5: $l_i^{(2)} = x_i - U_i^{(2)}$

Після цього ці значення знову піддаються обробці відповідно до етапів 2 і 3 метода до досягнення заданої точності у виділенні сезонної хвилі.

Інформаційна технологія реалізована у вигляді програмного модуля, робота якого ґрунтується на запропонованій тренд-сезонній моделі даних впорядкованих за часом значень із врахуванням функцій належності, асоційованих із нечіткими розділами, методі фільтрації компонент динамічного ряду із врахуванням крайових значень та методі визначення значущих чинників нечіткої регресійної моделі неоднорідних даних, який містить етап відбору підмножини значущих чинників.

Впроваджено інформаційну технологію у вигляді програмного засобу у діяльність ТОВ «Ендейвер», м.Полтава та в діяльність Головного управління

національної поліції Харківської області. Результати дисертаційної роботи впроваджені в навчальний процес кафедри програмної інженерії ХНУРЕ.

Ключові слова:нечіткий регресійний аналіз, тренд-сезонна модель, аналіз даних, тренд, неоднорідні послідовності, моделі соціальних явищ, часові ряди.

Список публікацій здобувача

1. A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta. A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features. Cybernetics and Systems Analysis, Springer US, 2016. V. 52, Issue 4, P. 641–646, DOI:10.1007/s10559-016-9867-5 (Входить до міжнародної наукометричної бази SCOPUS).

2. Зацеркляний М.М., Єрохін А.Л., Бабій А.С.,Турута О.П. Розробка методу виявлення сезоних коливань з застосуванням нечіткого згладжування на базі F-перетвореннння. Біоніка інтелекту, Харків: ХНУРЕ, 2011. №2011’2. С. 89 – 93

3. Бабій А.С., Зацеркляний М. М.. Автоматизація аналізу сезонних коливань рівня злочинності. Право і безпека. Харків: ХНУВС, 2005. Т4. № 3. С.163-166.

4. Бабій А.С., Зацеркляний М. М.. Аналіз тенденцій розвитку злочинності. Системи обробки інформації. Харків: ХУПС, 2007. №4. С. 153-155.

5. Зацеркляний М.М., Бабій А.С. Інформаційна система моделювання впливу чинників злочинності. Право і Безпека. Харків: ХНУВС, 2008. Т.7. № 2. С. 204-209.

6. Зацеркляний М. М., Бабій А.С. Попередній аналіз даних у системах обробки інформації про скоєні злочини. Право і Безпека. Харків: ХНУВС, 2009. № 1. С. 269-272.

7. Лановий О.Ф., Бабій А.С. Статистичний аналіз злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2006. №19. С. 24 – 30

8. Бабій А.С. Програмна система для аналізу злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2007. №19. С. 12 – 16

9. Бабій А.С. Автоматизація управління діяльністю правоохоронних органів. Державне управління та місцеве самоврядування: тези VII міжнародного наукового конгресу, 29-30 березня 2007 р. Харків:НАДУ, 2007. С. 20-22

10. Єрохін А.Л., Бабій А.С.,Турута О.П. Спеціальна інформаційна система для виклику екстрених служб в Україні. ХНУРЕ, Збірник праць IV міжнародної науково-практичної конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», 24-25 травня 2011. Харків: ХНУРЕ, 2011. С. 163

11. Бабій А.С. Побудова СППР для оцінювання злочинності. Збірник праць II міжнародної науково-технічної конференції «Інформаційні технології в навігації і управлінні», 16-17 липня 2011 р., Київ: «ДП ЦНДІ НіУ», 2011. С. 41

12. Зацеркляний М.М., Бабій А.С. Застосування методу найменших кутів для аналізу чинників злочинності. Материалы международной научно-технической конференции “Информационные системы и технологии”, Харьков: НТМТ, 2012, С. 37

13. Петров К.Е., Зацеркляний М.М., Бабій А.С. Оцінювання злочинності із врахуванням нечіткості. Спеціальна техніка у правоохоронній діяльності, Матеріали V Міжнародної науково-практичної конференції, Київ: НАВС, 2012, С.79

14. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta .Usage of F-transform to finding informative parameters of rhinomanometric signals. Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), 2015 Xth International. P. 129-132, DOI:10.1109/STC-CSIT.2015.7325449 (Входить до міжнародної наукометричної бази SCOPUS)

15. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta. Processing and analysis of rhinomanometric signals by F-transform approximation - 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). P. 314 - 317, DOI: 10.1109/DSMP.2016.7583566 (Входить до міжнародної наукометричної бази SCOPUS)

16. A. Yerokhin, O. Turuta, A. Babii, A, Nechyporenko, I. Mahdalina. Usage of phase space diagram to finding significant features of rhinomanometric signals. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), P. 70 - 72, DOI: 10.1109/STC-CSIT.2016.7589871 (Входить до міжнародної наукометричної бази SCOPUS)

ABSTRACT

Babii A.S. Models, methods and intelligent information technology for analysis of heterogeneous sequences. – Qualifying scientific work as a manuscript.

A thesis for obtaining the candidate degree in technical sciences in the speciality 05.13.06 «information technology». – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2017.

Solution of the actual scientific and practical problem of the development of models, methods and intelligent information technology for the analysis of heterogeneous data sequences to assess the current state of the domain for information-analytical system.

The object of this thesis research is the process of analisys data of heterogeneous sequences to assess the current state of the domain. The subject of the research is models, methods and intelligent information technology for the analysis of heterogeneous data sequences to assess the current state of the domain based on small datasets.

Research methods: Methods of the fuzzy set theory and regression analysis are used for research and development of method of important factor selection for fuzzy regression model. Methods of time series analysis and fuzzy approximation theory are used for development of time-series component filtering method. Mathematical statistic elements are used for experiments analysis and implementation.

The practical significance of received results. Information technology of heterogeneous data analysis is were improved at the basis of proposed models and methods. It differs from existing by giving abilty of additionally taking into account the information presented in the form of fuzzy data, during the construction of the domain model and the selection of significant factors, which provides an opportunity for analysis in the conditions of short data samples.

The method is proposed in the thesis research to determine significant factors of the fuzzy regression model of heterogeneous data; this method, in contrast to the actual methods, includes stages of factors selection in accordance with the criterion of

equal significance of angles of deviation between the vector of errors and vectors of variables, as well as selection of subsets of significant factors with coefficients that exceed the threshold value, that allows avoiding any over-fit of the fuzzy linear regression and receiving the subset of significant factors on the basis of the finite number of iterations:

Stage 1. In order to determine the membership function we use the approach based on the indirect method of membership function plotting.

Stage 2. We represent clear data about the factors in the form of the matrix of p values for explaining variables in *n* observations, and the data about centers of fuzzy magnitude $\tilde{Y}$ - in the form of values vector.

Stage 3. Let us use the approach based on the method of coefficients selection in accordance with the criterion of equal significance of angles of deviation between the vector of errors and vectors of variables. We define the initial assessed value, $\hat{\mu}_A = 0$, of the values vector for the dependant variable *y*.

Stage 4. Now calculate the correlation vector: $\hat{c} = X^T(y - \hat{\mu}_A)$.

Stage 5. We find the current set of indices, *A*, that fits to the distinctive features with the maximal absolute values of correlation: $A = \{ j : |\hat{c}_j| = \hat{C} \}$, here $\hat{C} = \max_{j=1,...,n} \{|\hat{c}_j|\}$.

Stage 6. Then find $s_j = sign(\hat{c}_j)$ for $j \in A$. Calculate matrixes X_A, ψ_A as follows: $X_A = \left[s_{j_1} x_{j_1}, ..., s_{j_{|A|}} x_{j_{|A|}} \right], j = (j_1, ..., j_{|A|}) \in A$, $\psi_A = (1_A^T \varsigma_A^{-1} 1_A)^{-\frac{1}{2}}$, here $s_j \in \{+1, -1\}$ and $|A|$- potency of set A (number of values for *A* set), $\varsigma = X_A^T X_A$, 1_A – unit matrix of $1 \times |A|$ size.

Stage 7. Now calculate vector $a = X^T u_A$, here $u_A = X_A w_{A'}$, $w_A = \psi_A \varsigma_A^{-1} 1_A$.

Stage 8. Calculate value $\hat{\gamma} = \min_{j \in A}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\psi_A - a_j}, \frac{\hat{C} + \hat{c}_j}{\psi_A + a_j} \right\}$, here minimum is taken

by all positive values for every j.

Stage 9. Find value $\hat{\mu}_A$ for the following iteration: : $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$.

Stage 10. The process shall be repeated n times (here n is the number of factors), starting with Stage 4. For every iteration Mallouse coefficient, *Cp* is calculated.

Stage 11. In order to design a fuzzy regression model we select a set of coefficients that corresponds to the minimal value of *Cp* coefficient.

Then the method is applied that solving the problem of linear regression model development, with the use of the selected set of factor variables.

The development of the trend-seasonal model of heterogeneous sequences has gone further; in contrast to actual models its trend component is given in the form of interpolated averaged values with regard to membership function, which is associated with every fuzzy class that permits to use this model for short series without any loss of the boundary data and, therefore, improve the modeling accuracy:

The model of the seasonal phenomenon is proposed on the basis of the decomposition approach, which may be represented in the form that is described in details in the second section: $X_i = U_i + V_i + \varepsilon_i$ *here i=1,…,N.*

Trend component, U_i, may be found with the help of the iterative use of F-transformation, that is, by making n fuzzy Ruspini partitions, and for each of them the centre of partition, i_k, membership function, A_k and values U_k, shall be determined, that represent points belonging to the trend of the time series: $U_k = \sum X_{t_i} A_k(t_i) / \sum A_k(t_i)$, $k = 1,...,n$.

If the trend value, Ui , must be found for values which are not the centers of fuzzy partition, then the values of F-component shall be used, which are the predecessor of the target value, U_{d0}, and the next one, U_{d1}, and then calculate $U_i^{(1)}$ using the interpolation:

We propose to represent the average seasonal wave in the form $V_j^{(1)} = \sum_{i=1}^{m} \bar{l}_{ij} / m$

here $\bar{l}_{ij}$ are average values of monthly deviations calculated by dividing l'_{ij}, specific monthly deviations, by the mean square deviation for one year.

The method of filtration of components of heterogeneous time sequences received its further development, in contrast to actual methods the initial sequence is broken into a finite number of fuzzy segments in order to find out the trend, and for each of these segments the averaged value is calculated with regard to the membership function associated with the fuzzy segment, that permits to take into account the boundary values of the series in order to find the trend component:

Stage 1. We divide the time series into *n* fuzzy Ruspini partitions of T_0 size. For this purpose determine *n* equally-spaced points with index t_k, here $k = 1..n$, that belong to these fuzzy segments, provided that $t = 1..N, t_k = 1 + h(k-1), k = 1,...n,$, when conditions $N > n, h = (N-1)/(n-1)$ are met.

Stage 2. Now determine *n* basis functions, $A_1 \ldots A_n$, that cover all parts of the times series and comply with the following requirements: A_k is a continuous function, A_k increases monotonically by $[t_{k-1}, t_k]$ and decreases monotonically by $[t_k, t_{k+1}]$, and every function meets the conditions $A_k : [1..N] \to [0,1], A_k(t_k) = 1$. Let us specify that $A_k(t) = 0$, if $t \notin (t_{k-1}, t_{k+1})$, in addition assume that $t_0 = t_1 = 1, t_{n+1} = t_n = N$ and equation $\sum A_k(t) = 1$ is true for all $t \in [1..N]$. According to the above conditions, in this case we shall use the triangular basis function.

Stage 3. Using basis functions we transform given time series X into a list of n real numbers, $[U_1 ... U_n]$, determined with the use of F-transformation.

Stage 4. The root-mean-square deviation, σ_i, is calculated for every year, *i*, and specific monthly (quarterly) deviations for the respective year shall be divided by this value.

Stage 5. Basing on these "normalized" deviations we calculate the mean seasonal wave, $V_j^{(1)} = \sum_{i=1}^{m} \bar{l}_{ij} / m$, at a first approximation.

Stage 6. The mean seasonal wave shall be multiplied by the root-mean-square deviation for every year and subtracted from the levels of the initial empirical series: $\overline{=}$
$x_{ij} = x_{ij} - V_j^{(1)} \cdot \sigma_i$.

Stage 7. Then this series undergoes fuzzy smoothing again; for monthly data the size of the Ruspini partition shall include four or six points depending on intensity of small oscillations). As a result we get a new estimate for trend $U_i^{(2)}$.

Stage 8. Let us calculate deviation of the initial empirical series $\{x_i\}$ from series $\{U^{(2)}\}$, received at Stage 5: $l_i^{(2)} = x_i - U_i^{(2)}$

Then these values undergo processing again in accordance with Stages 2 and 3 of the method until the required accuracy in distinguishing the seasonal wave is achieved.

Information technology is implemented in the form of a program module, it is based on the proposed trend-seasonal model of data of time-ordered values, taking into account membership functions associated with fuzzy sections, the method of filtering components of a dynamic series taking into account boundary values and a method for determining the important factors of a fuzzy regression model of heterogeneous data, that contains the stage of selecting a subset of significant factors.

Information technology were implemented as a software tool in the activities of LLC "Endeavor", Poltava and the activities of the Central department of the national police of the Kharkiv region. Results of dissertation work were implemented in the educational process of the department of software engineering NURE.

Keywords: fuzzy regression analysis, trend-season model, data analysis, trend, heterogeneous sequences, model of social phenomena, time series.

List of publications of the applicant

1. A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta. A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features. Cybernetics and Systems Analysis, Springer US, 2016. V. 52, Issue 4, P. 641–646, DOI:10.1007/s10559-016-9867-5 (Входить до міжнародної

наукометричної бази SCOPUS).

2. Зацеркляний М.М., Єрохін А.Л., Бабій А.С.,Турута О.П. Розробка методу виявлення сезоних коливань з застосуванням нечіткого згладжування на базі F-перетвореннння. Біоніка інтелекту, Харків: ХНУРЕ, 2011. №2011'2. С. 89 – 93.

3. Бабій А. С., Зацеркляний М. М.. Автоматизація аналізу сезонних коливань рівня злочинності. Право і безпека. Харків: ХНУВС, 2005. Т4. № 3. С.163-166.

4. Бабій А. С., Зацеркляний М. М.. Аналіз тенденцій розвитку злочинності. Системи обробки інформації. Харків: ХУПС, 2007. №4. С. 153-155.

5. Зацеркляний М. М., Бабій А. С. Інформаційна система моделювання впливу чинників злочинності. Право і Безпека. Харків: ХНУВС, 2008. Т.7. № 2. С. 204-209.

6. Зацеркляний М. М., Бабій А. С.. Попередній аналіз даних у системах обробки інформації про скоєні злочини. Право і Безпека. Харків: ХНУВС, 2009. № 1. С. 269-272.

7. Лановий О.Ф., Бабій А.С. Статистичний аналіз злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2006. №19. С. 24 – 30.

8. Бабій А.С. Програмна система для аналізу злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2007. №19. С. 12 – 16.

9. Бабій А.С. Автоматизація управління діяльністю правоохоронних органів. Державне управління та місцеве самоврядування: тези VII міжнародного наукового конгресу, 29-30 березня 2007 р. Харків:НАДУ, 2007. С. 20-22.

10. Єрохін А.Л., Бабій А.С.,Турута О.П. Спеціальна інформаційна система для виклику екстрених служб в Україні. ХНУРЕ, Збірник праць IV міжнародної науково-практичної конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», 24-25 травня 2011. Харків: ХНУРЕ, 2011. С. 163.

11. Бабій А.С. Побудова СППР для оцінювання злочинності. Збірник праць II міжнародної науково-технічної конференції «Інформаційні технології в навігації і управлінні», 16-17 липня 2011 р., Київ: «ДП ЦНДІ НіУ», 2011. С. 41.

12. Зацеркляний М.М., Бабій А.С. Застосування методу найменших кутів для аналізу чинників злочинності. Материалы международной научно-технической конференции “Информационные системы и технологии”, Харьков: НТМТ, 2012, С. 37.

13. Петров К.Е., Зацеркляний М.М., Бабій А.С. Оцінювання злочинності із врахуванням нечіткості. Спеціальна техніка у правоохоронній діяльності, Матеріали V Міжнародної науково-практичної конференції, Київ: НАВС, 2012, С.79.

14. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta .Usage of F-transform to finding informative parameters of rhinomanometric signals. Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), 2015 Xth International. P. 129-132, DOI:10.1109/STC-CSIT.2015.7325449 (Входить до міжнародної наукометричної бази SCOPUS).

15. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta. Processing and analysis of rhinomanometric signals by F-transform approximation - 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). P. 314 - 317, DOI: 10.1109/DSMP.2016.7583566 (Входить до міжнародної наукометричної бази SCOPUS)

16. A. Yerokhin, O. Turuta, A. Babii, A, Nechyporenko, I. Mahdalina. Usage of phase space diagram to finding significant features of rhinomanometric signals. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), P. 70 - 72, DOI: 10.1109/STC-CSIT.2016.7589871 (Входить до міжнародної наукометричної бази SCOPUS)

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

БД – база даних

IAC – інформаційно-аналітичні системи

IT – інформаційна технологія

СУБД – система управління базами даних

API – Application Program Interface

UML – Unified Markup Language

XML – eXtensible Markup Language

KDD – Knowledge Discovery in Databases

CRISP-DM – Cross-Industry Standard Process for Data Mining

SEMMA – list of sequencial steps (Sample, Explore, Modify, Model, and Assess) for implementation of data mining applications

REST – Representational State Transfer

HTTP – HyperText Transfer Protocol

FTP – File Transfer Protocol

SQL – Structured Query Language

ВСТУП

Актуальність теми.

Зміни, що відбуваються в сучасному суспільстві, спричинені різними факторами, в тому числі і швидким зростанням рівня проникнення інформаційних технологій в різні види діяльності людини.

Існуючі інформаційно-аналітичні системи та інструментальні засоби інтелектуального аналізу даних такі як IBM Cognos Analytics, Microsoft SRSS, MicroStrategy Analytics Platform, Oracle BI, SAP Business Warehouse, дозволяють здійснювати узагальнення великих масивів даних та формування інформації, на основі якої можуть прийматися рішення і оцінюватися поточний стан предметної області із застосуванням широкого діапазону моделей і методів обробки даних [1-4].

Разом з тим часто постає проблема, не охоплена існуючими реалізаціями інформаційно-аналітичних систем та інструментальних засобів інтелектуального аналізу даних, а саме обробка різнорідних даних, частина з яких виражена у вигляді чітких кількісних значень, отриманих в результаті вимірювань, а друга частина має суб’єктивну та нечітку природу.

Така задача може виникати, наприклад, в сфері наук про людину, її особисту поведінку, умови та процеси у суспільстві. Ці відомості можуть бути формалізовані у вигляді лінгвістичних змінних або нечітких множин чи інтервальних оцінок.

Дослідженням обробки даних різнорідної природи походження, узгодженню експертних оцінок, основним положенням теорії нечітких множин та її застосуванню для вирішення різних наукових задач були присвячені роботи таких вчених: Л. Заде[5], Т. Сааті[6], Е. Руспіні[7], І.Перфильєвої[8], Г.Танаки[9], П. Курран[10], Л. Г. Раскіна[11], Є.В. Бодянського[12] та інших.

Сучасні інформаційно-аналітичні системи, як правило, мають інтерфейси для створення додаткових користувацьких модулів, що дає можливість розширювати функціональність існуючих, в тому числі і користувацьких,

інформаційних систем, розроблених на основі поширених платформ інтелектуального аналізу даних.

Враховуючи невідповідність між можливостями існуючих інформаційних технологій обробки відомостей, представлених у вигляді неоднорідних послідовностей даних і розповсюдженості потреби в їх аналізі, виникає завдання відбору значущих чинників при проведенні нечіткого багатофакторного регресійного аналізу і завдання фільтрації неоднорідних часових послідовностей даних.

Таким чином, розробка моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей є актуальною науково-практичною задачею.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота виконана на кафедрі програмної інженерії Харківського національного університету радіоелектроніки відповідно до завдань НДР «Теорія, методи і моделі управління життєвим циклом інтелектуальних інформаційних середовищ регіональних соціо-економічних об'єктів» (№ ДР 0115U002430), де здобувач брав участь як виконавець окремих розділів.

Мета та задачі дослідження.

Метою дисертаційної роботи є розробка моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для підвищення ефективності оцінювання поточного стану предметних областей в інформаційно-аналітичних системах.

Для досягнення поставленої мети необхідно здійснити вирішення таких задач:

- провести дослідження існуючих моделей і методів аналізу неоднорідних послідовностей даних;
- вдосконалити тренд-сезонні моделі неоднорідних часових послідовностей;
- розробити метод відбору значущих чинників при побудові нечіткої

багатофакторної регресії для даних, що представлені у вигляді неоднорідних послідовностей;

- вдосконалити метод фільтрації компонент неоднорідних послідовностей даних;

- розробити інтелектуальну інформаційну технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметної області та виконати програмну реалізацію й впровадження результату дослідження при вирішенні практичних задач.

Об'єктом дослідження є процес аналізу неоднорідних послідовностей при оцінюванні поточного стану предметної області.

Предметом дослідження є моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей для оцінювання поточного стану предметної області на основі коротких вибірок даних.

Методи дослідження: при розробці та дослідженні методу визначення значущих чинників нечіткої регресійної моделі були використані методи теорії нечітких множин та регресійного аналізу. При розробці моделей і методів фільтрації компонент неоднорідних часових послідовностей були використані методи нечіткої апроксимації даних та аналізу динамічних рядів. При проведенні та аналізі результатів експериментальних досліджень були використані елементи математичної статистики.

Наукова новизна одержаних результатів.

В результаті проведення досліджень одержані такі нові результати:

1. Вперше запропоновано метод визначення значущих чинників нечіткої регресійної моделі неоднорідних послідовностей даних, який, на відміну від існуючих, містить етапи підбору коефіцієнтів за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних та відбору підмножини значущих чинників з коефіцієнтами, що перевищують порогове значення та дозволяє запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій.

2. Отримав подальший розвиток метод фільтрації компонент неоднорідних часових послідовностей, в якому, на відміну від існуючих, для виявлення тренду початкова послідовність ітеративно розбивається на скінчену кількість нечітких розділів, для кожного з яких розраховується усереднене значення із врахуванням функції належності, яка асоційована із нечітким розділенням, що дозволяє підвищити ефективність оцінювання зміни стану предметної області за рахунок фільтрації коливань різних періодів та виділення трендової складової.

3. Отримала подальший розвиток тренд-сезонна модель неоднорідних послідовностей, в якій, на відміну від існуючих моделей, трендова складова подається у вигляді інтерпольованих усереднених значень із врахуванням функції належності, яка асоційована із кожним нечітким розділенням, що дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень і тим самим підвищити ефективність моделювання стану предметних областей в інформаційно-аналітичних системах.

Практична значимість одержаних результатів: на основі запропонованих моделей та методів удосконалено інформаційну технологію аналізу даних неоднорідних послідовностей, яка, на відміну від існуючих технологій, надає можливість при побудові моделі предметної області додатково враховувати відомості подані у вигляді нечітких даних, та здійснювати відбір значущих чинників, що надає можливість проведення аналізу в умовах коротких вибірок даних.

Інформаційна технологія реалізована у вигляді програмного модуля, робота якого ґрунтується на запропонованій тренд-сезонній моделі даних впорядкованих за часом значень із врахуванням функцій належності, асоційованих із нечіткими розділами, методі фільтрації компонент динамічного ряду із врахуванням крайових значень та методі визначення значущих чинників нечіткої регресійної моделі неоднорідних даних, який містить етап відбору підмножини значущих чинників.

Впроваджено інформаційну технологію у вигляді програмного засобу у

діяльність ТОВ «Ендейвер»,м.Полтава (акт впровадження від 14.03.2017) та в діяльність Головного управління національної поліції Харківської області (акт впровадження від 20.12.2016).

Результати дисертаційної роботи впроваджені в навчальний процес кафедри програмної інженерії ХНУРЕ (акт впровадження від 15.03.2017).

Особистий вклад здобувача. Усі результати, що виносяться на захист отримані здобувачем особисто. У роботах опублікованих у співавторстві, здобувачу належать: у роботі [13] – метод відшукування значущих чинників на основі методу найменших кутів для побудови нечіткої регресійної моделі; у роботі [14] – метод виявлення сезонних коливань із застосуванням нечіткого згладжування на базі F-перетворення; у роботі [15] – модифікований метод аналізу сезонних коливань злочинності; у роботі [16] – підхід для аналізу тенденцій розвитку злочинності; у роботі [17] – інформаційна система моделювання впливу чинників злочинності; у роботі [18] – метод попереднього аналізу даних в системах обробки інформації про скоєні злочини; у роботі [19] – статистична модель аналізу злочинності; у роботі [20] – інформаційна система для виклику екстрених служб; у роботі [21] – модифікований метод визначення чинників злочинності на основі методу найменших кутів; у роботі [22] – метод оцінювання злочинності із врахуванням нечіткості; у роботі [23] – застосування F-перетворення для аналізу риноманометричних сигналів; у роботі [24] – застосування F-перетворення для апроксимації риноманометричних сигналів; у роботі [25] – застосування F-перетворення як один з підходів для відшукування значущих ознак риноманометричних сигналів;

Апробація роботи. Основні положення дисертаційної роботи доповідалися на таких міжнародних конференціях і форумах:

- XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), (Львів, 2016 р.);

- First International Conference on Data Stream Mining & Processing (DSMP), (Львів, 2016 р.);

- Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), (Львів, 2015 р.);

- V Міжнародній науково-практичній конференції «Спеціальна техніка у правоохоронній діяльності» (Київ, 2012 р.);

- Міжнародній науково-технічній конференції «Інформаційні системи і технології» (Харків, 2012 р.);

- II міжнародній науково-технічній конференції «Інформаційні технології в навігації і управлінні» (Київ, 2011 р.);

- IV міжнародній науково-практичній конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», (Харків, 2011р.);

- VII міжнародному науковому конгресі «Державне управління та місцеве самоврядування», (Харків, 2007 р.);

Публікації. За результатами дисертаційного дослідження опубліковано 16 наукових праць[13-28] (серед них 3 – одноосібних[26-28]), у тому числі 7 статей у наукових фахових виданнях України з технічних наук[14-19,27] і 1 стаття – за кордоном у виданні, що входить до міжнародної наукометричної бази Scopus[13], 8 тез у матеріалах міжнародних конференцій[20-26, 28] (з них 3 включено до наукометричної бази Scopus[23-25]).

Структура дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел зі 167 найменувань на 16 сторінках та 1 додаток на 4 сторінках, а також містить 25 рисунків і 3 таблиці. Загальний обсяг роботи складає 156 сторінок, включаючи 117 сторінок основного тексту.

В першому розділі проведено аналіз існуючих публікацій щодо моделей, методів та інформаційних технологій обробки великих масивів даних з різнорідною природою походження. Розглянуто основні проблеми, які виникають під час аналізу даних представлених у вигляді неоднорідних послідовностей. Розглянуто методи побудови нечітких регресійних моделей, умови їх використання та обмеження, що існують для застосування цих

моделей. Сформульовані мета та задачі дисертаційних досліджень.

Другий розділ присвячено розробці моделей неоднорідних послідовностей, зокрема вдосконаленої тренд-сезонної моделі. Описуються показники, що можуть використовуватися для аналізу неоднорідних послідовностей, методи вимірювання взаємозв'язку між неоднорідними послідовностями та моделі аналізу неоднорідних послідовностей.

Запропоновано тренд-сезонну модель із використанням елементів теорії нечітких множин.

Для аналізу багатовимірних даних впорядкованих сукупностей чітких та нечітких значень розроблена модель із врахуванням даних попарних порівнянь експертних оцінок для непрямого визначення параметрів функції належності.

Третій розділ присвячено розробці методів аналізу неоднорідних послідовностей. Запропонований метод відбору значущих чинників при побудові нечіткої багатофакторної регресії для даних, що представлені у вигляді неоднорідних послідовностей. Вдосконалений метод фільтрації компонент неоднорідних послідовностей даних.

Четвертий розділ присвячено розробці інтелектуальної інформаційної технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметної області та виконанню програмної реалізації і впровадження результату дослідження для вирішення практичних задач.

У висновках приведені основні результати дисертаційної роботи.

Список використаних джерел у даному розділі наведенj у повному списку використаних джерел під номерами: [1-28].

1 ДОСЛІДЖЕННЯ ІСНУЮЧИХ МОДЕЛЕЙ ТА МЕТОДІВ АНАЛІЗУ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

1.1 Аналіз особливостей неоднорідних послідовностей

Неперервне збирання та аналіз інформації різних типів та джерел походження дає можливість розробляти і приймати попереджувальні управлінські рішення, здійснювати повсякденний контроль за функціонуванням організації із головних напрямків її діяльності, своєчасно коригувати розміщення та використання наявних ресурсів і доступних засобів.

Потрібність, складність і трудомісткість цієї роботи вимагають пошуку нетрадиційних методів, нових заходів і використання всіх доступних можливостей.

Таким чином, на сьогоднішній день актуальними є інформаційні системи і технології, які інтегрують у собі сучасні методи взаємодії комп'ютер–особа, що приймає рішення, моделі та методи прийняття рішень, методи прогнозування та менеджменту процесів різної природи.

Мета даного розділу полягає в аналізі предметної області, визначенні задач, моделей і методів опрацювання інформації для подальшого створення інтелектуальної інформаційної технології для аналізу неоднорідних послідовностей.

Відповідно до [29], інформаційно-аналітичні системи (ІАС) забезпечують перетворення деталізованих даних в узагальнену інформацію, яка є придатною для прийняття рішень.

ІАС отримують інформацію з різних джерел і за допомогою моделей і методів аналізу даних надають відомості, що використовуються під час прийняття рішень людиною. ІАС можуть використовуватися для різних предметних областей.

Таким чином, ключовою характеристикою, що впливає на ефективність ІАС є моделі і методи аналізу даних, які використовуються для переробки

інформації.

Розглядаючи ті види даних, що обробляються ІАС та які досить часто подаються у вигляді неоднорідних послідовностей, можна виділити декілька досить розповсюджених прикладів відомостей такого роду: дані медичних обстежень та анамнезу, дані, зібрані з датчиків та опитувань користувачів або експертів, соціологічні дані та державні статистичні дані, у тому числі відомості про рівні злочинності та інше.

Але зрозуміло, що без певного розуміння природи явища та методики статистичного вимірювання показників, які представлені у вигляді кількісних характеристик, неможливе створення якісної моделі.

Наприклад, розглядаючи ІАС, призначені для обробки статистичних відомостей про функціонування соціуму, можна виділити для дослідження дані про скоєні злочини. Необхідно чітко розуміти природу отримання такої інформації та обставини, що можуть вплинути на процес обліку та вимірювання кримінологічних характеристик, які виражені у вигляді кількісних показників.

Злочинність у сучасному науковому розумінні – складне соціальне явище з ознаками системності, яке є наслідком взаємодії багатьох негативних за напрямком розвитку суспільства і особи чинників [30].

Під чинниками злочинності розуміються такі явища суспільного життя, що породжують злочинність, підтримують її існування, викликають її зростання, тобто детермінують злочинність [31].

Усі форми подання необхідної інформації потребують заходів, спрямованих на якісний відбір саме необхідних даних з усієї множини доступних відомостей.

Однією з проблем, яка суттєво може завадити ефективному використанню статистичних обліків, є неповне відображення кількісних показників. Такого роду випадки можуть бути спричинені різними факторами.

Як правило, інформація про невраховані відомості (наприклад, у випадку злочинності – латентну злочинність) є доступною серед експертного середовища і осіб, що безпосередньо стикаються з проблемою, яка

досліджується.

У такому разі, використання і отримання такої інформації потребує додаткових зусиль, незважаючи на наявність інформації про існування неврахованих даних.

Негативний ефект від приховування злочинів головним чином пов’язаний із тим, що інформація втрачає об’єктивність, що перешкоджає процесам управління державними органами, покликаними боротися із злочинністю. У свою чергу це знижує якість управлінських рішень стосовно діяльності правоохоронної системи, які були прийняті, ґрунтуючись на необ’єктивних статистичних даних.

Джерелом первинної інформації, яка в таких випадках повинна бути досліджена, можуть бути відомості, зафіксовані на основі офіційного обліку або в журналах, базах даних і комп’ютерних інформаційних системах. У певних випадках такі данні можна отримати шляхом окремо проведеного дослідження у статистичних картках, в анкетах опитування громадян або під час вивчення різних видів (кримінальних, адміністративних, цивільних) справ та інших матеріалів. Такі відомості формують значні масиви даних про вимірювані одиничні випадки досліджуваних сукупностей, що потребує засобів автоматизації обробки інформації.

Досить розповсюдженою є ситуація, коли дослідник отримує з різних джерел відомості, кожна з яких описує лише певну характеристику явища. Їх сумісне використання дозволяє отримати більш повну картину процесу, що досліджується.

Додаткові можливості надає розповсюдження таких технологічних явищ як “інтернет речей” і поява підходу до роботи з інформацією “великі дані”. Але для того, щоб скористатися ними, необхідно розв'язати не лише проблеми, пов'язані з великими об’ємами обробки накопичуваних даних та швидкістю їх отримання, але і з такою проблемою, як різнорідність інформації, що поступає до дослідника.

Вона може бути викликана комплексом причин, наприклад,

особливостями збору інформації, відмінностями методик, що викликають відповідно різницю в оцінці параметрів певного явища, зміною природи явища у різні періоди вимірювання, нестійкістю довжини періоду, наявністю періодичності, або помилками, які були зумовлені різними причинами [32].

Можна виділити три основні причини, які зумовлюють різнорідність даних - це умови збору, предмет дослідження та формат даних. Це схематично відображено на рисунку 1.1.

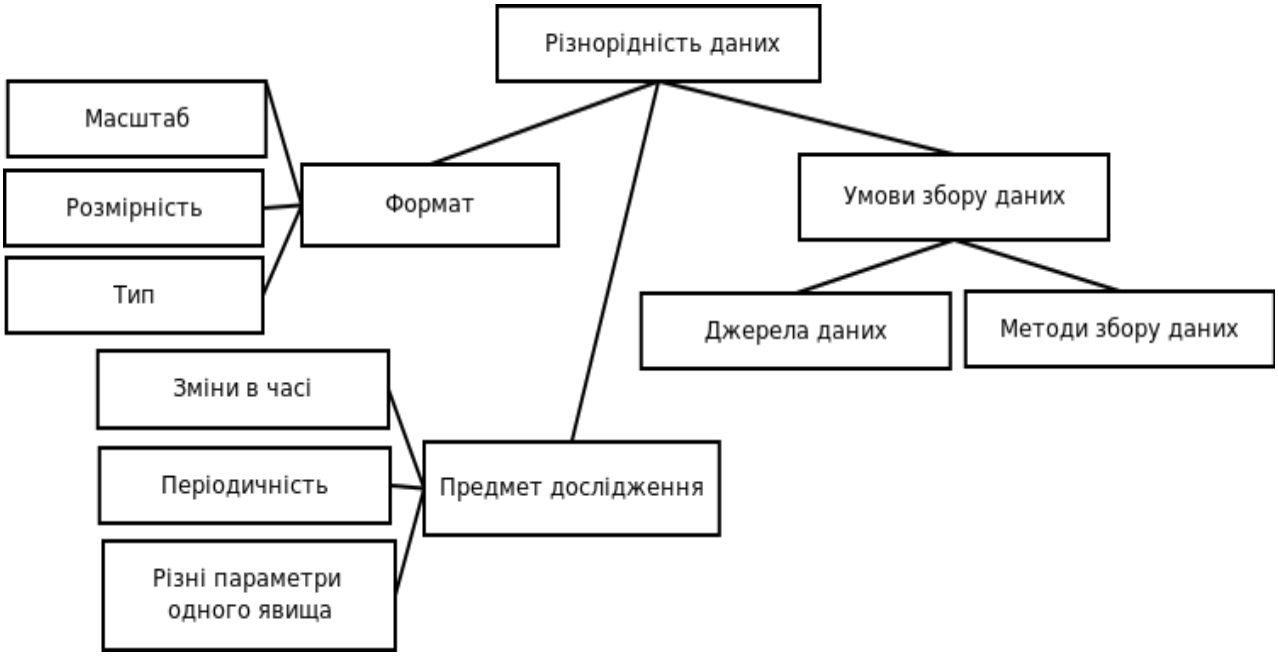

Рисунок 1.1 – Причини різнорідності даних

Розглянемо докладніше кожну групу обставин, складові якої є чинниками виникнення різнорідності.

