УДК 519.95 : 330.46

Грицюк П. М.,
доктор економічних наук Національного університету водного господарства та природокористування, м. Рівне,
Савич В. О.,
кандидат фізико-математичних наук Національного університету водного господарства та природокористування, м. Рівне

МАТЕМАТИЧНІ МЕТОДИ ОЦІНЮВАННЯ РОЗМІРНОСТІ СИСТЕМИ ЗЕРНОВИРОБНИЦТВА УКРАЇНИ

Розглянуті основні підходи до оцінювання розмірності системи зерновиробництва України.
Ключові слова: *часовий ряд, динамічна система, фазовий простір, розмірність вкладення.*

Рассмотрены основные подходы к оцениванию размерности системы зернопроизводства Украины.
Ключевые слова: *временной ряд, динамическая система, фазовое пространство, размерность вложения.*

Basic approaches to the estimation of dimension the grain production system of Ukraine are examined.
Keywords: *time series, dynamic system, phase space, embedding dimension.*

Постановка проблеми. Для ідентифікації складних систем, таких як метеорологічні, економічні, системи зерновиробництва, необхідно аналізувати обширну ретроспективну інформацію, отриману в результаті функціонування системи. Виміри досліджуваних параметрів проводять через фіксований інтервал часу T і представляють як часові ряди виду $x_1, x_2, \ldots x_n$. Основною метою аналізу часових рядів є отримання інформації про системи, що їх породили, і, насамперед, оцінка найголовнішої характеристики системи – її розмірності. Під розмірністю традиційно розуміють мінімальну кількість незалежних змінних, яка дозволяє однозначно описати установлену динаміку дисипативної системи. Відомо, що при розмірності, яка не перевищує 5–6, добре працюють різноманітні методи аналізу й прогнозу еволюції системи. При більших розмірностях застосування відповідних методів стає проблематичним [1; 2]. Зазвичай явний вигляд математичної моделі системи є невідомим, тому необхідно мати можливість оцінювати розмірність системи безпосередньо за спостереженнями. На сьогодні існують два підходи до дослідження часових рядів: статистичний (імовірнісний) та динамічний (теорія Такенса-Мане).

Аналіз останніх досліджень та публікацій. Система зерновиробництва належить до класу відкритих нелінійних систем, які зазвичай називають дисипативними. Такі системи є нерівноважними завдяки розсіюванню енергії, отримуваної ззовні. Для описання еволюції дисипативних систем вводять поняття фазової траєкторії. Фазова траєкторія – це крива у фазовому просторі, що відображає еволюцію системи. Сукупність фазових траєкторій, побудованих для різних початкових умов, складає фазовий портрет системи.

У більшості дисипативних систем сукупність фазових траєкторій притягується до деякої скінченновимірної підмножини фазового простору, називаної атрактором. Атрактор, який можна описати за допомогою скінченного набору деяких змінних, визначає властивості усталеного з часом коливного процесу. Коливання у системі зерновиробництва зумовлені коливаннями врожайності, причиною яких є циклічні зміни кліматичних умов. Таким чином, вивчення динаміки вихідної системи можна звести до вивчення її атрактора.

При аналізі часових рядів головною задачею є реконструкція динамічної системи, яка породила цей ряд. Згідно з теорією Такенса-Мане, задовільну геометричну картину атрактора системи можна отримати, якщо замість змінних, які характеризують поведінку системи (але значення яких нам невідомі), використати так звані “вектори затримок” $z_i = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}$ для одного виміряного параметра. Вперше цей підхід до аналізу часових рядів був математично обґрунтований у роботах Такенса [3]. Теорія Такенса дозволяє встановити динамічний характер системи. Систему в деякому наближенні можна вважати динамічною, якщо її розмірність не перевищує деякого критичного значення (зазвичай вибирають $D = 6$). Розмірність системи часто зіставляють із розмірністю вкладення атрактора – мінімальною розмірністю фазового простору, в який без самоперетинів може бути поміщений гладкий многовид, який цілком містить цей атрактор. Сьогодні найбільш популярним алгоритмом оцінювання розмірності вкладення D є алгоритм Грассбергера-Прокаччі [4]. Проте він є неефективним при роботі з короткими (до 104 значень) часовими рядами. Використовують і інші методи оцінювання розмірності систем: функціональний метод [2], метод “фальшивих найближчих сусідів” [5], метод власних значень [6]. Метою цієї роботи є визначення ефективності вказаних методів при роботі з короткими зашумленими часовими рядами врожайності.

