Статистичний аналіз активності тематичного контенту в мережі Інтернет для прогнозування розвитку інформаційних загроз

Statistical Analysis of the Activity of the Thematic Content on the Internet for Predicting the Development of Information Threats

Олексій Писарчук [1], **Олександр Лагодний** [1], **Юрій Міхєєв** [1]
Oleksii Pysarchuk, Oleksandr Lahodnyi, Yurii Mikhieiev

[1] *Zhytomyr Military Institute named after S. P. Koroljov*
22 Prospekt Myru, Zhytomyr, 10004, Ukraine

DOI: 10.22178/pos.25-2

LCC Subject Category: QA76.75-76.765

Received 14.07.2017
Accepted 10.08.2017
Published online 16.08.2017

Corresponding Author:
Oleksandr Lahodnyi
lov.82@ukr.net

Анотація. У статті наведено результати статистичного аналізу активності тематичного контенту в мережі Інтернет на сегменті реальних експериментальних даних інформаційних повідомлень. Проведений аналіз довів можливість розгляду частоти появи тематичного контенту, як монотонного міандрового процесу з адитивною випадковою складовою Результати аналізу забезпечили визначення виду залежності частоти появи тематичного контенту від часу, закону розподілу випадкової складової та отримати його статистичні характеристики. Отримані результати доцільно застосовувати на етапі прогнозування розвитку інформаційних загроз. Результати статистичного аналізу показали, що частота активності тематичного контенту в мережі Інтернет має нелінійний характер, є випадковим стаціонарним процесом з явно вираженим зростаючим або спадаючим міандровим трендом.

Ключові слова: контент; мережа Інтернет; статистичний аналіз; прогнозування; закон розподілу.

Abstract. The article presents the results of the statistical analysis of the thematic content of the Internet in the segment of real experimental data of information messages. The analysis proved the possibility of considering the frequency of appearance of thematic content as a monotone meander process with an additive random component. The results of the analysis provided the definition of the type of dependence of the frequency of thematic content occurrence from time, the law of the distribution and of the random component, and obtain its statistical characteristics. The obtained results should be used at the stage of forecasting the development of information threats. The results of the statistical analysis showed that the thematic content frequency on the Internet is nonlinear, is a random stationary process with a clearly pronounced rising or declining meander trend.

Keywords: content; internet; statistical analysis; prognostication; distribution law.

ВСТУП

Практика останніх локальних війн і збройних конфліктів, зокрема які відбулися в Грузії, Україні, Сирії, стан дипломатичних відносин, політичних рішень та виборів в Франції, США показують важливість інформаційної складової із залученням ресурсів мережі Інтернет, що негативно впливає на стан національної безпеки держави [7, 6, 3]. Більшість підходів щодо виявлення інформаційно-психологічного впливу (ІПсВ) у мережі Інтернет базуються на релевантному та семантичному аналізі тексту, в ході яких оцінюється тільки якісна складова контенту, в якому розміщена деструктивна інформація.

Зростання домінуючого контенту в мережі Інтернет становить загрозу для інформаційної безпеки держави і потребує невідкладних

рішень щодо протидії. Сучасні методи та способи ведення інформаційних війн вимагають адекватних контрзаходів протидії інформаційним загрозам (ІЗ), які можливо реалізувати за допомогою своєчасного виявлення та прогнозування розвитку ІЗ. Прогнозування активності тематичного контенту в мережі Інтернет дає вихідні дані на прийняття рішення, що є важливим заходом у реалізації протидії ІПсВ. Основними заходами з протидії ІЗ в мережі Інтернет (рис. 1) мають бути:

- моніторинг мережі Інтернет – процес збору статистичних даних контенту в обраній сфері життєдіяльності соціуму за обраною тематикою з ефективних Інтернет-сайтів цільового спрямування [5];

- виявлення ІЗ – процес розпізнавання в контенті ІЗ за сформованими ознаками [4];

- прогнозування розвитку ІЗ – застосування математичних операцій, моделей та методів апроксимації, екстраполяції часових рядів щодо побудови адекватних моделей опису досліджуваного процесу для кращої точності прогнозу.