У залежності від того, які були умови під час збору даних, ми можемо розрізняти, які були джерела отримання даних, методи, що використовувались для збору даних, їх точність, шкали вимірювання, рівень невизначеності та розподіл похибки, причини та закономірності, які викликали похибку в вимірюваннях.

Сам предмет дослідження також може бути джерелом і спричиняти різнорідність даних у випадку, якщо природа явища змінюється із часом, або вимірюються різні параметри одного і того самого явища, або вимірювання

параметру доступне лише через певні непрямі різні способи, періодичність та квазіперіодичність явища.

Формат даних, які утримуються є хіба не найбільш очевидним видом неоднорідності, наприклад, це тип даних, який зберігається, масштабування, розмірність даних. Що стосується типу даних, то інформація може бути подана у вигляді зображення, відео, звуку, одновимірних та багатовимірних рядів значень, упорядкованих у часі чи не впорядкованих у часі, тексту, лінгвістичних змінних тощо.

Різна розмірність даних полягає, наприклад, в отриманні і обробці як одновимірних даних, так і двовимірних або тривимірних. Різний масштаб та шкала оцінювання може спричиняти потребу у використанні різних типів даних або навіть способів збереження цих відомих відомостей точкових значень, нечітких, інтервальних.

Аналізу неоднорідних даних присвячена значна кількість публікацій. Наприклад, у роботі [33] пропонується застосування методів машинного навчання для об’єднання даних, отриманих з різних джерел у рамках однієї моделі, із застосуванням методу навчання розріджених ядерних функцій для групування ключових характеристик.

Тематика досліджень пов’язана з об’єднанням декількох наборів даних, отриманих із різних джерел, але описуючих одне явище, цікавила дослідників досить давно. З ранніх робіт можна відмітити [34] статті, присвячені паралельному факторному аналізу [35], кореляційному аналізу по декількох наборах даних [36].

Особливостям застосування методів, що дозволяють проводити одночасний аналіз різнорідних даних для окремих видів предметних областей, виконуючи інтеграцію, присвячений ряд наукових праць[37 – 40].

Відповідно до цих робіт, метою і основною задачею залучення таких даних є зменшення впливу недоліків одних джерел даних за рахунок збільшення впливу переваг інших джерел даних із розробкою відповідних математичних моделей і методів.

Також можна відмітити, що зважаючи на різноманітність джерел походження, кількість різних типів даних та їх комбінацій, різних варіантів наповнення та подання інформації, а також проблем пов’язаних із наявністю аномальних рівнів та неповних даних, перед дослідником постає дуже великий об’єм досліджень з досить широким колом питань.

У роботі [41] запропонований огляд основних характеристик неоднорідних послідовностей, що надходять з сенсорів, які породжують труднощі їх обробки:

1) Неточність даних. Деякий рівень неточності завжди притаманний інформації, отриманій за допомогою вимірювання.

2) Кореляція між вхідними даними, яка може бути викликана умовами збирання відомостей. Неврахування такої інформації може призвести до викривлених результатів.

3) Асоціація даних, пов’язаних з різними джерелами, та їх зведення до одного синхронізованого за часом потоку даних.

4) Наявність відхилень та хибних даних. Неоднозначність та невідповідність умов для вимірювання можуть викликати значні відхилення та появу помилкових даних.

5) Калібрація (вирівнювання) даних, отриманих у різних локальних умовах на різному базовому рівні до одного, нормалізованого базового рівня.

6) Суперечливість даних. Особливо це стосується систем визначення причинно-наслідкових зв’язків.

7) Різна модальність даних, яка може виражатись у тому, що система збирає якісно різні (аудіо, відео, символьно-цифрові) набори даних або однакові за своєю природою набори.

8) Структура обробки даних, яка може бути реалізована, як централізовано, так і децентралізовано.

9) Операційний час. Обсяг, який охоплюється датчиками, може бути дуже великим, із значною кількістю різних видів даних і часових інтервалів вимірювання.

10) Статичність чи динамічність явища. Явище може бути постійним або змінюватися із часом, зокрема, наскільки важливо використовувати саме свіжі дані і з якою частотою поновлювати їх.

11) Розмірність даних. Розмірність даних у деяких випадках може бути зменшена, припускаючи певний рівень втрати інформації, що може використовуватися для зменшення об’єму, необхідного для передачі даних, або обмеження обчислювальних можливостей.

Алгоритми обробки різнорідних послідовностей повинні використовувати надмірність та різну природу походження даних для зменшення впливу негативних характеристик.

Варто доповнити перелік характеристик, наведений вище та опублікований у роботі [41], ще одним пунктом, пов’язаним з природою джерела походження даних, а саме:

- інформація про значимі обставини вимірювання чи характеристики об’єкту дослідження, отримана від опитування людини або групи людей, яка може бути подана у вигляді нечітких даних.

Таким чином, з'являється потреба в моделях і методах інтеграційного аналізу, який дозволяє використовувати джерела даних з різнорідною природою походження, а саме отримані шляхом застосування класичних вимірювань та відомостей, отриманих від людини.

1.2 Аналіз існуючих моделей неоднорідних послідовностей

Основною метою аналізу неоднорідних послідовностей є використання різних джерел походження даних таким чином, щоб після їх комбінування точність моделі була більша, ніж у випадку використання цих даних окремо[42].

Підходи та моделі обробки різнорідних послідовностей з точки зору інтеграції даних розглянуті в роботі [43]. Їх можна розділити на напрямки в залежності від підходу, на якому ґрунтується процес прийняття рішення із

використанням даних, отриманих з різних джерел:

- Байєсівська модель;
- нечітка модель;
- нейромережева модель;
- семантична модель;
- абдуктивна модель;
- модель Демпстера – Шафера;

Також можливі і комбінації цих методів для вирішення окремих задач. Такі напрямки виведення можна схематично зобразити на рисунку 2.

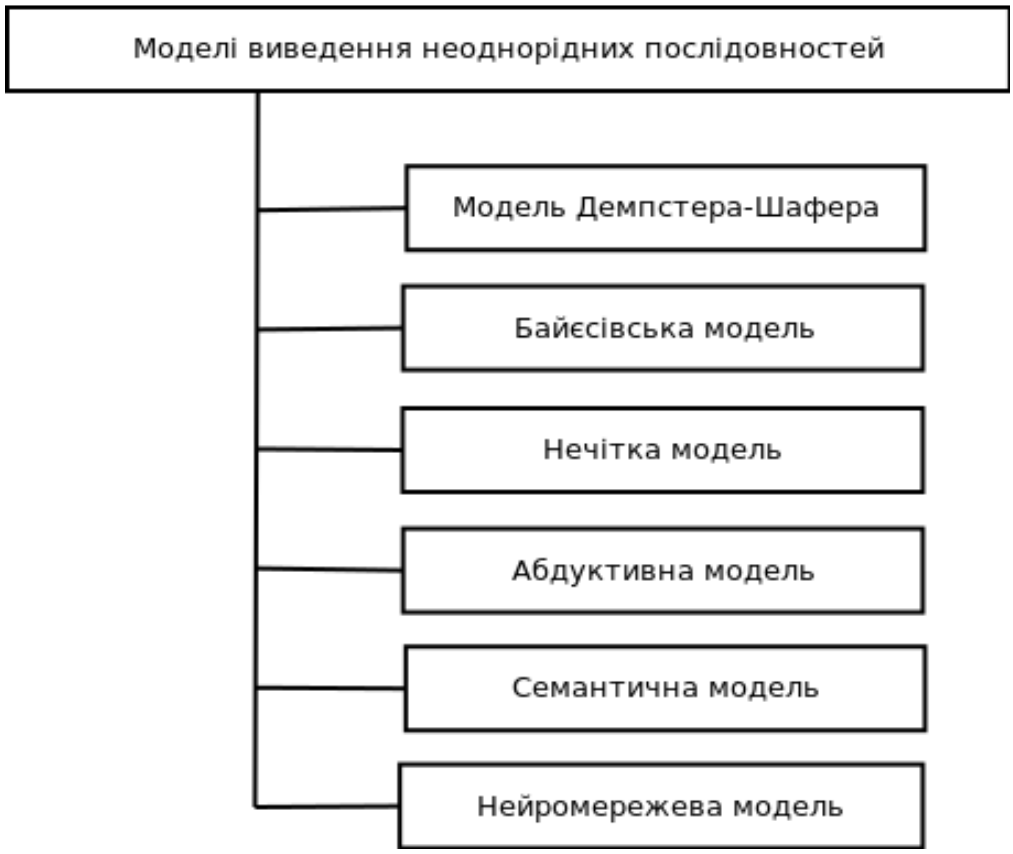

Рисунок 1.2 – Основні моделі виведення неоднорідних послідовностей

Так, розглядаючи кожну з моделей докладніше, можна відмітити їх особливості. Обробка та поєднання даних неоднорідних послідовностей на основі Байєсівської моделі використовує закони теорії ймовірності. Невизначеність представлена в такому випадку в ймовірнісному розумінні і може відповідно приймати значення від 0 до 1. Виведення в такому разі

ґрунтується на основі теореми Байєса.

Застосування моделей аналізу неоднорідних послідовностей може здійснюватися на різних етапах обробки даних.

У роботі [44] пропонується комбінування застосування нейронних мереж і Байєсівського виведення для прийняття рішень. У роботі [45] цей підхід застосовується для вирішення задач, пов’язаних із інтелектуальним поєднанням даних неоднорідних послідовностей, що надходять з різних сенсорів, а саме лазерного сенсора, відеокамери і радара для підвищення точності моделювання, яке використовувалося для вирішення задач, пов’язаних із управлінням транспортним засобом.

У роботі [46] була запропонована концепція під назвою «вимірювання псевдо-інформації», за допомогою якої комбінування інформації, отриманої з різних джерел на основі Байєсівського підходу, було розширено до широкого переліку можливих формулювань, і були запропоновані нові формули для комбінування інформації. Даний підхід застосовувався для побудови карт за допомогою сонарних сенсорів.

У роботі [47] пропонується застосовувати даний підхід для вирішення задач класифікації, ґрунтуючись на даних неоднорідних послідовностей.

У роботі [48] пропонується використовувати Байєсівське виведення для задач локалізації пристроїв, у роботі [49] пропонується використовувати модель поєднання сенсорів у вигляді Байєсівської мережі і комбінувати її з застосуванням методу Монте-Карло марківських ланцюгів для задач класифікації стану інтелектуального агента.

У роботі [50] описаний підхід на основі ймовірнісного виведення застосовувався для визначення помилкових вимірювань і їх подальших виправлень.

У роботі [51] був запропонований розподілений алгоритм, що використовує Байєсівське виведення для визначення наявності пропущених наборів даних, що не були отримані з тимчасово неактивних сенсорів упродовж вимірювання.

Підходи на основі виведення Демпстера-Шафера базуються на роботах [52, 53], які узагальнювали Байєсівьску теорію із введенням таких понять як «довіра» і «правдоподібність», і відповідно їх оцінюванням.

У роботі [54] запропоновано використання цієї теорії для задач представлення неповних знань, оцінювання і зміни «довіри», а також доказування.

Відповідно до роботи [55], було запропоновано використання цього підходу до аналізу даних, отриманих з сенсорів. Однією з переваг застосування була гнучкість, що дозволяла використовувати дані з різними рівнями масштабування.

Також у роботі [56] було запропоновано об’єднання в одному алгоритмі використання різних методів для одночасного процесу маршрутизації та виявлення значущих подій у мережі.

Проблемі маршрутизації була присвячена робота [57], у якій пропонується алгоритм перебудови топології мережі у зв’язку із збоями на основі виведення Демпстера – Шафера.

У роботі [58] пропонується використання цього підходу для комбінування різнорідної інформації, поданої у вигляді неоднорідних послідовностей для побудови моделі динамічного оцінювання операційної картини оточуючого середовища для безпілотних літальних апаратів.

Підходи для обробки даних із застосуванням обчислювального інтелекту розглянуті в роботі [12].

Нейромережеві методи для аналізу неоднорідних послідовностей даних розглянуті в роботі [59].

У роботі [60] запропонований підхід до об’єднання різнорідних даних, який використовує нейронну мережу у якості асоціативної пам’яті, що дозволяє використати неповні і різнорідні дані для класифікації на прикладі автоматичного розпізнавання цілі.

У роботі [61] запропонована модель KBNNF (Knowledge-Based Neural Network Fusion) для обробки різнорідних послідовностей, отриманих з різних

сенсорів.

Абдуктивне виведення дозволяє вибрати гіпотезу, яка є найбільш вірогідною щодо пояснення певного факту [62].

Як правило цей підхід використовується в контексті ймовірнісного визначення причинно-наслідкових зв'язків [63]. Застосування цінової абдукції разом з нейронною мережею, яка створюється для абдукційної моделі, описано в [64].

Комбінація абдуктивного виведення з моделями на основі нечіткої логіки дозволяє об'єднати дані, які отримані в реальному часі з якісними оцінками, зібраними в експертів [65].

Такі моделі використовуються для визначення причинно-наслідкових зв'язків в різних сферах людської діяльності [66-68].

Семантичні моделі поєднання неоднорідних послідовностей, як правило, пов'язані з інтерпретацією даних. Це дозволяє зменшити об'єм інформації на різних етапах переробки даних.

Як правило, їх застосування двохетапне – побудова бази знань і виведення за допомогою співставлення зразків. Приклад побудови таких моделей наведено в роботах [69].

Концепцію семантичного злиття неоднорідних даних запропоновано в роботі [70]. Основна ідея полягає в застосуванні формальних мов для збору і аналізу даних таким чином, щоб можна було порівнювати відомості, які надходять від сенсорів, з відомостями, які зберігаються в базі знань.

Виходячи з цього, розширяється інструментарій аналізу відомостей за рахунок використання моделей і методів, орієнтованих на обробку формальних мов.

Підходи, що ґрунтуються на основі нечіткої логіки, використовують ідеї, викладені в роботі [5]. У роботі Бенона [71] пропонується модель, що за допомогою нечіткої логіки дозволяє моделювати невизначеність.

В роботі [72] пропонується модель, що дозволяє будувати нечітке виведення на основі нечітких тверджень. Відповідно до цього підходу, всі

кількісні вимірювання пропонується подати у вигляді нечітких значень із функцією належності.

Нечіткі правила застосовуються до фаззифікованих даних і повертають нечіткий результат, який дефаззифікується у подальшому. Подібний підхід був запропонований у роботі [73].

Моделі нечіткого виведення застосовувалися дослідниками в роботах [74, 75] для задач дослідження геометричних характеристик кластерів сенсорів і визначення місцезнаходження їх центру.

В роботі [76] розглядається проблема позиціонування сенсорів із використанням неповних, нечітких і усереднених даних, отриманих з сенсорів. Також дослідження в цьому напрямку [77] присвячувалося питанням нечіткої оптимізації.

Окремо варто виділити підходи до поєднання різнорідних даних, представлених у вигляді нечітких значень за допомогою нечіткого регресійного аналізу.

Основи даного підходу були покладені в роботі [9], де було запропоновано нечітку регресійну модель. Дану модель пропонувалося використовувати в тих випадках, якщо дані з невизначеністю або судження людини з'являлись і суттєво впливали на вимірювання чіткого числового значення.

Порівнюючи нечіткий регресійний аналіз з класичними ймовірністними моделями, можна відмітити, що різниця між очікуваними значеннями і прогнозованими відображає не випадкову складову, похибку в статистичному значенні (як в класичних моделях), а невизначеність у структурі досліджуваної системи, що виражається через нечіткість параметрів [78] досліджуваного явища.

Такими чином, запропонована модель мінімізує нечіткість моделі, мінімізуючи сумарне розкидання нечітких коефіцієнтів для всіх доступних рівнів даних.

Відповідно до роботи [79], нечіткий регресійний аналіз доцільно

застосовувати у випадках:

- малої кількості вимірювань;
- складностей визначення розподілів змінних;
- невизначеності у наявності стійких причинно-наслідкових відносин між залежними і незалежними змінними;
- неоднозначності у рівнях вимірювання;
- викривленнях внесених лінеаризацією;

У роботі [80] запропоновано нечітку регресійну модель на основі інтервальної гібридної нейро-фаззі штучної нейронної мережі та її застосування до задач менеджмента.

Проблеми різних типів нечітких регресійних моделей розглядаються в публікації [81].

У роботі [82] пропонується порівняльний аналіз множинної нечіткої регресійної моделі і методу найменших квадратів. У роботі [83] пропонується нечітка регресійна модель для багатовимірних нечітких даних.

Також окремим питанням застосування і розробки нечітких регресійних моделей присвячені публікації[84 - 86].

Для неоднорідних послідовностей, які можуть бути подані у вигляді динамічних рядів, підхід на базі нечіткої логіки запропоновано в роботі [8].

Інформаційно-аналітичні системи, які використовуються для аналізу предметних областей, що описуються даними, представленими у вигляді коротких неоднорідних послідовностей, потребують розвитку моделей і методів, які б мали зменшені вимоги до мінімальної довжини рядів даних, які необхідні для проведення аналізу.

Одним з перспективних напрямків розвитку є застосування теорії нечітких множин у комбінації із класичними методами.

Ідея побудови даної моделі полягає в розбитті всіє множини даних на нечіткі розбиття Руспіні[7], із подальшим застосуванням функцій належності. Таким чином, за допомогою двох трансформацій (прямої і зворотньої) відбувається апроксимація даних.

Апроксимаційним властивостям даного перетворення присвячені роботи [87 - 91].

1.3 Дослідження існуючих методів аналізу неоднорідних послідовностей

Проблеми, що виникають під час аналізу неоднорідних послідовностей, можна умовно розбити на декілька підгруп, в залежності від особливостей даних та тих проблем, що виникають під час обробки інформації.

Ці групи методів можуть вирішувати проблеми, що зв’язані з неточністю, неповнотою, мультиколінеарністю та неспівставністю даних. Схематичне зображення цих складнощів відображено на рис. 1.3.

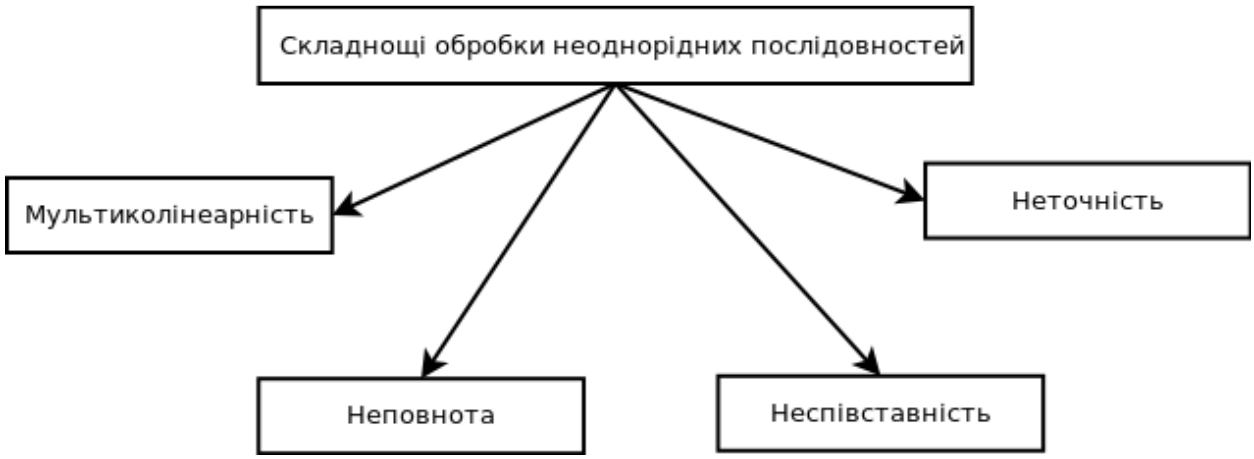

Рисунок 1.3 – Проблеми обробки неоднорідних послідовностей

Ймовірнісні методи, пов’язані із застосуванням декількох основних підходів, представлених в літературі. Наприклад, дослідження, опубліковані в роботах [92, 93], присвячені застосуванню фільтра Калмана для комбінування даних, отриманих з різних джерел.

Недоліком такого роду підходу є суттєва чутливість цього методу до наявності аномальних рівнів даних.

У роботах [94, 95] пропонується застосовувати послідовний метод Монте-Карло та його модифікації у поєднанні з ланцюгами Маркова та рекурсивні

реалізації для апроксимації ймовірностей. Разом з тим, необхідно відмітити, що такий підхід зберігає чутливість до аномальних рівнів даних.

Також у рамках ймовірнісного підходу для розв'язку задачі аналізу неоднорідних послідовностей можуть застосовуватися сіткові методи [96]. Їх недоліком є значне зростання обчислювальної складності не лише у випадку зростання розмірності задачі, але і у випадках масштабування (наприклад, зростання точності). Відповідно, для розв'язку такого роду труднощів доводиться застосовувати розподілені алгоритми розрахунків.

У цьому напрямку, з точки зору аналізу неоднорідних послідовностей з недостатньою точністю вимірювання, варто виділити групу методів альтернативних ймовірнісному. Вони ґрунтуються на теорії Демпстера-Шафера [52,53].

Основна ідея полягає в тому, щоб використовуючи функції довіри і правдоподібності, характеризувати різнорідні послідовності даних і комбінувати їх із застосуванням правила Демпстера [52,54]. Головним недоліком даного підходу є висока обчислювальна складність, алгоритми зменшення якої запропоновано в роботах [97-99].

Альтернативним є підхід на основі теорії можливостей, запропонованої Заде [100], і в подальшому розвиненою [101,102].

Переваги і недоліки застосування методів на основі теорії можливостей у порівняння із ймовірнісним і на основі теорії Демпстера-Шафера розглянуто в роботі [103].

У роботі [104] запропоновано використання теорії можливостей на прикладі моделювання знання експертів про числові параметри в області надійності. Цей метод запропоновано використовувати за умов неоднорідності та неповноти даних.

Було описано представлення відомостей, отриманих від експертів, з використанням математичного апарату теорії можливостей, що, на думку авторів, краще відображає невизначеність експертів.

Також можна відмітити групу методів, що ґрунтуються на теорії

наближених множин [105]. В основу покладений підхід, який полягає в представлені неточних даних у вигляді граничних областей, обробляючи їх таким чином, щоб рівень невизначеності зменшувався.

У роботі [106] пропонується метод аналізу неоднорідних послідовностей, який полягає в комбінуванні даних, отриманих із різних джерел. Запропонований метод відбору вхідних даних, отриманих з сенсорної мережі за допомогою оцінювання релевантності сенсорів (наборів даних отриманих з нього). Одним з недоліків даного підходу є достатньо неочевидний спосіб вибору рівня гранулярності, який залежить від природи досліджуваних даних.

У роботі [107] показано основну перевагу застосування теорії наближених множин, яка полягає в тому, що методи, які ґрунтуються на її постулатах, не потребують додаткових знань про явище, таких, як вид розподілу або функція належності. Натомість вони використовують відомості про внутрішню структуру даних(гранулярність).

Досить перспективною є розробка гібридних методів обробки неоднорідних послідовностей, які поєднують у собі переваги кожної з методик, що комбінуються.

Наприклад, у роботі [108] розглядається узагальнення теорії нечітких і наближених множин. У роботі [109] – метод, заснований на використанні нечітких множин і теорії Демпстера-Шафера.

Методи, що використовують основні постулати теорії випадкових множин [110], описані в роботах [111, 112]. З переваг можна відзначити достатньо широкий діапазон проблем, які можуть бути вирішені за допомогою цього підходу і можливість інтерпретувати результати. З недоліків варто відзначити обчислювальну складність.

Іншим підходом до вирішення проблеми недостатньої точності даних є застосування нечіткого підходу. Ключовим аспектом такого підходу є застосування функцій належності[5].

Досить часто такий підхід комбінують з іншими методами обробки неоднорідних послідовностей. Окремо варто відмітити роботи, присвячені

застосуванню нечіткого регресійного аналізу.

У роботі [113] запропоновано використовувати гібридний нечіткий аналіз найменших квадратів і досліджена доцільність такого використання.

Для розв’язку задачі регресійного аналізу в роботі [9] запропоновано використовувати методи лінійного програмування.

Можна відмітити ряд основних недоліків, що виникають під час використання методів лінійного програмування:

- недостатнє обґрунтування співвідношення між рішенням задачі лінійної оптимізації сумарної «нечіткості» результату побудови регресійної моделі і мінімізації сумарної похибки моделі по відношенню до навчальної вибірки, наприклад, у роботі [78] пропонується мінімізувати відстань між нечіткими значеннями які були на виході моделі і навчальною вибіркою, що в результаті зводилося до розв’язання нелінійної оптимізаційної задачі;
- модель занадто чутлива до появи аномальних рівнів [114];
- нечітка лінійна регресія має тенденцію до появи мультиколінеарності із збільшенням кількості факторів, що враховуються під час побудови моделі [115];

Для розв'язку проблеми скорочення кількості факторних змінних для побудови нечітких моделей у [116] пропонується використовувати метод шагового регресійного аналізу.

В якості критеріїв вибору факторів для задачі побудови нечіткої лінійної регресійної моделі в роботі [117] пропонується використовувати критерій Фішера. Відповідно до даного критерію здійснюється послідовне додавання або видалення ознак. Суттєвим недоліком даного підходу є неможливість отримання оптимального рівняння регресії у випадку кореляції між факторними змінними.

Таким чином, значуща змінна може бути ніколи не включена до рівняння, а другорядні змінні можуть бути додані.

Розглядаючи аналіз неоднорідних послідовностей з точки зору обробки даних поданих у вигляді часових рядів, варто відзначити можливість

застосування нечітких методів.

У роботі [118] запропоновано підхід, який використовує F-перетворення для обробки послідовностей даних із сезонністю.

Даний метод використовує запропоноване в роботі [8] F-перетворення для побудови моделі явища із врахуванням сезонності. Варто відмітити, що виконується лише одне згладжування і даний підхід є неітеративним.

Таким чином виникає потреба в розвитку даного методу для випадків, коли одноразове згладжування не дає необхідної точності моделювання.

Стосовно проблеми мультиколінераності, або взаємної кореляції незалежних змінних, необхідно відмітити, що вона досить суттєво впливає на якість моделювання.

Методи, що дозволяють подолати цю проблему, можна умовно розділити на дві групи: методи вилучення рівнів даних і методи, що усувають вплив мультиколінеарності на процес обробки рівнів неоднорідних послідовностей.

До першої групи можна віднести методи із простим вилученням даних [119] або зміною даних через дослідження процесу вимірювання [120], асоціюванням [121].

Друга група методів, що дозволяє позбутися впливу ефекту мультиколінеарності або суттєво його зменшити, полягає в проведені певних трансформацій з даними, наприклад, описаний у публікації [122] метод перетину коваріацій, який вимагає застосування математичного апарату нелінійної оптимізації. Розвитком цього підходу став метод швидкого перетину коваріацій [123].

У роботі [124] описаний метод побудови внутрішнього еліпсоїду для усунення ефекту мультиколінеарності.

Робота ІАС пов’язана із обробкою даних, отриманих із різних джерел та збережених для проведення перетворень у вигляд придатний для прийняття рішень.

Розглядаючи проблему неповноти даних, що були зареєстровані під час збору інформації, можна відзначити, що дані можуть вважатися неповними за

рахунок появи аномальних рівнів, невчасного отримання або неотримання інформації в певні моменти часу або появі суперечностей у даних.

Досить часто застосовуються статистичні методи, які, в залежності від об'єктивних обставин вимірювання, мають певні обмеження і умови використання [125].

У роботах [126,127] розглядаються методи на основі навчання «без вчителя», на основі застосування дискримінаційної функції, яка дозволяє виміряти «схожість» між двома послідовностями даних.

Інший підхід полягає в тому, що оцінка аномалії для всього часу спостережень проводиться не з точки зору міри «схожості», а з точки зору ймовірності того, що даний елемент є аномальним. До цієї групи методів найчастіше відносять марківські ланцюги та кінцеві автомати.

Такі методи розглядалися в роботах [128,129]. Такі підходи також можуть застосовуватися «без вчителя»

У роботі [130] запропонований підхід до виявлення аномальних рівнів на основі OLAP.

Також досить поширеним є застосування методів на основі навчання «з учителем» для виявлення аномальних рівнів [131,132].

Проблема обробки неспівставних даних, які можуть отримуватися з різних сенсорів або джерел інформації в певному сенсі перетинається з проблемою взаємодії людини і комп'ютерної системи.

Приклади такого роду досліджень із створенням наборів даних, отриманих від людини та програмно-апаратної системи освітлені в [133] .

У роботі [134] розглянуте використання методу на основі теорії Демпстера-Шафера для поєднання відомостей сгенерованих апаратними засобами та людьми. Проблемі оцінювання неточності в лінгвістичних даних присвячена робота [135].

У роботі [136] запропонована людино-центричний підхід до поєднання неспівставних даних неоднорідних послідовностей. Методи, що реалізовані, ґрунтуючись на цьому підході, застосовують гібридне поєднання

обчислювальних можливостей комп'ютерної техніки і людини.

Особливості застосування відомостей, отриманих від короткочасної співпраці тимчасових груп експертів, разом із наявними даними вимірювання або статистичного узагальнення існуючих масивів інформації, із використанням інтелектуальних методів обробки даних, є перспективним напрямком сучасних наукових досліджень.

1.4 Постановка задачі дослідження.

Постановку задачі дисертаційного дослідження можна подати у такому вигляді:

Відповідно до існуючих особливостей предметної області, інформація про неоднорідні послідовності зберігається в реляційній базі даних у вигляді атрибутів $(L_1 : D_1, L_2 : D_2, L_3 : D_3 ..., L_k : D_k)$, де L_1 – дата/час вимірювання або події, з доменом D_1, який містить всі допустимі значення дати/часу, L_2 – вид джерела інформації, що застосовувалося, з доменом D_2 який містить всі внесені в систему джерела інформації, та інших відомостей $L_3 ... L_k$ з доменами $D_3 ... D_k$ що стосуються події або вимірювання.

1) З існуючої множини даних про дати та час події або вимірювання, а також пов'язані з цим відомості необхідно сформувати динамічний ряд кількісних характеристик складної системи або окремих її складових за певний період часу T, який складається з n чітких або нечітких рівнів(наприклад з врахуванням відомостей, отриманих від експертів та поданих у вигляді лінгвістичних змінних): $Y = y_1, y_2, y_3 ... y_n$, або $\tilde{Y} = \tilde{y}_1, \tilde{y}_2, \tilde{y}_3 ... \tilde{y}_n$, де $\tilde{y}_i = (y_i, v_i)_L, i = 1...n$ і кількісні характеристики чинників, які можуть бути представлені як у вигляді чітких, так і нечітких значень(з врахуванням експертних відомостей про значення чинника): $X = x_1, x_2, x_3 ... x_n$, або $\tilde{X} = \tilde{x}_1, \tilde{x}_2, \tilde{x}_3 ... \tilde{x}_n$, де $\tilde{x}_i = (x_i, c_i)_L, i = 1...p$.

2) З множин X та Y необхідно знайти аномальні рівні x_a і y_a, такі, які мають суттєві відхилення від інших рівнів зразка та викликані об'єктивними та суб'єктивними причинами або короткочасною дією невідомого фактора.

3) Для кожного з цих динамічних рядів необхідно створити моделі і методи, які б дозволяли вибрати модель сезонності та визначити:

- трендову компоненту U_i,

- циклічну компоненту C_i,

- випадкову компоненту ε_i.

4) Вибрати модель та створити довгостроковий та короткостроковий прогноз для кількісних показників, тобто знайти y_{n+1}, за відомих $y_1, y_2, y_3, \ldots, y_n$

5)Створити модель залежності кількісних показників стану складної системи від факторів, які впливають на неї, тобто $Y = f(X, B, E)$, або для нечітких множин $\tilde{Y} = f(\tilde{X}, \tilde{B}, \tilde{E})$, де $B, \tilde{B}$ - вектор невідомих параметрів, $E, \tilde{E}$ - вектор помилок.

Необхідно здійснити відбір значущих чинників, тобто відбір із множини стовпчиків X вибрати такі, які б найбільше впливали на конкретний показник стану системи.

6)З множини методів, доступних для реалізації кожної з цих задач, обрати такі, які б були адекватними для рішення конкретної задачі користувача:

Необхідно знайти таке відображення R^s, яке описує зв'язки і представлено формулою (1.1)

$$R^s = (F, C, D, P, M, K), \qquad (1.1)$$

де C – множина цілей, що досягається системою обробки накопичених неоднорідних даних;

D – множина звітів, відповідей та інших відомостей, що є результатом роботи інформаційної технології;

P – множина послідовностей перетворення інформації з різних джерел,

що реалізується методами;

M – множина програмних реалізацій методів аналізу та прогнозування неоднорідних даних;

K – множина елементів обчислювальних засобів спеціалізованої інформаційної системи;

F – множина функцій (моделей), що виконує інформаційна технологія:

$$F = (f_1, f_2, f_3, f_4, f_5), \qquad (1.2)$$

де f_1 – набір функцій формування динамічних рядів;

f_2 – набір функцій перевірки рівнів динамічного ряду на аномальність;

f_3 – набір функцій для побудови моделей;

f_4 – набір функцій для проведення аналізу;

f_5 – набір функцій для визначення адекватності створених моделей для різних показників стану складної системи.

У зв’язку з цим, метою дисертаційної роботи є розробка моделей, методів та інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для підвищення ефективності оцінювання поточного стану предметних областей в інформаційно-аналітичних системах.

Для досягнення поставленої мети визначені такі задачі дисертаційної роботи:

- провести дослідження існуючих моделей і методів аналізу неоднорідних послідовностей даних;
- розробити моделі неоднорідних послідовностей даних;
- розробити метод відбору значущих чинників при побудові нечіткої багатофакторної регресії для даних, що представлені у вигляді неоднорідних послідовностей;
- вдосконалити метод фільтрації компонент неоднорідних послідовностей даних;

- розробити інтелектуальну інформаційну технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметної області та виконати програмну реалізацію й впровадження результату дослідження при вирішенні практичних задач.

Список використаних джерел у даному розділі наведено у повному списку використаних джерел під номерами: [5,7,8,9,12,29-136].

2 РОЗРОБКА МОДЕЛЕЙ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

2.1 Визначення показників для аналізу неоднорідних послідовностей

Неоднорідні послідовності досить часто вимірюються і подаються у вигляді динамічних рядів.

Як правило, послідовність спостережень одного показника чи ознаки, впорядкована в залежності від послідовно зростаючих чи спадаючих значень часу. Досить розповсюдженим є аналіз рядів динаміки за допомогою графічного підходу, коли використовуючи наочне подання і досвід в певній предметній області, дослідник проводить первинний аналіз даних.

Для візуалізації подання рядів динаміки можна використовувати, як хронологічні таблиці, так і графіки.

Але у випадку автоматизації процесів аналізу даних і для зменшення рівня залучення людини в процес дослідження використовуються певні показники, що характеризують зібрані відомості статистично.

Неоднорідні послідовності можуть бути подані у вигляді рядів динаміки, які можуть бути умовно розділені на групи за різними характеристиками, відповідно до [137]. Наприклад, за видом узагальнюючих показників, які застосовуються в динамічному ряді, їх можна розбити на групи, які складаються з рядів абсолютних, відносних і середніх величин.

У залежності від часових характеристик поданих даних ряди динаміки поділяються на моментні та інтервальні. Також в залежності від повноти охоплення часу вони можуть бути повні і неповні.