Мета і завдання дослідження. Об'єктом дослідження у нашій роботі є ряд врожайності зернових в Україні за 1950–2012 роки [7]. Дослідження автора [8] дозволяють віднести систему зерновиробництва до хаотичних динамічних систем. При цьому варто мати на увазі, що поведінка подібних систем визначається складним комплексом чинників як динамічної, так і стохастичної природи.

Виклад основного матеріалу. Згідно з теоремою Такенса, низка спостережень однієї змінної містить інформацію про поведінку системи в цілому. Основою досліджень динаміки системи є фазова траєкторія, відтворена з одновимірного часового ряду. Нехай відомо, що стан системи повністю описується *D* змінними: $x_1(t), x_2(t),\ldots,x_D(t)$. З інтервалом *T* проводять вимірювання довільної змінної – наприклад $x_1(t)$. Тоді замість послідовності, яка складається з *D* величин

$x_1(t), x_2(t),\ldots,x_D(t)$. (1)

можна розглядати таку:

$$x_1(t),\ x_1(t+T),\ x_1(t+2T),\ldots,x_1(t+(D-1)\cdot T). \tag{2}$$

Тобто в кожний момент часу стан всієї системи може бути описаний *D* значеннями однієї змінної, взятими із зсувом *T*. Відновлення фазового портрета системи, яка має розмірність *D*, за низкою спостережень однієї змінної $\{x_i\}$ полягає у формуванні послідовності фазових векторів $\{y_i\}$ за таким правилом:

$$y_j=(x_j,x_{j+1},x_{j+2},\ldots x_{j+D-1})^T. \tag{3}$$

Тут знак *T* означає транспонування вектора-рядка у вектор-стовпець.

Методи досліджень. Критично важливим для адекватного моделювання системи та прогнозування її еволюції є встановлення істинного значення її розмірності *D*. Для визначення розмірності системи за наявним часовим рядом використаємо декілька різних методів.

Перший метод ґрунтується на виявленні існування функціональної залежності [2]. Якщо розмірність системи дорівнює *D*, то це означає, що у часовому ряду $\{x_i\}$ кожне значення x_k є функцією вектора фазового простору, побудованого за *D* попередніми значеннями ряду:

$x_k=f(y_k)$ де $y_k=(x_{k-1},x_{k-2},\ldots,x_{k-D})$.

Нехай шукана функція є неперервною. Тоді, згідно з означенням неперервності, отримаємо: якщо $|y_k-y_n|\to 0$, то і $|x_k-x_n|\to 0$. Відстань між векторами будемо розглядати як евклідову, тобто:

$$|y_k-y_n|=\sqrt{\frac{1}{D}\sum_{i=1}^{D}(x_{k-i}-x_{n-i})^2}. \tag{4}$$

Тоді алгоритм оцінювання розмірності системи за часовим рядом, що містить *N* значень, буде таким.

1. Приймаємо *D = 1*.

2. Для всіх *k* (від *D* до *N*) виконуємо таке:

2.1. Формуємо вектор $y_k=(x_{k-1},x_{k-2},\ldots,x_{k-D})^T$.

2.2. Знаходимо вектор $y_n=(x_{n-1},x_{n-2},\ldots,x_{n-D})^T$ (*n* в межах від *D* до *N* і $n\neq k$) такий, що відстань $|y_k-y_n|$ є мінімальною для всіх можливих *n*. Позначаємо цю відстань ρ_k.

2.3. Визначаємо величину $r_k=|x_k-x_n|$.

3. Будуємо графік залежності r_k від ρ_k.

4. Повторюємо кроки 2–3 для інших послідовних значень *D* (2,3,4,5,...).

Якщо в даних існує деяка функціональна залежність, то точки графіка, який будується на кроці 3, будуть групуватися поблизу початку координат. Починаючи з деякого значення розмірності *D* розкид точок практично перестає змінюватися. Цю величину можна прийняти за розмірність системи.