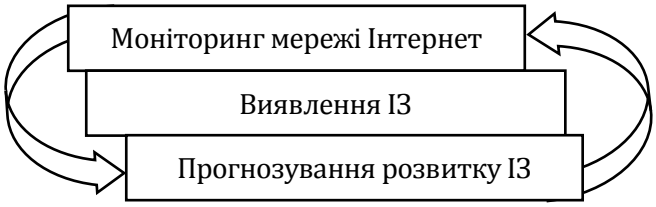

Рисунок 1 – Заходи протидії інформаційним загрозам у мережі Інтернет

Рівень ІЗ пропонується оцінювати через величину активності тематичного контенту в мережі Інтернет за частотою його появи в дискретні проміжки часу у відкритих джерелах інформації (ВДІ) – форумах, блогах, сайтах соціальних мереж, інформаційних сайтах. Це дає змогу отримувати кількісні характеристики досліджуваного процесу, але залежність частоти появи тематичного контенту в мережі Інтернет від часу має невідомі характеристики і є невідомим процесом. Тоді виникає необхідність у проведення статистичного аналізу активності тематичного контенту в мережі Інтернет для: визначення виду процесу; знаходження закону розподілу випадкової складової та його числових характеристик. Надалі отримані залежності та характеристики будуть вихідними даними для прогнозування активності тематичного контенту в мережі Інтернет з метою розрахунку сил та засобів протидії ІЗ.

Тому актуальним завданням є здійснення статистичного аналізу активності тематичного контенту під час проведення противником інформаційної війни в мережі Інтернет.

Аналіз досліджень і публікацій свідчить про те, що питанням статистичного аналізу загалом приділена увага в працях як вітчизняних так і закордонних науковців [10, 2, 12, 11, 1, 8]. Статистичний аналіз широко використовують у економічному аналізі, маркетинговій справі, соціальній, медичній та інших сферах життєдіяльності. Аналіз джерел показує факт застосування статистичного згладжування, але не доведено, що процес дійсно випадковий з реальними числовими характеристиками. Похибки прогнозування активності тематичного контенту є наслідком відсутності проведення статистичного аналізу і в свою чергу неможливістю адекватного вибору закону і моделі зміни досліджуваного процесу та відповідного алгоритму згладжування. Таким чином, на даний час існує потреба у застосуванні відповідного математичного апарату до активності тематичного контенту в мережі Інтернет для пошуку закономірностей, тенденцій, кількісних характеристик процесу. Після виявлення закону розподілу та його числових характеристик можливо проводити прогнозування розвитку активності тематичного контенту в мережі Інтернет, який завдяки статистичному аналізу буде мати адекватні та кращі прогностичні дані.

Тому *метою статті* є проведення статистичного аналізу активності тематичного контенту в мережі Інтернет для прогнозування розвитку інформаційних загроз.

РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ

Планування проведення інформаційних війн в мережі Інтернет передбачає попереднє проведення інформаційно-психологічних атак та інформаційно-психологічних акцій в критичних сферах життєдіяльності соціуму. Часові інтервали даних заходів можуть тривати від одного дня до місяця. Активність тематичного контенту в мережі Інтернет пов’язана з частотою появи інформаційних повідомлень, яка залежить від таких факторів:

- одночасна поява інформаційних повідомлень у всіх сферах життєдіяльності соціуму;

- почергова поява інформаційних повідомлень у критичних сферах життєдіяльності соціуму;

- інформаційні повідомлення можуть розміщуватися на замовлення;

- активність появи інформаційних повідомлень зростає після виступу авторитетних осіб, політиків;

- цілеспрямовано розміщувати інформаційні повідомлення можуть підрозділи інформаційно-психологічних операцій.

Необхідно на перших етапах підготовки проведенню інформаційної війни проводити прогнозування розвитку ІЗ з метою подальшої протидії. Для забезпечення даного завдання необхідно провести статистичний аналіз активності тематичного контенту в мережі Інтернет щоб виявити основні закономірності та тенденції процесу в умовах обмеженої кількості інформації.

Таким чином в результаті моніторингу ВДІ в мережі Інтернет отримуємо рівноточну і рівнодискретну вибірку параметрів процесу частоти активності тематичного контенту (1):

$$\overline{Y} = |y_i|^T, i = 1, 2, \dots n, \quad (1)$$

де y_i – значення кількості контенту за обраною тематикою;

$1, 2, \dots, n$ – дискретні проміжки часу в які було отримано значення.

Вимірювання проводяться в мережі Інтернет за пошуковою фразою в дискретні проміжку часу і фіксуються у відповідній базі даних. Дослідження закономірностей контенту було проведено на основі отриманих параметрів досліджуваного процесу в головних, на нашу думку, сферах життєдіяльності соціуму (економічна, політична, екологічна, військова, релігійна, соціальна) на зрізі 20 новин з ІЗ в кожній сфері і по 220 точок (дискрет) в них. На (рис. 2) наведено графік залежності частоти появи контенту однієї з новин «Економічне блокування території Криму» у тематиці «Політичні новини».