Відповідно, розрахувавши з даних статистичні показники можна отримати похідні ряди, які можуть складатися, наприклад, з рядів темпів приросту, коефіцієнтів, індексів, середніх значень, відхилень, дисперсій і т.д. Це можуть бути, як узагальнюючі функції, які в результаті дозволяють згрупувати дані і отримати певну характеристику такої групи, так і такі, що зберігають часову складову динамічного ряду.

Схематично зобразити форми подання динамічних рядів можна на рисунку 2.1.

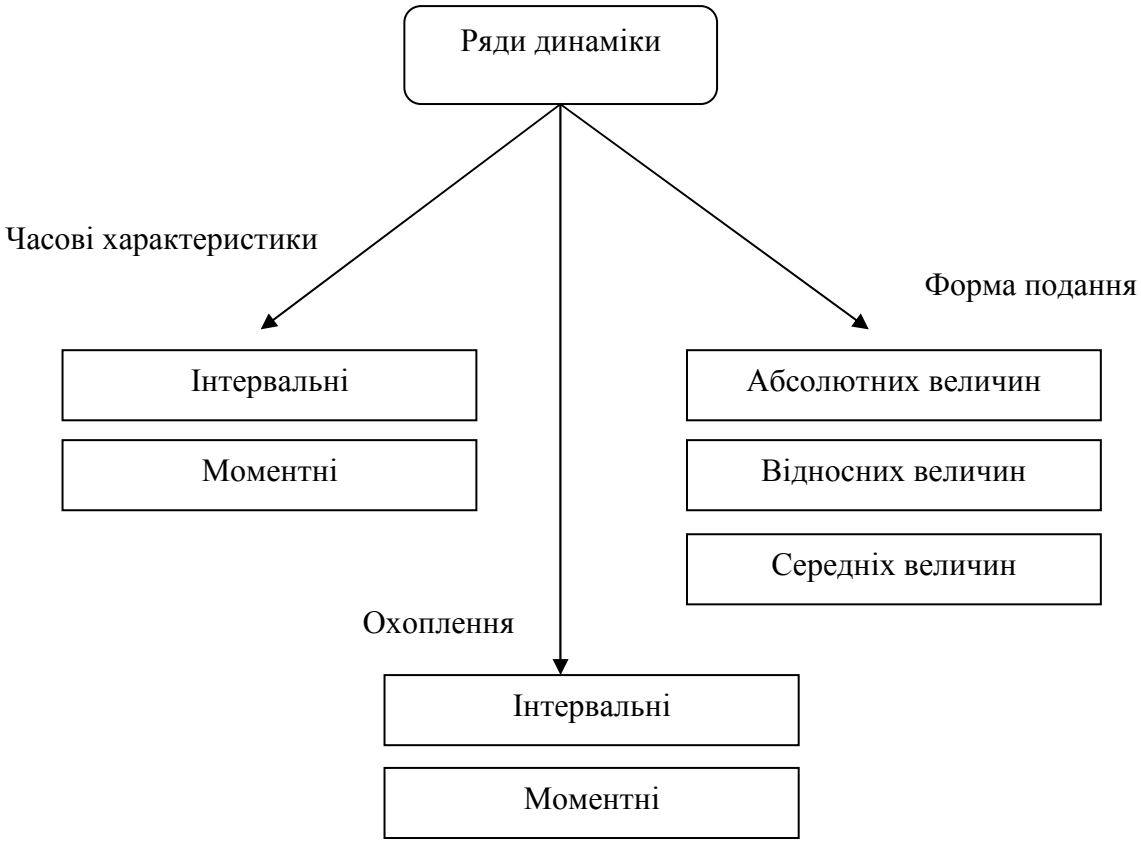

Рисунок 2.1 - Структурна схема рядів динаміки

Розглядаючи неоднорідні послідовності необхідно відмітити, що початкові дані отримані з різних джерел, можуть бути представлені в різних формах подання.

Відповідно, першою задачею дослідника, що постає на початку проведення аналізу, є зведення часових рядів поданому вигляді до одного способу відображення інформації.

Окремо варто відмітити невідповідність інтервалів часових шкал, особливо це стосується ситуацій, коли певні характеристика вимірюються різними установами. У такому разі, варто подати динамічні ряди застосовуючи єдину часову шкалу.

Застосувавши запропоноване подання даних на єдиній часовій шкалі, дані представлені в кожному з вимірювань, можна подати в такому вигляді:

Нехай існує динамічний ряд, який складається з n рівнів:

$$X = x_1, x_2, x_3, \ldots, x_n \qquad (2.1)$$

У загальному випадку кожний рівень часового ряду x_i, i=1, 2, ..., n будемо вважати таким, що складається у вигляді чотирьох елементів:

- трендової компоненти U_i, i= 1, 2, ..., n;
- сезонної компоненти V_i, i= 1,2, ..., n;
- циклічної компоненти C_i, i= 1, 2, ..., n;
- випадкової компоненти ε_i, i = 1, 2, ..., n.

Розглядаючи такий підхід до декомпозиції динамічного ряду, як правило застосовується спрощений варіант(2.2) описаний в [138]:

$$X = f(U, V, \varepsilon) \qquad (2.2)$$

Ідея такого подання полягає в тому, що застосовується моделювання явища за допомогою виділення окремих складових, наприклад, трендова складова характеризує основну тенденцію зміни явища у часі.

А початковий динамічний ряд може бути відновлений із застосуванням функції від змінних, кожна з яких характеризує внесок означених складових. Ця функція може бути адитивною або мультиплікативною.

Розглядаючи інші складові, варто відзначити, що коливання, які мають строго періодичний чи близький до нього характер і завершуються протягом одного року подаються за допомогою сезонної складової V_i.

У випадках, коли період коливань складає декілька років, у динамічному ряді присутня циклічна компонента C_i.

Після вилучення з моделі трендової, сезонної і циклічної компонент, які є регулярними, або систематичними компонентами динамічного ряду, складова частина динамічного ряду, яка залишається, є нерегулярною компонентою, поданою в моделі у вигляді випадкової компоненти ε_i.

Відштовхуючись від того, що випадкові відхилення супроводжують будь яке соціальне явище, ця компонента є обов'язковою складовою частиною будь-якого динамічного ряду.

Досліджуючи залишки після вилучення систематичних компонент динамічного ряду, розглядаючи їх рамках класичної теорії помилок як випадкову компоненту ряду, можна відзначити, що вона має такі властивості:

- відповідність нормальному закону розподілу;
- рівність математичного сподівання нулю;
- незалежність значень рівнів, тобто відсутність істотної автокореляції.

Таким чином, перевіряючи залишкову складову після декомпозиції і моделювання компонент динамічного ряду на відповідність властивостям випадкової компоненти, можна з’ясувати властивості побудованої моделі.

Перевірка випадковості коливань рівнів залишкової послідовності означає перевірку гіпотези про правильність вибору виду тренда. Для дослідження випадковості відхилень від тренда маємо набір різниць

$$\varepsilon_i = x_i - U_i \left(i = 1, 2, \ldots, n\right). \qquad (2.3)$$

Характер цих відхилень вивчається за допомогою ряду критеріїв, до яких, зокрема, відноситься критерій серій, критерій піків (поворотних точок) [139].

Перевірка відповідності розподілу випадкової компоненти нормальному закону розподілу виконується лише наближено за допомогою дослідження показників асиметрії (γ_1) та ексцесу (γ_2), в тому випадку, якщо динамічні ряди короткі [140].

Перевірка рівності математичного сподівання випадкової компоненти нулю, якщо вона розподілена за нормальним законом, здійснюється на основі t-критерію Стьюдента [139].

Перевірка незалежності значень рівнів випадкової компоненти, тобто перевірка відсутності істотної автокореляції в залишковій послідовності, може здійснюватися, зокрема, з використанням *d*-критерію Дарбіна-Уотсона[140].

У випадку якщо всі зазначені чотири описані перевірки повертають позитивний результат, робиться висновок про випадковість залишкової компоненти.

Динамічний ряд можна характеризувати значною кількістю числових величин (рисунок 2.2).

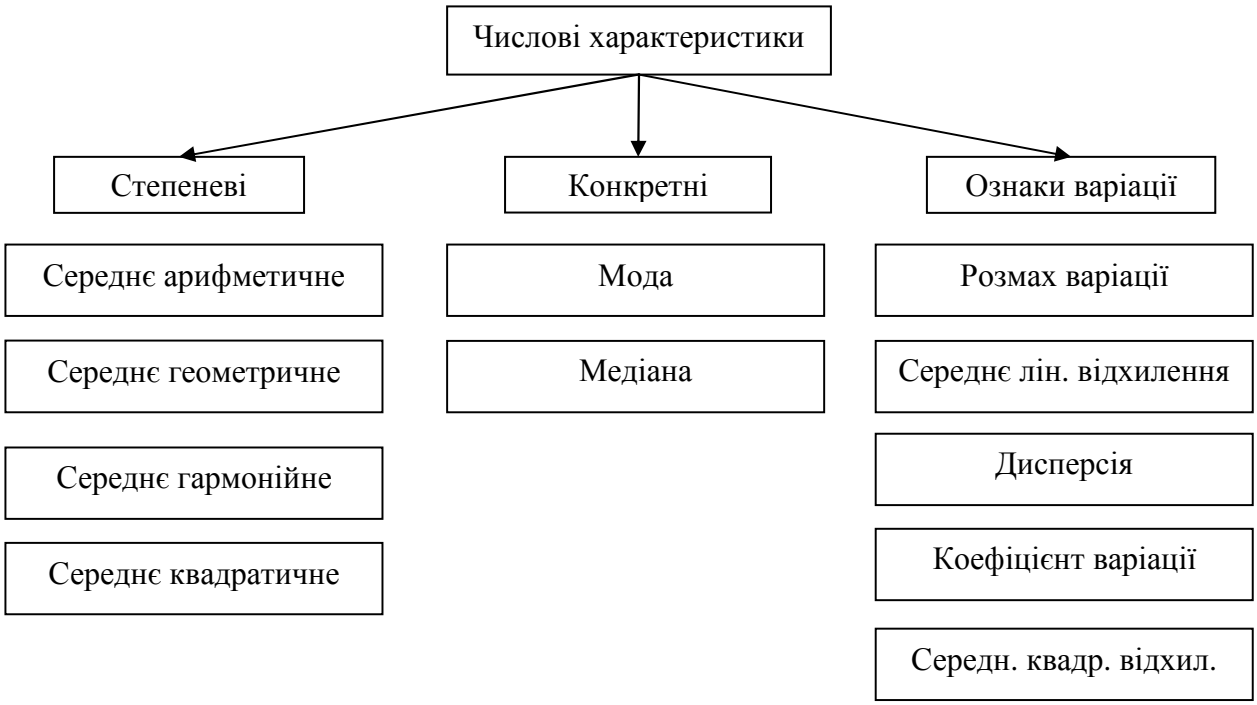

Рисунок 2.2 – Структурна схема числових характеристик ряду

Для узагальнення і групування значень динамічного ряду досить широко використовуються середні величини. За кількісною варійованою ознакою можна охарактеризувати сукупність за допомогою середніх значень.

Але важливо відмітити, що застосування середніх величин дозволяє розкрити лише загальну тенденцію досліджуваного явища і тільки в тому випадку, якщо вона отримується з великого числа фактів однорідної сукупності.

У випадку неоднорідних послідовностей, процес розрахунку середніх значень відбувається після застосування методів, що дозволяють провести інтеграційний аналіз.

При недотриманні цих умов застосування середніх показників може потягти за собою хибні висновки. Прикладом може бути розрахунок середньої заробітної платні, коли в одну сукупність зараховують екстремальні значення як великі так і малі, розрив у рівнях яких може досягати великих значень. В таки випадках доцільно застосовувати підходи на основі кластеризації, описані в [141]

Клас степеневих середніх включає до себе декілька видів середніх статистичних величин. Загальний вид формули (2.4) для степеневої середньої[142] такий:

$$\bar{x} = \sqrt[m]{\frac{\sum_{i=1}^{n} x_i^m n_i}{\sum_{i=1}^{n} n_i}} \quad , \qquad (2.4)$$

де $\bar{x}$ - середня степеня m;

x_i – варіанти (змінювані значення ознаки);

n_i – частота варіанта (сума всіх частот дорівнює обсягу вибірки, тобто $\sum_{i=1}^{n} n_i = N$);

m – показник степеня середньої величини.

Таким чином при m = 0 отримаємо середню геометричну, при m = 1 – середню арифметичну, при m = 2 середню квадратичну, при m = -1 – середню гармонійну.

Всі середні величини, в тому числі середня арифметична, середня геометрична та інші — це функції агрегації, оскільки вони, базуючись на значеннях виміряних величин, відображають загальне значення, яке властиве всій сукупності оброблених одиниць у цілому.

Поряд із абстрактними середніми під час аналізу неоднорідних послідовностей можуть використовуватися конкретні середні. До таких

середніх відносяться мода і медіана.

Дисперсія характеризує рівень розсіювання вимірюваної величини і є одним з показників, що застосовуються під час статистичного аналізу даних [143]. Також значення дисперсії визначає необхідний обсяг вибіркової сукупності. Для того щоб вибірка була репрезентативною, необхідно забезпечити більший обсяг даних у випадку більших значень дисперсії.

Дисперсія кількісної ознаки знаходиться за формулою [139]:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{\sum_{i=1}^{n} n_i}. \qquad (2.5)$$

Середнє квадратичне відхилення також є загальноприйнятою мірою варіації ознаки. Воно позначається символом σ і може бути розраховано на основі середнього квадрата відхилень, як корінь квадратний із дисперсії.

Застосування структурних показників у часі дає можливість виявити існування тенденцій для складових частин явищ, спираючись на розраховані значення, за допомогою яких можна здійснювати аналіз неоднорідних послідовностей.

Використання узагальнюючих показників дозволяє сформувати набори даних для проведення багатофакторного аналізу у випадках, якщо присутня неоднорідність даних на рівні різної періодичності вимірювання.

2.2. Розробка підходів для виявлення взаємозв'язків між рівнями динамічного ряду.

Виявлення закономірностей розвитку будь-якого явища, в тому числі і причин його виникнення, змін в часі є метою аналізу. На основі цієї інформації можна передбачувати розвиток ситуації в залежності від відомих факторів.

Відповідно, для досягнення цієї мети необхідно вивчати явища у

взаємозв’язку з іншими об’єктивно існуючими явищами і процесами.

Розглядаючи вивчення закономірностей, що виникають під час розвитку суспільних явищ, необхідно відмітити що вони формуються під впливом значної кількості причин. Точно невідомо, якою мірою кожна з них впливає на величину явища. Такого роду зв'язки є кореляційними. Кореляційна залежність - статистичний взаємозв’язок двох і більше випадкових величин. При цьому зміни значень однієї або багатьох з цих величин супроводжуються систематичними змінами значень іншої або інших величин [144].

Крім того розрізняють також зв'язки прямі і зворотні. Якщо зі зростанням факторної ознаки результатна ознака зростає, то це прямий зв'язок. Якщо зі збільшенням факторної ознаки результатна зменшується, то це зворотний зв'язок.

За аналітичним виразом кореляційні зв’язки можна розділити на прямолінійні і нелінійні. При цьому кореляційні зв'язки одержують лише наближений аналітичний вираз, характерний тільки для певних умов.

Також варто відзначити, що наявність кореляційного зв’язку не завжди тягне за собою наявність причинно-наслідкового зв’язку.

В роботі [145] зібрані приклади аналізу наборів даних, які мають високі значення коефіцієнтів кореляційного зв’язку щодо очевидно непов'язаних між собою феноменів.

Інформацію про міру статистичного взаємозв'язку між двома явищами можна одержати за допомогою кореляційного аналізу, результатом проведення якого є числове значення.

Парний кореляційний аналіз є найбільш поширеним способом визначення рівня неповного, статистичного зв’язку. Для цього між ознакою-наслідком і ознакою-фактором проводять розрахунок коефіцієнта кореляції для одного фактору.

Варто зауважити, що інтерпретація результатів одно факторного кореляційного аналізу повинна носити обережний характер. Числове значення отриманого коефіцієнту не може обґрунтовувати наявність причинно-

наслідкового зв’язку між явищами. Воно лише надає додаткову інформації про силу статистичного зв’язку у випадку, якщо причинно наслідковий зв’язок визначений.

Розглядаючи статистичні методи можна виділити ряд підходів, що дозволяють виявити і оцінити характер взаємозв’язку. В залежності від даних які є в наявності та підходу можна зобразити їх схематично (рис. 2.3).

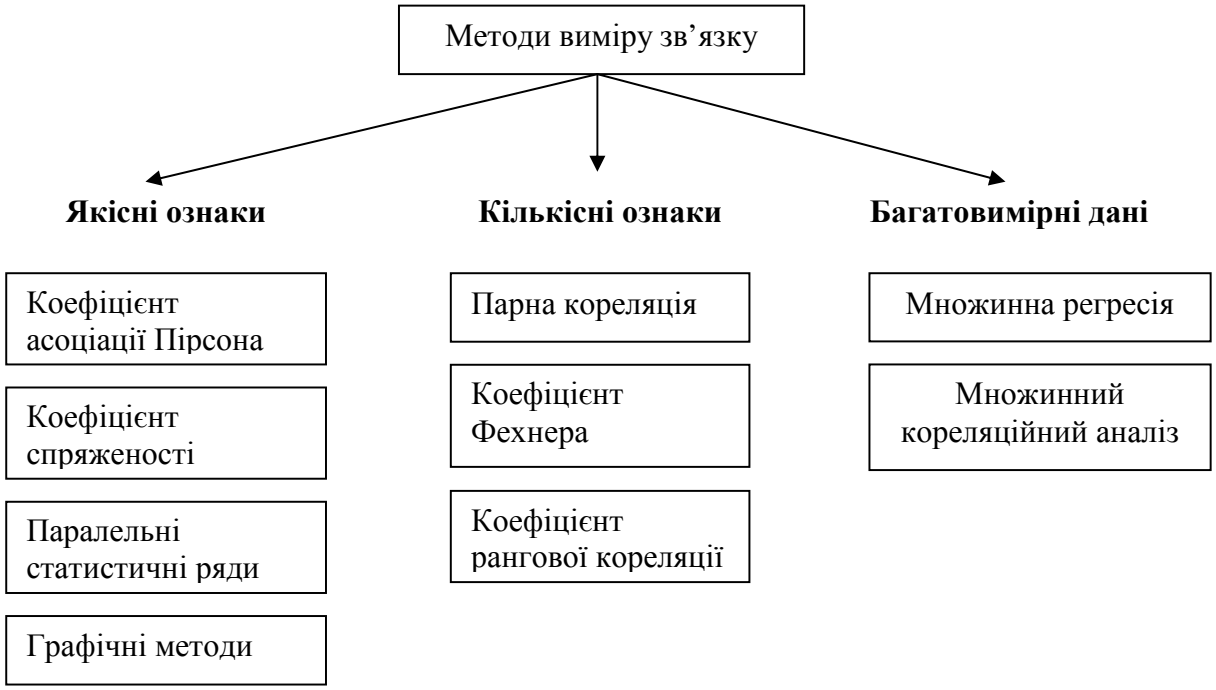

Рисунок 2.3 – Методи виміру зв’язку

Така кореляція дозволяє відносно адекватно виміряти виявлений зв'язок між двома рядами даних за допомогою лінійного[140] коефіцієнта парної кореляції (2.6):

$$R = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right] \left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2 \right]}}, \qquad (2.6)$$

що дорівнює (2.7):

$$\frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right]}}, \qquad (2.7)$$

де R – лінійний коефіцієнт кореляції між двома статистичними рядами X і Y; X_i - рівні ряду X, Y_i - рівні ряду Y.

Наступним питанням, що постає перед дослідником є перевірка значимості одержаного коефіцієнта кореляції. В цьому випадку, ґрунтуючись на певному рівні довіри (як правило, 0.95 або вище) перевіряють, чи можна за конкретним значенням робити висновок про наявність взаємозв’язку між досліджуваними ознаками. Для цього використовують спеціальні статистичні критерії.

Відповідно до цих критеріїв визначається критичне значення, виходячи з кількості рівнів порівнюваних рядів і необхідної ймовірності оцінки. Такі критерії розраховуються для кожного конкретного випадку застосування коефіцієнта кореляції.

Якщо розрахункове значення дорівнює критичне чи перевищує його, то статистичний зв'язок визнається значимим. У протилежному випадку такий зв'язок між ознаками не враховується таким, що має статистичне значення.

У випадку коли необхідно виміряти тісноту зв’язку між ознакою наслідком і рядом ознак-факторів одночасно, застосовується багатофакторний кореляційний аналіз.

Як правило в ході такого аналізу можуть бути розраховані такі коефіцієнти [139]:

- окремі коефіцієнти кореляції;

- множинні коефіцієнти кореляції;

- множинний коефіцієнт детермінації;

- сукупний коефіцієнти множинної кореляції;

- коефіцієнт множинної детермінації;

та інші показники, які дозволяють отримати біль точну оцінку впливу різних факторів на досліджуване явище.

Для виміру зв'язку між якісними ознаками в статистиці широко використовуються коефіцієнт спряженості А.А. Чупрова та коефіцієнт асоціації К. Пірсона [146].

Застосування паралельних статистичних рядів грає важливу роль у виявленні зв'язків. Із їх співставлення[140] розпочинається розрахунок однофакторних, багатофакторних чи інших коефіцієнтів кореляці. Також паралельні ряди є відносно самостійним і важливим методом виявлення кореляційної залежності.

Застосування такого подання інформації дає можливість не лише визначити взаємозалежну зміну двох і більше явищ, а й відстежити зміни одного явища в динамічному ряді або ряді розподілів.

Розглядаючи підходи до визначення взаємозв’язків серед неоднорідних послідовностей необхідно відмітити поширену ситуацію відновлення сукупності з набору невеликих вимірювань.

Ці експерименти та заміри можуть бути зроблені стосовно одного явища але в різних умовах, які зумовлюють різнорідність даних. До таких умов відносяться:

- умови експерименту;

- відмінності в методиці вимірювання;

- різниця в обладнанні

та інші відмінності, що можуть спровокувати виникнення неоднорідності послідовностей.

В такому випадку поєднуючи послідовності пропонується застосовувати

підхід описаний в роботі [147]: нехай існують дві групи випадкових даних, і $B_{s,i}$ де $s=1,2$ визначає групу вимірювань (або експеримента) із довжиною $n_1 \, i \, n_2$.

Приклад такої неоднорідної вибірки зображений на рисунку 2.4

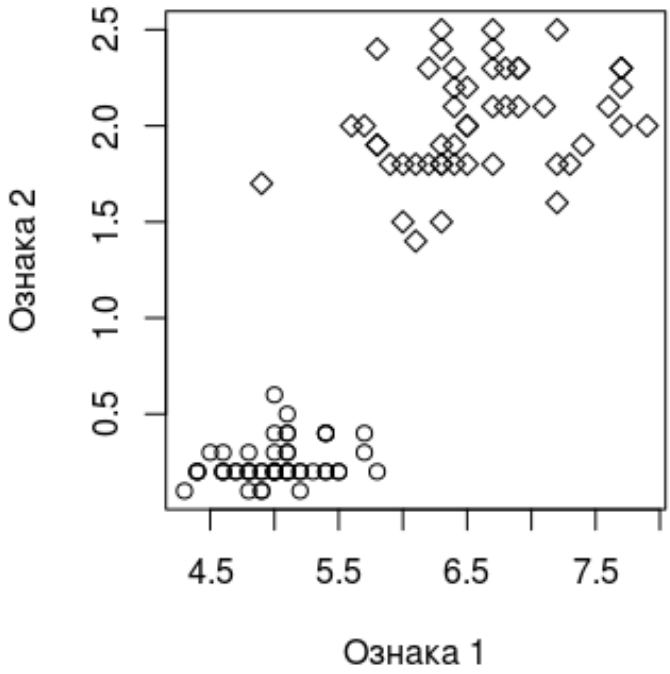

Рисунок 2.4 – Неоднорідна послідовність вимірювань з двома групами даних

Припустимо, що ці дві випадкові величини розподілені однаково, але незалежно таким чином, що кожна з груп має різні математичні очікування (*M*) та дисперсію(*D*).

Подамо ці дві групи у вигляді таких рівнянь(2.8, 2.9):

$$A_{s,i} = M\left(A,s\right) + D\left(s,i\right), \qquad (2.8)$$

$$A_{s,i} = M\left(A,s\right) + D\left(s,i\right), \qquad (2.9)$$

де $i = 1,\ldots,n_s; n = n_1 + n_2$.

Також, коваріація і кореляція між випадковими змінними відрізняються в залежності від групи.

В такому випадку, якщо сукупність складається з двох груп з різними значеннями математичного очікування і варіації, а $n \to \infty$, кореляція відповідно до [147] може бути обчислена за формулою 2.10:

$$r_{A,B} \xrightarrow{p} \frac{\lambda cov\left(A_{1,}B_1\right)+\left(1-\lambda\right)cov\left(A_{2,}B_2\right)}{\delta_A\delta_B}+$$

$$+\frac{\lambda\left(1-\lambda\right)(M\left(A_2\right)-M\left(A_1\right))(M\left(B_2\right)-M\left(B_1\right))}{\delta_A\delta_B}=\tau_{A,B}, \qquad (2.10)$$

де λ обчислюється з формули (2.11). Фактично це відношення розмірів двох виборок:

$$\lambda=\frac{n_1}{n_1+n_2}, \qquad (2.11)$$

де $\lambda\in(0,1)$.

Значення δ_Aі δ_Bрозраховуються з формул, використовуючи значення математичного очікування і дисперсії для окремих груп. Для кожної з різнорідних груп буде мати вигляд (2.12, 2.13):

$$\delta_A^2=\lambda D(A,1)+(1-\lambda)D(A,2)+\lambda(1-\lambda)\big(M(A,2)-M(A,1)\big)^2, \qquad (2.12)$$

$$\delta_B^2=\lambda D(B,1)+(1-\lambda)D(B,2)+\lambda(1-\lambda)\big(M(B,2)-M(B,1)\big)^2. \qquad (2.13)$$

з формул (2.10, 2.11) випливає, що границя $\tau_{A,B}$ залежить від:

- відношення розмірів виборок;
- значень коефіцієнта міжгрупових коваріацій;
- дисперсії;
- математичного очікування;

Застосування цієї формули можливо з деякими застереженнями:

1) У випадку коли статистичний зв’язок між величинами відсутній, сумарна оцінка $\tau_{A,B}$ буде зміщеною, причому у видку коли $\lambda = 0.5$ зміщення буде максимальне.

2) У випадку негомогенності кореляційного зв’язку, результат застосування формули (2.10) також буде повертати зміщений результат.

3) У випадку застосування в умовах парадоксу Сімпсона [148], результат проведення кореляційного аналізу для об’єднання двох вибірок, які мають позитивний кореляційний зв’язок, може бути негативним. Це відбувається у випадку, якщо математичні очікування зсунуті в обох вибірках з різними знаками.

Також одним з підходів до вирішення задачі визначення статистичного зв’язку є оцінка кореляції за допомогою тренда [140]. Він полягає в тому, що якщо значення однієї змінної попередньо впорядкувати, то значення другої змінної теж буде мати тенденцію до впорядкування.

У разі ж відсутності кореляційного зв’язку, випадковий характер розташування значень другої змінної після проведення впорядкування першої, не буде змінюватися.

Відповідно, для цього можуть застосовуватися такі критерії як критерій Кенуя, Кокса-Стюарта або знаковий кореляційний критерій Нелсона та інші.

У випадку, якщо одна з характеристик подана у вигляді відомостей, що можуть бути ранжовані, а друга характеристика має такі властивості, що дозволяють подати її лише у вигляді розбиття на дві групи значень за якісною ознакою, варто розглядати використання точково-бісектрального коефіцієнта кореляції.

У випадку, якщо доводиться досліджувати кореляцію послідовності випадкових величин, яка характеризується наборами рангів пропонується застосовувати коефіцієнт конкордації Кендала-Бебінгтона-Сміта [149].

Можна відзначити два обмеження, які виникають під час застосування коефіцієнту конкордації:

- відсутня можливість розрахувати узгодженість експертних оцінок по

кожній змінній окремо;

- коефіцієнт конкордації вимірює узгодженість експертних оцінок у сенсі їх корельованості;

У випадку аналізу двох груп експертів і відповідно аналізу кореляції послідовностей оцінок отриманих від них, застосовується коефіцієнт конкордації Шукені-Фролі [150], який є аналогом коефіцієнта Кендала-Бебингтона Сміта, але є спеціалізованим для цієї задачі.

2.3 Моделі неоднорідних послідовностей.

Розглянемо більш докладно побудову моделей неоднорідних послідовностей.

Як уже зазначалося, динамічний ряд подається у вигляді: x_1, x_2, x_3, …, x_n, де x_i, $i = 1..n$ - рівні динамічного ряду, які впорядковані за зростанням часу i.

Розглядаємо часовий ряд, відповідно до формули 2.2 отримуємо подання динамічного ряду у вигляді, представленому формулою (2.14):

$$X_i = U_i + V_i + \varepsilon_i , \qquad (2.14)$$

де вважаємо, що $i = 1,...,N$, а також:

U_i – тренд,

V_i - сезонна компонента,

ε_i - випадкова компонента,

N - число рівнів спостереження.

Щодо U_i вважається, що ця компонента згенерована деякою гладкою функцією, міра гладкості якої заздалегідь невідома.

Сезонна компонента V_i має період T_0, таким чином (T_0=12 для ряду місячних даних; T_0=4 - для ряду квартальних даних) що $V_{i+T_0} = V_i$. Крім того,

відомо, що $N=mT_0$, де m - ціле число, а саме число років, поданих у часовому ряді.

Початкові дані часового ряду можна подати у вигляді матриці $\{x_{ij}\}$ розміром $m\times T_0$.

Тоді вираз (2.14) можна сформулювати у вигляді, представленому формулою (2.15):

$$X_{ij} = U_{ij} + V_{ij} + \varepsilon_{ij}, i = \overline{1,m},\ j = \overline{1,T_0}. \qquad (2.15)$$

Зазвичай в даних можуть бути присутні аномальні рівні даних. Для побудови моделей явища необхідно внести такі модифікації, щоб мінімізувати вплив аномальних рівнів.

Відповідно до розглянутих в першому розділі підходів до обробки аномальних рівнів, зважаючи на особливість даних рядів(а саме їх короткий розмір) пропонується застосувати ймовірнісні підходи перед розробкою моделей неоднорідних послідовностей для виявлення аномальних рівнів даних.

Даний етап розробки моделей неоднорідних послідовностей (виявлення та усунення аномальних рівнів даних) грає важливу роль під час проведення попереднього аналізу часових рядів.

Під аномальним рівнем вважаємо таке значення рівня часового ряду, яке виходить за межі досліджуваних можливостей системи і яке, у випадку продовження знаходження у складі рівнів ряду може істотно впливати на значення основних статистичних характеристик.

Аномальні спостереження можуть бути спричинені різними обставинами. Це можуть бути похибки, які викликані певними технічними причинами і потребують виявлення і вилучення.

Іноді, аномальні рівні можуть бути спричинені чинниками, які насправді носять об’єктивний характер і проявляються досить рідко. Такі чинники теж підлягають аналізу, а в певних випадках, їх варто включати до моделі, для того, щоб виявити вплив факторів, які проявляють себе нечасто.

Розглядаючи аномальні рівні як викиди, що суттєво відрізняються від загального рівня даних динамічного ряду, ми застосовуємо статистичні моделі динамічного ряду для їх виявлення.

Порівнюючи статистичні критерії [139] для роботи з неоднорідними послідовностями, наприклад:
критерій Шовене (2.16),

$$K = \frac{|x_i - \bar{x}|}{s}, \tag{2.16}$$

де значення $\bar{x}$ та s знаходимо за формулами (2.17, 2.18):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.17}$$

$$s = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2; \tag{2.18}$$

критерій найбільшого абсолютного відхилення (2.19) :

$$\tau = \frac{\max\limits_{i \le 1 \le n} |x_i - \bar{x}|}{s}, \tag{2.19}$$

де s розраховується відповідно до формули (2.20):

$$s = \sqrt{\frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}{n}}; \tag{2.20}$$

та критерій Ірвіна (2.21), можна відзначити різницю в результатах виявлення аномальних рівнів даних.

$$\lambda_i = \frac{|x_i - x_{i-1}|}{\sigma_x};\ i=2,\ \ldots,\ n, \qquad (2.21)$$

де σ_x - середнє квадратичне відхилення, яке розраховується відповідно до формули (2.22):

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (2.22)$$

Таким чином, виходячи з формулювань цих методів, критерій найбільшого абсолютного відхилення і критерій Шовене мають тенденцію до виявлення як аномальні найбільші чи найменші значення рівнів динамічного ряду. Таким чином, для виявлення аномальних рівнів в рядах неоднорідних послідовностей пропонується застосовувати критерій Ірвіна.

Також важливим є питання про наявність тренда, яке повинно бути вирішене під час попереднього аналізу перед побудовою моделі неоднорідної послідовності. Для цього перевіряють гіпотезу про наявність тренда за допомогою статистичних критеріїв. Перевірка гіпотези про наявність тренда здійснюється відповідно до підходу описаного в [140].

Початковий часовий ряд $x_1, x_2, x_3, \ldots, x_n$ розбивається на дві приблизно однакові за кількістю рівнів частини: у першій частині n_1 перших рівнів ряду, у другій - n_2 інших рівнів ($n_1 + n_2 = n$). Для кожної з цих частин обчислюються середні значення (2.23, 2.24) для кожної з груп:

$$\bar{x}_1 = \frac{\sum_{t=1}^{n_1} x_t}{n_1}; \qquad (2.23)$$

$$\bar{x}_2 = \frac{\sum_{t=n_1+1}^{n} x_t}{n_2}; \qquad (2.24)$$

і дисперсії (2.25, 2.26). Які використовуються для з'ясування гіпотези про однорідність дисперсій.

$$\sigma_1^2 = \frac{\sum_{t=1}^{n_1}(x_t - \bar{x}_1)^2}{n_1 - 1}, \qquad (2.25)$$

$$\sigma_2^2 = \frac{\sum_{t=n_1+1}^{n}(x_t - \bar{x}_2)^2}{n_2 - 1}. \qquad (2.26)$$

F-критерій Фішера[140] використовується для з'ясування чи дійсно дві частини одного ряду даних мають однакові дисперсії. Застосування цього критерію полягає в порівнянні розрахункового значення(2.27) цього критерію

$$F = \begin{cases} \dfrac{\sigma_1^2}{\sigma_2^2}, & якщо\ \sigma_1^2 > \sigma_2^2, \\ \dfrac{\sigma_2^2}{\sigma_1^2}, & якщо\ \sigma_1^2 < \sigma_2^2 \end{cases} \qquad (2.27)$$

з критичним значенням F_α, яке залежить від заданого рівня значимості α.

Гіпотеза про однорідність дисперсій є статистично значимою, якщо розрахункове значення *F* менше табличного F_α. В іншому випадку, гіпотеза про однорідність дисперсій відхиляється.

Після цього перевіряється гіпотеза про відсутність тренда з використанням *t*-критерію Стьюдента [140]. Розрахункове значення цього критерію знаходиться за формулою (2.28):

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad (2.28)$$

де σ - середнє квадратичне відхилення різниці середніх(2.29):

$$\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}} \quad . \tag{2.29}$$

В умовах недостатньої кількості даних, гіпотеза про відсутність тренду перевіряється за допомогою критерія Велча[139]. Це актуально для інформаційно-аналітичних систем, які працюють з короткими рядами даних.

Розрахункове значення критерію знаходиться за формулою(2.30):

$$t = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{\sqrt{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)}} \tag{2.30}$$

Таким чином, розглянувши питання про виявлення та усунення аномальних рівнів даних та перевіривши гіпотезу про наявність чи відсутність трендової складової, отримуємо додаткову інформацію для побудови моделі.