Розрахунки, виконані методом функціональної залежності, дозволяють оцінити значення розмірності як *D = 5*. Справді, як видно з рис.1, при переході від значення *D = 4* до значення *D = 5* розкид точок значно зменшується, після чого практично не змінюється. Зауважимо, що для підвищення якості оцінок, рекомендується працювати з нормованими рядами. Ми використовували стандартний спосіб нормування початкового ряду

$$\tilde{x}_i=\frac{x_i-\bar{x}}{\sigma_x}. \tag{5}$$

Другий метод оцінювання розмірності є одним з найбільш поширених. Він був запропонований на початку 80-х Грассбергером і Прокаччі [4]. Метод полягає в обчисленні для декількох послідовних векторів розмірності *D* кореляційного інтеграла $C_D(r)$:

$$C_D(r)=\frac{1}{N^2}\sum_{i,j=1}^{N}\Theta(r-|y_i-y_j|). \tag{6}$$

Тут $\Theta(x)$ – функція Хевісайда, яка дорівнює 1 при *x > 0* і дорівнює 0 у всіх інших випадках, *N* – кількість векторів, сформованих на основі досліджуваного ряду. Іншими словами, кореляційний інтеграл $C_D(r)$ дорівнює кількості пар векторів розмірності D, відстань між якими не перевищує *r*, щодо загальної кількості пар векторів. Показано, що при малих *r* кореляційний інтеграл веде себе як степенева функція:

$$C_D(r) = Ar^{d_D}, \tag{7}$$

тобто він є пропорційним до деякого степеня (залежного від розмірності, для якої проводилися обчислення).

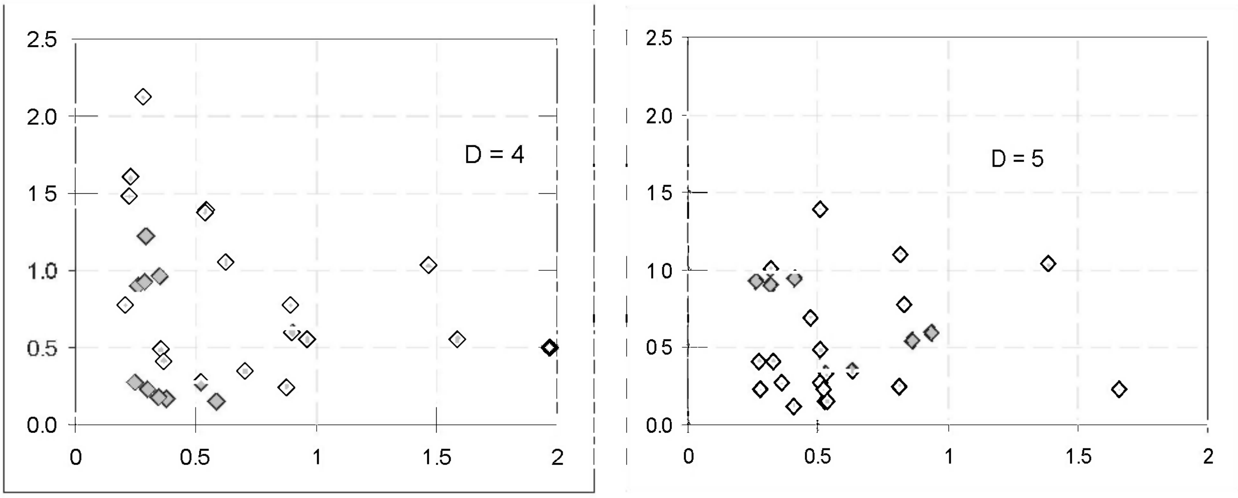

Рис. 1. Оцінка розмірності системи методом функціональної залежності.

У двічі логарифмічних координатах (тобто у випадку, коли по осі абсцис відкладається $\ln r$, а по осі ординат $\ln C_D$) графік повинен мати лінійну ділянку. Із збільшенням розмірності D нахил цієї ділянки також збільшується. Починаючи з деякого значення D показник степеня d_D перестає (в межах точності) змінюватися. Це значення D можна прийняти за розмірність системи. Величина d_D називається кореляційною розмірністю і також є важливою характеристикою хаотичних атракторів. Для отримання якісних оцінок за методом Грасбергера-Прокаччі необхідно мати ряд великої довжини. Про це свідчить порівняння результатів дослідження ряду врожайності зернових України (63 значення, рис. 2, верхній лівий рисунок) та ряду врожайності озимої пшениці США (125 значень, рис. 2, верхній правий рисунок). Досягнути кращої якості при малій довжині ряду дозволяє введення нової характеристики близькості фазових векторів – кутової відстані. Ця характеристика дозволяє виявити подібні (гомогенні) вектори для нестаціонарних рядів і, таким чином, збільшити кількість "найближчих сусідів", що сприяє підвищенню якості оцінок (рис. 2, нижній рисунок). Кутова відстань d_φ між двома векторами фазового простору $\{x_{n-D+1}, x_{n-D+2}, \ldots, x_n\}$ та $\{x_{k-D+1}, x_{k-D+2}, \ldots, x_k\}$ визначається за формулою