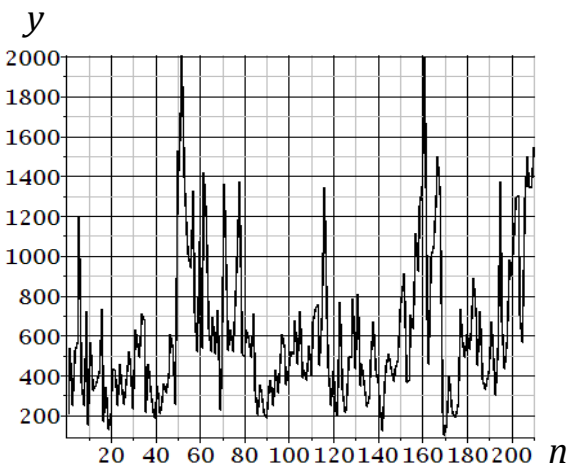

Рисунок 2 – Залежність значень частоти появи контенту в мережі Інтернет

За виглядом кривої, яка описує досліджуваний процес активності тематичного контенту в мережі Інтернет, можна стверджувати, що це стохастичний процес. Аналіз решти новин після опрацювання масиву даних, який становив у сумі 26400 значень показав, що в кожній з них присутній зростаючий або спадаючий тренд, він є нелінійним і на нього накладається випадкова складова.

Висунемо гіпотезу про те, що зміна частоти появи тематичного контенту в часі – випадковий стаціонарний процес, в якому присутня систематична складова у вигляді тренду та випадкова складова у вигляді коливань навколо тренду, яка розподілена за невідомим законом і має свої ймовірнісні характеристики. Математична (статистична) модель, якою можна описати процес (рис. 2) має наступний вигляд (2):

$$y_t = f(t) + \xi_t, t = 1, 2, \dots n, \quad (2)$$

де $f(t)$ – тренд (міандрова модель);

ξ_t – випадкова складова.

Закон розподілу невідомої величини ξ_t розраховуємо після виділення тренду і тим самим наближаємо до значень y_t. Побудуємо гістограму щільності розподілу появи тематичного контенту в мережі Інтернет для даного часового ряду (рис. 3).

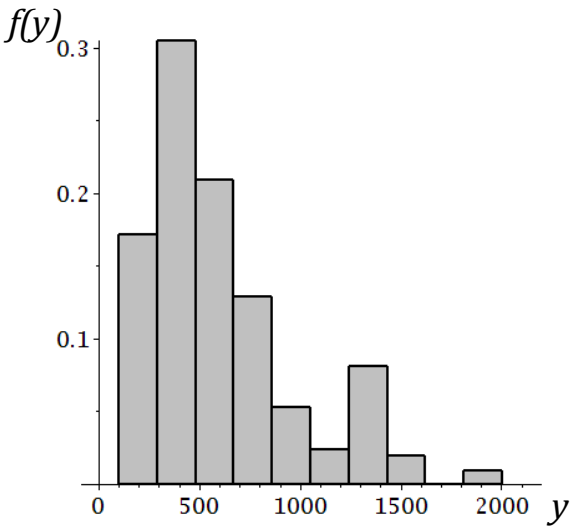

Рисунок 3 – Гістограма щільності розподілу появи тематичного контенту

За формою гістограми можливо зробити припущення про те, що в часовому ряді присутня випадкова складова, яка розподілена за логарифмічно-нормальним закон.

Здійснимо статистичний аналіз для визначення характеристик та закономірності випадкової величини, а саме: виділимо тренд; виявимо закон розподілу випадкової складової та розрахуємо характеристики; визначимо вид процесу.

Розв'язання даної задачі можливе з використанням наступного алгоритму:

Виділення систематичної складової (тренду) з часового ряду, на яку накладається випадкова складова, проводиться за рахунок апроксимації з використанням методу найменших квадратів (МНК) [9].

Розраховують характеристики закону розподілу випадкової складової з оцінок математичного сподівання m_y, дисперсії D_y, середнього квадратичного відхилення σ_y, достовірності апроксимації R^2 моделі згладжування, що виступають в якості контрольованих параметрів, вирази (3)–(6).

$$m_y = \frac{1}{n}\sum_{i=1}^{n}\Delta_i, \quad (3)$$

$$D_y = \frac{1}{n-1}\sum_{i=1}^{n}\left(\Delta_i - m_y\right)^2, \quad (4)$$

$$\sigma_y = \sqrt{D_y}, \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}, \quad (6)$$

де $\Delta_i = |y_i - \hat{y}_i|$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; y_i – виміряне значення;

$\hat{y}_i$ – оцінка виміряного значення [13].