Аналіз тренд-сезонних моделей досить суттєво залежить від якості виявлення тренду. Пропонується застосувати для моделювання елементи теорії нечітких множин [5], та F-перетворення яке описано в публікації [8].

Традиційний підхід до визначення сезонності полягає в тому, щоб за допомогою статистичних методів отримати відповідь про наявність сезонності в певному чіткому переліку місяців.

Це відповідає класичному підходу до представлення множин, який полягає в тому, що існує певна функція належності $f_A(x)$, представлена у вигляді (2.31), яка б визначала, чи належить x до множини A.

$$f_A(x) = \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases}. \tag{2.31}$$

Відповідно до підходу запропонованому в теорії нечітких множин, функція належності може приймати не лише два значення 0 та 1, а всі значення з діапазону [0,1] характеризуючи таким чином належність елемента до множини. Таким чином 1 будуть значення які повністю належать множині, а значення 0 – які повністю не належать множині.

В даному випадку важливо відмітити, що на відміну від ймовірності, нечіткість вимірює ступінь з якою певний об’єкт належить множині, а не ймовірність його належності (у випадку з ймовірністю, після того як наступить подія, об’єкт або буде належати множині, або ні).

В цьому випадку нечіткі множини дозволяють досить вдало передавати відомості отримані від людини і виражені у вигляді лінгвістичних змінних[14].

Одним з механізмів розділення даних є нечіткі розбиття Руспіні [7], зображені на рисунку 2.5

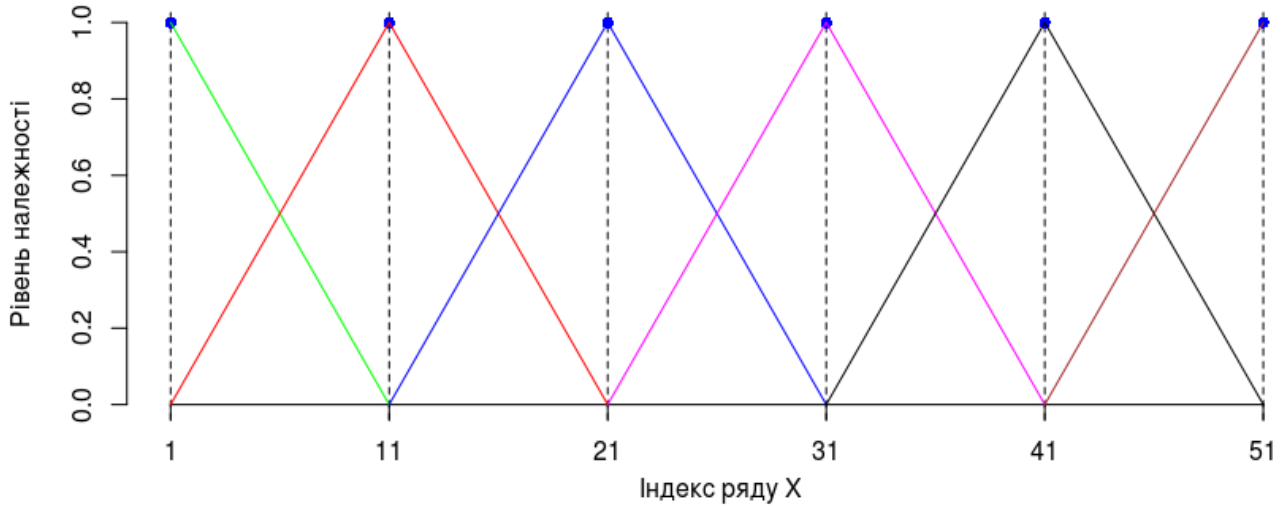

Рисунок 2.5 – Розбиття на шість нечітких розділів.

Відповідно, за допомогою прямого та зворотного F-перетворення[8] можна апроксимувати значення, які належать цим нечітким розбиттям і отримати компоненти, які характеризують значення функції в кожному з центрів нечітких розбиттів.

Застосування F-перетворення поєднує в собі переваги класичних інтегральних трансформацій і апроксимаційних систем нечіткого виводу, які

ґрунтуються на нечітких правилах виводу. Також перевагою цього підходу є можливість легкої інтерпретації результатів і візуалізації.

F-перетворення може бути прямим і зворотнім. За допомогою зворотного перетворення можна відновити функцію з F-компонент отриманих після застосування прямого F-перетворення. Це можна відобразити на рисунку 2.6

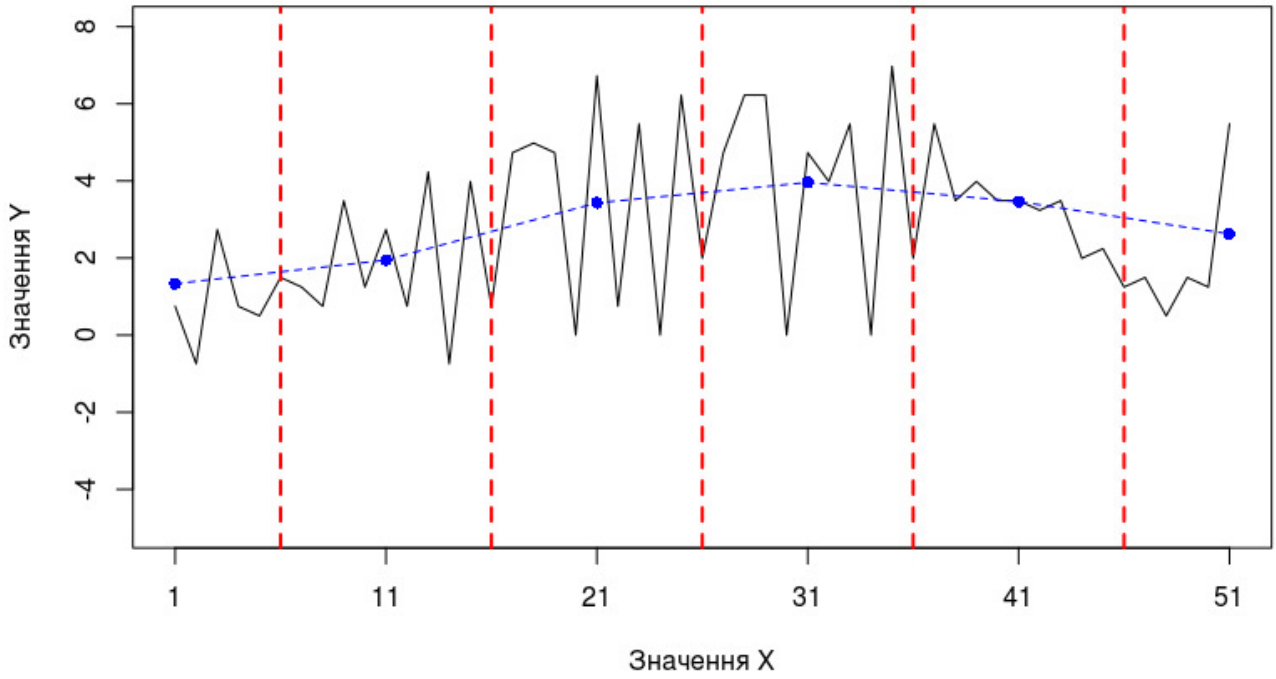

Рисунок 2.6 – Апроксимація динамічного ряду за допомогою F-перетворення.

В загальному випадку, розрахунок F-компонент здійснюється відповідно до формули (2.32) наведеної в [8]:

$$F_k = \frac{\int_a^b f(x) A_k(x) dx}{\int_a^b A_k(x) dx}, \qquad (2.32)$$

де A_k – функція належності, F_k – F-компонента – функції f.

Застосовуючи цей підхід, пропонується використовувати для адитивної тренд-сезонної моделі (2.14) обчислення і подання трендової компоненти із використанням F-перетворення[15].

Трендова компонента U_i ітеративно відшукується з використанням F-компоненти, тобто створивши n нечітких розбиття Руспіні, для кожного з них буде визначений центр розбиття i_k, функція належності A_k і значення U_k, що представляє собою точки, які належать тренду динамічного ряду(2.33):

$$U_k = \frac{\sum X_{t_i} A_k(t_i)}{\sum A_k(t_i)}, k = 1,...,n, \qquad (2.33)$$

У випадку, якщо значення тренду U_i необхідно відшукати для значень які відмінні від центрів нечіткого розбиття, тоді використовуються значення F-компоненти, що передує шуканому - U_{d0}та наступне за ним U_{d1}, після чого розраховують $U_i^{(1)}$ (2.34):

$$U_i^{(1)} = U_{d0} + \frac{U_{d1} - U_{d0}}{d1 - d0}(i - d0) \qquad (2.34)$$

.

Середню сезонну хвилю (2.35) можна представити у вигляді

$$V_j^{(1)} = \frac{\sum_{i=1}^{m} \bar{l}_{ij}}{m}, \text{ де } \bar{l}_{ij} = \frac{l_{ij}^{'}}{\sigma_i}, \qquad (2.35)$$

де $l_{ij}^{'}$ - окремі місячні відхилення

Для аналізу багатовимірних даних впорядкованих сукупностей чітких та нечітких значень розроблена модель із врахуванням даних попарних порівнянь експертних оцінок для непрямого визначення параметрів функції належності. Запропоновано подати дані попарних порівнянь експертних оцінок у вигляді матриці M, таким чином, що $M = \{m_{ij}\}$, де кожен елемент m_{ij} відображає відношення ступеня належності характеристики двох значень до нечіткої

множини S, $\mu_S(x)$ – відображає рівні належності елементів x до нечіткої множини S і для потреб нечіткої лінійної регресійної моделі апроксимується трикутною функцією належності.

Нехай $\tilde{Y}$ - нечіткі дані неоднорідних послідовностей, Xj={xij}, j=1..n, $i=1..m$ - чіткі фактори. Тоді в загальному випадку рівняння регресії має вигляд(2.36):

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + ... + \tilde{A}_n X_n, \qquad (2.36)$$

де $\tilde{Y}_i = (y_i, e_i)$, $i=1..m$ - нечітка величина з центром y_i і шириною e_i.

$\tilde{A}_j = (a_j, c_j)$, $j=0..n$ - нечітка величина з центром a_i і шириною c_i.

В даній постановці задача згідно [9] зводиться до задачі лінійного програмування:

мінімізувати функцію (2.37)

$$S = c_0 + \sum_{j=1}^{n} c_j \sum_{i=1}^{m} x_{ij}, \qquad (2.37)$$

за умов

$$c_0 \geq 0, c_j \geq 0, \; j=1..n,$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} + (1-h)\left[c_0 + \sum_{j=1}^{n} c_j \left| x_{ij} \right| \right] > y_i + (1-h)e_i,$$

$$a_0 + \sum_{j=1}^{n} a_j x_{ij} - (1-h)\left[c_0 + \sum_{j=1}^{n} c_j \left| x_{ij} \right| \right] < y_i - (1-h)e_i, i=1..m$$

де h –коефіцієнт чіткості $h \in \{0, 1]$. Результатом розв'язку цієї задачі є a_j та c_j, тобто нечіткі коефіцієнти.

Нехай відомості про скоєні злочини та чинники злочинності подані у вигляді нечітких значень.

В такому випадку, відповідно до [79], нечітка регресійна модель злочинності прийме вигляд (2.38):

$$\tilde{Y} = (\tilde{A}_0 \oplus \tilde{A}_1 \otimes \tilde{X}_1 \oplus ... \oplus \tilde{A}_j \otimes \tilde{X}_j) = \tilde{A} \otimes \tilde{X}\,, \qquad (2.38)$$

де $\tilde{y}_i = (y_i, e_i)_L$ - компоненти вектора $\tilde{Y}$, нечітка величина з цетром y_i і шириною e_i;

$\tilde{x}_{ij} = (x_{ij}, v_{ij})_L, j = 1...p$ - компоненти вектора X_j, нечітка величина з центром x_{ij} і шириною v_{ij};

$\tilde{A}_j = (a_j, c_j)_L, j = 0..p$ - нечітка величина з центром a_j і шириною c_j.

$L(x)$ - функція належності, така що

1) $L(x) = L(-x)$;

2) $L(0) = 1, L(1) = 0$;

3) L зростає на $[0, \infty)$;

4) L може обернена на $[0,1]$.

Відповідно до моделі, одержуємо таку задачу математичного програмування: мінімізувати (2.39)

$$S = \sum_{i=1}^{n} \max_{1 \le j \le p} (|a_j| v_{ij}, |x_{ij}| c_j) - |L^{-1}(h)| e_i\,, \qquad (2.39)$$

за умови:

$$\left| y_i - \sum_{j=1}^{p} a_j x_{ij} \right| \le |L^{-1}(h)| \max_{1 \le j \le p} (|a_j| v_{ij}, |x_{ij}| c_j) - |L^{-1}(h)| e_i$$

та виконання умови $c_j \ge 0, \partial e i = 1,...,n; j = 1,...,p$.

Результатом розв'язку цієї задачі є a_j та c_j, тобто нечіткі коефіцієнти.

Оскільки кількість змінних в даній моделі значна, може виникнути явище мультиколінеарності. Для перевірки наявності мультиколінеарності, скористаємось таким підходом [78]:

Обчислюємо коефіцієнт нечіткої кореляції (2.40) для X_l, X_k, де $k \neq l, k = 1..n, l = 1..n$:

$$R_{x_k,x_l} = \frac{\sum_{i=1}^{n}(x_{ik} - \overline{x}_{ik})(x_{il} - \overline{x}_{il})}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_{ik})^2 \cdot \sum_{i=1}^{n}(x_{il} - \overline{x}_{il})^2}} \qquad (2.40)$$

і знаходимо $\tilde{R}_{x_k,x_l} = [R^-_{x_k,x_l}, R^+_{x_k,x_l}]$ де: $R^-_{x_k,x_l} = \min\{R_{x_k,x_l} \mid x_{ik} \in \tilde{x}_{ik}, x_{il} \in \tilde{x}_{il}\}$,

$R^+_{x_k,x_l} = \max\{R_{x_k,x_l} \mid x_{ik} \in \tilde{x}_{ik}, x_{il} \in \tilde{x}_{il}\}$, і $x^L_{ik} \le x_{ik} \le x^R_{ik}$, $x^L_{il} \le x_{il} \le x^R_{il}$, а також значення $\overline{x}_k \, i \, \overline{x}_l$ знаходяться за формулами (2.41, 2.41):

$$\overline{x}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n}, \qquad (2.41)$$

$$\overline{x}_l = \frac{\sum_{i=1}^{n} x_{il}}{n}. \qquad (2.42)$$

Для розрахунку $R^+_{x_k,x_l}$, використовуються x^L_{ik}, x^L_{il} - ліві значення кожної змінної x_k, x_l, а для розрахунку $R^-_{x_k,x_l}$ - x^R_{ik}, x^R_{il} - праві значення кожної змінної.

Таким чином, застосування цієї формули дозволяє обчислити коефіцієнти кореляції для нечітких рядів даних і перевірити наявність кореляційного зв'язку між факторами, які подані у вигляді неоднорідних послідовностей із нечіткими значеннями рівнів даних.

Розв'язавши задачу задачу математичного програмування (2.39)

отримуємо коефіцієнти нечіткої регресійної моделі. Це дозволяє інформаційно-аналітичним системам обробляти неоднорідні послідовності даних, представлених у вигляді нечітких значень.

Таким чином з'являється можливість поєднання інформації, отриманої за допомогою вимірювання або статистичних обліків, із відомостями, отриманими від експертів.

2.4 Висновки.

Даний розділ присвячений розробці моделей неоднорідних послідовностей. Для цього були визначені показники, що можуть застосовуватися для аналізу неоднорідних послідовностей, розроблені підходи для виявлення взаємозв'язків між неоднорідними послідовностями, та запропоновані моделі неоднорідних послідовностей.

В результаті проведення аналізу способів моделювання, подання та опрацювання інформації, отриманої у вигляді даних гетерогенного походження, отриманих від людино-машинних систем і представлених у вигляді неоднорідних послідовностей, було встановлено, що:

1. Для дослідження зібраних відомостей представлених у вигляді інформаційних повідомлень про появу події чи витягів з інформаційних систем обліку, які отримуються нерівномірно у часі і формують неоднорідні послідовності даних, для зменшення рівня залучення людини в процес дослідження, доцільно подавати їх у вигляді динамічного ряду і використовувати показники, що характеризують зібрані відомості статистично.

2. Для дослідження взаємозв'язків між кількісними показниками пропонується використовувати існуючі методи кореляційного аналізу та множинної регресії. Для визначення зв'язку між якісними (атрибутивними) ознакам доцільно застосовувати коефіцієнт спряженості Чупрунова і асоціації Пірсона.

3. У випадку утворення ряду неоднорідних послідовностей, який

відображає сукупність даних, що була отримана об'єднанням певної групи послідовності коротких вимірювань, пропонується використання способу розрахунку коефіцієнту кореляції із врахуванням наявної різниці в дисперсіях та математичних сподіваннях різних груп в середині однієї вибірки.

4. Пропонується застосовувати спосіб розрахунку коефіцієнту кореляції із врахуванням наявної різниці в дисперсіях та математичних сподіваннях різних груп в середині однієї вибірки за таких застережень:

- у випадку коли статистичний зв'язок між величинами відсутній, сумарна оцінка кореляції $\tau_{A,B}$ буде зміщеною, причому у випадку коли коефіцієнт $\lambda = n_1 / (n_1 + n_2)$ буде дорівнювати 0.5, зміщення оцінки буде максимальне;

- у випадку негомогенності кореляційного зв'язку, результат застосування формули розрахунку сумарної оцінки кореляції $\tau_{A,B}$ буде повертати зміщений результат;

- у випадку застосування в умовах парадоксу Сімпсона, тобто у випадку, якщо математичні очікування зсунуті в обох вибірках з різними знаками), результат проведення кореляційного аналізу для об'єднання двох вибірок, які мають позитивний кореляційний зв'язок може бути негативним.

5. Для виявлення аномальних рівнів при попередньому аналізі неоднорідних послідовностей, пропонується використовувати критерій Ірвіна.

6. Для попередньої перевірки гіпотези про наявність чи відсутність тренду серед рівнів значень неоднорідної послідовності, представленої у вигляді динамічного ряду, пропонується використовувати критерій Стьюдента, або за умови малої кількості даних, критерій Велча.

7. Для обробки даних неоднорідних послідовностей поданих у вигляді чітких значень факторних змінних та нечітких значень факторів відгуку, пропонується нечітка модель регресії, яка дозволяє використовувати дані з різної природою походження в рамках однієї моделі.

8. Отримала подальшого розвитку тренд-сезонна модель неоднорідних послідовностей, в котрій, на відміну від існуючих тренд-сезонних моделей,

трендова складова подається у вигляді інтерпольованих усереднених значень із врахуванням функції належності, яка асоційована із кожним нечітким розділом, що дозволяє застосовувати дану модель для коротких вибірок без втрати крайових значень.

Список використаних джерел у даному розділі наведено у повному списку використаних джерел під номерами: [5, 7, 8, 9, 14, 15, 78, 79, 137 - 150].

3 РОЗРОБКА МЕТОДІВ АНАЛІЗУ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

3.1 Розробка підходу для узгодження експертних оцінок

Існуючі підходи до аналізу динамічних рядів , як правило, оперують кількісними даними, але досить часто існують додаткові якісні характеристики які можуть бути включені до моделі.

Ця особливість безпосередньо пов’язана з таким аспектом функціонування інформаційно-аналітичних систем, як отримання неоднорідних послідовностей у вигляді наборів даних із джерел з різнорідною природою походження і проведення перетворення, що приводять їх до узагальненої структури і формату.

Потреба в такому комбінуванні даних виникає при зборі інформації, що стосуються соціальних та соціально-економічних систем. Під впливом різних чинників, зібрані статистичні показники не повністю відображають справжній стан явища.

Яскравим прикладом таких даних є інформація про латентну злочинність. Латентна злочинність - це сукупність передбачених кримінальним законом діянь, які з різних причин не враховані органами внутрішніх справ, прокуратурою, службою безпеки, судом [151].

В таких випадках інформація про латентність злочинності може бути отримана лише на основі соціологічних або експертних досліджень. Результати таких досліджень досить часто є неузгодженими.

Чинники, які детермінують злочинність, також є невизначеними, виключенням є тільки чинник часу.

Також, в якості прикладу можна навести дослідження рівня оплати праці в умовах тонізації економіки.Використання показників отриманих в таких умовах для економічних розрахунків та планування витрат є досить суперечливим.

У зв’язку з цим при аналізі даних такого роду виникають задачі:

1. Моделювання динамічного ряду із врахуванням експертних оцінок кількісних показників за умови чітко визначених факторів.

2. Моделювання динамічного ряду з урахуванням експертних оцінок кількісних показників результату і чинників, які його детермінують .

Для розв'язування цих задач доцільно скористатися математичним апаратом теорії нечітких множин. Експертні відомості про рівні динамічного ряду в такому випадку потребують узгодження.

На основі нечіткої регресії можна враховувати експертні оцінки в статистичній моделі рівнів динамічного ряду. Такі дані, одержані від різних експертів, досить часто є неузгодженими, іноді суперечливими.

В залежності від особливостей предметної області, що досліджується, для обробки та узгодження експертних відомостей можуть бути використані різні підходи.

Для потреб створення нечітких регресійних моделей, доцільно використати непрямий метод побудови функцій належності (метод попарних порівнянь) запропонований в роботі Сааті [6].

Пропонується подати дані попарних порівнянь експертних оцінок у вигляді матриці *M*, таким чином, що $M = \{m_{ij}\}$, де кожен елемент m_{ij} відображає відношення ступеня належності характеристики двох значень до нечіткої множини *S*.

Ця множина може бути представлена в такому вигляді (3.1)

$$S = \{[x, \mu_s(x)]\}, \qquad (3.1)$$

де $x \in X$, X скінчена множина значень, яка складається з *n* елементів.

$\mu_s(x)$ – відображає рівні належності елементів множини *x* до нечіткої множини *S*. Таким чином, необхідно отримати значення функції належності, із врахуванням виконанная умови $0 \leq \mu_s(x) \leq 1$, для кожного з елементів множини x із використанням матриці попарних порівнянь рівня належності елементів $M = \{m_{ij}\}$, де кожний елемент m_{ij} розраховується за формулою (3.2):

$$m_{ij} = \frac{\mu_S(x_i)}{\mu_S(x_j)}, \tag{3.2}$$

тобто елементами матриці є відношення значень функцій належності. При цьому вважаємо, що матриця *M* має такі властивості:

1) Ця матриця обернено-симетрична, тобто для будь-якого елемента матриці m_{ij} виконується умова (3.3)

$$m_{ij} = \frac{1}{m_{ji}}. \tag{3.3}$$

2) Для будь-якої трійки елементів матриці з індексами *i, j, k* , тобто m_{ij}, m_{ik}, m_{jk} , їх значення підпорядковані співвідношенню $m_{ik} = m_{ij} m_{jk}$, тобто матриця є кардинально узгодженою.

Припустимо, що значення функції належності для кожного з елементів множини *X* відомі, тоді позначимо $\mu_S(x_i) = w_i, i = \overline{1,n}$, вектор значень коефіцієнтів, що відшукуються прийме вигляд (3.4).

$$W^T = (w_1 \quad w_2 \quad \dots \quad w_n). \tag{3.4}$$

Виходячи з цього, матрицю попарних порівнянь *M*, яка зберігає відомості визначені експертами, можна подати у вигляді (3.5):

$$M = \begin{pmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \dots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \dots & \frac{w_2}{w_n} \\ \dots & \dots & \dots & \dots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \dots & \frac{w_n}{w_n} \end{pmatrix}. \tag{3.5}$$

В такому разі, значення вектора W можна знайти виходячи з відношення (3.6), зазначеного в роботі [152]:

$$w_i = \frac{1}{C}\sum_{j=1}^{n} m_{ij} = \frac{\sum_{j=1}^{n} m_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}}, \qquad (3.6)$$

Дійсно, для довільного i-го рядка одержуємо (3.7):

$$\sum_{j=1}^{n} m_{ij} = \sum_{j=1}^{n} \frac{w_i}{w_j} = w_i \sum_{j=1}^{n} \frac{1}{w_j} = C w_i, i = \overline{1,n}. \qquad (3.7)$$

Отже, власний вектор W з точністю до константи можна розрахувати безпосередньо за елементами матриці M.

Одержаний вектор $W^T = (w_1 \quad w_2 \quad \dots \quad w_n)$ є власним вектором матриці M, що відповідає власному числу, яке дорівнює n.

Співвідношення (3.6) дозволяє визначити значення функції належності $\mu_S(x_i) = w_i$ для переліку значень x_i, $i = \overline{1,n}$, тоді, коли матриця M є n-ідемпотентною.

Але, у випадку узгоджень експертних оцінок, отриманих від різних експертів, матриця M, яка містить результати попарних порівнянь значимості оцінок,в більшості випадків такою не є.

Тому вектор W^T оцінює вагові коефіцієнти з похибкою, при чому ця похибка тим більша, чим більше матриця M відрізняється від n-ідемпотентної.

В зв'язку з цим у роботі [11] пропонується відшукувати не матрицю M, а злагоджену матрицю M', яка мінімально відрізняється від матриці M в розумінні найменших квадратів.

Відповідно до попередньо введених позначень:

- $M=(m_{ij}),\ i=\overline{1,n},\ j=\overline{1,n}$ - початкова матриця попарних порівнянь;

- $M'=(m'_{ij}),\ i=\overline{1,n},\ j=\overline{1,n}$ - злагоджена матриця попарних порівнянь, яка відшукується

- λ_{ij} – множники Лагранжа

Тоді елементи злагодженої матриці $M'=(m'_{ij})$, яка відшукується, є розв'язком такої системи рівнянь(3.8), запропонованої в [6]:

$$\begin{aligned}
&M_1M'^{T}_{1}+B_1M'^{T}_{2}+B_1M'^{T}_{3}+\ldots+B_1M'^{T}_{n}=P_1\\
&B_2M'^{T}_{1}+M_2M'^{T}_{2}+B_2M'^{T}_{3}+\ldots+B_2M'^{T}_{n}=P_2\\
&\qquad\qquad\qquad\qquad\qquad\qquad\ldots\\
&B_nM'^{T}_{1}+B_nM'^{T}_{2}+B_nM'^{T}_{3}+\ldots+M_nM'^{T}_{n}=P_n,
\end{aligned}\qquad (3.8)$$

де матриця B_s розраховується за формулами (3.9):

$$B_s=\begin{pmatrix}
\frac{\lambda_{s1}}{n} & 0 & 0 & 0\\
0 & \frac{\lambda_{s2}}{n} & 0 & 0\\
0 & 0 & \ddots & 0\\
0 & 0 & 0 & \frac{\lambda_{sn}}{n}
\end{pmatrix},\qquad (3.9)$$

а матриця P_s може бути розрахована за (3.10)

$$P_s=\begin{pmatrix}
2m_{s1}+\lambda_{s1}\\
2m_{s2}+\lambda_{s2}\\
\ldots\\
2m_{sn}+\lambda_{sn}
\end{pmatrix},M'^{T}_{s}=(m'_{s1},m'_{s2}\ldots m'_{sn})^T,\qquad (3.10)$$

де $s=1..n$. Для розв'язку одержаної системи лінійних рівнянь може бути

застосований будь-яким відомий метод (Гаусса, Жордана-Гаусса тощо).

Одержана система лінійних алгебраїчних рівнянь може бути розв'язаною чисельно Разом із тим специфічна структура системи (3.8) дозволяє одержати розв'язок у явному вигляді (3.11):

$$\begin{aligned} M'^{T}_{j} &= (B_j^{-1}M_j - I)^{-1}\left[\left(B_1^{-1}M_1 - I\right)M'^{T}_{1} - B_1^{-1}P_1 + B_j^{-1}P_j\right] = \\ &= (B_j^{-1}M_j - I)^{-1}(B_1^{-1}M_1 - I)M'^{T}_{1} + \\ &+ (B_j^{-1}M_j - I)^{-1}(B_j^{-1}P_j - B_1^{-1}P_1) = C_{1j}M'^{T}_{1} + D_{1j},\ j = \overline{2,n}, \end{aligned} \tag{3.11}$$

при цьому, M'^{T}_{1}, визначається з (3.12):

$$M'^{T}_{1} = \left(M_1 + \sum_{j=2}^{n} C_{1j}\right)^{-1}\left(P_1 - \sum_{j=2}^{n} D_{1j}\right). \tag{3.12}$$

Таким чином, остаточне визначення матриці M' залежить від визначення значення множників Лагранжа. Нелінійна система рівнянь, яка при цьому виникає може бути розв'язана лише чисельно.

В результаті ітераційної процедури [152] одержується вектор функцій належності (3.13),

$$W^T \approx M^T = (\mu_A(x_1),...,\mu_A(x_n)) \tag{3.13}$$

Нечітка множина, що описується функцією належності, не є нормованою, тому множина приводиться до нормалізованої за формулою (3.14):

$$\tilde{\mu}_A(x_i) = \frac{\mu_A(x_i)}{\max\limits_i \mu_A(x_i)}. \tag{3.14}$$

Вказана процедура узгодження експертних оцінок використовується в даній роботі для побудови функцій належності нечітких рівнів динамічного ряду.

Функція належності нечіткої величини може приймати різний вигляд, але для потреб побудови лінійної регресійної моделі вона апроксимується за допомогою трикутної функції належності (рис.3.1)

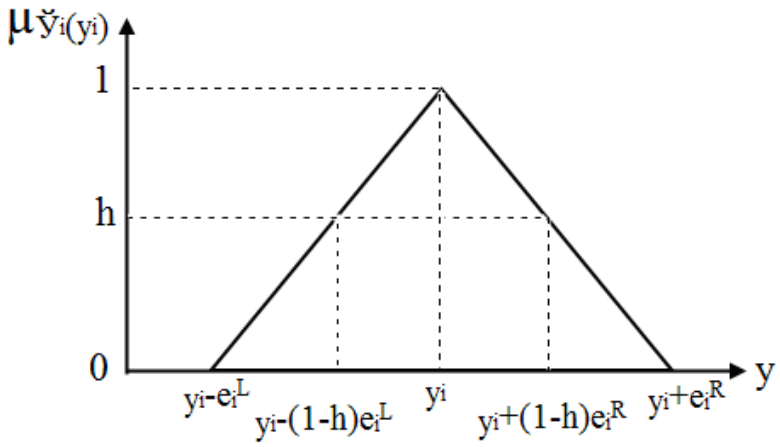

Рисунок 3.1 – Трикутна функція належності нечіткої величини.

Таким чином, якісні характеристики експертів, подані у вигляді функцій належності можуть бути застосовані для корегування даних. Подані в такому вигляді відомості потребують методів для їх аналізу і обробки.

3.2 Розробка методу визначення значимих чинників нечіткої регресійної моделі

В залежності від того, які дані доступні користувачу інформаційно-аналітичної системи (виключно чіткі, чи також наявні неоднорідні послідовності подані у вигляді нечітких змінних) необхідно обрати підхід для обробки даних.

Відштовхуючись від того, що проведення інтелектуального і оперативного аналізу, підготовка результатів оцінювання стану предметної

області у вигляді придатному для прийняття рішення є ключовими задачами ІАС, виникає потреба у розвитку методів інтелектуального аналізу даних, представлених у вигляді неоднорідних послідовностей.

Фактично, у разі подання неоднорідних послідовностей у вигляді нечітких даних, необхідно внести певні корективи до класичної регресійної моделі, яка застосовується виходячи з певних припущень і підходів до інтерпретації.

Розглянемо ймовірнісний підхід, тобто випадок застосування статистичної моделі, а саме певного математичного співвідношення між рівнями динамічного ряду і чинниками, які визначають його значення.

У такому разі виходять з того, що рівень послідовності даних, яка досліджується, є реакцією, а чинники, які його детермінують - фактори.

Позначимо через ψ_l функцію, яка зв'язує реакцію y_l з факторами x_{li}, де $i = \overline{1..n}$ Зв'язок між рівнями факторів і реакцією системи подається у вигляді співвідношень(3.15)

$$y_l = y_l\left(x_1, x_2, \ \ldots\ , x_n\right), l = \overline{1,m} \quad (3.15)$$

тобто у вигляді існуючої сукупності реакцій, які залежать від визначеної сукупності факторів, а вид залежностей ψ_l заздалегідь невідомий. Тому використовуються наближені співвідношення (3.16):

$$\overline{y}_l = \varphi_l\left(x_1, x_2, \ \ldots\ , x_n\right), l = \overline{1..m} \quad (3.16)$$

Основними вимогами, які висуваються до сукупності факторів, є сумісність і незалежність.

Таким чином, вважається, що всі їхні комбінації здійсненні, а також існує можливість встановити значення фактора на будь-якому рівні незалежно від рівнів інших факторів.

Надалі вважаємо, що існує функція $y=\psi(x_1, x_2, \dots, x_n)$, яка визначає рівень значення відгуку і залежить від факторів $x_1, x_2, \dots, x_n$. Також до властивостей функції віднесемо можливість диференціації по кожному з аргументів необхідну кількість разів.

Виходячи з цих припущень, її можна подати у вигляді збіжного ряду Маклорена (3.17):

$$y = b_0 + \sum_{1\le i\le n} b_i x_i + \sum_{1\le i\le j\le n} b_{ij} x_i x_j + \cdots + \sum_{i_1, i_2, \dots, i_n} b_{i_1, i_2, \dots, i_n} x_1^{i_1} x_2^{i_2} \dots x_n^{i_n} + \cdots \qquad (3.17)$$

На основі експериментальних даних можна дістати оцінку цього співвідношення (3.18):

$$\eta = \bar{b}_0 + \sum_{1\le i\le n} \bar{b}_i x_i + \sum_{1\le i\le j\le n} \bar{b}_{ij} x_i x_j + \cdots + \sum_{i_1, i_2, \dots, i_n} \bar{b}_{i_1, i_2, \dots, i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}, \qquad (3.18)$$

тобто виходячи з даного припущення, ми можемо перейти до поліноміальної статистичної моделі. Така модель з певною точністю апроксимує реальну залежність рівнів відгуку від факторів, які його детермінують.

В деяких випадках можна обмежити ступінь поліномів в моделі до двох, виходячи з припущення, що взаємодії між елементами моделі вищих порядків такі, що ними можна знехтувати (3.19):

$$\eta = \bar{b}_0 + \sum_{i=1}^{n} \bar{b}_i x_i + \sum_{i=1}^{n} \bar{b}_{ii} x_i^2 + \sum_{i\ne j}^{n} \bar{b}_{ij} x_i x_j, \qquad (3.19)$$

чи, навіть поліномами першого ступеня (3.20):

$$\eta = \bar{b}_0 + \sum_{i=1}^{n} \bar{b}_i x_i . \qquad (3.20)$$

Функція реакції крім поліноміальної функції може подаватися і більш складною залежністю від факторів.

У деяких випадках ці залежності можна привести до лінійного вигляду. Проте невраховані взаємодії факторів і припущення здійснені для оцінки співвідношення можуть істотно впливати на цю функцію.

В такому разі, невраховані фактори варто подати випадковою величиною ε, закон розподілу якої необхідно встановити на основі наявних експериментальних даних.

Тоді поліноміальна модель приймає вигляд(3.21):

$$\eta = \bar{b}_0 + \sum_{1\le i\le n} \bar{b}_i x_i + \sum_{1\le i\le j\le n} \bar{b}_{ij} x_i x_j + \cdots + \sum_{i_1,i_2,\ldots,i_n} \bar{b}_{i_1,i_2,\ldots,i_n} x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} + \varepsilon, \qquad (3.21)$$

або для лінійного випадку(3.22):

$$\eta = \bar{b}_0 + \sum_{i=1}^{n} \bar{b}_i x_i + \varepsilon. \qquad (3.22)$$

Таким чином, при побудові моделі необхідно враховувати, що основними вимогами, які висуваються до сукупності факторів, є сумісність і незалежність, а невраховані фактори варто подати випадковою величиною ε, закон розподілу якої потрібно встановити на основі наявних експериментальних даних.

