УДК 004.056:681.3

Регресійний аналіз тенденцій розвитку кібератак

С. О. Аксьончиков, І. В. Ємельянова, К. Д. Маркова, І. І. Сватовський
Харківский національний університет імені В. Н. Каразіна, Україна

В статті приведено обґрунтування проведення регресійного аналізу тенденцій розвитку кібератак. Виділені фактори, що, найімовірніше, найбільше позначаються на коливаннях інтенсивності атак. Проведені кореляційний та регресійний аналіз факторів впливу на кількість кібератак на web-ресурси з використанням статистики за останні кілька років. Проведена перевірка адекватності отриманих моделей і доведена можливість використання отриманих даних для покращення захисту web-ресурсів. Зроблені висновки щодо отриманих результатів і обґрунтовані шляхи підвищення достовірності результатів у подальших дослідженнях.

Ключові слова: *кібератака, web-техології, регресійний аналіз, кореляційний аналіз, кореляція* .

В статье приведено обоснование проведения регрессионного анализа тенденций развития кибератак. Выделены факторы, которые, вероятнее всего, сказываются на колебаниях интенсивности атак. Проведены корреляционный и регрессионный анализ факторов влияния на количество кибератак на web-ресурсы с использованием статистики за последние несколько лет. Проведена проверка адекватности полученных моделей и доказана возможность использования полученных данных для улучшения защиты web-ресурсов. Сделаны выводы о полученных результатах и обоснованы пути повышения достоверности результатов в дальнейших исследованиях.

Ключевые слова: *кибератака, web-технологии, регрессионный анализ, корреляционный анализ, корреляция.*

The article gives a justification for conducting a regression analysis of trends in the development of cyber attacks. Factors which are likely to affect most the fluctuations in the intensity of attacks have been selected. Correlation and regression analysis of factors which influence the number of cyber attacks on web-resources have been carried out by using the statistics for a few previous years. The adequacy of the obtained models has been checked and the possibility of using the obtained data to improve the protection of web-resources has been proved. Conclusions regarding the obtained results are made and the ways of increasing the reliability of the results in further research are substantiated.

Key words: *Cyber Attack, Web-technology, Regression Analysis, Correlation Analysis, Correlation.*

1 Вступ

Сучасне суспільство користується всіма перевагами використання інформаційних технологій, які грають вирішальну роль практично в будь-який людській діяльності. Очевидно, що в цих умовах значення кібербезпеки для сучасного суспільства надзвичайно зростає [1]. На сьогоднішній день кібербезпека перестала бути проблемою, яка турбує лише фахівців в цій області. Інциденти в сфері кібербезпеки позначаються на життєдіяльності споживачів інформаційних і багатьох інших послуг, яким наразі також добре відомо про вірусні атаки (WannaCry, Petya.A та інші) чи кібератаки, націлені на різноманітні об’єкти інфраструктури систем електронних комунікацій чи управління технологічними процесами (Stuxnet, Flame, BlackEnergy та інш.) [2].

Велике занепокоєння визивають як тяжкі техніко-економічні наслідки кібератак, так і тенденції до зростання їх кількості і різноманітності, що відображається статистичною звітністю в оглядах з кібербезпеки відомих світових компаній [6-11]. Деякі з цих кібератак націлені на веб-ресурси, зокрема, на веб-ресурси державних підприємств, адже мають політичне або економічне підґрунтя.

Численність елементів, що становлять кіберпростір [1,3], велика кількість взаємозв'язків між ними, можливість застосування спеціальних технік управління діями цих елементів по типу бот-мереж, наприклад, визначають можливості розвитку загроз, які присутні в інформаційному просторі. При цьому все наростаюча інтенсивність кібератак походить від масштабності світового кіберпростору. Складні атаки типу APT (Advanced Persistent Threats) мають комплексну структуру, різноманітні механізми своєї реалізації та спираються на можливість використання різних напрямків поширення інформації. Використання методів соціальної інженерії дозволяє знаходити найбільш продуктивні методи організації атак [3]. У кіберпросторі, як прогнозується спеціалістами, можуть розвиватися все більш небезпечні і складні загрози, що робить актуальною задачу їх всебічного аналізу та використання результатів її вирішення для ефективної протидії існуючим і можливим в майбутньому кіберзагрозам.