На стаціонарність часовий ряд перевіряється завдяки знаходження статистичних характеристик: математичне сподівання (середнє) $m\{y_t\} = a = const$, дисперсія (середнє квадратичне відхилення) $D\{y_t\} = \sigma^2 = const$. Коли дані характеристики не залежать від моменту часу то це доказує стаціонарність процесу.

Знаходження закону розподілу випадкової складової реалізується за допомогою графічного відображення часового ряду у формі гістограми.

Адекватність роботи алгоритму досліджувалась при наявності у вимірах стаціонарного процесу випадкової складової з відомими характеристиками. За тестовими даними на вхід розробленого алгоритму подавалася модель з відомими параметрами:

$$y(t) = 815 - 86t + 4{,}4t^2 - 0{,}08t^3 + \\ + 7\times10^{-4}t^4 - 3\times10^{-6}t^5 + 5\times10^{-9}t^6.$$

Потім проводилось зашумлення ξ з характеристиками $m_y \cong 0$, $\sigma_y = 3$, що є контрольованими параметрами. Після накладання випадкової складової на модель ми отримали нову вибірку вимірів з наступними характеристиками $m_y \cong 0$, $\sigma_y = 76$, рис. 4а. Працездатність алгоритму розрахунків з використанням МНК підтвердилась на практичних розрахунках та побудові гістограм, де $m_y \cong 0$, $\sigma_y = 2{,}8$, рис. 4б. Дані розрахунки свідчать про те, що у статистичних даних частоти поя-

ви тематичного контенту в мережі Інтернет присутня випадкова складова, що обумовлено похибками вимірювання і виявляється після виділення тренду за допомогою розробленого алгоритму. Таким чином розроблений алгоритм можливо використати для вирішення поставлених завдань та провести статистичний аналіз активності тематичного контенту в мережі Інтернет.