Відштовхуючись від цього, розглядаємо лінійну статистичну багатофакторну регресійну модель спостережень, яка має має вигляд(3.23):

$$y_i = \theta_1 x_{i1} + \ldots + \theta_p x_{ip} + \varepsilon_i \ , \ i = 1,\ldots,n, \ \ n \ge p, \qquad (3.23)$$

де y_i - значення пояснювальної змінної в i-му спостереженні;

x_{ij} - відоме значення j-ої пояснюючої змінної в i-му спостереженні;

θ_j - невідомий коефіцієнт при j -ій пояснюючій змінній;

ε_i - випадкова складова (похибка) в i-му спостереженні.

Для того, щоб дана модель була адекватною, повинні виконуватися вимоги до вектору похибки, а саме: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ - незалежні випадкові величини, які мають однаковий нормальний розподіл $N(0, \sigma^2)$ з нульовим математичним сподіванням і дисперсією $\sigma^2 > 0$.

Також, як правило, до пояснюючих змінних включається змінна, яка тотожно дорівнює одиниці і є першою пояснюючою змінною, так що $x_{i1}=1$, $i=1, \ldots, n$.

Виходячи з вищеописаного можна виділити такі складнощі у застосуванні лінійної багатофакторної статистичної моделі:

- невизначеність у значеннях рівнів результату або факторів;

- необгрунтованість результатів при недостатній кількості спостережень;

- складнощі в оцінюванні та перевірці припущень стосовно параметрів розподілу випадкової складової моделі, навіть за наявності необхідної кількості спостережень, тобто однаковий нормальний розподіл $N(0, \sigma^2)$ з нульовим математичним сподіванням і дисперсією;

- складність оцінювання моделі у випадку невизначеності із взаємозалежністю між рівнями відгуку і детермінуючими змінними;

- неточність лінійної моделі за рахунок відкидання з розрахунку складових взаємодій вищих порядків;

Відповідно до нечітких підходів до побудови нечітких лінійних моделей і в залежності від того, яким чином подана невизначеність в рівнях даних може бути застосована одна з двох лінійних регресійних моделей:

1) для нечітких рівнів відгуку і чітких факторів (3.24):

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + \ldots + \tilde{A}_n X_n, \qquad (3.24)$$

де позначення визначені таким чином:

- $\tilde{Y}$ нечіткі дані неоднорідних послідовностей; $\tilde{Y}_i = (y_i, e_i)$, $i = 1..m$ -

нечітка величина з центром y_i і шириною e_i;

- $X_j=\{x_{ij}\}, j=1..n, i=1..m$ - чіткі фактори;

- $\tilde{A}_j=(a_j,c_j),\ j=0..n$ - нечітка величина з центром a_i і шириною c_i.

2) Для нечітких рівнів відгуку і нечітких рівнів факторів пропонується використовувати рівняння (3.25):

$$\tilde{Y}=(\tilde{A}_0\oplus\tilde{A}_1\otimes\tilde{X}_1\oplus...\oplus\tilde{A}_j\otimes\tilde{X}_j)=\tilde{A}\otimes\tilde{X}\quad , \qquad (3.25)$$

де позначення визначені таким чином:

- $\tilde{Y}$ нечіткі дані неоднорідних послідовностей, $\tilde{Y}_i=(y_i,e_i),\ i=1..m$ - нечітка величина з центром y_i і шириною e_i.

- $\widetilde{X}$ нечіткі дані неоднорідних послідовностей, $\tilde{x}_{ij}=(x_{ij},v_{ij})_L, j=1...p$ - компоненти вектора X_j, нечітка величина з центром x_{ij} і шириною v_{ij};

- $\tilde{A}_j=(a_j,c_j),\ j=0..n$ - нечітка величина з центром a_i і шириною c_i.

Також відзначимо, що існують вимоги до функції належності нечіткої величини. Вони були розглянуті в другому розділі.

Підходи до оцінювання параметрів нечіткої моделі також були розглянуті в другому розділі. Задача нечіткого регресійного аналізу розглядалась в роботі - [79], де пропонувалося використовувати для розв'язку методи лінійного програмування. Основні недоліки, які виникають під час використання такого підходу були розглянуті в першому розділі.

Варто відмітити, що із особливостями методів оцінювання параметрів нечіткої регресійної моделі, на основі лінійного програмування пов’язано недостатнє обґрунтування співвідношення між рішенням задачі лінійної оптимізації значення сумарної «нечіткості» і мінімізації сумарної похибки моделі по відношенню до навчальної вибірки [78], чутливість до аномальних значень [114], висока ймовірність появи мультиколінеарності із збільшенням кількості факторів [115].

В першому розділі були розглянуті підходи до розв'язання цих проблем, а саме скорочення кількості факторних змінних, відкидаючи незначущі (або включаючи до моделі лише значущі) за допомогою шагового методу регресійного аналізу [116].

В якості критеріїв вибору факторів для задачі побудови нечіткої лінійної регресійної моделі в такому разі використовується критерій Фішера [117]. Послідовне додавання або відкидання ознак здійснюється відповідно до заданого критерію. В першому розділі був описаний суттєвий недолік цього підходу, а саме неможливість отримання оптимального рівняння регресії у випадку кореляції між факторними змінними. В такому випадку, значуща змінна може бути ніколи не включена до рівняння, а другорядні змінні можуть бути додані. Таким чином виникає проблема побудови нечіткої лінійної регресійної моделі із врахуванням значущих змінних.

Розглянемо метод визначення значимих чинників нечіткої регресійної моделі для аналізу даних неоднорідних послідовностей [13,17,19].

Етап 1. Для визначення функції належності скористаємось підходом описаним в пункті один третього розділу. Створена в такому випадку функція буде мати такий вигляд (3.26):

$$\mu_{\tilde{Y}_i} = \max\left\{1 - \frac{y - y_i}{e_i}, 0\right\}, \qquad (3.26)$$

де y_i - центр нечіткої величини, e_i - розкидання значень нечіткої величини.

Етап 2. Подамо чіткі дані про фактори у вигляді матриця значень p пояснюючих змінних у *n* спостереженнях (3.27):

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad (3.27)$$

а дані про центри нечіткої величини $\tilde{Y}$ у вигляді (3.28)

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} . \qquad (3.28)$$

Ці данні пропонується використати для відшукування значущих чинників нечіткої регресійної моделі.

Для цього замість шагової регресії [116] та застосування коефіцієнта Фішера [117] пропонується використати метод найменших кутів[153].

Етап 3. Задамо початкову оцінку $\hat{\mu}_A = 0$ вектора значень залежної змінної y .

Етап 4. Обчислимо вектор кореляцій (3.29):

$$\hat{c} = X^T (y - \hat{\mu}_A) . \qquad (3.29)$$

Етап 5. Знайдемо поточний набір індексів A, що відповідає ознакам із найбільшими абсолютними значеннями кореляцій (3.30):

$$A = \{ j : |\hat{c}_j| = \hat{C} \} , \text{ де } \hat{C} = \max_{j=1,...,n} \{ |\hat{c}_j| \} . \qquad (3.30)$$

Етап 6. Знайдемо $s_j = sign(\hat{c}_j)$ для $j \in A$. Розрахуємо матриці X_A, ψ_A таким чином(3.31):

$$X_A = \left[s_{j_1} x_{j_1}, ..., s_{j_{|A|}} x_{j_{|A|}} \right], j = (j_1, ..., j_{|A|}) \in A , \ \psi_A = (1_A^T \varsigma_A^{-1} 1_A)^{-\frac{1}{2}} , \qquad (3.31)$$

де $s_j \in \{+1, -1\}$ і $|A|$- потужність множини A (кількість значень множини A)

і $\varsigma = X_A^T X_A$, 1_A – одинична матриця розміру $1\times|A|$.

Етап 7. Розрахуємо вектор $a = X^T u_A$, де $u_A = X_A w_{A'}$, $w_A = \psi_A \varsigma_A^{-1} 1_A$.

Етап 8. Розрахуємо значення $\hat{\gamma}$ згідно формули (3.32)

$$\hat{\gamma} = \min_{j\in A}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{\psi_A - a_j}, \frac{\hat{C} + \hat{c}_j}{\psi_A + a_j} \right\}. \qquad (3.32)$$

де мінімум береться по всім додатнім значенням для кожного j.

Етап 9. Знаходимо значення $\hat{\mu}_A$ для наступної ітерації: $\hat{\mu}_{A+} = \hat{\mu}_A + \hat{\gamma} u_A$.

Етап 10. Процес повторюється n раз (де n – кількість факторів), починаючи з етапу 4. Для кожної ітерації обчислюється коефіцієнт *Ср* Маллоуза [154].

Етап 11. Для побудови нечіткої регресійної моделі обираємо набір коефіцієнтів, який буде відповідати мінімальному значенню коефіцієнта *Ср*

Після цього застосовується метод розв'язку задачі побудови лінійної регресійної моделі в загальному випадку використовуючи лише обраний набір факторних змінних(3.33):

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X'_1 + \ldots + \tilde{A}_k X'_k, \qquad (3.33)$$

де $X'_j=\{x_{ij}\}$, $j=1..k$, $i=1..m$ - чіткі фактори, обрані за допомогою обрані за допомогою запропонованого методу.

3.3 Розробка методу фільтрації компонент неоднорідних послідовностей

Дослідження сезонних коливань виникає в ситуаціях, коли існують певні обставини, що виникають циклічно, незалежно від року і суттєво впливають на значення вимірюваного явища.

Варто зазначити, що інформаційно-аналітичні системи забезпечують

перетворення інформації у вигляд, який може бути використаний для прийняття рішень.

У випадку аналізу часових рядів, в тому числі представлених у вигляді неоднорідних послідовностей даних, важливим є обмеженням на мінімальну довжину ряду, який необхідний для визначення аналізу і існує потреба в розвитку методів для зменшення вимог до розміру вибірки.

Вивчення сезонності важливе для побудови моделей аналізу і прогнозування багатьох соціально-економічних явищ.

Відповідно до матеріалу розглянутого в першому розіділі, існує два основних підходи до виділення і врахування сезонності в моделях прогнозування:

- на основі моделей, що походять від ідеї, запропонованої Боксом-Дженкінсом[155];

- на основі декомпозиційного підходу до моделей, що полягає в адитивному або мультиплікативному поєднанні трендової, циклічної, сезонної компоненти та похибки[156];

Відповідно до розглянутого у другому розділі, пропонується модель сезонного явища на основі декомпозиційного підходу, яку можна представити у вигляді, який докладно описаний в другому розділі: $X_i=U_i+V_i+\varepsilon_i$, *де* $i=1,\ldots,N$.

Відштовхуюсь від запропонованої моделі, розробимо метод фільтрації компонент неоднорідної послідовності.

Для цього скористаємось ідеями ітеративної процедури, викладеними в роботі [157] і модифікуємо їх використовуючи метод апроксимації на основі F-перетворення [8].

На відміну від існуючого підходу до аналізу сезонної компоненти із застосуванням F-перетворення [118], пропонується ітеративна процедура, яка дозволяє поступово підвищувати точність, на кожній ітерації підвищуючи точність виділення складових динамічного ряду.

Відповідно до немодифікованого методу Четверикова [157] виділення складових відбувається таким чином:

На першому кроці динамічний ряд вирівнюють із застосуванням зваженого тринадцятичленного ковзного середнього. Всі ваги такого ковзного середнього дорівнюють одиниці, окрім першого і останнього членів, які беруться із вагою 0.5, відповідно до формули (3.34).

$$X_i = \frac{X_{i-6} \cdot 0.5 + \sum_{j=i-5}^{n=i+5} X_j + X_{i+6} \cdot 0.5}{12} \qquad (3.34)$$

Недоліком такого підходу є відсутність можливості вирівняти перші та останні шість значень по місяцях.

Як правило, зазначені незгладжені рівні або відкидаються в вирівняному ряді, або вирівнюються екстраполюванням значень отриманого згладженого ряду.

У випадку коротких рядів даних, або якщо дослідження сфокусовано саме на цих ділянках, дана проблема є досить суттєвою.

На другому кроці місячні відхилення «нормуються». Для цього розбивши динамічний ряд на декілька груп по роках, знаходять середньоквадратичне відхилення емпіричного ряду від вирівняного для кожної групи. Після чого місячні відхилення кожного року діляться на відповідні середньоквадратичні відхилення.

На третьому кроці знаходять попередню середню сезонну хвилю. Для цього розраховують значення середньої арифметичної з «нормованих» відхилень по місяцях.

На четвертому кроці отримують ряд, позбавлений попередньої хвилі, шляхом розрахунку попередніх значень середньої сезонної хвилі, а саме добутку середньої попередньої сезонної хвилі на середньоквадратичне відхилення кожного року.

На п’ятому кроці змінюють період згладжування зваженою ковзною середньою і застосовують її до отриманого на попередньому кроці ряду.

Як правило, для другого етапу беруть п’ять або сім значень динамічного ряду. Наприклад, формула для згладжування за сімома місяцями прийме такий вигляд(3.35):

$$X_i = \frac{X_{i-3} \cdot 0.5 + \sum_{j=i-2}^{n=i+2} X_j + X_{i+3} \cdot 0.5}{6} \qquad (3.35)$$

Після цього знаходять відхилення членів ряду від результатів цього вирівнювання і переходять до кроку два, виконуючи всі описані дії із зменшеним періодом, відшукуючи таким чином середню сезонну хвилю.

Останнім кроком у методиці стає вилучення «остаточної» сезонної хвилі. Для цього перемножують значення виділеної сезонної хвилі на множник, який враховує ступінь виразності впливу середньої сезонної хвилі та розкид відхилень у кожному році. Цей множник називають коефіцієнтом напруженості сезонної хвилі [157].

Внесемо модифікації до цього підходу, на основі застосування F-перетворення [14] для апроксимації динамічного ряду. З моделі $X_i=U_i+V_i+\varepsilon_i$, *де* $i=1,\ldots,N$, необхідно виділити сезонну компоненту V_i і проаналізувати її динаміку[15].

Така процедура обумовлена декомпозиційним підходом до аналіза сезонності. Часовий ряд, який буде оброблятися також позначимо через $\{x_i\}$.

На першому кроці, замість застосування ковзної середньої виконаємо такі дії:

1) Розділимо динамічний ряд на n однакових за розміром нечітких розбиттів Руспіні.

2) Визначимо n базисних функцій $A_1 \ldots A_n$, які покривають всі частини динамічного ряду, та відповідають ряду вимог, розглянутих нижче.

3) Проведемо пряме F-перетворення та отримаємо F-компоненти, необхідні для апроксимації трендової складової.

4) За допомогою зворотного F-перетворення отримаємо місячні значення трендової складової.

На п’ятому кроці, замість зменшення розміру зваженої ковзного середнього проведемо зменшення розмірів нечітких розбиттів руспіні відповідно вдвічі.

Таким чином, після внесених модифікацій до методу фільтрації компонент динамічного ряду, а саме з використанням методу F-перетворення для виділення трендової складової, можна відзначити, що у випадку коротких рядів даних не втрачаються крайові значення у вирівняному ряді.

Таким чином, можемо подати метод фільтрації компонент неоднорідних послідовностей у вигляді таких етапів:

Eman 1. Розділимо динамічний ряд на n нечітких розбиттів Руспіні розміром T0. Для цього визначимо n рівновіддалених точок з індексом t_k, де $k = 1..n$, які належать до цих нечітких частин, причому $t = \overline{1..N}$ і визначається за формулою (3.37)

$$t_k = 1 + h(k-1), \tag{3.36}$$

де виконуються умова (3.37):

$$N > n,\ h = \frac{N-1}{n-1}. \tag{3.37}$$

Eman 2. Визначимо n базисних функцій $A_1 ... A_n$, які покривають всі частини динамічного ряду та відповідають таким умовам:

1) Функція A_k неперервна;

2) Функція A_k монотонно зростає на $[t_{k-1,} t_k]$ і монотонно спадає на $[t_k, t_{k+1}]$;

3) Для кожної функції виконуються умови (3.38):

$$A_k : [1..N] \to [0,1], A_k(t_k) = 1. \tag{3.38}$$

4) Визначимо, що $A_k(t)=0$ якщо $t \notin (t_{k-1}, t_{k+1})$, при цьому вважаємо, що $t_0 = t_1 = 1$ і $t_{n+1} = t_n = N$

5) Для всіх $t \in [1, N]$ виконується рівняння (3.39)

$$\sum A_k(t) = 1. \tag{3.39}$$

Виходячи з перерахованих вище умов, для даного випадку скористаємося базисною функцією поданою у вигляді (3.40)

$$A_k(t) = \begin{cases} \dfrac{t - t_{k-1}}{t_k - t_{k-1}}, if: t_{k-1} \le t \le t_k \\ \dfrac{t_{k+1} - t}{t_{k+1} - t_k}, if: t_k \le t \le t_{k+1} \\ 0, в\ інших\ випадках \end{cases}. \tag{3.40}$$

Етап 3. Використовуючи базисні функції перетворимо даний динамічний ряд X у кортеж з n дійсних чисел $[U_1,...,U_n]$, які визначаються за формулою (3.41):

$$U_k = \frac{\sum X_{t_i} A_k(t_i)}{\sum A_k(t_i)}, k = 1,..,n. \tag{3.41}$$

F – компоненти, тобто U_k, представляють собою точки, які належать тренду динамічного ряду.

Для одержання значень тренду в інших точках скористаємося лінійною інтерполяцією між двома найближчими точками з відомими значеннями.

Нехай відомі значення U_{d0} та U_{d1}, тоді відшукуване значення $U_i^{(1)}$, де $\mathrm{d0} < i < d1$ можна знайти за формулою (3.42):

$$U_i^{(1)} = U_{d0} + \frac{U_{d1} - U_{d0}}{d1 - d0}(i - d0). \qquad (3.42)$$

У результаті дістається попередня оцінка тренда $U_i^{(1)}$і відхилення емпіричного ряду від вирівняного (3.43)

$$l_i' = l_i = x_i - U_i, \qquad (3.43)$$

Або поданий у вигляді (3.44)

$$l_i' = x_{ij} - U_{ij}^{(1)}, i = 1..m; j = 1..T_0. \qquad (3.44)$$

Етап 4. Для кожного року i обчислюється σ_i - середнє квадратичне відхилення (3.45):

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{T_0} l_{ij}'^2 - \frac{\left(\sum_{j=1}^{T_0} l_{ij}'\right)^2}{T_0}}{T_0 - 1}}, \qquad (3.45)$$

на яке діляться окремі місячні (квартальні) відхилення відповідного року, для проведення нормування отриманих таким чином відхилень (3.46), які в подальшому застосовуються для обчислення середньої сезонної хвилі на наступному етапі:

$$\bar{l}_{ij} = \frac{l_{ij}'}{\sigma_i}. \qquad (3.46)$$

Етап 5. З «нормованих» таким чином відхилень обчислюється в першому

наближенні середня сезонна хвиля(3.47):

$$V_j^{(1)} = \frac{\sum_{i=1}^{m} \bar{l}_{ij}}{m} \qquad (3.47)$$

Етап 6. Середня сезонна хвиля множиться на середнє квадратичне відхилення кожного року і віднімається від рівнів початкового емпіричного ряду (3.48):

$$\overline{=}_{x_{ij}} = x_{ij} - V_j^{(1)} \cdot \sigma_i \qquad (3.48)$$

Утворений за допомогою виконання цих етапів динамічний ряд позбавлений сезонної хвилі, одержаної в першому наближенні.

Етап 7. Цей ряд знову піддається нечіткому згладжуванню. Для місячних даних розмір розбиття Руспіні повинен буде складати чотири або шість точок, в залежності від того, на скільки інтенсивні дрібні коливання.

У результаті одержується нова оцінка тренда $U_i^{(2)}$.

Етап 8. Розрахуємо відхилення початкового емпіричного ряду $\{x_i\}$ від ряду $\{U_i^{(2)}\}$, одержаного в п. 5, відповідно до формули(3.49):

$$l_i^{(2)} = x_i - U_i^{(2)} \qquad (3.49)$$

Після цього ці значення знову піддаються обробці відповідно до пунктів 2 і 3 метода для виявлення наступного наближення середньої сезонної хвилі і далі цей процес ітеративно повторюється до досягнення заданої точності у виділенні сезонної хвилі.

Таким чином, запропонований модифікований метод фільтрації компонент неоднорідної послідовності, на основі моделей запропонованих в

другому розділі дозволяє підвищити точність моделі, у випадку коротких рядів даних, за рахунок того, що не втрачаються перші та останні крайові значення у вирівняному ряді.

3.4 Висновки

Даний розділ присвячений розробці методів аналізу неоднорідних послідовностей. Для цього був запропонований підхід для узгодження експертних оцінок, розроблені метод визначення значимих чинників нечіткої регресійної моделі та метод фільтрації компонент неоднорідних послідовностей.

На основі викладеного в першому розділі огляду і аналізу існуючих підходів до моделювання даних гетерогенного походження, отриманих від людино-машинних систем і представлених у вигляді неоднорідних послідовностей, в цьому розділі були розроблені методи для побудови моделей, запропонованих в другому розділі. В результаті дослідження було встановлено, що:

1. Для розв’язку задач моделювання динамічного ряду із врахуванням експертних оцінок кількісних показників динамічного ряду пропонується використовувати підхід на основі непрямого методу побудови функцій належності.

2. Вперше запропонований метод визначення значущих чинників нечіткої регресійної моделі неоднорідних даних, який на відміну від існуючих, містить етапи підбору коефіцієнтів за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних та відбору підмножини значущих чинників з коефіцієнтами, що перевищують порогове значення, що дозволяє запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій.

3. Отримав подальшого розвитку метод фільтрації компонент неоднорідних часових послідовностей, у якому на відміну від існуючих для

виявлення тренду початкова послідовність розбивається на скінчену кількість нечітких розділів, для кожного з яких розраховується усереднене значення із врахуванням функції належності, яка асоційована із нечітким розділом, що дозволяє враховувати крайові значення ряду для виділення трендової складової.

Список використаних джерел у даному розділі наведено у повному списку використаних джерел під номерами: [6, 8, 11, 13, 14, 15, 17, 19, 78, 79, 114, 115, 116, 117, 118, 151-157].

4 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ АНАЛІЗУ НЕОДНОРІДНИХ ПОСЛІДОВНОСТЕЙ

4.1 Розробка інтелектуальної інформаційної технології аналізу неоднорідних послідовностей

Основною метою розробки інтелектуальної інформаційної технології аналізу неоднорідних послідовностей є поєднання моделей і методів, розглянутих у попередніх розділах, із технічними та програмними засобами для збору та обробки інформації, які дозволяють відобразити користувачу результат проведеного аналізу або передати дані у вигляді придатному для подальшого використання іншими програмними засобами.

Вважаємо, що дослідження неоднорідних послідовностей проводиться користувачем із застосуванням запропонованої інформаційної технології в рамках одного з поширених підходів до інтелектуального аналізу даних KDD [158], CRISP-DM [159] або SEMMA [160].

Розглянемо, яким чином застосовується інформаційна технологія аналізу неоднорідних послідовностей у випадку застосування користувачем цих підходів та місце розроблених моделей і методів для випадків застосування широко розповсюджених стандартів інтелектуальної обробки даних.

У випадку, якщо обрано KDD (Knowledge Discovery in Databases), основне призначення якого полягає в дослідженні відомостей, що зберігаються в базах даних, у такому разі, розроблена інформаційна технологія може бути використана, в тій чи іншій мірі, із врахуванням запропонованих моделей і методів, на етапах:

- попередньої обробки даних, а саме за допомогою розрахунку розглянутих у підрозділі 2.1 показників для аналізу неоднорідних послідовностей та критеріїв виявлення аномальних рівнів даних, узгодження експертних оцінок (підрозділ 3.1);

- перетворення за допомогою запропонованого методу фільтрації компонент неоднорідних послідовностей (підрозділ 3.3), моделі неоднорідних послідовностей (підрозділ 2.3);

- інтелектуального аналізу даних з використанням методу визначення значимих чинників нечіткої регресійної моделі (підрозділ 3.2) та підходів для виявлення взаємозв'язків між рівнями динамічного ряду (підрозділ 2.2).

Етап вибірки реалізується за допомогою засобів СУБД, а на етапі інтерпретації результатів користувач приймає остаточне рішення про використання запропонованих моделей.

Застосування у випадку використання KDD[158] можна схематично зобразити на рис. 4.1. Цей процес також може бути ітеративним із залученням користувача на кожному з етапів обробки.

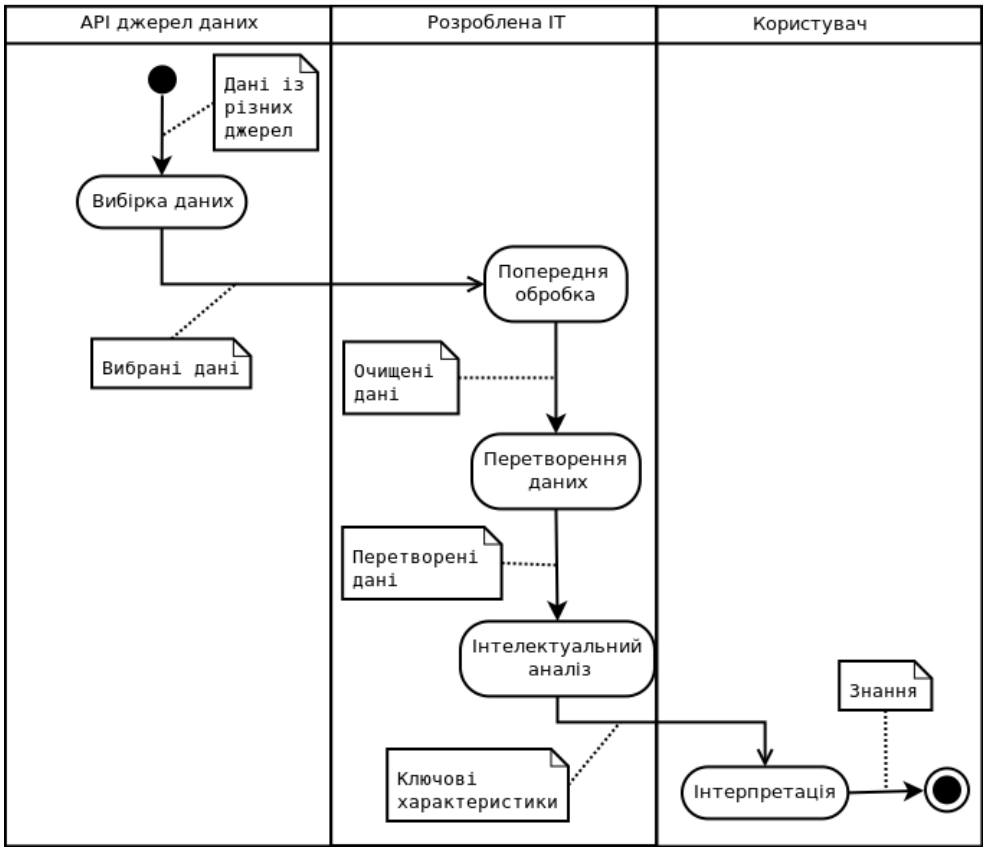

Рисунок 4.1 – Застосування розробленої інформаційної технології в KDD

У випадку, якщо користувач застосовує для аналізу даних підхід CRISP-DM (Cross-industry standard process for data mining), тобто міжіндустріальний

стандарт процесу інтелектуальної обробки даних, який полягає в ітеративному виконанні елементів циклу, який складається з шести елементів.

Складові цього стандарту [159] та можливості застосування запропонованої інформаційної технології зображені на рисунку 4.2.

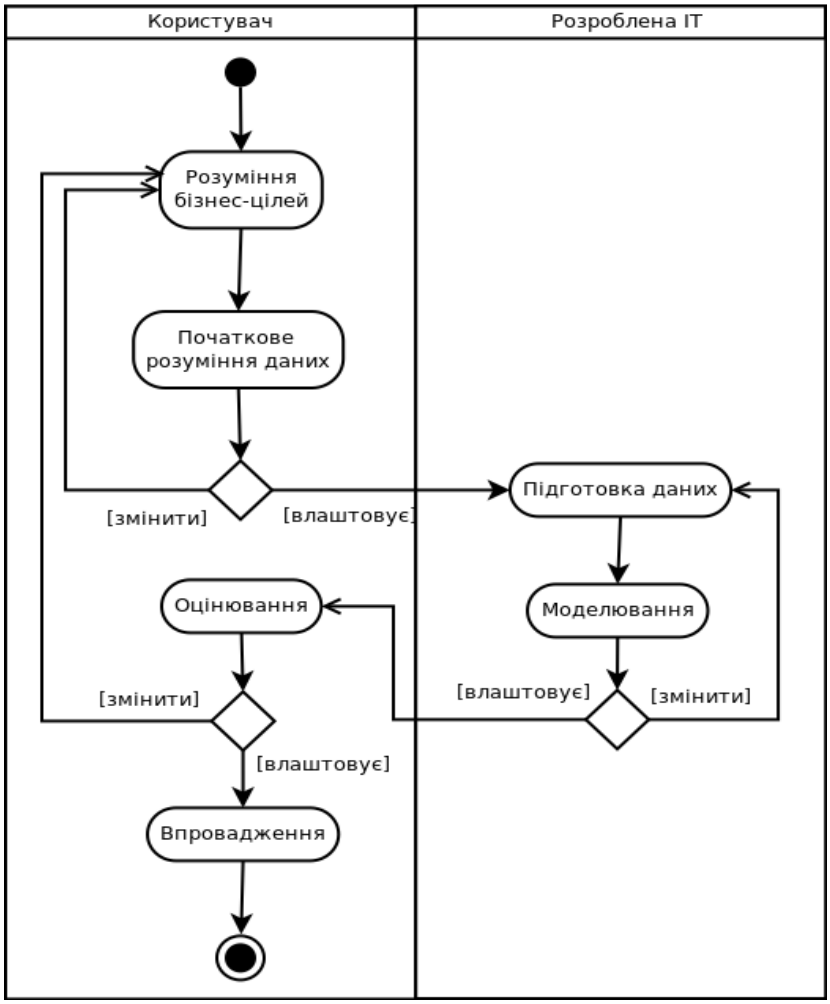

Рисунок 4.2 – Застосування розробленої інформаційної технології в CRISP-DM

Для випадку, якщо обрано CRISP-DM, розроблена інформаційна технологія може бути застосована із використанням запропонованих моделей і методів на етапах:

- початкового розуміння даних, а саме за допомогою розрахунку показників для аналізу неоднорідних послідовностей (розділ 2.1), узгодження експертних оцінок (підрозділ 3.1);

- підготовки даних, за допомогою запропонованого методу фільтрації

компонент неоднорідних послідовностей (підрозділ 3.3) та критеріїв виявлення аномальних рівнів даних(підрозділ 2.1)

- моделювання, застосовуючи моделі неоднорідних послідовностей (підрозділ 2.3), та моделювання взаємозв'язків між рівнями динамічного ряду (підрозділ 2.2);

- оцінювання, а саме використання методу визначення значимих чинників нечіткої регресійної моделі (підрозділ 3.2), оцінювання рівня сезонності із використанням методу фільтрації компонент неоднорідних послідовностей для виявлення сезонної складової (підрозділ 3.3);

У випадку застосування користувачем підходу SEMMA [160](Sample, Explore, Modify, Model, Assess) моделі і методи, які застосовуються в розробленій інформаційній технології, та описані в попередніх розділах, можуть бути використані таким чином(рис. 4.3):

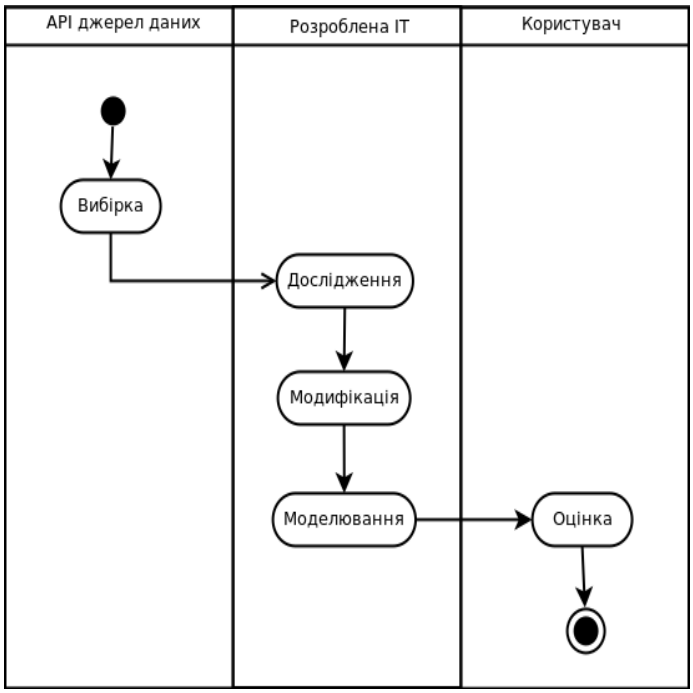

Рисунок 4.3 – Застосування розробленої інформаційної технології в SEMMA

- на етапі дослідження даних можуть застосовуватися показники для аналізу неоднорідних послідовностей (розділ 2.1) та узгодження експертних оцінок (підрозділ 3.1), критерії виявлення аномальних

рівнів даних (підрозділ 2.1);

- на етапі модифікації даних, за допомогою запропонованого методу фільтрації компонент неоднорідних послідовностей (підрозділ 3.3) та визначення значимих чинників нечіткої регресійної моделі (підрозділ 3.2)

- на етапі моделювання, застосовуючи моделі неоднорідних послідовностей (підрозділ 2.3), та моделювання взаємозв'язків між рівнями динамічного ряду (підрозділ 2.2);

Розглянемо докладно, що може виступати в якості джерела даних в рамках запропонованої інформаційної технології.

В узагальненому вигляді це відомості, які були отримані від людини, наприклад, оціночні характеристики явища, корекції помилкових рівнів вимірювання рівнів даних відповідно до розуміння людиною характеристик процесів, що відбуваються під час вимірювання або статистичної обробки та відомості, які є результатом технічних вимірювань або статистичних обліків.

Також у випадку, якщо джерела походження однорідні, особливості явища можуть біти причиною нестаціонарності отриманих даних.

Беручи до уваги підходи, що склалися в напрямку зберігання інформації, дані неоднорідних послідовностей для задач, які вирішуються запропонованою інформаційною технологією, можуть бути отримані з таких джерел:

1) Сховища даних, що зберігають інформацію із використанням реляційного підходу до управління базами даних. Такі системи управління базами даних досить поширені і широко використовуються в діяльності різного роду організацій.

2) Сховища, які зберігають інформацію за допомогою нереляційного підходу до управління даними. Використання такого способу керування може бути викликане різними причинами.

3) Особливо поширене застосування такого роду підходів у випадку отримання великих об'ємів потокових даних або зберігання значної кількості неструктурованих даних в умовах жорстких обмежень на швидкість виконання операцій запису.

4) Структуровані відомості, які зберігаються у вигляді файлів. Як правило, об’єм таких файлів порівняно невеликий, але досить розповсюджений.

Розглянемо етапи інформаційної технології аналізу неоднорідних послідовностей.

На першому етапі функціонування інформаційної технології отримуються початкові відомості, які в подальшому використовуються для формування рядів даних.

Дані можуть бути отримані або шляхом запиту до одного з сховищ даних (на основі реляційної або не реляційної СУБД) від клієнта у вигляді SQL або із застосуванням REST запита.

У випадку REST запиту, він перенаправляється до бази даних чи сервіса, який підтримує WEB API для отримання даних.

Отримані на цьому етапі дані зберігаються. Також інформаційна технологія підтримує можливість самостійного внесення даних користувачем. Як правило, це відбувається за допомогою завантаження даних із файлового ресурсу.

Після отримання відомостей з віддалених джерел та самостійного внесення інформації (за потребою) користувач переходить до наступного етапу.

Другий етап інформаційної технології полягає в формуванні динамічних рядів даних. В залежності від того, яка інформація є в наявності, користувач може вводити як кількісні значення показників та факторів, що зумовлюють значення величини, так і дані подані у вигляді нечітких значень.