$$d_\varphi = \sqrt{1-\cos^2\varphi_k}, \quad \cos\varphi_k = \frac{\sum_{i=1}^{D} x_{n-D+i}\cdot x_{k-D+i}}{\sqrt{\sum_{i=1}^{D} x_{n-D+i}^2 \cdot \sum_{i=1}^{D} x_{k-D+i}^2}}. \tag{8}$$

Аналіз проведених розрахунків дозволяє дійти висновку: розмірність системи зерновиробництва в Україні можна оцінити як D = 5 (суцільна лінія на рис. 2).

Третій метод оцінювання розмірності іноді називають методом "фальшивих сусідів" [5]. Ідею методу може пояснити такий приклад. Нехай задана просторова тривимірна лінія. Виберемо на ній дві точки, відстань між якими є достатньо великою.

Водночас може виявитися, що відстань між проекціями цих точок на одну із координатних площин є дуже малою. Такі точки називають "фальшивими сусідами". Тоді задачу оцінки розмірності системи можна сформулювати так. Необхідно підібрати таке число D, щоб при реконструкції фазового простору розмірності D не існувало близьких сусідів таких, щоб при переході до розмірності D+1 відстань між ними значно збільшується. Для оцінок найчастіше використовують евклідове визначення відстані. Ми використали таку методику розрахунків. Перш за все шляхом комп'ютерних експериментів ми підібрали оптимальну кількість точок – найближчих сусідів, при якій ефекти оцінки розмірності проявляються найбільш чітко. Було встановлено, що доцільно досліджувати 100 пар точок – найближчих сусідів. При розмірності D було вибрано 100 пар векторів відстань між якими найменша. Потім здійснювався перехід до розмірності D+1 і розраховувалася відстань між тими ж парами векторів. Визначався коефіцієнт збільшення відстані для кожної пари та середній коефіцієнт збільшення k. Результати розрахунку методом "фальшивих сусідів" подані на рис. 3. Як видно з рисунка, при переході від розмірності $D = 5$ до розмірності $D = 6$ середнє значення відстані між найближчими сусідами практично не змінюються. Це дозволяє дійти висновку, що шукана розмірність $D = 5$.