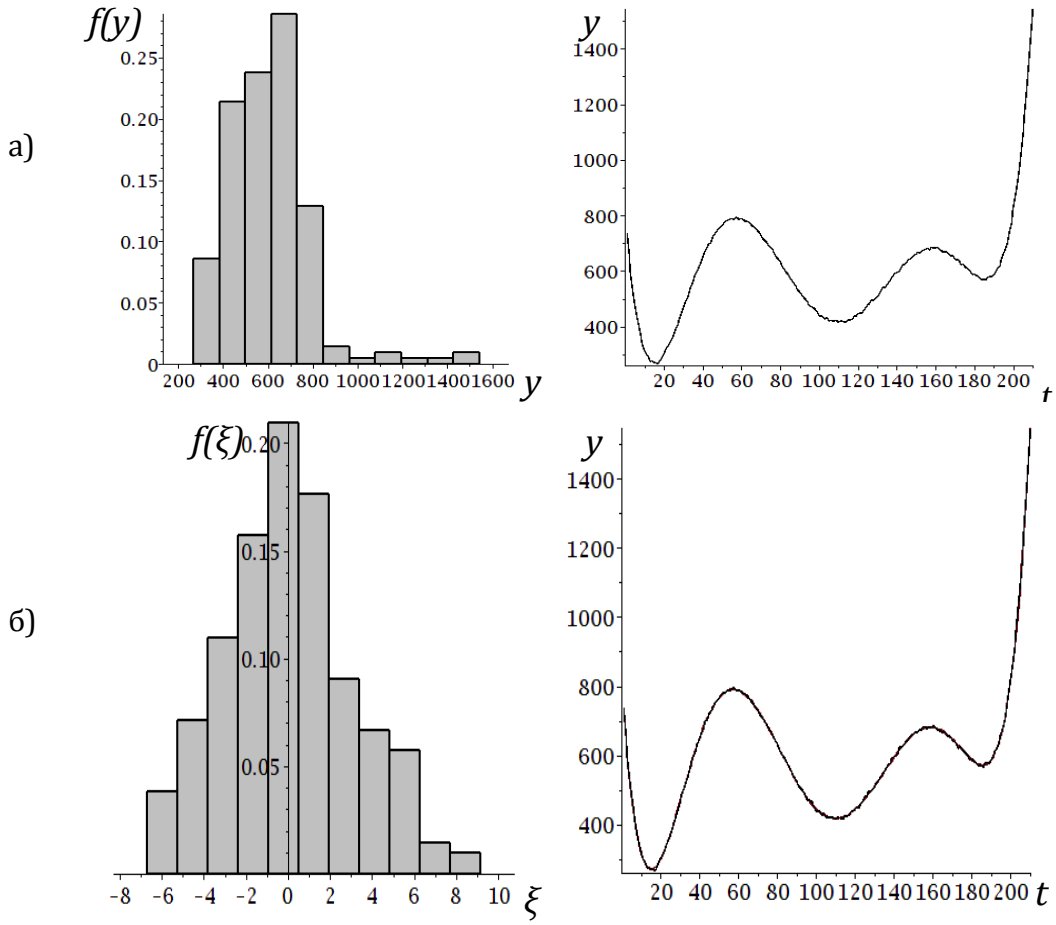

Рисунок 4 – Гістограма щільності розподілу часового ряду та розподілу випадкової складової: а) – виміряні дані до обробки; б) – випадкова складова у часовому ряді після обробки

Практичний приклад застосування

Нехай під час моніторингу мережі Інтернет та виявлення ІЗ отримано вибірку вимірів за пошуковою фразою – «Економічне блокування території Криму» за контентом К1 в тематиці «Політичні новини». Обсяг вибірки для контенту складає 210 значень за період з 14.09.2015 р. по 15.10.2015 р. (тобто 30 діб (7 вимірів на добу)) у дискретні проміжки часу з інтервалом 2 год. $t_1 = 8:00 \ldots t_7 = 20:00$.

Розрахунок характеристик закону розподілу випадкової складової проводився для знаходження параметрів часового ряду до і після обробки даних розробленим алгоритмом. Характеристики закону розподілу часового ряду на вході алгоритму мають наступні значення $m_y \cong 600{,}43$, $\sigma_y = 371{,}51$. Після проведення розрахунків за розробленим алгоритмом отримано результати закону розподілу випадкової складової, які наведені в табл. 1, де ρ – степінь полінома. З отриманих результатів табл. 1 видно, що чим більше степінь апроксимуючого полінома, тим кращі контрольовані параметри.

Виділення систематичної складової проводилося за допомогою згладжуванням часового ряду за МНК поліномами до шостого порядку включно, рис. 5. Наведені графіки дають можливість узагальнити результати розрахунків та наочно показують, що в часовому ряді присутня систематична складова (тренд) і із збільшенням порядку апроксимуючого полінома збільшується адекватність кривої, яка описує досліджуваний процес.

Таблиця 1 – Статистичні характеристики закону розподілу випадкової складової часового ряду «Економічне блокування території Криму»

Характеристики	Поліном					
	$\rho=1$	$\rho=2$	$\rho=3$	$\rho=4$	$\rho=5$	$\rho=6$
m_y	274,20	272,60	260,31	259,48	259,47	240,92
D_y	55912,16	55304,83	51150,82	51462,73	51464,63	43908,88
σ_y	236,45	235,16	226,16	226,85	226,85	209,54
R_y^2	0,05	0,06	0,13	0,13	0,13	0,26