На цьому етапі дані про нечіткі значення рівнів чинників та/або відгуку можуть задаватися за допомогою внесення самих значень та функції належності до нечіткої множини, так і за допомогою внесення результатів експертних опитувань стосовно значення певної лінгвістичної змінної.

В подальшому з цих даних, за допомогою методики описаної в підрозділі 3.1, формуються функції належності нечіткої величини.

Така технологія схематично відображена на діаграмі активності (рисунок 4.4).

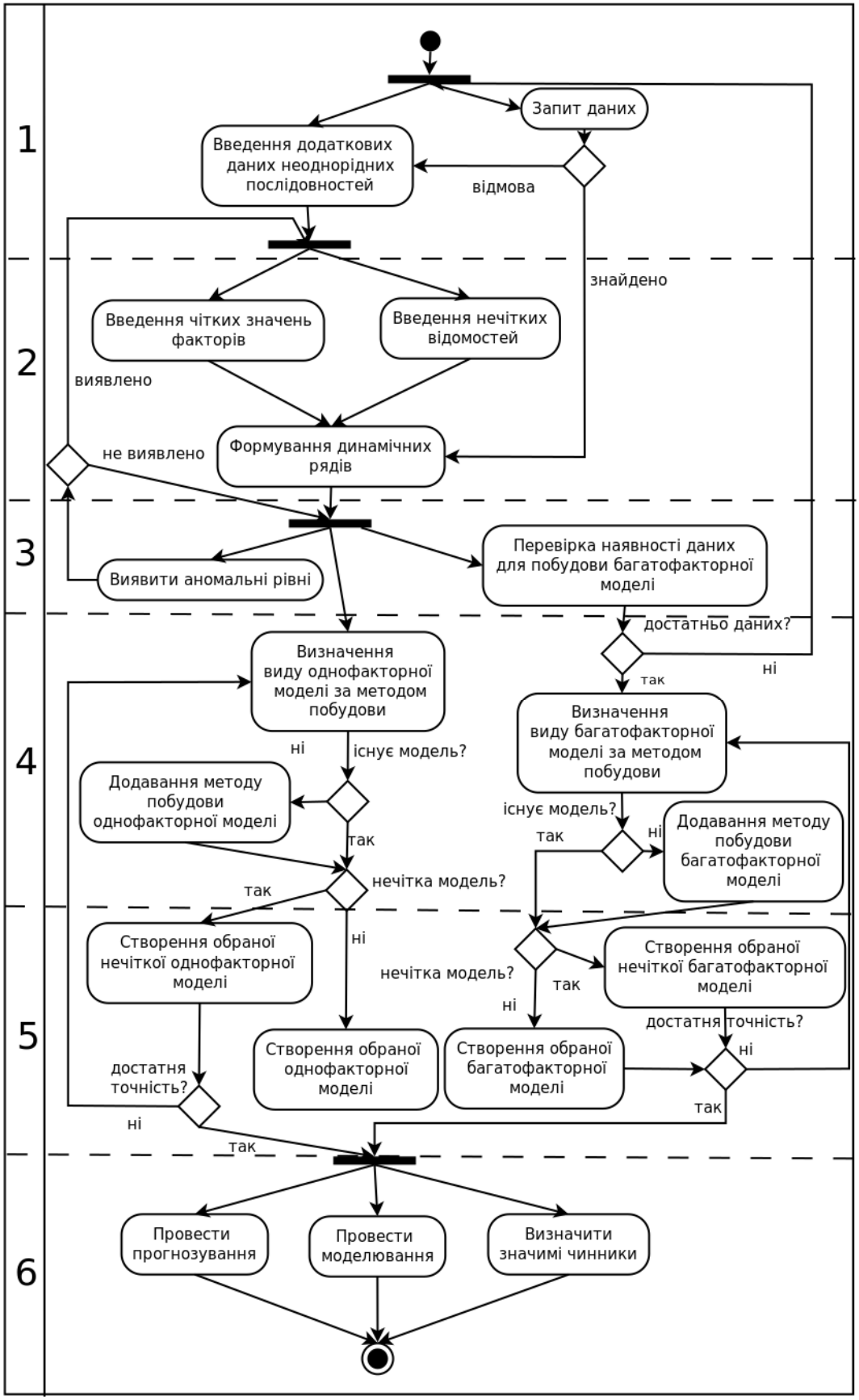

Рисунок 4.4 – Схема етапів інформаційної технології

На третьому етапі, таким чином сформовані дані перевіряються на

наявність аномальних рівнів та на достатню кількість значень для багатофакторного аналізу за допомогою підходів, викладених в підрозділі 2.1.

У випадку, якщо кількість даних не достатня користувач повертається до етапу запиту даних з додаткових джерел даних, у зв'язку з неможливістю побудови багатофакторної моделі на основі неоднорідних послідовностей за недостатньою кількістю значень рядів даних.

У випадку, якщо знайдені рівні даних з підозрою на аномальність, користувач обирає один із запропонованих системою варіантів подальшої обробки таких даних, які ґрунтуються на відкиданні цього рівня, або заміні його на інше значення. Після цього відбувається перехід до наступного етапу роботи інформаційної технології.

Четвертий етап роботи полягає у виборі способу побудови моделі. Користувачу пропонується здійснити вибір способу побудови моделі та можуть бути підключені додаткові модулі, що реалізують моделі. Дані модулі можуть створюватися окремо та розширяти можливості системи.

При роботі із додатковим модулем, користувач повинен визначити тип задачі з переліку типових задач, які виникають під час аналізу даних. В процесі роботи вибір користувача зберігається і задачу можна вибрати із переліку тих проблем, які уже розв'язувалися.

Після формулювання задачі на екрані монітора з'являється схема алгоритму розв'язування поставленої задачі. Тут зображується стандартний алгоритм, тобто алгоритм, який уже використовувався при розгляді подібної задачі. Алгоритм будується за схемою "І-АБО", надаючи користувачу можливість вибору однієї з декількох альтернатив.

Також користувачу надається можливість внести певні зміни в даний алгоритм (інший метод розв'язування певної підзадачі чи інша модель конкретного процесу).

Для цього потрібно на опції, яка визначає стандартну операцію, натиснути праву кнопку мишки, задати назву потрібного методу чи моделі і вказати місце, де знаходиться відповідна програмна одиниця (база знань чи

файл). Такі методи та моделі додаються в базу знань і в подальшому фігурують в альтернативах.

Аналогічно формується алгоритм розв'язування задачі в тому випадку, коли за результатами моделювання виникає потреба у зміні моделі досліджуваного процесу. Після цього автоматично підбираються потрібні модулі і встановлюються між ними потрібні інформаційні зв'язки з використанням таблиць.

При формулюванні невідомої задачі програмна система разом із користувачем створює алгоритм її розв'язування у режимі "запитання-відповідь". Такий алгоритм можна подати у вигляді графа.

В даній роботі для реалізації такої системи пропонується застосувати підхід, який забезпечує можливість моделей, що застосовуються в системах, адаптуватися до конкретної, специфічної реальності в результаті діалогу з користувачем, і можливість системи інтерактивного генерування моделей. Під час роботи четвертого етапу можуть застосовуватися як моделі, розглянуті в розділі 2.3, так і підходи для виявлення взаємозв'язків викладені в розділі 2.2.

П'ятий етап роботи інформаційної технології полягає в застосуванні моделей, розглянутих у другому розділі за допомогою методів, викладених в розділі 3.

За допомогою розроблених рішень на даному етапі можуть створюватися як одно факторні, так і багатофакторні моделі, чіткі або нечіткі, у випадку, якщо наявні дані, які дозволяють це зробити. Також для випадку часових послідовностей можна перевіряти динамічний ряд на сезонність і створювати моделі на основі сезонності із застосуванням методу, описаного в 3.3 навіть для коротких рядів неоднорідних даних.

Шостий етап інформаційної технології дозволяє провести моделювання із використанням даних, підготованих на попередніх етапах, з використанням параметрів, розрахованих для моделей визначених у підрозділі 2.3 за допомогою методів, описаних у підрозділах 3.2 і 3.3.

Відповідно до побудованих моделей можуть бути визначені значимі чинники та здійснене прогнозування з використанням як багатофакторних, так і однофакторних моделей.

Подана таким чином інформаційна технологія може бути використана в рамках одного із загальновизнаних підходів до аналізу даних, розглянутих вище. Реалізація інтелектуальної інформаційної технології здійснюється у вигляді модуля, написаного мовою програмування Python.

4.2 Розробка модуля обробки неоднорідних послідовностей

Моделі і методи аналізу неоднорідних послідовностей даних, накопичених користувачами, пропонується реалізувати у вигляді програмного модуля, який би втілював викладені вище етапи інформаційної технології.

Варто зазначити, що необхідність у реалізації полягає в тому, що велика кількість джерел різнорідної інформації надає відомості, які потребують швидкої обробки та прийняття рішень стосовно набору методів, що будуть застосовуватися для інтелектуального аналізу даних [17].

Запропонована інформаційна технологія реалізується програмною системою, створеною у вигляді модуля аналізу неоднорідних даних. Ключові функції модуля реалізовані таким чином, що вони є сумісними з програмним забезпеченням, реалізованим за допомогою мови програмування Python.

У даній дисертаційній роботі для опису інформаційної технології використовуються діаграми UML [161].

Перш за все, необхідно визначити «сутності» (актори), які взаємодіють в процесі функціонування. Такі «сутності» та пов’язані з ними «сценарії використання» формують діаграму прецедентів. Пропонується їх подати таким чином:

– «Користувач» - особа, яка здійснює взаємодію з програмною системою та використовує її для аналізу неоднорідних послідовностей, приймає остаточне рішення про доцільність використання створеної моделі;

– «Модуль» - програмна система, яка реалізує обробку неоднорідних послідовностей за допомогою моделей і методів, викладених у попередніх розділах;

– «Внутрішня БД» - джерело структурованої інформації у вигляді бази даних, розташованої локально на файловій або в локальній мережі користувача

– «Зовнішня БД» - джерело структурованої інформації у вигляді бази даних, розташованої поза межами локальної мережі користувача. Така структурована інформація може зберігатися у сховищі, як із використанням реляційного так і не реляційного підходу до збереження та управління даними;

– «Мережеве з API» - джерело даних, яке має стандартний протокол запитів та отримання даних за допомогою прикладного програмного інтерфейсу, з визначеним протоколом доступу до даних та набором функцій для автоматизації їх отримання. Такого роду джерела даних досить часто дозволяють здійснювати обмін даними на основі підходу REST, який дозволяє ідентифікувати ресурси, отримувати відомості, визначені розробником, та робити більш складні запити в межах інтерфейсу, що підтримується сервером;

– «Мережеве без API» - джерело даних, яке не має стандартного протоколу запитів та отримання даних за допомогою прикладного програмного інтерфейсу та потребує використання додаткових засобів автоматизації для отримання даних. Як правило, такого роду дані зберігаються або у вигляді файлових ресурсів, доступних для передачі мережею за допомогою протоколів передачі даних HTTP та FTP. У деяких випадках такі дані розташовані у вигляді таблиць, які є частиною веб-сторінки. В такому разі автоматизація отримання даних може бути застосована лише із використанням додаткових модулів, які можуть бути під’єднані до розробленого безпосередньо для цього джерела.

Розглянуті актори здійснюють взаємодію в рамках функціонування інформаційної технології. За допомогою актора «Модуль» втілюються етапи інформаційної технології, запропонованої в підрозділі 4.1, включаючи і ті з них,

які потребують зовнішніх з’єднань з використанням мережі.

За допомогою діаграми прецедентів відобразимо типову взаємодію користувача та середовища програмного забезпечення (рис. 4.5).

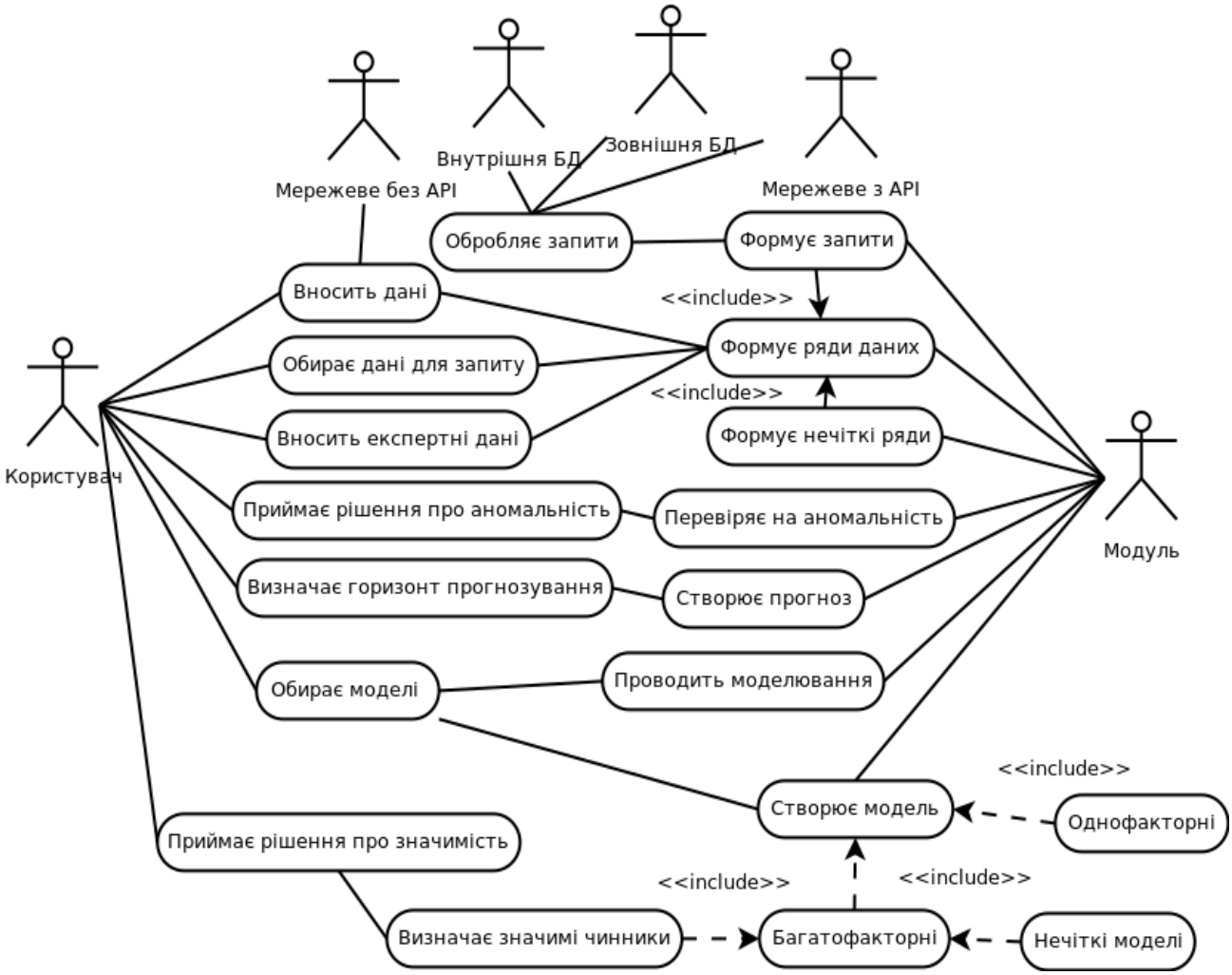

Рисунок 4.5 – Діаграма прецедентів

На діаграмі зображені дії, що виконуються акторами: користувачем описаної в попередньому підрозділі інформаційної технології, джерелами даних, які можуть бути реалізованими сторонніми розробниками у вигляді систем управління базами даних або веб-служб і модулем обробки неоднорідних послідовностей. «Користувач» обирає відомості, які потрібно отримати з різних джерел. Наприклад, дані можуть розташовуватися локально на файловій системі користувача або в локальній мережі організації.

В іншому випадку дані можуть розташовуватися у мережі в

структурованому або неструктурованому вигляді. Існують джерела даних що надають API, або доступ до мережевої СУБД.

У такому разі дані будуть представлені у структурованому вигляді, хоча і потребуватимуть певної попередньої обробки даних. У випадку коли мережевих джерело не надає API для отримання даних, користувач самостійно вносить ці дані або використовує автоматизовані системи отримання даних з неструктурованих джерел[26,27].

Як правило, такі дані потребують додаткової обробки. Досить часто вони є неузгодженими і потребуватимуть окремих процедур для створення структури збереження інформації. Також користувач може самостійно вносити дані та відомості, отримані від експертів, наприклад, дані про коригування значень рівнів динамічних рядів, або факторів, які визначають поведінку того чи іншого явища.

Користувач також приймає остаточне рішення про належність рівня до категорії «аномальні» під час перевірки даних на аномальність.

На етапі моделювання має можливість обирати вид моделі серед запропонованих та приймати рішення щодо значущості чинника. Обирає дані для створення прогнозу на визначений період часу.

Актор «Мережеве без API» відображає взаємодію з мережевими джерелами неструктурованої інформації, які не надають API для доступу до даних.

Актори «Внутрішня БД» , «Зовнішня БД» , «Мережеве з API» - джерела даних, які мають стандартний протокол запитів та отримання даних. Відповідно вони можуть обробляти запити сформовані до локальних або віддалених джерел даних.

Модуль дозволяє сформувати ряди даних, які включають в себе як чіткі, так і нечіткі відомості, провести попередню обробку даних, перевірити дані на аномальність, побудувати модель, провести моделювання поведінки системи за різних умов і створити прогноз.

Моделі можуть бути, як однофакторні так і багатофакторні, чіткі, нечіткі.

Для багатофакторних моделей проводиться процедура визначення значущих чинників за допомогою методів викладених вище.

На рисунку 4.6 показані основні стани, в яких може знаходитись програмна система(модуль) під час роботи.

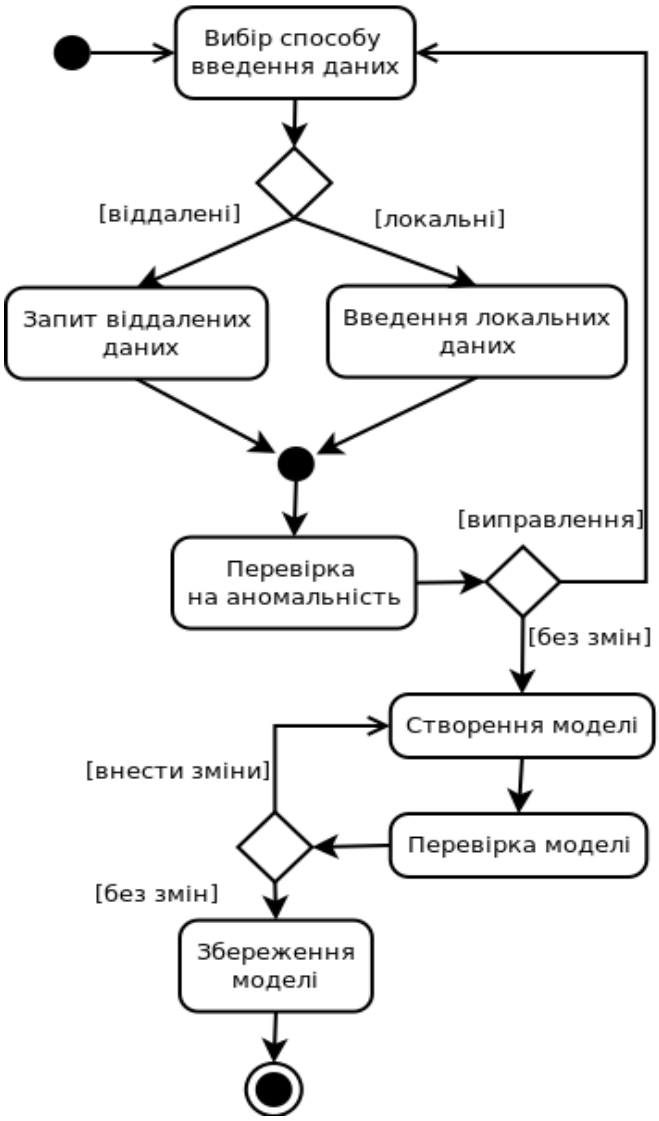

Рисунок 4.6 – Діаграма станів модуля аналізу неоднорідних послідовностей

На початку роботи програмна система(модуль) знаходиться в стані «Вибір способу введення даних», тобто в стані очікування вибору користувачем виду даних, які будуть отримуватися. У випадку, якщо користувач обирає введення даних з локальних джерел, модуль переходить в стан «Введення локальних даних». Дані можуть вводитися за допомогою

запиту, тоді модуль переходить в стан проведення запиту до віддалених даних.

Якщо запит не можливо виконати (з причин недоступності серверу, обмеження прав доступу та інших причин, що не дозволяють джерелу обробити запит), тоді модуль переходить в стан «Вибір способу введення даних».

Після того як всі дані зібрано, модуль переходить в стан «Перевірка на аномальність», звідки можуть відбутися переходи до «Вибір способу введення даних», якщо дані не введені, або користувач хоче змінити окремі значення масивів даних.

Якщо аномальних рівнів не виявлено або якщо аномальні рівні виявлені, користувачу видається підсумкове повідомлення про варіанти обробки цих рівнів, після чого робота з цим блоком завершується.

За вибором користувача модуль може перейти в стан «Створення моделі». В процесі моделювання програмна система(модуль) пропонує користувачу варіанти моделей та їх параметри.

Якщо це задовольняє користувача, робота з підсистемою моделювання завершується, якщо ні – модуль переходить в режим «Перевірка моделі». Виконується моделювання, і користувач може переглянути результати і зберегти модель, або змінити модель, в такому випадку модуль переходить в стан «Створення моделі».

Для розробки модуля пропонується застосувати таку властивість об’єктно-орієнтованого програмування, як поліморфізм і подати програмні компоненти, що реалізують виконання окремих підзадач програмної системи, у вигляді об’єктів. З метою забезпечення можливості використання інших мов програмування, наприклад C/C++, розширення функціоналу програмного засобу також може здійснюватися за допомогою бібліотечних файлів (*.dll або *.so - в залежності від платформи).

Розглянемо підхід, який забезпечує можливість розширення і поновлення програмної системи(модуля) у випадку потреби внести зміни в реалізацію або додати нові методи.

Для опису та відображення цього підходу скористаємось діаграмою класів UML (рис. 4.7).

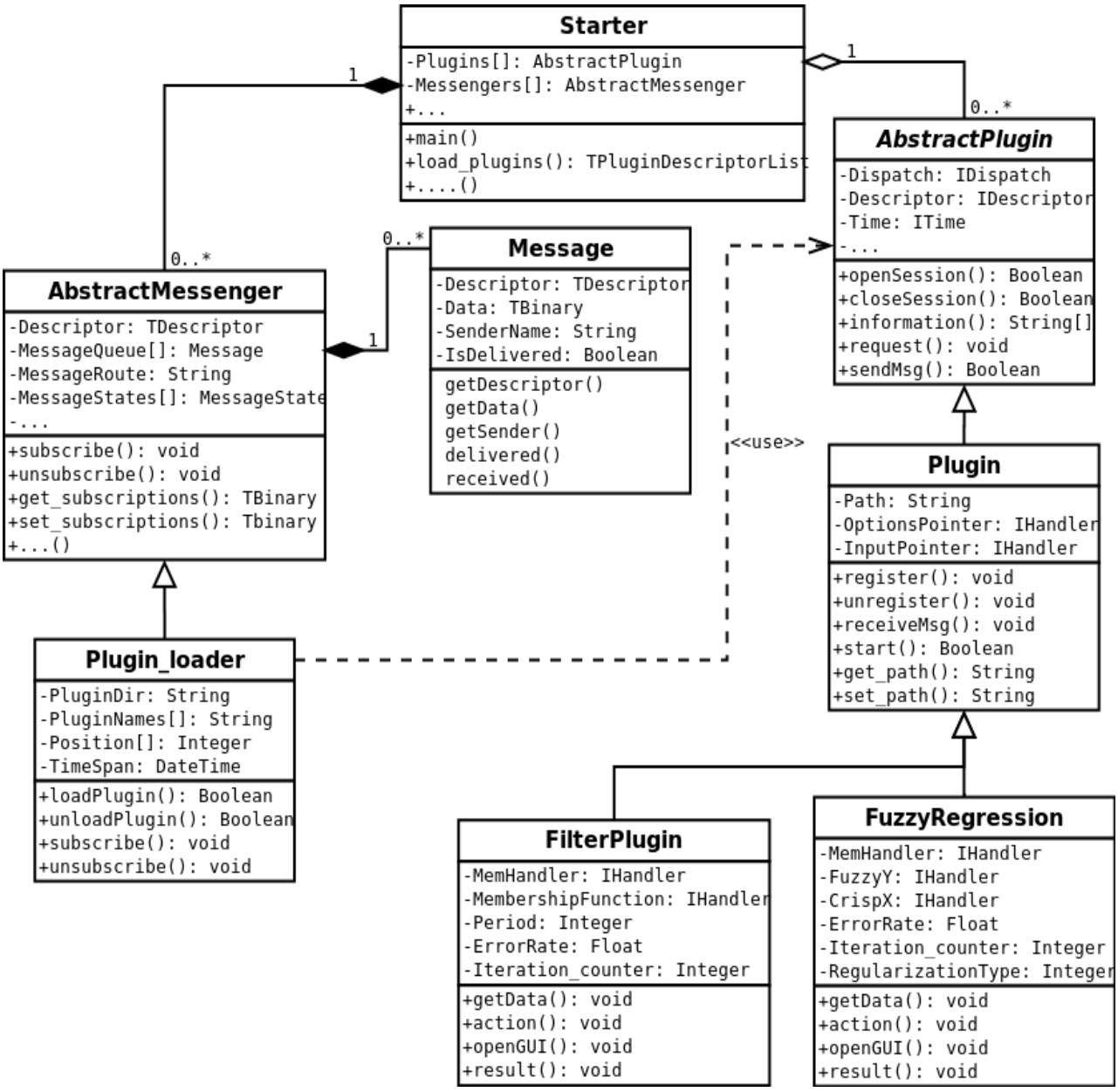

Рисунок 4.7 - Діаграма структури класів компонентної моделі.

Застосування такого підходу дозволяє суттєво розширювати функціональність програмного забезпечення за допомогою під'єднання нових компонент. Зазначений підхід застосовується, наприклад, у випадках, коли необхідно додати можливість отримання даних з мережевих джерел, які підтримують власний прикладний програмний інтерфейс для отримання даних.

Класс Plugin_loader є класом, який контролює та використовує в роботі абстрактний клас AbstractPlugin, який реалізовується класом Plugin. Функції цього класу реалізовані таким чином, щоб вони відповідали вимогам класу Plugin_loader і реалізовували абстрактні методи класу AbstractPlugin.

За допомогою повідомлень, що реалізовані у вигляді класу Message, завантажувач компонент Plugin_loader і класи похідні від AbstractPlugin обмінюються інформацією та виконують команди.

Похідними від класу Plugin є всі інші класи, що реалізують конкретні методи обробки та аналізу даних. Наприклад, клас FilterPlugin реалізує функції, які потім можуть використовуватися похідними від нього класами, які забезпечують різні методи фільтрації даних, представлених у вигляді часових рядів. Функція getData() дозволяє визначити дані, що будуть оброблятися, функція action() проводить операцію фільтрації, функція result() дозволяє отримати результати фільтрації.

Реалізовані за допомогою такого підходу компоненти можуть мати власний графічний інтерфейс, реалізацію методу, що отримує дані з програмної системи(модуля) та повертає її у вигляді, який очікується іншими об’єктами.

Дані компоненти можуть також розширювати способи введення даних, наприклад використання додаткових драйверів завантаження даних з СУБД, які можуть використовуватися користувачами в своїй діяльності.

Розглядаючи логічну послідовність діяльності системи, можна відзначити, що підсистема моделювання формує альтернативи розв’язку задач з попередньо визначеного кола моделей.

Після цього кожна з них завантажується з бази моделей і транслюється.

Такий підхід до створення альтернатив під час моделювання дозволяє розглянути різні моделі і обрати таку, яка найбільше задовольняє дослідника

Кожний параметр моделі характеризується певним набором даних, що описує тип значення, звідки значення дістається, назву для користувача, назву для використання параметру під час виклику із системи тощо.

Трансляція внутрішнього подання моделі у виконуване, виконується

транслятором моделі. Транслятор одержує і використовує інформацію про те, звідки дістаються дані і як вони опрацьовуються. Діаграма дій підсистеми моделювання зображена на рисунку 4.8.

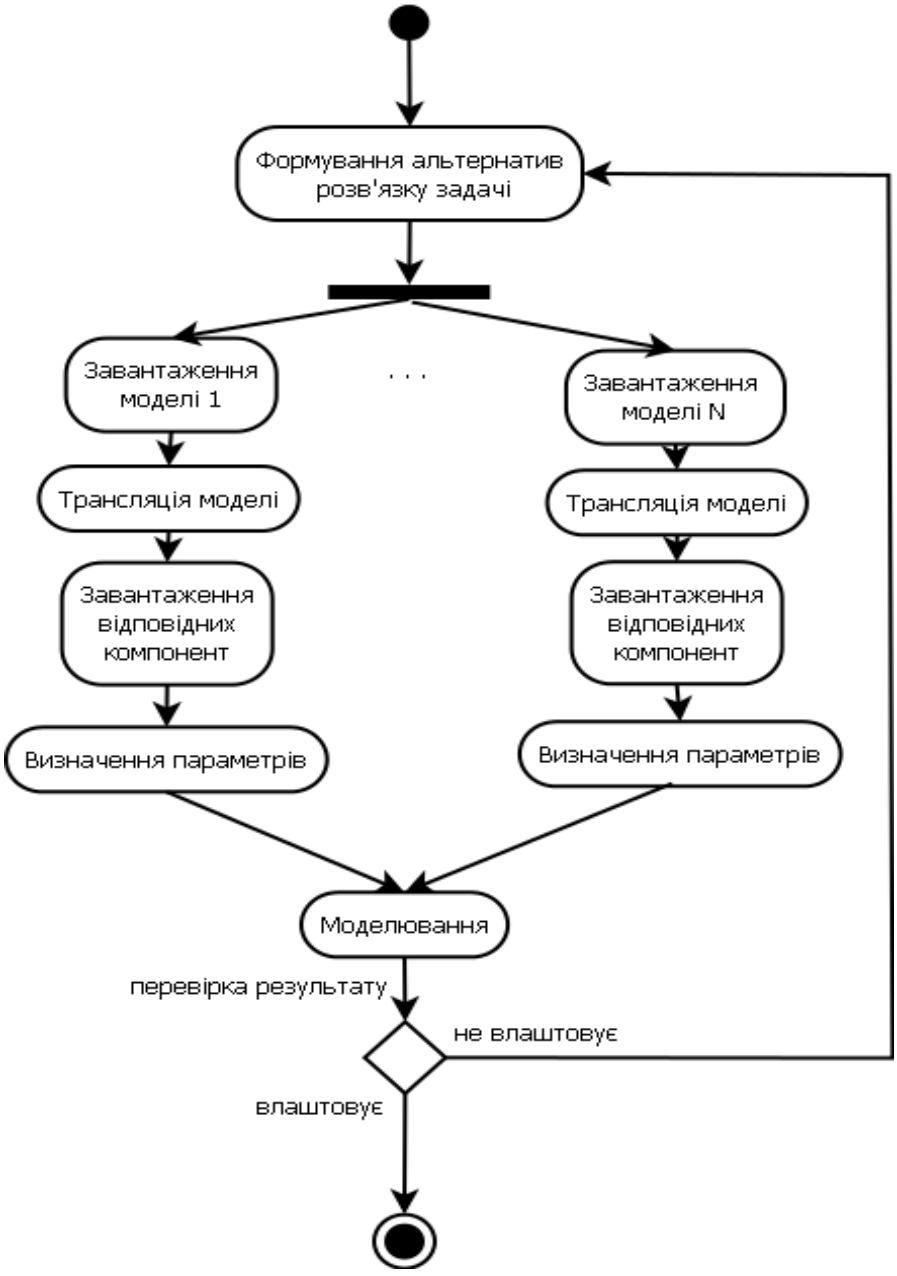

Рисунок 4.8 - Діаграма дій підсистеми моделювання.

Для створення опису елементів бази моделей у роботі використовується розширювана мова розмітки (XML). Кожна з моделей подається у вигляді певної ієрархічної структури, що дозволяє визначити, які компоненти можуть бути включені до неї [162]. Ці відомості зберігаються в реляційній базі даних.

Таким чином, пропонується підхід який дозволяє здійснювати вибір методу обробки даних в залежності від того, які дані подані до входу інформаційної системи.

Для цього завантажуються відповідні компоненти, що є програмною реалізацією цих моделей і визначаються параметри моделей.

Потім проводиться моделювання, і користувачу пропонуються найбільш точні моделі. Користувач обирає, чи влаштовує його отриманий результат.

Процес завантаження моделі з бази моделей можна подати у вигляді діаграми послідовності опрацювання моделі, що зображена на рис. 4.9.

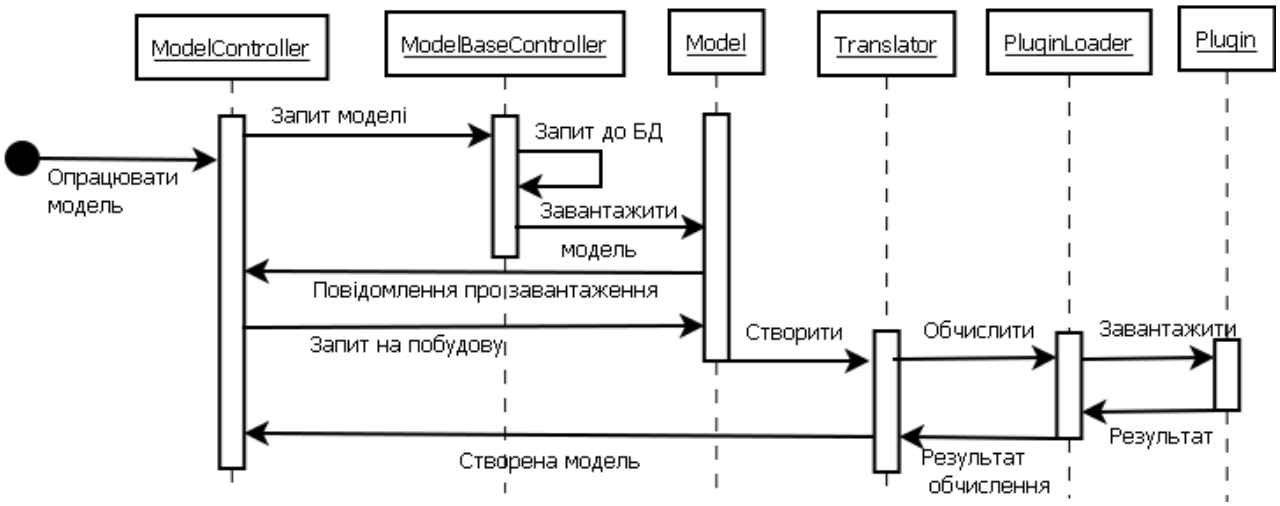

Рисунок 4.9 - Діаграма послідовності опрацювання моделі.

Розглянемо послідовність завантаження обраної моделі із бази. Контролер ModelController отримує запит на опрацювання моделі. Після цього завантажується контролер бази моделей ModelBaseController, який здійснює запит до бази даних і повертає результат.

З отриманого результату завантажується модель Model. Отримана модель надсилає повідомлення до контролера, що відповідає за завантаження моделі. Моделі обираються зі сховища збережених моделей.

Дана модель зберігається у вигляді, що дозволяє виділити які компоненти потрібні для виконання моделі, і послідовність та набір параметрів для побудови моделі.

Ці дані передаються до транслятора, що здійснює завантаження за допомогою PluginLoader. Цей об’єкт завантажує компонент Plugin з числа доступних відповідно до рисунку 4.7.