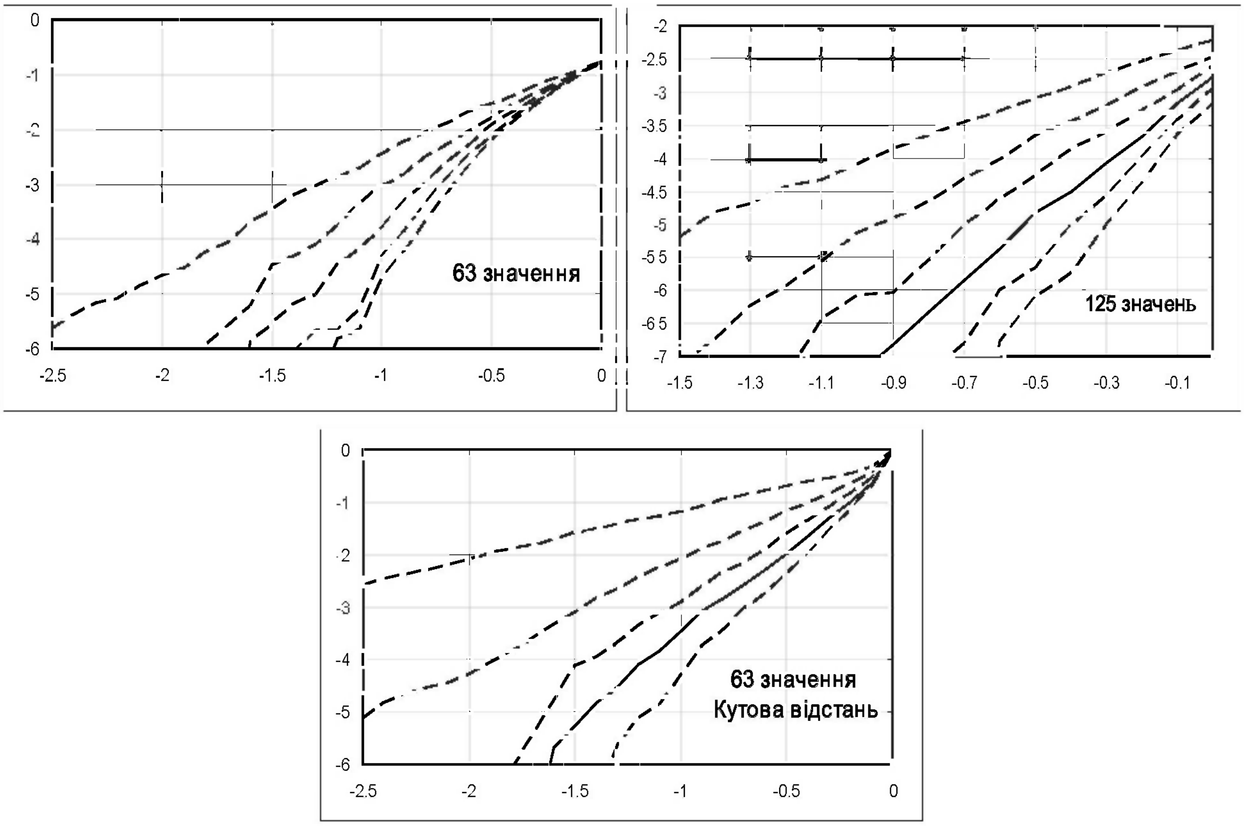

Рис. 2. Дослідження розмірності системи методом Грассбергера-Прокаччі

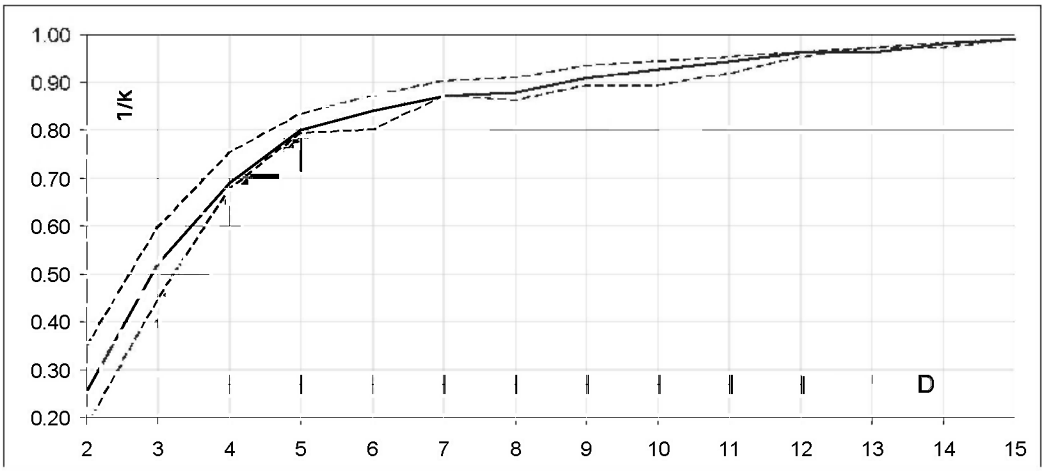

Рис. 3. Визначення розмірності вкладення методом "фальшивих сусідів".
Нижня лінія: кількість сусідів KS=50, середня лінія KS=100, верхня лінія KS=200.

Четвертий метод [6] відрізняється від розглянутих вище. Для його реалізації спочатку приведемо ряд до нульової середньої, тобто проведемо заміну

$$x_i \to x_i - \frac{1}{N}\sum_{j=1}^{N} x_j \quad (9)$$

Після цього сформуємо із отриманого ряду матрицю спостережень M за таким правилом:

$$M = \begin{bmatrix} x_1 & x_2 & \dots & x_D \\ x_2 & x_3 & \dots & x_{D+1} \\ \dots & \dots & \dots & \dots \\ x_{N-D+1} & x_{N-D+2} & \dots & x_N \end{bmatrix} \quad (10)$$

У рядках цієї матриці стоять компоненти реконструйованих за часовим рядом векторів. Розмірність реконструкції ($D = 10$) виберемо завідомо більшою від очікуваної розмірності системи. На основі матриці M сформуємо матрицю коваріації C:

$$C = M^T M \quad (11)$$

і обчислимо її власні значення $\lambda_1, \lambda_2, \ldots, \lambda_D$. Як правило, в результаті декілька власних значень будуть помітно більшими від інших, які є приблизно однаковими. При деякому номері власного значення величини власних значень стабілізуються ("виходять на плато"). Цей номер і буде розмірністю системи. Результати розрахунків методом власних значень подані на рис. 4.