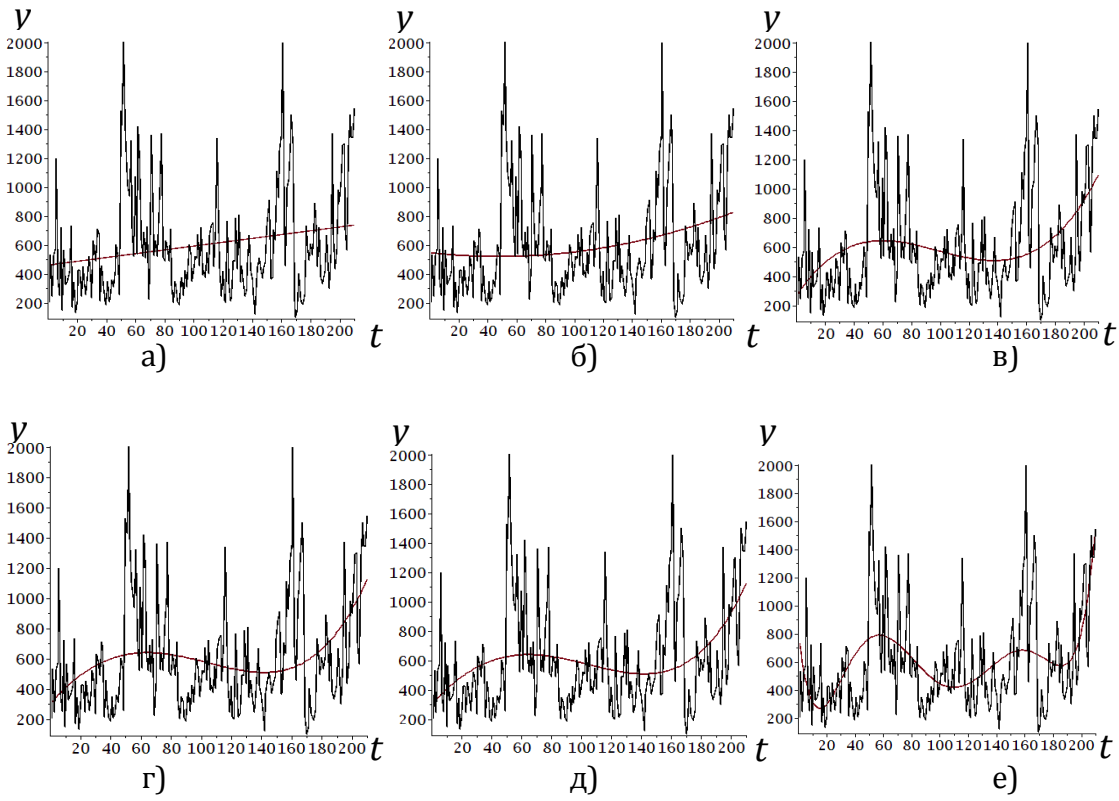

Рисунок 5 – Згладжування часового ряду поліномами:
а – лінійний; б – квадратичний; в – кубічний; г – 4-го порядку; д – 5-го порядку; е – 6-го порядку

Виявлення закону розподілу випадкової складової є наступним кроком статистичної обробки даних. Від того, якому закону розподілу підпорядкований часовий ряд залежить тактика подальшого аналізу. Якщо гістограма має єдиний чітко виражений максимум і є приблизно симетричною, то припускають, що випадкова величина розподілена за нормальним законом. Також при висуненні гіпотези про закон розподілу випадкової величини враховуються апріорні дані про закони розподілу, що притаманні обраній області дослідження.

Після згладжування часового ряду за розробленим алгоритмом отримані наступні гістограми щільності розподілу випадкової складової у статистичних даних, рис. 6. Після проведення згладжування статистичних даних за допомогою МНК закон розподілу випадкової складової є логарифмічно-нормальний, а зі збільшенням порядку полінома вигляд гістограми має виражену форму та наближається до нормального розподілу.