Вона повертає транслятору посилання на функцію що запускає компонент. Транслятор використовує ці посилання для створення та обчислення параметрів моделі. Після цього результат моделювання(Створена модель) повертається до ModelController.

4.3 Експериментальна перевірка інтелектуальної інформаційної технології аналізу неоднорідних послідовностей

Розглянемо використання запропонованої інформаційної технології аналізу неоднорідних послідовностей для розв’язування конкретних задач, що виникають в різних сферах суспільного життя.

Розглянемо приклад з сфери досліджень медичних даних. Розробці системного підходу до визначення коефіцієнту гідродинамічного опору носової порожнини присвячена робота [163].

Для оцінювання результатів риноманометричних досліджень прийнято розраховувати такі параметри:

- коефіцієнт R100;
- коефіцієнт R75;
- коефіцієнти k1 и k2 згідно формули Рехрера [164];
- коефіцієнти R2(V2) по модели Бромса [165];
- коефіцієнт λ [166].

В ринологічній практиці превалює концепція оцінювання обструкції носового дихання, що ґрунтується на розрахунку носового спротиву R150.

Але необхідно відмітити, що для окремих пацієнтів можна спостерігати зсув меж ступенів обструкції.

Як правило, при прийнятті рішення оперують не точеними значеннями а певними інтервалами, які характеризуються деякими функціями належності.

Вихідні дані подаються у вигляді нечітких значень. Масив вихідних значень спротиву R_{150}, позначимо як $\tilde{Y}_{i,} i=\overline{1,n}$, тоді функція належності і-го коэффициенту буде мати вигляд описаний формулою (3.26). Таким чином, ми можемо представити вхідні дані відповідно в таблиці 1.

Таблиця 4.1 – Вхідні дані моделі носового дихання.

Ім'я змінної	Вид вимірювання
X_1	Значення $k_{1,}$ Па с/м^3
X_2	Значення $k_{2,}$ Па с^2/м^6
X_3	Значення R_{100}, Па с/м^3
X_4	Значення R_{75}, Па с/м^3
X_5	Значення R_2 (V_2), градус
X_6	Діаметр ноздрі, мм

Для даної задачі пропонується скористатися поданням значень коефіцієнта R_{150} у вигляді нечітких даних. В роботі [13] Нечипоренко А.С. пропонує скористатися нечіткою регресійною моделлю для розрахунку коефіцієнтів R_{150}.

Для розрахунку параметрів цієї моделі скористаємся підходами описаними в попередніх розділах. На першому кроці необхідно виділити значущі фактори нечіткої регресійної моделі.

Траєкторії зміни значень коефіцієнтів при факторах при відборі значущих за допомогою запропонованого методу можна зобразити на рисунку 4.10.

Можна виділити два значимих фактора, а саме x_3 и x_5, які в подальшому можуть бути використані для побудови нечіткої регресійної моделі, виходячи з того, які з коефіцієнтів перевищують значення коефіцієнта Маллоуза Ср.

В підрозділі 2.3 запропонована нечітка регресійна модель з чіткими вхідними даними та нечітким виходом. Таким чином, необхідно розрахувати нечіткі коефіцієнти нечіткої лінійної регресійної моделі за допомогою методу описаного в підрозділі 3.2.

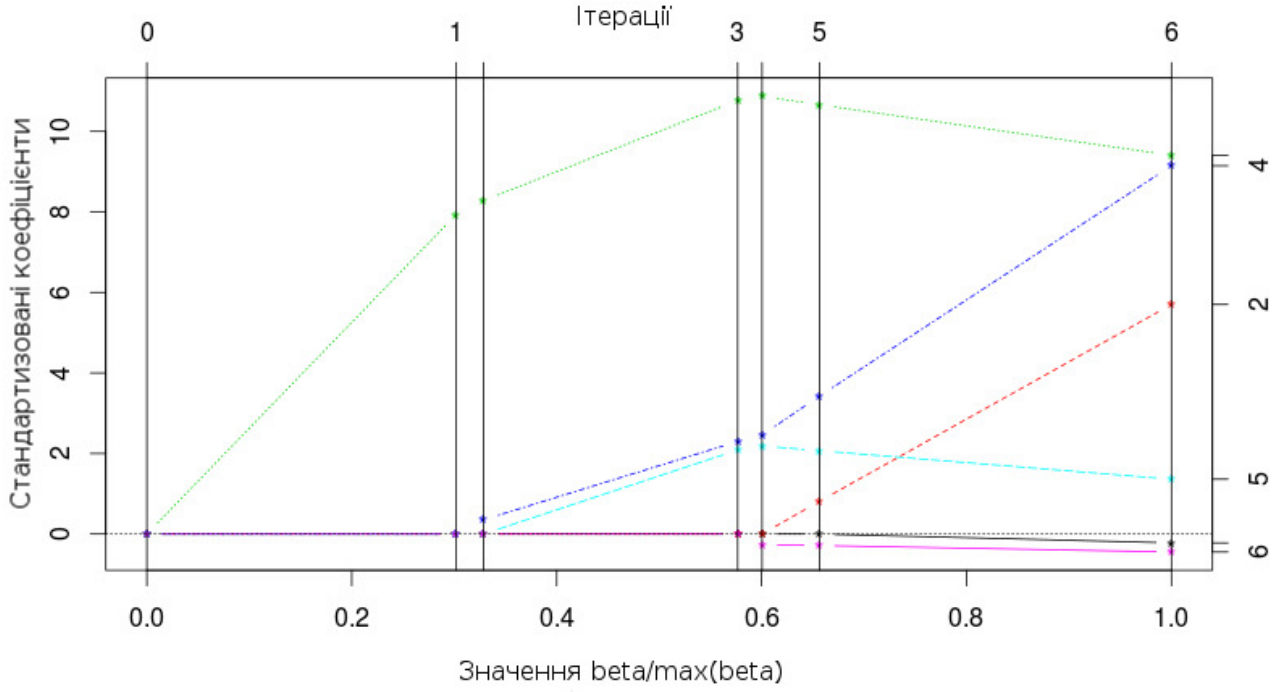

Рисунок 4.10 – Траєкторії зміни коефіцієнтів при відборі значущих ознак.

Для цього необхідно знайти значення нечітких коефіцієнтів при чітких значеннях рівнів фактора.

В загальному вигляді їх можна подати таким чином:

$$\tilde{Y} = \tilde{A}_0 + \tilde{A}_1 X_1 + \tilde{A}_2 X_2 + \tilde{A}_3 X_3 + \tilde{A}_4 X_4 + \tilde{A}_5 X_5 + \tilde{A}_6 X_6, \qquad (4.2)$$

де $\tilde{A}_j = (a_j, c_j)$, $j = 0..n$ - нечітка величина з центром a_i і шириною c_i.

Таким чином, скориставшись запропонованим методом побудови нечіткої регресійної моделі отримуємо таке рівняння, в якому вказані параметри нечітких коефіцієнтів для чітких значень факторів (4.3):

$$Y = (0.14, 0.001) X_3 + (0.001, 0.01) X_5 + (0.48, 0.2). \qquad (4.3)$$

Кількість чинників, які ввійшли до рівняння менша ніж початкова, за рахунок того, що під час застосування методу описаного в підрозділі 3.2 були

обрані значущі фактори.

Для тестового набору даних можна розрахувати похибку. Значення відхилення розрахункових значень від тестових відображенні на рис. 4.11.

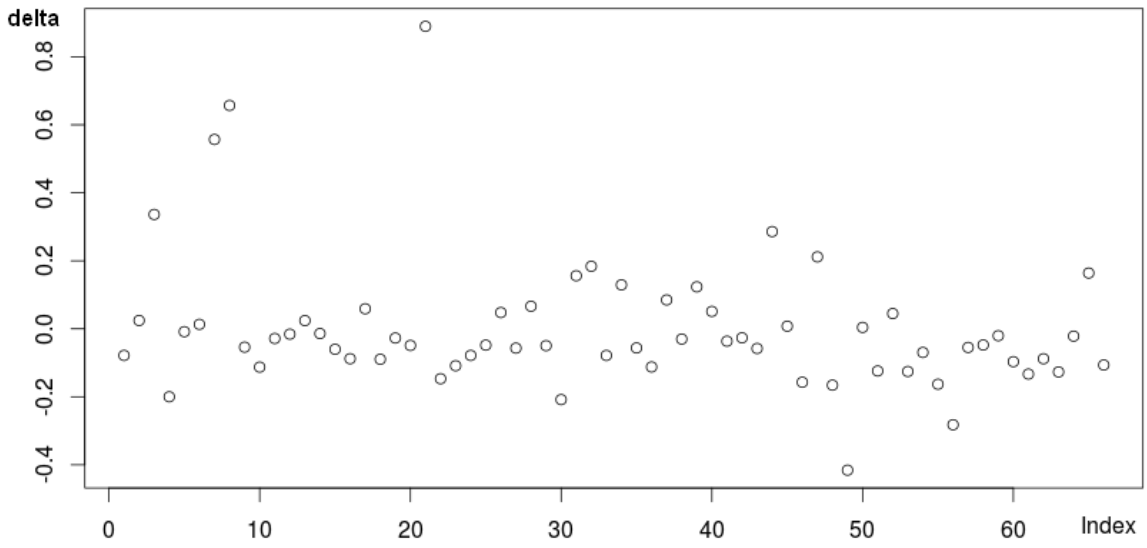

Рисунок 4.11 – Відхилення розрахункових значень від тестових

Кількість помилок для запропонованої моделі склала для навчальної вибірки 1%, кількість помилкових значень по тестовій вибірці склала 3,4%.

Для моделі, яка включає всі фактори, кількість посилок склала відповідно 1% і 60% для навчальної та тестової вибірки, що не дозволяє використовувати таку модель на практиці.

Для нечіткої регресійної моделі, яка побудована із використанням крокового регресійного аналізу із використанням критерію Фішера, значення кількості помилок для навчальної вибірки склало 1%, а по тестовій вибірці 4,1%.

Таким чином запропонований метод дозволяє зменшити кількість вхідних параметрів моделі, що дозволяє запобігти перетренованості моделі.

Розглянемо моделі, запропоновані в підрозділі 2.3, та метод, запропонований в підрозділі 3.3, на прикладі дослідження даних про злочинність.

На відміну від методів відбору ознак за допомогою покрокової регресії, з

використанням F-критерію не потрібно задавати рівень значимості. Отримані результати дозволяють виділити два значимих коефіцієнта, що впливають на рівень обструкції.

Розглянемо іншу задачу, пов'язану з моделюванням сезонної хвилі. Значна частина людської діяльності так чи інакше пов'язана з сезонністю. Не винятком є і певні види злочинів. Розглянемо такий тип злочинів як пограбування.

Наявні дані про кількість скоєних злочинів (пограбувань) представлено в таблиці 4.2.

Таблиця 4.2 – Дані про скоєні пограбування, наростаючим підсумком (тис. злочинів)

	2015	**2016**	**2017**
Січень	X	2	1,5
Лютий	2,9	4,2	2,9
Березень	4,3	6,4	X
Квітень	5,8	8,5	X
Травень	7,6	10,7	X
Червень	9,4	13,4	X
Липень	11,9	16,4	X
Серпень	14,2	19,4	X
Вересень	16,1	21,4	X
Жовтень	17,5	23,9	X
Листопад	19,4	25,6	X
Грудень	22,1	27,2	X

Досить часто з об'єктивних причин відсутня можливість порівнювати відомості, що сформовані в періоди, що досить суттєво відрізняються за базою статистичного вимірювання, методикою збору або структурою об'єкту дослідження в рамках однієї моделі без додаткового залучення експертів для коригування отриманих відомостей.

Дані наростаючим підсумком доступні, починаючи з сумарних даних для січня-лютого 2015, відповідно довжина сформованого динамічного ряду місячних даних лише 24 значення.

Статистична інформація по кількості скоєних пограбувань, сформована станом на березень 2017 року починаючи з березня 2015 року [167]. Такий ряд є досить коротким для побудови тренд-сезонної моделі за допомогою стандартного підходу.

У випадку, якщо тренд виділяється за допомогою ковзного середнього, відомості представлені крайовими значеннями не можуть враховуватися при розрахунку сезонної хвилі, що знижує точність моделі. Приклад виділення тренду за допомогою ковзного середнього з періодом згладжування 12 зображено на рисунку 4.12.

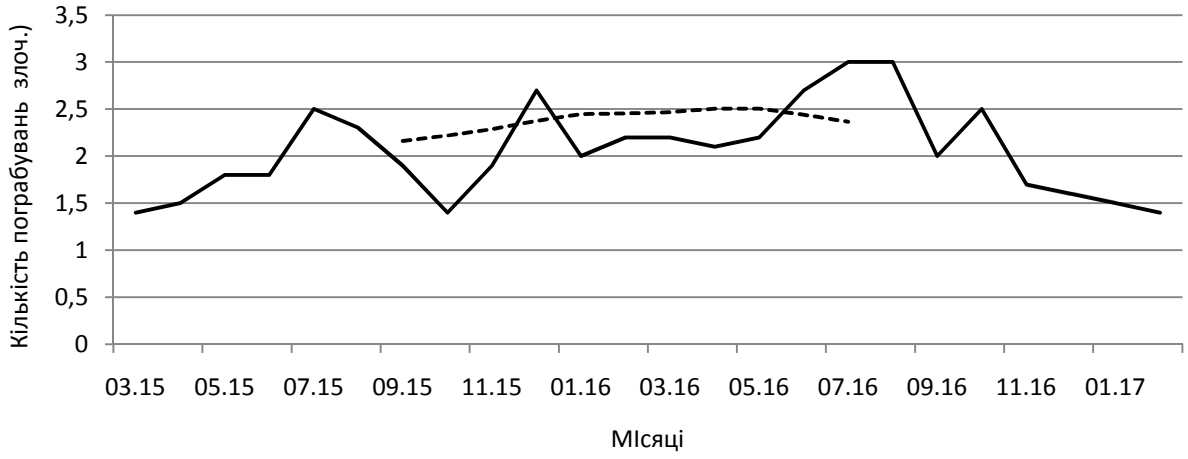

Рисунок 4.12 – Виділення тренду за допомогою ковзного середнього

Суцільною лінію на малюнку позначені існуючі рівні дані, переривчастою – виділена лінія тренда за допомогою ковзного середнього.

Виділена середня сезонна хвиля в такому випадку буде відображати лише центральну ділянку, для якої в наявності трендова складова, сформована за допомогою ковзного середнього.

Застосування лінійної регресійної моделі для виявлення тренду, як правило є недоцільним, і тягне за собою зменшення точності моделі, у порівнянні з методами виявлення тренду, які не роблять припущення про лінійність.

Застосувавши ітеративний підхід, запропонований в підрозділі 3.3, ми

можемо виявити трендову складову без втрати крайових значень. Така лінія тренда зображена на рис. 4.13

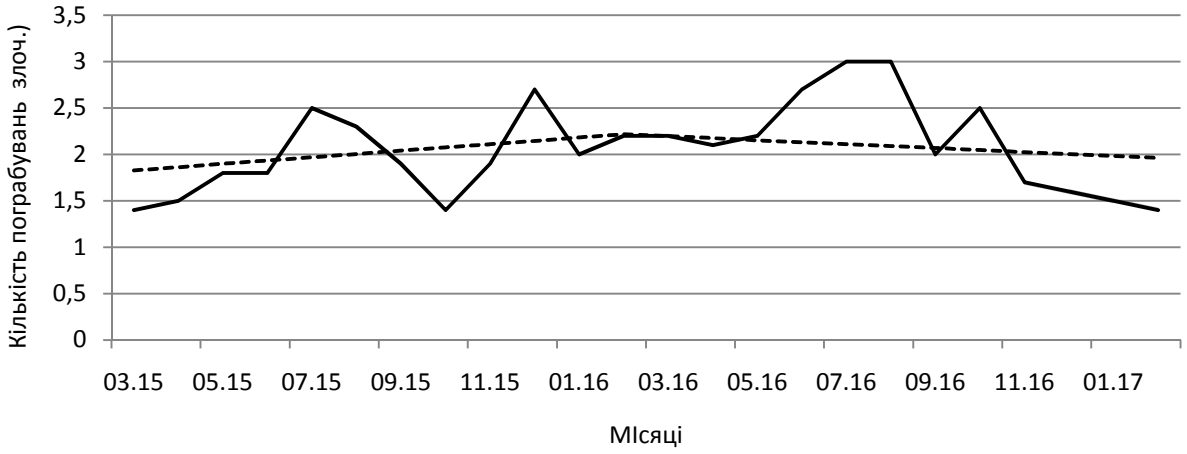

Рисунок 4.13 – Виділення тренду за допомогою F-перетворення

Суцільною лінію на рисунку позначені існуючі рівні дані, переривчастою – виділена лінія тренда за допомогою методу F-перетворення.

Таким чином, під час розрахунку середньої сезонної хвилі за допомогою нечіткого перетворення використовується більше точок, що дозволяє врахувати значення, які б не були використані при розрахунку за допомогою ковзного середнього.

Це пояснюється особливостями розрахунку ковзного середнього, які полягають в тому, що половина періоду розрахункових значень на початку та в кінці ряду не може бути розрахована, у зв’язку з недостатньою кількістю даних. Підходи на основі екстраполяції лінії тренда на ці значення, або за допомогою циклічного заповнення потребують обґрунтування для застосування в кожному конкретному випадку.

Також порівнюючи запропонований в розділі 3.3 підхід, можна відмітити, що розрахунок на основі F – перетворення дозволяє врахувати нелінійність тренду, на відміну від моделей на основі лінійної регресії.

Варто відзначити, що запропонований метод дозволяє ітеративно наближатися до очікуваної лінії тренду, що надає можливість більш точно

виділяти коливання меншого періоду у порівнянні з неітеративним, тобто одноразовим виявленням тренду із застосуванням F-перетворення, або ковзного середнього.

В розділі 3 пропонується застосовувати ітеративний метод виявлення сезонної хвилі, відповідно для даного прикладу сезонна хвиля буде мати вигляд зображений на рис. 4.14.

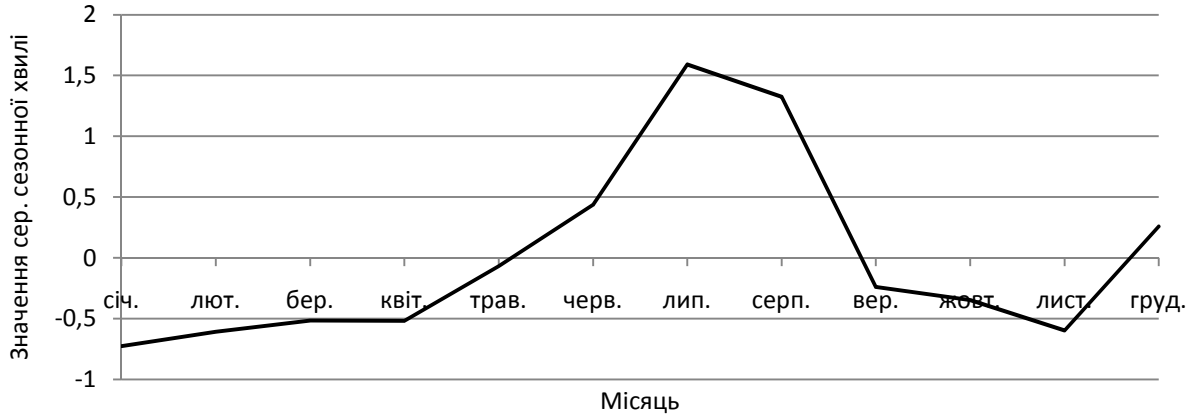

Рисунок 4.14 – Середня сезонна хвиля для кількості скоєних пограбувань

Аналіз сезонної хвилі дозволяє виділити важливу інформацію для планування роботи правоохоронних органів, стосовно попередження скоєння пограбувань в період коли сезонна складова починає діяти і відповідно більша кількість сил та засобів повинна бути задіяна для розслідування та розкриття скоєних злочинів.

Таким чином тренд сезонна модель скоєних пограбувань буде складатися з масиву F-компонент, які визначають трендову складову, та сезонної компоненти, представленої у вигляді середньої сезонної хвилі та коефіцієнту напруженості сезонної хвилі.

За допомогою розрахунку значень зворотнього F-перетворення можна розрахувати значення трендової складової, використавши метод розглянутий в підрозділі 3.2 можна розрахувати значення сезонної складової, сформувавши таким чином модель розглянуту в розділі 2.3.

Тренд-сезонна модель кількості скоєних пограбувань зображена на рис. 4.15. Суцільною лінією на малюнку позначені існуючі дані, розривчастою – дані запропонованої тренд-сезонної моделі.

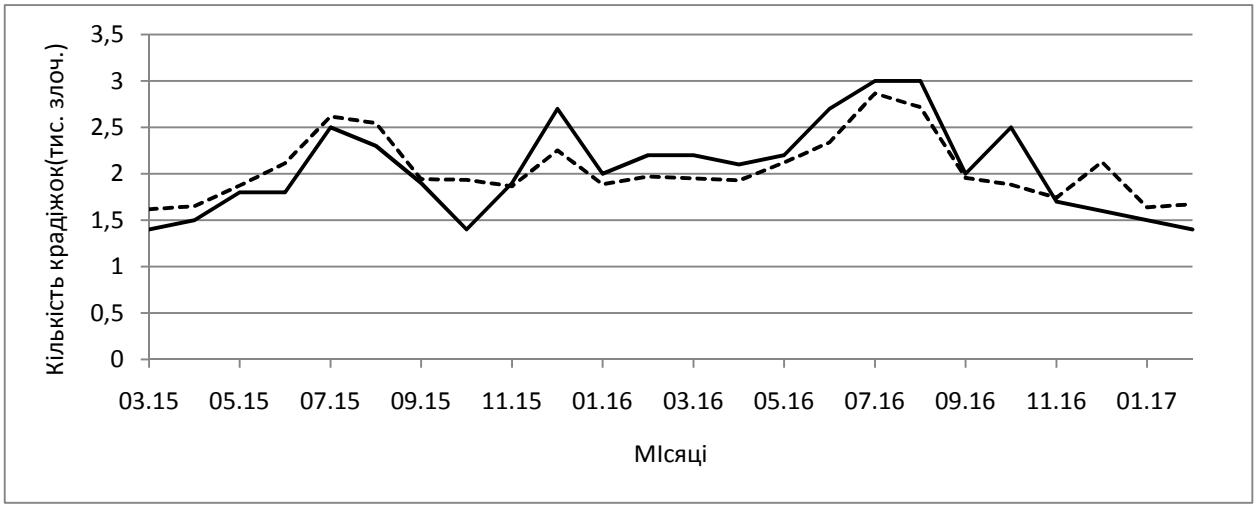

Рисунок 4.15 – Тренд-сезонна модель кількості скоєних пограбувань

Розроблену інтелектуальну інформаційну технологію пропонується використовувати для інформаційно-аналітичних систем в різних предметних галузях.

Однією з ключових функціональних складових діяльності такого типу систем є перетворення інформації, що надходить з джерел даних, із різним рівнем деталізації, в узагальнений вигляд, придатний для прийняття рішень.

В залежності від виду предметної області ефект від рішення, прийнятого на основі підготовленої інформації, може виражатися в різних вимірюваних одиницях – зменшенні вартості витрат, скороченні часу реагування на скоєний злочин, збільшенні прибутку від надання послуг та іншому матеріальному або грошовому виміру.

Позначимо ефект від рішення прийнятого на основі використання:

- тренд-сезонної моделі через E_S;

- багатовимірної нечіткої регресійної моделі – E_F.

В такому разі ефективність застосування зазначених моделей і методів, може бути розрахована як відношення:

- E_S / C_S для підзадачі аналізу наявності сезонної складової для річних даних (або періодичної складової для даних вимірювань з періодом відмінним від річного), де C_S - час витрачений на збирання даних вибірки, який можна подати у вигляді $C_S = k_S N_S$, де k_S - коефіцієнт витрат на збирання одного елементу даних, N_S - кількість елементів вибірки;

- E_F / C_F для підзадачі нечіткого регресійного аналізу, де C_F - це витрати на виправлення помилково прийнятих рішень, які можна подати у вигляді $C_F = k_F N_F$ де k_F - коефіцієнт витрат на виправлення помилкового рішення, N_F - кількість помилкових рішень.

Розглянемо побудову тренд-сезонної моделі в умовах коротких вибірок даних. Для використання класичних методів визначення сезонної складової, що ґрунтуються на застосуванні ковзного середнього, необхідно мінімум 36 значень, з яких за рахунок особливостей застосування ковзного середнього відкидають 12 елементів вибірки (по 6 елементів на початку і в кінці) і досліджують сезонну складову із використанням 24 значень.

Для запропонованого методу фільтрації трендової складової на основі F-перетворення та відповідної тренд-сезонної моделі, необхідно 24 значення. Це дозволяє не очікувати збирання статистичної інформації за додаткові 12 місяців, і відповідно зменшити час і витрати на формування початкових даних C_S на 33% та підвищити ефективність в 1,5 рази в умовах коротких вибірок даних.

Розглянемо побудову нечіткої регресійної моделі для розрахунку коефіцієнтів R_{150}. Кількість помилок для тестової вибірки для побудови нечіткої регресійної моделі без проведення визначення значущих факторів складає 60%. Кількість помилкових рішень для запропонованого методу побудови нечіткої регресійної моделі складає 3,4%, а для застосування методу на основі крокового нечіткого регресійного аналізу 4,1%.

Таким чином, запропонований метод дозволяє зменшити кількість помилок на тестовій вибірці на 17% і підвищити ефективність в 1,2 рази у

порівнянні із методом на основі крокового нечіткого регресійного аналізу.

Узагальнені дані для розрахунку ефективності вказані в таблиці 4.3.

Таблиця 4.3. Узагальнені дані зміни показників витрат та ефективності за рахунок використання запропонованих методів.

	Витрати	Ефективність
Метод фільтрації компонент неоднорідних послідовностей	-33%	1,5
Метод визначення значимих чинників нечіткої регресійної моделі	-17%	1,2

4.4 Висновки

Даний розділ присвячений розробці інтелектуальної інформаційної технології аналізу неоднорідних послідовностей. За допомогою поєднання моделей і методів, розглянутих в другому і третьому розділі, беручи до уваги особливості предметної області викладені в першому розділі, із врахуванням можливостей технічних і програмних засобів збору і обробки інформації були запропоновані етапи інформаційної технології.

Як практичне втілення цих етапів, був запропонований модуль обробки неоднорідних послідовностей, який дозволяє відобразити користувачу результат проведеного аналізу або передати дані у вигляді придатному для подальшого використання іншими програмними засобами. Проведена експериментальна перевірка інтелектуальної інформаційної технології аналізу неоднорідних послідовностей. В результаті дослідження було встановлено, що:

1. Доцільно застосовувати запропоновану інтелектуальну інформаційну технологію неоднорідних послідовностей в рамках одного з поширених підходів до інтелектуального аналізу даних KDD, CRISP-DM або SEMMA.

2. Розроблена інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей даних. Ця технологія дозволяє гнучко налаштовуватися до конкретної, специфічної задачі в результаті діалогу з користувачем, за допомогою того, що використовується можливість інтерактивного генерування моделей, дозволяючи таким чином використовувати декілька альтернативних моделей, які складають функціональне наповнення інформаційно-аналітичної системи.

3. Пропонується отримувати дані неоднорідних послідовностей для задач, які вирішуються запропонованою інформаційною технологією, з таких джерел:

- сховища даних, що зберігають інформацію із використанням реляційного підходу до управління базами даних;
- сховища, які зберігають інформацію за допомогою нереляційного підходу до управління даними, у випадках у випадку отримання великих об’ємів потокових даних або зберігання значної кількості неструктурованих даних в умовах жорстких обмежень на швидкість виконання операцій запису;
- структуровані відомості, які зберігаються у вигляді файлів.

4. Запропонована практична реалізація викладених етапів інформаційної технології у вигляді програмного модуля, який би втілював описані в попередніх розділах моделі і методи, що застосовуються під час аналізу неоднорідних послідовностей. Ключові функції модуля реалізовані таким чином, що вони є сумісними з програмним забезпеченням створеним за допомогою мови програмування Python.

5. У ході дослідження було проведено ряд експериментів з аналізу неоднорідних послідовностей із побудовою моделей та за допомогою методів викладених в попередніх розділах із застосуванням запропонованої інформаційної технології. Результати, які були отримані в ході експериментів свідчать, що:

- під час побудови тренд-сезонної моделі рівнів злочинності із умовою наявності лише короткого ряду даних, для вилучення трендової складової за

допомогою запропонованого методу фільтрації компонент неоднорідних послідовностей враховувалися крайові значення ряду, що дозволило виділити сезонну хвилю і відповідно зменшити час і витрати на формування початкових даних на 33% та підвищити ефективність в 1,5 рази в умовах коротких вибірок даних;

- для нечіткої регресійної моделі, застосованої для моделювання показників носового дихання за допомогою розробленого методу визначення значимих чинників нечіткої регресійної моделі, була визначена множина значущих факторів, що дозволило запобігти перенавчанню нечіткої лінійної регресії та отримати підмножину значущих чинників за скінченну кількість ітерацій і в результаті зменшити кількість помилок на тестовій вибірці на 17% і підвищити ефективність в 1,2 рази у порівнянні із методом на основі крокового нечіткого регресійного аналізу

Список використаних джерел у даному розділі наведено у повному списку використаних джерел під номерами: [13, 17,26 ,27, 158 - 167].

ВИСНОВКИ

У дисертаційній роботі розв’язана нова актуальна науково-практична задача розробки моделей, методів та створення на їх основі інтелектуальної інформаційної технології аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметних областей в інформаційно-аналітичних системах. В результаті виконання роботи отримані нові наукові та практичні результати.

1. Дослідження сучасних моделей і методів аналізу неоднорідних послідовностей даних для задач оцінювання поточного стану предметної області показало, що існує потреба в створенні методу побудови багатофакторної нечіткої регресійної моделі неоднорідних послідовностей даних із урахуванням значущих чинників. Також на основі проведених досліджень було виявлено, що тренд-сезонні адитивні моделі неоднорідних часових послідовностей, які ґрунтуються на використанні ковзного середнього, потребують вдосконалення пов’язаного із використанням крайових значень часового ряду в умовах коротких вибірок даних.

2. Отримала подальший розвиток тренд-сезонна модель неоднорідних часових послідовностей шляхом утворення нечітких розділів із асоційованими функціями належності, які враховуються при поданні трендової складової у вигляді інтерпольованих усереднених значень. Це дозволяє застосовувати вдосконалену модель для коротких вибірок без втрати крайових елементів неоднорідної послідовності даних.

3. Вперше запропонований метод визначення значущих чинників нечіткої регресійної моделі даних неоднорідних послідовностей на основі відбору підмножини значущих чинників з коефіцієнтами, які перевищують порогове значення. Коефіцієнти підбирається за критерієм рівнозначності кутів відхилення між вектором похибки і векторами змінних. Запропонований метод дозволяє отримати підмножину значущих чинників за скінченну кількість ітерацій та запобігти перенавчанню нечіткої лінійної регресії.

4. Отримав подальший розвиток метод фільтрації компонент неоднорідних часових послідовностей шляхом застосування ітеративного розбиття початкової послідовності на скінчену кількість нечітких розділів з кожним з яких асоційована функція належності, яка враховується при отриманні усереднених значень для кожного з центрів нечітких розділів, за допомогою яких подається трендова складова. Значення, які знаходяться на ділянках поза центрами нечітких розділів розраховуються за допомогою інтерполяції. Це дозволяє відфільтровувати коливання різних періодів при виділенні трендової складової і тим самим підвищити ефективність оцінювання зміни стану предметної області.

5. На основі запропонованих моделей і методів аналізу неоднорідних послідовностей даних було розроблено інтелектуальну інформаційну технологію аналізу неоднорідних послідовностей даних для оцінювання поточного стану предметних областей в інформаційно-аналітичних системах. Застосування запропонованої інтелектуальної інформаційної технології для моделювання показників носового дихання дозволило зменшити кількість помилок на тестовій вибірці на 17% і підвищити ефективність в 1,2 рази. Використання для дослідження сезонної складової дозволило зменшити час і витрати на формування початкових даних на 33% та підвищити ефективність в 1,5 рази.

6. Проведено впровадження моделі, методів та інтелектуальної інформаційної технології при вирішенні практичних задач в діяльність ТОВ «Ендейвер», м. Полтава, та в діяльність Головного управління національної поліції Харківської області, а також в навчальний процес кафедри програмної інженерії ХНУРЕ, що підтверджено відповідними актами впровадження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] P.-T. Chung and S. H. Chung, “On data integration and data mining for developing business intelligence”, in *Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island*, 2013, pp. 1–6.

[2] O. Ali, P. Crvenkovski, and H. Johnson, “Using a business intelligence data analytics solution in healthcare”, in *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual*, 2016, pp. 1–6.

[3] M. H. ur Rehman, V. Chang, A. Batool, and T. Y. Wah, “Big data reduction framework for value creation in sustainable enterprises”, *International Journal of Information Management*, vol. 36, no. 6, pp. 917–928, 2016.

[4] A. Nagy and J. Tick, “Improving transport management with big data analytics”, in *Intelligent Systems and Informatics (SISY), 2016 IEEE 14th International Symposium on*, 2016, pp. 199–204.

[5] L. Zadeh, “Fuzzy logic and soft computing: Issues, contentions and perspectives”, в *Proc. IIZUKA*, 1994, vol 94, pp 1–2.

[6] T. L. Saaty, “Measuring the Fuzziness of Sets”, *Journal of Cybernetics*, vol 4, no 4, pp 53–61, 1974.

[7] E. H. Ruspini, “A new approach to clustering”, *Information and control*, vol 15, no 1, pp 22–32, 1969.

[8] I. Perfilieva, “Fuzzy transforms: Theory and applications”, *Fuzzy sets and systems*, vol 157, no 8, pp 993–1023, 2006.

[9] H. Tanaka, S. Uejima, and K. Asai, “Linear regression analysis with fuzzy model”, *IEEE Transaction Systems Man and Cybermatics*, vol 12, no 6, pp 903–907, 1982.

[10] P. J. Curran and A. M. Hussong, “Integrative data analysis: the simultaneous analysis of multiple data sets”, *Psychological methods*, vol 14, no 2, p 81, 2009.

[11] Л. Раскин и О. Серая, *Нечеткая математика*, [Fuzzy math], Парус. Харьков, 2008.

[12] Y. Bodyanskiy, “Computational Intelligence Techniques for Data Analysis”, in *Leipziger Informatik-Tage*, 2005, pp. 15–36.

[13] A. Yerokhin, A. Babii, A. Nechyporenko, and O. Turuta, “A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features”, *Cybernetics and Systems Analysis*, vol 52, no 4, pp 641–646, 2016.

[14] М. М. Зацеркляний, А. Л. Єрохін, А. С. Бабій і О. П. Турута, “Розробка методу виявлення сезоних коливань з застосуванням нечіткого згладжування на базі F-перетворення”, *Біоніка інтелекту*, no 2, pp 89–93, 2011.

[15] А. С. Бабій і М.М. Зацеркляний, “Автоматизація аналізу сезонних коливань рівня злочинності”, *Право і безпека*, vol 4, no № 3, pp 163–166, 2005.

[16] А. С. Бабій і М.М. Зацеркляний, “Аналіз тенденцій розвитку злочинності”, *Системи обробки інформації*, no 4, pp 153–155, 2007.

[17] М. М. Зацеркляний і А.С. Бабій, “Інформаційна система моделювання впливу чинників злочинності”, *Право і Безпека*, vol 7, no № 2, pp 204–209, 2008.

[18] М. М. Зацеркляний і А. С. Бабій, “Попередній аналіз даних у системах обробки інформації про скоєні злочини”, *Право і Безпека*, no 1, pp 269–272, 2009.

[19] А. С. Бабій і О. Ф. Лановий, “Статистичне моделювання злочинності”, *ВІсник НТУ ХПІ*, no 19, pp 24–30, 2006.