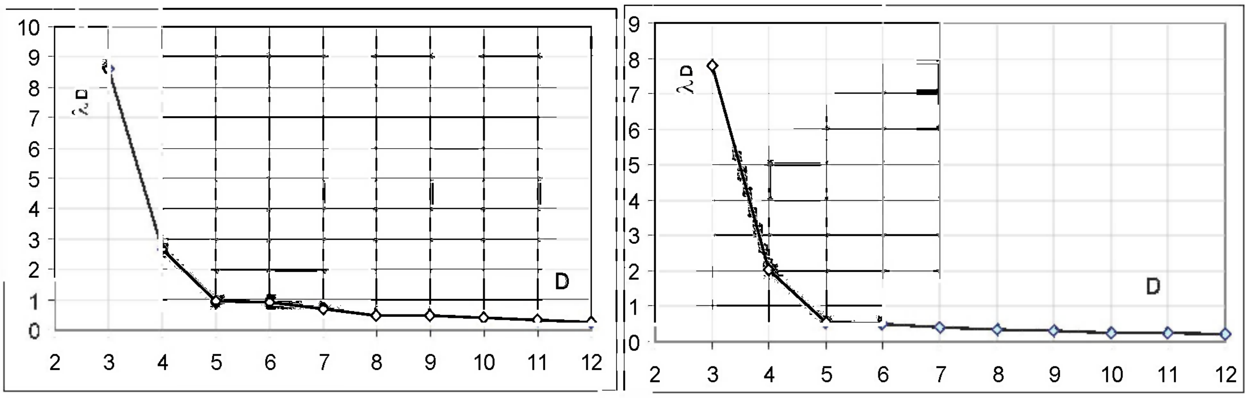

Рис. 4. Визначення розмірності вкладення методом власних значень

Як видно з рисунка, при $D > 5$ величини власних значень стабілізуються. Отже, значення $D = 5$ і є значенням розмірності системи зерновиробництва в Україні. Лівий рисунок побудований за часовим рядом врожайності 1950–2012 років, правий містить лише частину ряду від 1950 року до 1990 року. Лівий рисунок дозволяє побачити як вплинули на розмірність системи зерновиробництва зміни, які відбулися в економіці України після 1990 року (ускладнення системи, зменшення її детермінованості) .

Висновки. У цій роботі виконано аналіз системи зерновиробництва України на предмет встановлення її розмірності. Дослідження виконувалися методами функціональної залежності, Грассбергера-Прокаччі, "фальшивих сусідів" і методом власних значень. Метод Грассбергера-Прокаччі є дуже чутливим до довжини ряду і погано працює при довжинах $N < 10^4$. Найкращі результати для коротких часових рядів, які містять елементи шуму, показали методи власних значень та "фальшивих сусідів". Всі методи дозволяють оцінити розмірність системи зерновиробництва України як $D = 5$. Це означає, що досліджувану нами систему можна повністю описати за допомогою п'яти параметрів. Щоправда вигляд моделюючої системи рівнянь при цьому залишається невідомим. Застосовану нами методику можна застосувати до аналізу інших економічних, біологічних та метеорологічних систем.

Література:

1. Лоскутов А. Ю., Михайлов А. С. Основы теории сложных систем. – М.: РХД, 2007. – 612 с.
2. Лоскутов А. Ю. Анализ временных рядов. Курс лекций. – М.: МГУ. – 113 с.
3. Takens F. Detecting strange attractors in turbulence // Dynamical Systems and Turbulence. Lect. Notes in Math. – Berlin, Springer Verlag, 1981. – Vol. 898. – Pp. 366–382.
4. Grassberger P., Procaccia I. Measuring the strangeness of strange attractors. Physica 9D, 189, 1983.
5. M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, Phys. Rev. A 45, 3403 (1992). 6. Theil, H. (1971). Principles of Econometrics. John Wiley & Sons, New York.
7. Електронний ресурс http://www.ukrstat.gov.ua/ .
8. Hrytsyuk P.M. Evidence for Low Dimensional Chaos in Grain Production System of Ukraine. Material of the International Symposium RA08, Riga-Jurmala. – Riga : RTU, 2008. – P. 34–37.