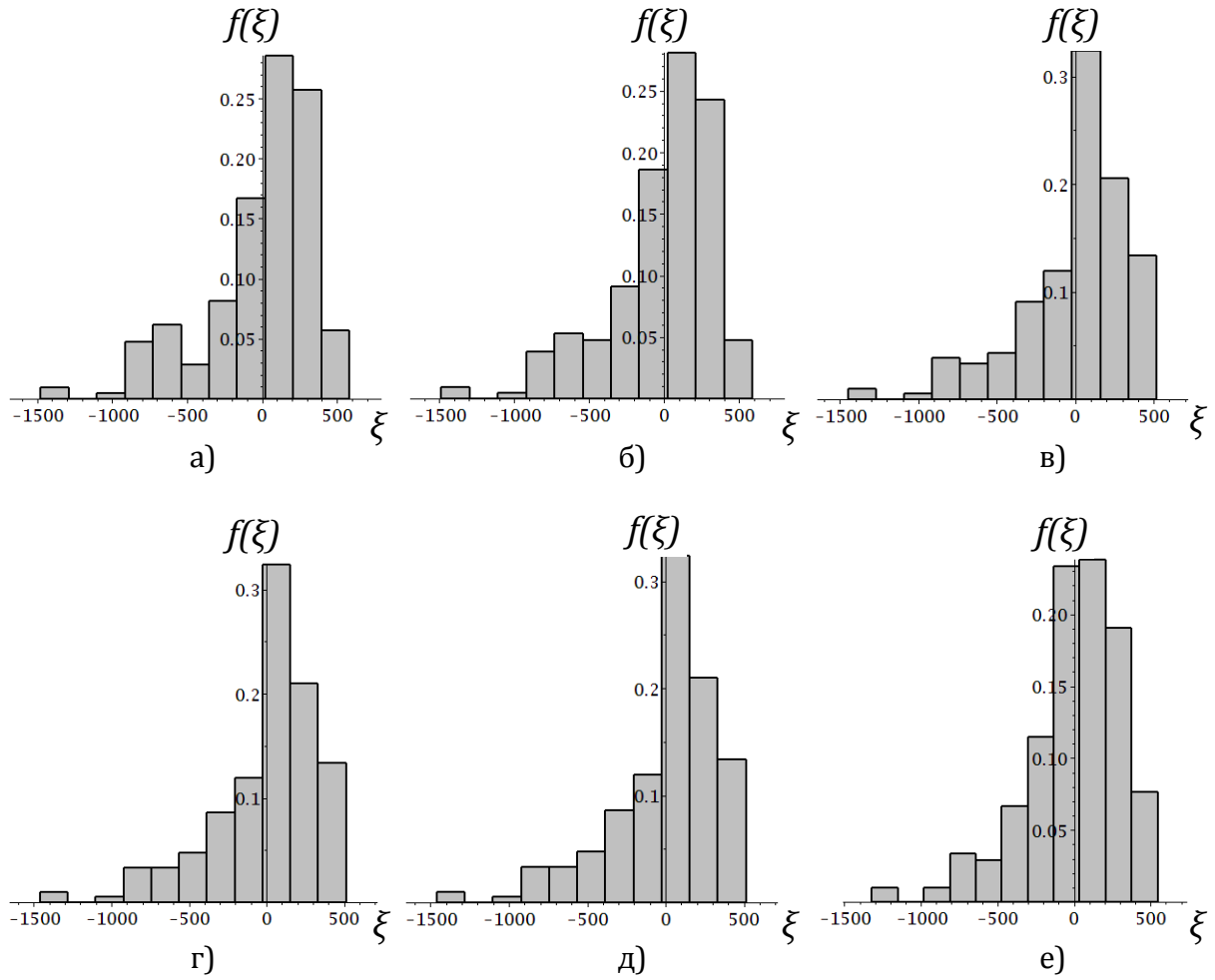

Рисунок 6 – Гістограми щільності розподілу випадкової складової після згладжування поліномами: а – лінійний; б – квадратичний; в – кубічний; г – 4-го порядку; д – 5-го порядку; е – 6-го порядку

ВИСНОВКИ

1. Залежність частоти появи тематичного контенту в мережі Інтернет від часу має нелінійний і складний характер, що не дає можливості точно знайти модель процесу при відсутності відповідного алгоритму розрахунків. Динаміку тематичного контенту в мережі Інтернет під час проведення противником інформаційної війни можливо відслідковувати за частотою його появи у ВДІ і вести розрахунки, основані на математичній статистиці.

2. Випадкова складова є обов'язковою компонентою часового ряду, яка визначає стохастичний характер його елементів y_i і розподілена за логарифмічно-нормальним або нормальним законом. Вид закону розподілу випадкової складової змінюється в залежності від виду апроксимуючої функції і зі збільшенням її кривизни наближається до нормального закону розподілу. Після проведення розрахунків, виділення тренду і виявлення закон розподілу випадкової складової можна використовувати отримані дані для прогнозування.

3. Достовірність апроксимації часового ряду частоти появи тематичного контенту в мережі Інтернет збільшується зі збільшенням порядку полінома.

4. Проведення статистичного аналізу дозволяє зменшити вплив випадкової складової і наблизити часовий ряд до систематичної складової.

5. Часовий ряд є стаціонарним, оскільки має систематичну складову та не змінюються статистичні характеристики з часом.

Перспективою подальших досліджень є використання міандрових та біфуркаційних моделей з перенесенням на них знайдених параметрів поліноміальних моделей за допомогою диференціальних перетворень та методу балансів диференціальних спектрів. Даний підхід є перспективою для пошуку адекватних моделей опису активності тематичного контенту в мережі Інтернет з подальшим його прогнозуванням.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ / REFERENCES

1. Boks, Dzh., & Dzhenkins, G. (1974). *Analiz vremennyih ryadov, prognoz i upravlenie* [Analysis of time series, forecast and management] (Vol. 1). Moscow: Myr (in Russian)
[Бокс, Дж., Дженкинс, Г. (1974). *Анализ временных рядов, прогноз и управление* (Кн. 1). Москва: Мир].

2. Buryachok, V. L., Tolyupa, S. V., Anosov, A. O., Kozachok, V. A., & Lukova-Chuiko, N. V. (2015). *Systemnyj analiz ta pryjnyattya rishen v informacijnij bezpeci* [System analysis and decision-making in information security]. Kyiv: DUT (in Ukrainian)
[Бурячок, В. Л., Толюпа, С. В., Аносов, А. О., Козачок, В. А., & Лукова-Чуйко, Н. В. (2015). *Системний аналіз та прийняття рішень в інформаційній безпеці*. Київ: ДУТ].