[20] А. Л. Єрохін, А. С. Бабій, і О. П. Турута, “Спеціальна інформаційна система для виклику екстрених служб в Україні”, в *Збірник праць IV міжнародної науково-практичної конференції*, Харків, 2011, p 163.

[21] М. М. Зацеркляний і А. С. Бабій, “Застосування методу найменших кутів для аналізу чинників злочинності”, в *Материалы международной научно-технической конференции “Информационные системы и технологии”*, Харьков, 2012, p 37.

[22] К. Е. Петров, М. М. Зацеркляний і А. С. Бабій, “Оцінювання злочинності із врахуванням нечіткості”, в *Спеціальна техніка у правоохоронній діяльності, Матеріали V Міжнародної науково-практичної конференції*, Київ, 2012, p 79.

[23] A. Yerokhin, A. Nechyporenko, A. Babii, and O. Turuta, “Usage of F-transform to finding informative parameters of rhinomanometric signals”, в *Scientific and Technical Conference« Computer Sciences and Information Technologies»(CSIT), 2015 Xth International*, 2015, pp 129–132.

[24] A. Yerokhin, A. Nechyporenko, A. Babii, and O. Turuta, “Processing and analysis of rhinomanometric signals by F-transform approximation”, в *Data Stream Mining & Processing (DSMP), IEEE First International Conference on*, 2016, pp 314–317.

[25] A. Yerokhin, O. Turuta, A. Babii, A. Nechyporenko, and I. Mahdalina, “Usage of phase space diagram to finding significant features of rhinomanometric signals”, в *Scientific and Technical Conference “Computer Sciences and Information Technologies (CSIT), 2016 XIth International*, 2016, pp 70–72.

[26] А. С. Бабій, “Побудова СППР для оцінювання злочинності”, в *Збірник праць II міжнародної науково-технічної конференції Інформаційні технології в навігації і управлінні*, Харків, 2011, p 41.

[27] А. С. Бабій, “Програмна система для аналізу злочинності”, *Вісник НТУ ХПІ*, no 19, pp 12–16, 2007.

[28] А. С. Бабій, “Автоматизація управління діяльністю правоохоронних органів”, в *Державне управління та місцеве самоврядування: тези VII міжнародного наукового конгресу*, Харків, 2007, pp 20–22.

[29] А. Г. Додонов, С. Р. Коженевский, Д. В. Ландэ и В.Г. Путятин, “Компьютерные информационные системы и хранилища данных.Толковый словарь”, Київ, 2013, с 554.

[30] В. Кудрявцев, “Основные причины организованной преступности в России”, *Вестник Российской академии наук*, no 9, p 794, 1999.

[31] В. В. Лунеев, *Преступность XX века: мировые, региональные и российские*

тенденции. Wolters Kluwer Russia, 2005.

[32] J. Thomas and L. Sael, “Overview of integrative analysis methods for heterogeneous data”, в *Big Data and Smart Computing (BigComp), 2015 International Conference on*, 2015, pp 266–270.

[33] N. Subrahmanya and Y. C. Shin, “Sparse multiple kernel learning for signal processing applications”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 32, no 5, pp 788–798, 2010.

[34] L. R. Tucker, “Some mathematical notes on three-mode factor analysis”, *Psychometrika*, vol 31, no 3, pp 279–311, 1966.

[35] J. D. Carroll and J.-J. Chang, “Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition”, *Psychometrika*, vol 35, no 3, pp 283–319, 1970.

[36] J. R. Kettenring, “Canonical analysis of several sets of variables”, *Biometrika*, vol 58, no 3, pp 433–451, 1971.

[37] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, “Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey”, *Proceedings of the IEEE*, vol 98, no 10, pp 1692–1715, 2010.

[38] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion: A review of the state-of-the-art”, *Information Fusion*, vol 14, no 1, pp 28–44, 2013.

[39] I. Van Mechelen and A. K. Smilde, “A generic linked-mode decomposition model for data fusion”, *Chemometrics and Intelligent Laboratory Systems*, vol 104, no 1, pp 83–94, 2010.

[40] M. Turk, “Multimodal interaction: A review”, *Pattern Recognition Letters*, vol 36, pp 189–195, 2014.

[41] D. Lahat, T. Adali, and C. Jutten, “Multimodal data fusion: an overview of methods, challenges, and prospects”, *Proceedings of the IEEE*, vol 103, no 9, pp 1449–1477, 2015.

[42] W. Elmenreich, “Sensor fusion in time-triggered systems”, Technische Universitaat Wien, Institut four Technische Informatik, Vienna,

Austria, 2002.

[43] E. F. Nakamura, A. A. Loureiro, and A. C. Frery, “Information fusion for wireless sensor networks: Methods, models, and classifications”, *ACM Computing Surveys (CSUR)*, vol 39, no 3, p 9, 2007.

[44] H. Pan, Z.-P. Liang, T. J. Anastasio, and T. S. Huang, “A hybrid NN-Bayesian architecture for information fusion”, в *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 1998, vol 1, pp 368–371.

[45] C. Cou, T. Fraichard, P. Bessiere, and E. Mazer, “Multi-sensor data fusion using Bayesian programming: An automotive application”, в *Intelligent Robots and Systems, 2002.* 2002, vol 1, pp 141–146.

[46] B. Moshiri, M.R. Asharif, and R. HoseinNezhad, “Pseudo information measure: A new concept for extension of Bayesian fusion in robotic map building”, *Information Fusion*, vol 3, no 1, pp 51–68, 2002.

[47] A. Tsymbal, S. Puuronen, and D. W. Patterson, “Ensemble feature selection with the simple Bayesian classification”, *Information fusion*, vol 4, no 2, pp 87–100, 2003.

[48] M. L. Sichitiu and V. Ramadurai, “Localization of wireless sensor networks with a mobile beacon”, в *Mobile Ad-hoc and Sensor Systems, 2004 IEEE International Conference on*, 2004, pp 174–183.

[49] R. Biswas, S. Thrun, and L. J. Guibas, “A probabilistic approach to inference with limited information in sensor networks”, в *Information Processing in Sensor Networks, 2004. IPSN 2004.*, 2004, pp 269–276.

[50] B. Krishnamachari and S. Iyengar, “Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks”, *IEEE Transactions on Computers*, vol 53, no 3, pp 241–250, 2004.

[51] G. Hartl and B. Li, “infer: A Bayesian inference approach towards energy efficient data collection in dense sensor networks”, в *Distributed Computing Systems, 2005. ICDCS 2005*, 2005, pp 371–380.

[52] A. P. Dempster, “A generalization of Bayesian inference”, *Classic works of the dempster-shafer theory of belief functions*, vol 219, pp 73–104, 2008.

[53] G. Shafer and others, *A mathematical theory of evidence*, vol 1. Princeton university press Princeton, 1976.

[54] G. M. Provan, "The validity of Dempster-Shafer belief functions", *International Journal of Approximate Reasoning*, vol 6, no 3, pp 389–399, 1992.

[55] M. A. Fischler, "An inference technique for integrating knowledge from disparate sources", *Multisensor integration and fusion for intelligent machines and systems*, p 309, 1995.

[56] A. J. Pinto, J. M. Stochero, and J. F. de Rezende, "Aggregation-aware routing on wireless sensor networks", в *IFIP International Conference on Personal Wireless Communications*, 2004, pp 238–247.

[57] E. F. Nakamura, F. G. Nakamura, C. M. Figueiredo and A. A. Loureiro, "Using information fusion to assist data dissemination in wireless sensor networks", *Telecommunication Systems*, vol 30, no 1, pp 237–254, 2005.

[58] B. Yu, K. Sycara, J. Giampapa, and S. Owens, "Uncertain information fusion for force aggregation and classification in airborne sensor networks", в *AAAI-04 Workshop on Sensor Networks*, 2004.

[59] M. W. Roth, "Survey of neural network technology for automatic target recognition", *IEEE Transactions on neural networks*, vol 1, no 1, pp 28–43, 1990.

[60] R. H. Baran, "A collective computation approach to automatic target recognition", в *Proceedings of the International Joint Conference on Neural Networks*, 1989, vol 1, pp 39–44.

[61] L. Yiyao, Y. Venkatesh, and C. C. Ko, "A knowledge-based neural network for fusing edge maps of multi-sensor images", *Information Fusion*, vol 2, no 2, pp 121–133, 2001.

[62] C. S. Peirce, "Abduction and induction", *Philosophical writings of Peirce*, vol 11, 1955.

[63] L. M. De Campos, J. A. Gamez, and S. Moral, "Partial abductive inference in Bayesian belief networks-an evolutionary computation approach by using problem-specific genetic operators", *IEEE Transactions on Evolutionary*

Computation, vol 6, no 2, pp 105–131, 2002.

[64] A. M. Abdelbar, E. A. Andrews, and D. C. Wunsch, “Abductive reasoning with recurrent neural networks”, *Neural Networks*, vol 16, no 5, pp 665–673, 2003.

[65] J. R. Aguero and A. Vargas, “Inference of operative configuration of distribution networks using fuzzy logic techniques-Part II: extended real-time model”, *IEEE Transactions on Power Systems*, vol 20, no 3, pp 1562–1569, 2005.

[66] J. Keppens, Q. Shen, and B. Schafer, “Probabilistic abductive computation of evidence collection strategies in crime investigation”, в *Proceedings of the 10th international conference on artificial intelligence and law*, 2005, pp 215–224.

[67] V. Kumar and U. B. Desai, “Image interpretation using Bayesian networks”, *IEEE Transactions on pattern analysis and machine intelligence*, vol 18, no 1, pp 74–77, 1996.

[68] R. J. Mooney, “Integrating abduction and induction in machine learning”, в *Abduction and Induction*, Springer, 2000, pp 181–191.

[69] D. S. Friedlander, *Semantic information extraction*. CRC Press, Boca Raton, 2005.

[70] D. Friedlander and S. Phoha, “Semantic information fusion for coordinated signal processing in mobile sensor networks”, *The International Journal of High Performance Computing Applications*, vol 16, no 3, pp 235–241, 2002.

[71] G. Banon, “Distinction between several subsets of fuzzy measures”, *Fuzzy sets and systems*, vol 5, no 3, pp 291–305, 1981.

[72] V. Novák, I. Perfilieva, and J. Mockor, *Mathematical principles of fuzzy logic*. Springer Science & Business Media, 2012.

[73] C. C. Lee, “Fuzzy logic in control systems: fuzzy logic controller”, *IEEE Transactions on systems, man, and cybernetics*, vol 20, no 2, pp 404–418, 1990.

[74] I. Gupta, D. Riordan, and S. Sampalli, “Cluster-head election using fuzzy logic for wireless sensor networks”, в *Communication Networks and Services Research Conference, 2005. Proceedings of the 3rd Annual*, 2005, pp 255–260.

[75] M. N. Halgamuge, S. M. Guru, and A. Jennings, “Energy efficient cluster

formation in wireless sensor networks", в *Telecommunications, 2003. ICT 2003. 10th International Conference on*, 2003, vol 2, pp 1571–1576.

[76] X. Cui, T. Hardin, R. K. Ragade, and A. S. Elmaghraby, "A swarm-based fuzzy logic control mobile sensor network for hazardous contaminants localization", в *Mobile Ad-hoc and Sensor Systems, 2004 IEEE International Conference on*, 2004, pp 194–203.

[77] H. Shu and Q. Liang, "Fuzzy optimization for distributed sensor deployment", в *Wireless Communications and Networking Conference, 2005 IEEE*, 2005, vol 3, pp 1903–1908.

[78] P. Diamond, "Fuzzy least squares", *Information Sciences*, vol 46, no 3, pp 141–157, 1988.

[79] A. F. Shapiro, "Fuzzy regression models", *Penn State University*, vol 6, p 12, 2005.

[80] S. Yeylaghi, M. Otadi, and N. Imankhan, "A new fuzzy regression model based on interval-valued fuzzy neural network and its applications to management", *Beni-Suef University Journal of Basic and Applied Sciences*, 2017.

[81] J. de Andrés-Sánchez, "Fuzzy Regression Analysis: An Actuarial Perspective", в *Fuzzy Statistical Decision-Making*, Springer, 2016, pp 175–201.

[82] A. Ubale and S. Sananse, "A comparative study of fuzzy multiple regression model and least square method", *International Journal of Applied Research*, vol 2, no 7, pp 11–15, 2016.

[83] J. Chachi and S. M. Taheri, "Multiple Fuzzy Regression Model for Fuzzy Input-Output Data", *Iranian Journal of Fuzzy Systems*, vol 13, no 4, pp 63–78, 2016.

[84] Y. H. O. Chang and B. M. Ayyub, "Fuzzy regression methods–a comparative assessment", *Fuzzy sets and systems*, vol 119, no 2, pp 187–203, 2001.

[85] Q. Cai, D. Zhang, W. Zheng, and S. C. Leung, "A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression", *Knowledge-Based Systems*, vol 74, pp 61–68, 2015.

[86] L. S. Riza, C. N. Bergmeir, F. Herrera, and J. M. Benítez Sánchez, "frbs: Fuzzy rule-based systems for classification and regression in R", 2015.

[87] I. Perfilieva and R. Valášek, “Fuzzy transforms in removing noise”, *Computational Intelligence, Theory and Applications*, pp 221–230, 2005.

[88] L. Stefanini, “F-transform with parametric generalized fuzzy partitions”, *Fuzzy Sets and Systems*, vol 180, no 1, pp 98–120, 2011.

[89] M. Holčapek and T. Tichỳ, “A smoothing filter based on fuzzy transform”, *Fuzzy sets and systems*, vol 180, no 1, pp 69–97, 2011.

[90] G. Patanè, “Fuzzy transform and least-squares approximation: analogies, differences, and generalizations”, *Fuzzy Sets and Systems*, vol 180, no 1, pp 41–54, 2011.

[91] I. Perfilieva, V. Novák, and A. Dvořák, “Fuzzy transform in the analysis of data”, *International Journal of Approximate Reasoning*, vol 48, no 1, pp 36–46, 2008.

[92] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[93] S. J. Julier and J. K. Uhlmann, “A new extension of the Kalman filter to nonlinear systems”, в *Int. symp. aerospace/defense sensing, simul. and controls*, 1997, vol 3, pp 182–193.

[94] N. De Freitas, C. Andrieu, P. Højen-Sørensen, M. Niranjan, and A. Gee, “Sequential Monte Carlo methods for neural networks”, в *Sequential Monte Carlo methods in practice*, Springer, 2001, pp 359–379.

[95] B. A. Berg and A. Billoire, *Markov chain monte carlo simulations*. Wiley Online Library, 2008.

[96] L. D. Stone, R. L. Streit, T. L. Corwin, and K. L. Bell, *Bayesian multiple target tracking*. Artech House, 2013.

[97] A. Benavoli, B. Ristic, A. Farina, M. Oxenham, and L. Chisci, “An approach to threat assessment based on evidential networks”, в *Information Fusion, 2007 10th International Conference on*, 2007, pp 1–8.

[98] H. H. S. Ip and H. Tang, “Parallel evidence combination on a SB-tree architecture”, в *Intelligent Information Systems, 1996., Australian and New Zealand Conference on*, 1996, pp 31–34.

[99] M. Bauer, “Approximation algorithms and decision making in the Dempster-Shafer theory of evidence—An empirical study”, *International Journal of Approximate Reasoning*, vol 17, no 2–3, pp 217–237, 1997.

[100] L. A. Zadeh, “Fuzzy sets as a basis for a theory of possibility”, *Fuzzy sets and systems*, vol 100, pp 9–34, 1999.

[101] D. Dubois and H. Prade, “Possibility theory in information fusion”, в *Information Fusion, 2000*, vol 1, p PS6–P19.

[102] D. Dubois and H. Prade, *Possibility theory: an approach to computerized processing of uncertainty*. Springer Science & Business Media, 2012.

[103] H. Borotschnig, L. Paletta, and A. Pinz, “A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition”, *Computing*, vol 62, no 4, pp 293–319, 1999.

[104] D. Dubois and H. Prade, “Possibility theory and data fusion in poorly informed environments”, *Control Engineering Practice*, vol 2, no 5, pp 811–823, 1994.

[105] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*, vol 9. Springer Science & Business Media, 2012.

[106] J. Peters, S. Ramanna, A. Skowron, J. Stepaniuk, and Z. Suraj, “Sensor fusion: A rough granular approach”, в *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, 2001, vol 3, pp 1367–1371.

[107] L. Yong, X. Congfu, and P. Yunhe, “A new approach for data fusion: implement rough set theory in dynamic objects distinguishing and tracing”, в *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, 2004, vol 4, pp 3318–3322.

[108] D. S. Yeung, D. Chen, E. C. Tsang, J. W. Lee, and W. Xizhao, “On the generalization of fuzzy rough sets”, *IEEE Transactions on fuzzy systems*, vol 13, no 3, pp 343–361, 2005.

[109] O. Basir, F. Karray, and H. Zhu, “Connectionist-based Dempster-Shafer evidential reasoning for data fusion”, *IEEE Transactions on Neural Networks*, vol 16, no 6, pp 1513–1530, 2005.

[110] D. Kendall, “Foundations of a theory of random sets”, *Stochastic geometry*,

vol 3, no 9, 1974.

[111] R. Mahler, “Random sets: Unification and computation for information fusion-a retrospective assessment”, в *Proceedings of the Seventh International Conference on Information Fusion*, 2004, vol 1, pp 1–20.

[112] B. T. Vo, B.N. Vo, and A. Cantoni, “The cardinalized probability hypothesis density filter for linear Gaussian multi-target models”, в *Information sciences and systems, 2006 40th annual conference on*, 2006, pp 681–686.

[113] Y. H. O. Chang, “Hybrid fuzzy least-squares regression analysis and its reliability measures”, *Fuzzy Sets and Systems*, vol 119, no 2, pp 225–246, 2001.

[114] G. Peters, “Fuzzy linear regression with fuzzy intervals”, *Fuzzy sets and Systems*, vol 63, no 1, pp 45–55, 1994.

[115] K. J. Kim, H. Moskowitz, and M. Koksalan, “Fuzzy versus statistical linear regression”, *European Journal of Operational Research*, vol 92, no 2, pp 417–434, 1996.

[116] K. Y. Chan, H. K. Lam, T. S. Dillon, and S. H. Ling, “A Stepwise-Based Fuzzy Regression Procedure for Developing Customer Preference Models in New Product Development”, *IEEE Transactions on Fuzzy Systems*, vol 23, no 5, pp 1728–1745, 2015.

[117] V. Milea, R. J. Almeida, U. Kaymak, and F. Frasincar “A fuzzy model of a European index based on automatically extracted content information”, в *(CIFEr), 2011 IEEE Symposium on*, 2011, pp 1–8.

[118] V. Novák, M. Štěpnička, A. Dvořák, I. Perfilieva, V. Pavliska, and L. Vavříčková, “Analysis of seasonal time series using fuzzy approach”, *International Journal of General Systems*, vol 39, no 3, pp 305–328, 2010.

[119] S. P. McLaughlin, R. J. Evans, and V. Krishnamurthy, “Data incest removal in a survivable estimation fusion architecture”, в *Proc. of the International Conference on Information Fusion*, 2003, pp 229–236.

[120] L. Y. Pao and M. Kalandros, “Algorithms for a class of distributed architecture tracking”, в *American Control Conference, 1997*, vol 3, pp 1434–1438.

[121] S. P. McLaughlin, R. J. Evans, and V. Krishnamurthy, “A graph theoretic approach to data incest management in network centric warfare”, в *Information Fusion, 2005 8th International Conference on*, 2005, vol 2, p 8–pp.

[122] S. J. Julier and J. K. Uhlmann, “A non-divergent estimation algorithm in the presence of unknown correlations”, в *American Control Conference, 1997. Proceedings of the 1997*, 1997, vol 4, pp 2369–2373.

[123] W. Niehsen, “Information fusion based on fast covariance intersection filtering”, в *Information Fusion, 2002*, vol 2, pp 901–904.

[124] Y. Zhou and J. Li, “Robust decentralized data fusion based on internal ellipsoid approximation”, *IFAC Proceedings Volumes*, vol 41, no 2, pp 9964–9969, 2008.

[125] S. Walfish, “A review of statistical outlier methods”, *Pharmaceutical technology*, vol 30, no 11, p 82, 2006.

[126] S. Budalakoti, A. N. Srivastava, and M. E. Otey, “Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety”, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol 39, no 1, pp 101–113, 2009.

[127] K. Sequeira and M. Zaki, “ADMIT: anomaly-based data mining for intrusions”, в *Proceedings of the eighth ACM SIGKDD*, 2002, pp 386–395.

[128] C. C. Michael and A. Ghosh, “Two state-based approaches to program-based anomaly detection”, в *Computer. ACSAC’00*, 2000, pp 21–30.

[129] S. Salvador and P. Chan, “Learning states and rules for detecting anomalies in time series”, *Applied Intelligence*, vol 23, no 3, pp 241–255, 2005.

[130] X. Li and J. Han, “Mining approximate top-k subspace anomalies in multi-dimensional time-series data”, в *Proceedings of the 33rd international conference on Very large data bases*, 2007, pp 447–458.

[131] W. Lee, S. J. Stolfo, and others, “Data mining approaches for intrusion detection”, в *USENIX Security Symposium*, 1998, pp 79–93.

[132] S. Tian, S. Mu, and C. Yin, “Sequence-similarity kernels for SVMs to detect anomalies in system calls”, *Neurocomputing*, vol 70, no 4, pp 859–866, 2007.

[133] M. A. Pravia, R. K. Prasanth, P. O. Arambel, C. Sidner, and C.-Y. Chong, “Generation of a fundamental data set for hard/soft information fusion”, в *Information Fusion, 2008 11th International Conference on*, 2008, pp 1–8.

[134] K. Premaratne, M. N. Murthi, J. Zhang, M. Scheutz, and P. H. Bauer, “A Dempster-Shafer theoretic conditional approach to evidence updating for fusion of hard and soft data”, в *Information Fusion, 2009*, pp 2122–2129.

[135] A. Auger and J. Roy, “Expression of uncertainty in linguistic data”, в *Information Fusion, 2008 11th International Conference on*, 2008, pp 1–8.

[136] D. L. Hall, M. D. McNeese, D. B. Hellar, B. J. Panulla, and W. Shumaker, “A cyber infrastructure for evaluating the performance of human centered fusion”, в *Information Fusion, 2009. FUSION’09.*, pp 1257–1264.

[137] Б. Грабовецький, *Теоретико-методологічні основи аналізу і прогнозування тенденції змін технікоекономічних показників в системі АПК*, ВНТУ. Вінниця, 2011.

[138] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[139] А. И. Кобзарь, *Прикладная математическая статистика*. Москва: Физматлит, 2006.

[140] R. A. Johnson, D. W. Wichern, and others, *Applied multivariate statistical analysis*, vol. 4. Prentice-Hall New Jersey, 2014.

[141] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A comparative study of efficient initialization methods for the k-means clustering algorithm” , *Expert systems with applications*, vol. 40, no. 1, pp. 200–210, 2013.

[142] F. J. Gravetter and L. B. Wallnau, *Statistics for the behavioral sciences*. Cengage Learning, 2016.

[143] P. Giudici, *Applied data mining: statistical methods for business and industry*. John Wiley & Sons, 2005.

[144] А. Готман, “Теория вероятностей и математическая статистика”, *Международный журнал прикладных и фундаментальных исследований*, no 7, 2011.

[145] T. Vigen, *Spurious correlations*. Hachette Books, 2015.

[146] Н. Е. Булетова, Г. В. Кузибецкая, Е. В. Демичева, И. А. Злочевский и С. Н. Демянчук, *Статистические методы исследования макроэкономических явлений и процессов*. Scientific magazine" Kontsep, 2014.

[147] U. Hassler and T. Thadewald, “Nonsensical and biased correlation due to pooling heterogeneous samples”, *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol 52, no 3, pp 367–379, 2003.

[148] E. H. Simpson, “The interpretation of interaction in contingency tables”, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp 238–241, 1951.

[149] M. G. Kendall and B. B. Smith, “The problem of m rankings”, *The annals of mathematical statistics*, vol 10, no 3, pp 275–287, 1939.

[150] W. R. Schucany and W. H. Frawley, “A rank test for two group concordance”, *Psychometrika*, vol 38, no 2, pp 249–258, 1973.

[151] В. Оболенцев, *Латентна злочинність: проблеми теорії та практики попередження*. Х.: Вид. СПД ФО Вапнярчук НМ, 2005.

[152] О. Серая, *Многомерные модели логистики в условиях неопределенности*, Харьков, ФОП Стеценко И. И., 2010. с. 512.

[153] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, and others, “Least angle regression”, *The Annals of statistics*, vol 32, no 2, pp 407–499, 2004.

[154] S. G. Gilmour, “The interpretation of Mallows’s C_p-statistic”, *The Statistician*, pp 49–56, 1996.

[155] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[156] М. Кендалл і А. Стюарт, *Многомерный статистический анализ и временные ряды*. Наука, 1976.

[157] Н. С. Четвериков, *Статистические исследования:(Теория и практика)*. Москва: Наука, 1975.

[158] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, “From data

mining to knowledge discovery: an overview", *Advances in knowledge discovery and data mining*, vol 21, pp 561–572, 1996.

[159] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models", *The Knowledge Engineering Review*, vol 21, no 1, pp 1–24, 2006.

[160] A. I. R. L. Azevedo, and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", *IADS-DM*, pp 182–185, 2008.

[161] D. Mouheb et al., "Unified Modeling Language," in *Aspect-Oriented Security Hardening of UML Design Models*, Springer, 2015, pp. 11–22.

[162] N. Klarlund, T. Schwentick, and D. Suciu, "XML: model, schemas, types, logics, and queries", в *Logics for Emerging Applications of Databases*, Springer, 2004, pp 1–41.

[163] V. V. Chmovzh, A.S. Nechyporenko, and O.G. Garyuk, "System approach to finding hydrodynamic resistance coefficient of a nasal cavity", *Computer science, information technology, automation journal*, no 4, pp 8–15, 2016.

[164] F. Roher, "Der Stromungswiderstand in der menschlichen Atemwegen und der Einfluss der unregelmassigen Verzweigung es Bronchial-systems auf der Atmungsverlauf in vershiedenen Lungenbezinken", *Arch Ges Physiol*, vol 162, pp 225–229, 1915.

[165] P. Broms, "Rhinomanometry. III. Procedures and criteria for distinction between skeletal stenosis and mucosal swelling", *Acta oto-laryngologica*, vol 94, no 3–4, pp 361–370, 1982.

[166] G. Mlynski and A. Beule, "Diagnostik der respiratorischen Funktion der Nase", *Hno*, vol 56, no 1, pp 81–99, 2008.

[167] Державна служба статистики України, "Соціально-економічне становище України",*Офіційний сайт державної служби статистики України*, [Online],Available: http://www.ukrstat.gov.ua/druk/soc_ek/2017/arh_2017_u.htm, [Accessed: Mar. 5, 2017].

ДОДАТОК А

А.1 Список публікацій здобувача

1. A. L. Yerokhin, A. S. Babii, A. S. Nechyporenko, O. P. Turuta. A Lars-Based Method of the Construction of a Fuzzy Regression Model for the Selection of Significant Features. Cybernetics and Systems Analysis, Springer US, 2016. V. 52, Issue 4, P. 641–646, DOI:10.1007/s10559-016-9867-5 (Входить до міжнародної наукометричної бази SCOPUS).

2. Зацеркляний М.М., Єрохін А.Л., Бабій А.С.,Турута О.П. Розробка методу виявлення сезоних коливань з застосуванням нечіткого згладжування на базі F-перетвореннння. Біоніка інтелекту, Харків: ХНУРЕ, 2011. №2011’2. С. 89 – 93

3. Бабій А.С., Зацеркляний М.М. Автоматизація аналізу сезонних коливань рівня злочинності. Право і безпека. Харків: ХНУВС, 2005. Т4. № 3. С.163-166.

4. Бабій А.С., Зацеркляний М. М.. Аналіз тенденцій розвитку злочинності. Системи обробки інформації. Харків: ХУПС, 2007. №4. С. 153-155.

5. Зацеркляний М.М., Бабій А.С. Інформаційна система моделювання впливу чинників злочинності. Право і Безпека. Харків: ХНУВС, 2008. Т.7. № 2. С. 204-209.

6. Зацеркляний М. М., Бабій А.С. Попередній аналіз даних у системах обробки інформації про скоєні злочини. Право і Безпека. Харків: ХНУВС, 2009. № 1. С. 269-272.

7. Лановий О.Ф., Бабій А.С. Статистичний аналіз злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2006. №19. С. 24 – 30

8. Бабій А.С. Програмна система для аналізу злочинності. Вісник НТУ ХПІ, Харків: НТУ «ХПІ», 2007. №19. С. 12 – 16

9. Бабій А.С. Автоматизація управління діяльністю правоохоронних органів. Державне управління та місцеве самоврядування: тези VII

міжнародного наукового конгресу, 29-30 березня 2007 р. Харків:НАДУ, 2007. С. 20-22

10. Єрохін А.Л., Бабій А.С.,Турута О.П. Спеціальна інформаційна система для виклику екстрених служб в Україні. ХНУРЕ, Збірник праць IV міжнародної науково-практичної конференції «Наука і соціальні проблеми суспільства: інформатизація і інформаційні технології», 24-25 травня 2011. Харків: ХНУРЕ, 2011. С. 163

11. Бабій А.С. Побудова СППР для оцінювання злочинності. Збірник праць II міжнародної науково-технічної конференції «Інформаційні технології в навігації і управлінні», 16-17 липня 2011 р., Київ: «ДП ЦНДІ НіУ», 2011. С. 41

12. Зацеркляний М.М., Бабій А.С. Застосування методу найменших кутів для аналізу чинників злочинності. Материалы международной научно-технической конференции “Информационные системы и технологии”, Харьков: НТМТ, 2012, С. 37

13. Петров К.Е., Зацеркляний М.М., Бабій А.С. Оцінювання злочинності із врахуванням нечіткості. Спеціальна техніка у правоохоронній діяльності, Матеріали V Міжнародної науково-практичної конференції, Київ: НАВС, 2012, С.79

14. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta .Usage of F-transform to finding informative parameters of rhinomanometric signals. Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), 2015 Xth International. P. 129-132, DOI:10.1109/STC-CSIT.2015.7325449

15. A. Yerokhin ,A. Nechyporenko, A. Babii, O. Turuta. Processing and analysis of rhinomanometric signals by F-transform approximation - 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). P. 314 - 317, DOI: 10.1109/DSMP.2016.7583566

16. A. Yerokhin, O. Turuta, A. Babii, A, Nechyporenko, I. Mahdalina. Usage of phase space diagram to finding significant features of rhinomanometric signals. 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT), P. 70 - 72

А.2 Акти впровадження результатів дисертаційної роботи

«ЗАТВЕРДЖУЮ»
В.о. проректора з науково-методичної роботи Харківського національного університету радіоелектроніки

_____ Рубан І.В.

«15» 03 2017 р.

АКТ
про впровадження в навчальний процес результатів дисертаційної роботи на тему:«Моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей» аспіранта кафедри програмної інженерії Харківського національного університету радіоелектроніки
Бабія Андрія Степановича

Комісія у складі:
Голови: завідувача кафедри програмної інженерії к.т.н., проф. Дудар З.В.
Членів комісії: заступника начальника відділу організації методичної роботи ХНУРЕ, к.т.н., доц. Шубін І.Ю., доцента кафедри програмної інженерії, к.т.н., доц. Турути О.П.

встановила, що результати наукових досліджень реалізовано в навчальному процесі Харківського національного університету радіоелектроніки на кафедрі програмної інженерії (протокол засідання кафедри ПІ №12 від 07.02.2017).

Розглянувши матеріали роботи та організацію навчального процесу на кафедрі комісія відзначає, що при проведенні лекційних занять та лабораторних робіт з курсу «Теорія ймовірностей і математична статистика та емпіричні методи програмної інженерії» використані такі результати дисертаційної роботи:

- метод визначення значущих чинників нечіткої регресійної моделі, який заснований на використанні методу найменших кутів;
- метод фільтрації компонент динамічного ряду, в якому використано згладжування на основі F-перетворення.

Завідувач кафедри програмної інженерії З.В. Дудар

Заст. нач. відділу організації методичної роботи І.Ю. Шубін

Доцент кафедри програмної інженерії О.П. Турута

ЗАТВЕРДЖУЮ
Начальник управління
інформаційної підтримки та
координації поліції 102
ГУНП в Харківській області
полковник поліції Д.Ю. Узлов

"20" грудня 2016 р.

АКТ

про впровадження результатів дисертаційної роботи
«Моделі, методи та інтелектуальна інформаційна технологія
аналізу неоднорідних послідовностей»
старшого викладача кафедри програмної інженерії
Харківського національного університету радіоелектроніки
Бабія Андрія Степановича

В період з 13.12.2016 по 17.12.2016 комісія у складі:
Голови - заступника начальника УІПКП 102, Власова О.В.
та членів комісії - начальника відділу , Григоровича О.Б.
- головного спеціаліста, Боровика Р.В.

розглянула результати використання матеріалів кандидатської дисертаційної роботи в діяльність управління інформаційної підтримки та координації поліції 102 ГУНП в Харківській області по аналізу динамічних рядів показників злочинності.

Запропонована тренд-сезонна модель динамічного ряду злочинності та метод згладжування на основі F-перетворення на етапі виявлення тренду може застосовуватися для аналізу сезонної компоненти показників скоєних злочинів, що дозволяє позбутися крайових ефектів. Також запропонований метод побудови моделей злочинності із виділенням значимих соціально-економічних факторів з застосуванням теорії нечітких множин та врахуванням експертних оцінок. Це дозволяє збільшити кількість джерел даних, за рахунок використання відомостей, які можуть бути представлені у вигляді нечітких змінних.

Комісія підтверджує доцільність впровадження запропонованих матеріалів в діяльність управління інформаційної підтримки та координації поліції 102 ГУНП в Харківській області.

Даний акт не дає підстав для будь-яких фінансових розрахунків, винагороди.

Голова комісії Власов О.В.
члени комісії Григорович О.Б.
Боровик Р.В.

ТОВ «Ендейвер»
36011, м. Полтава,
вул. Пушкіна, 28, оф.4
тел./факс +38(0532)569987
e-mail: office@endv.com.ua

«Endevour» LLC
36011, Poltava city
office 4, 28 Pushkin street
tel/fax +38(0532)569987
e-mail: office@endv.com.ua

№ 03/14-01
від 14.03.2017 р.

ЗАТВЕРДЖУЮ
Директор ТОВ "Ендейвер"
"14" березня 2017 р.

АКТ

про впровадження результатів дисертаційної роботи
«Моделі, методи та інтелектуальна інформаційна технологія аналізу неоднорідних послідовностей»
аспіранта кафедри програмної інженерії
Харківського національного університету радіоелектроніки
Бабія Андрія Степановича

Комісія у складі:
голови - директор, Піддубний Дмитро Ігорович
та членів комісії - начальник лабораторії НК, Гоголь Микола Миколайович
- фінансовий директор, Василів Людмила Василівна

розглянула результати використання матеріалів кандидатської дисертаційної роботи Бабія А.С. в рамках розв'язання задач, що пов'язані з аналізом даних про показники роботи бригад при проведенні установки ізоляційних мостів в свердловині.

Дисертаційне дослідження спрямоване на підвищення ефективності обробки відомостей отриманих з джерел різнорідної природи походження. Розроблені та всебічно досліджені в дисертаційній роботі моделі та методи, зокрема метод визначення значущих чинників при побудові нечіткої регресійної моделі, створюють достатні передумови для розширення діапазону джерел даних, що застосовуються для аналізу інформації про показники роботи бригад.

Комісія підтверджує впровадження інформаційної технології реалізованої у вигляді програмного засобу для аналізу показників роботи бригад, що дозволило зменшити рівень похибки оцінювання характеристик виконаної роботи на 5%.

Даний акт не дає підстав для будь-яких фінансових розрахунків, винагороди.

Голова комісії Піддубний Д. І.

члени комісії Гоголь М.М.

Василів Л.В.