3. Danyk, Yu. (Ed.). (2016). *Osnovy kibernetychnoyi bezpeky* [Fundamentals of cybernetic security]. Zhytomyr: ZhNAEU (in Ukrainian)
[Даник, Ю. Г. (Ред.). (2016). *Основи кібернетичної безпеки*. Житомир: ЖНАЕУ].

4. Danyk, Yu., Pysarchuk, O., Lahodnyj, O., & Gajdarly, G. (2016). Fasetna systema klasyfikaciyi informacijnyx zagroz vyznachenij cilovij audytoriyi v kibernetychnomu prostori [Facet system of classification of information threats to a specific target audience in cybernetic space]. *Weapons and Military Equipment, 3*(11), 46–51 (in Ukrainian)
[Даник, Ю. Г, Писарчук, О. О., Лагодний, О. В., & Гайдарли, Г. С. (2016). Фасетна система класифікації інформаційних загроз визначеній цільовій аудиторії в кібернетичному просторі. *Озброєння та військова техніка, 3*(11), 46–51].

5. Danyk, Yu., Pysarchuk, O., Lahodnyj, O., & Vyporxonyuk, O. (2016). Matematychna model bagatokryterijnogo ocinyuvannya efektyvnosti internet-sajtiv cilovogo spryamuvannya [Mathematical model of multicriterion evaluation of the effectiveness of Internet sites of the target direction]. *Visnyk ZhDTU, 1*(76), 114–120 (in Ukrainian)
[Даник, Ю. Г, Писарчук, О. О., Лагодний, О. В., & Випорхонюк, О. В. (2016). Математична модель багатокритерійного оцінювання ефективності інтернет-сайтів цільового спрямування. *Вісник ЖДТУ, 1*(76), 114–120].

6. Doktryna informacijnoyi bezpeky Ukrayiny [Doctrine of Information Security of Ukraine] (Ukraine), 25 February 2017, No 47/2017. Retrieved July 1, 2017, from http://zakon3.rada.gov.ua/laws/show/47/2017 (in Ukrainian)
[Доктрина інформаційної безпеки України (Україна) 25 лютого 2017, №47/2017. Актуально на 01.07.2017. URL: http://zakon3.rada.gov.ua/laws/show/47/2017].

7. Gorbulin, V. (2017). *Svitova gibrydna vijna: Ukrayinskyj front* [The World Hybrid War: The Ukrainian Front]. Kyiv: NISD (in Ukrainian)
[Горбулін, В. П. (Ред.). (2017). *Світова гібридна війна: український фронт*. Київ: НІСД].

8. Gorbulin, V. P., Dodonov, O. G., & Lande, D. V. (2009). *Informacijni operaciyi ta bezpeka suspilstva: zagrozy, protydiya, modelyuvannya* [Information Operations and Society Safety: Threats, Opposition, Modeling]. Kyiv: Intertexnologiya (in Ukrainian)
[Горбулін, В. П., Додонов, О. Г., & Ланде, Д. В. (2009). *Інформаційні операції та безпека суспільства: загрози, протидія, моделювання*. Київ: Інтертехнологія].

9. Kharchenko, V. P., & Pysarchuk, O. O. (2015). *Nelinijne ta bagatokryterialne modelyuvannya procesiv u systemax keruvannya ruxom* [Nonlinear and multicriterial modeling of processes in traffic control systems]. Kyiv: Instytut obdarovanoyi dytyny(in Ukrainian)
[Харченко, В. П., & Писарчук, О. О. (2015). *Нелінійне та багатокритеріальне моделювання процесів у системах керування рухом*. Київ: Інститут обдарованої дитини].

10. Lande, D. V., & Furashev, V. M. (2012). *Osnovy informacijnogo i socialno-pravovogo modelyuvannya* [Fundamentals of information and socio-legal modeling]. Kyiv: PanTot (in Ukrainian)
[Ланде, Д. В., & Фурашев, В. М. (2012). *Основи інформаційного і соціально-правового моделювання*. Київ: ПанТот].

11. Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2016). *Introduction to time series analysis and forecasting*. Hoboken: Wiley.

12. Panasenko, L. I., & Tkach, I. M. (2013). *Voyenno-ekonomichnyj analiz* [Military-economic analysis]. Kyiv: NUOU im. Ivana Chernyaxovskogo (in Ukrainian)
[Панасенко, Л. І., & Ткач, І. М. (2013). *Воєнно-економічний аналіз*. Київ: НУОУ ім. Івана Черняховського].

13. Venttsel, E. S. (1969). *Teoriya veroyatnostey* [Theory of Probabilities]. Moscow: Nauka (in Russian)
[Вентцель, Е. С.(1969). *Теория вероятностей*. Москва: Наука].