2 Обгрунтування напряму досліджень

Тенденції розвитку кібератак змінюються рік від року і на перший погляд є досить випадковими. Ми можемо провести їх аналіз по звітах авторитетних в цій сфері компаній, таких як Positive Technologies [6] та Cisco [9], що регулярно публікуються кожен квартал року. Звіти та огляди по кібератакам містять досить детальні статистичні дані, однак не дають можливості знайти причинно-наслідковий зв’язок між різними факторами, параметрами (векторами) та наслідками кібератак, лише поглянувши на них.

Для того, щоб дослідити закономірності підготовки і проведення кібератак та виявити зв’язок між ними та впливом різноманітних факторів, на наш погляд, необхідно провести кореляційний та регресійний аналіз таких даних, а конкретно – факторів, що характеризують кібератаки. Названі види аналізу є прикладом найбільш важливих і популярних кількісних методів математичного моделювання, метою яких є встановлення факту наявності або відсутності певного зв'язку (кореляції) між випадковими величинами або процесами та виявлення функціонального зв'язку, тобто виду цієї залежності [4,5].

Для того, щоб визначити основні тенденції розвитку кібератак, необхідно перевірити кореляцію між кількістю виявлених атак і чинниками (факторами), що імовірно можуть позначатися на коливаннях інтенсивності атак. Для проведення кореляційного аналізу були вибрані фактори, які, на наш погляд, найбільш імовірно впливають на основні тенденції розвитку кібератак:

N – кількість виявлених атак на веб-ресурси, млн;

N_1 – кількість нових виявлених видів (технік) атак, млн;

N_2 – географічний фактор 1 (розповсюдженість атак);

N_3 – географічний фактор 2 (потерпання від атак по світу);

N_4 – поширеність технологій атак;

N_5 – поширеність платформи web-ресурсів, на які здійснювались атаки;

N_6 – сфера життєдіяльності, що найбільше потерпала від атак.

Статистичні дані, які використані у даній роботі, було запозичено зі звітів по кібератакам за 2015-2017 роки від компаній Positive Technologies [6], Kasperky [7], McAfee [8], Cisco [9] та TrendMicro [10], а також з рейтингів веб-технологій, зпівставлених Stack Overflow [11]. Для аналізу в якості кількісного відображення рейтингу країни використовувався її порядковий номер в рейтингу країн за економічним розвитком [12]. Дані також були розподілені по кварталам з 2015 по 2017 роки.

3 Проведення кореляційного аналізу впливу факторів на розвиток кібератак

Для числової оцінки можливого зв'язку між двома випадковими величинами: Y (із середнім M_y і середньоквадратичним відхиленням S_y) і - X (із середнім M_x і середньоквадратичним відхиленням S_x) прийнято використовувати [4] так званий вибірковий лінійний коефіцієнт кореляції:

$$r^* = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - a^*)(b_i - b^*)}{S[a]S[b]}, \qquad (3.1)$$

де a^*, b^* - оцінки математичних очікувань $M[a]$ і $M[b]$;

$$S[a] = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(a_i - a^*)^2}\ ,\ S[b] = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(b_i - b^*)^2} \qquad (3.2)$$

Коефіцієнт кореляції приймає значення від -1 до +1 в залежності від тісноти зв'язку між даними випадковими величинами. Якщо коефіцієнт кореляції дорівнює нулю, то X і Y називають некорельованими [5].

Нами було проведено кореляційний аналіз факторів $N_1 \ldots N_6$ з фактором N (кількість виявлених атак). Результати зроблених розрахунків зведено у таблицю коефіцієнтів кореляцій (табл.3.1).

Таблиця 3.1 Коефіцієнти кореляції

Позначення фактору	Значення коефіцієнту кореляції
N_1	0,720521
N_2	-0,7367
N_3	0,773986
N_4	0,328211
N_5	0,422164
N_6	0,600009

По результатам аналізу даних в табл. 3.1 можна зробити висновок, що тісно корельованим з кількістю атак є фактор виявлення нових шкідливих технологій, а також географічні фактори. Саме ці фактори було використано на наступному етапі аналізу даних по кібератакам.

4 Застосування регресійного аналізу впливу факторів на розвиток кібератак

Окрім встановлення факту наявності або відсутності певного зв'язку (кореляції) між випадковими величинами або процесами метою дослідження може бути і безпосереднє визначення функціонального зв'язку, тобто виду цієї залежності. Поставлена мета досягається регресійним аналізом значень факторів і відгуку [13,14]. Для цього необхідно знайти коефіцієнти та рівняння регресії [15,16], загальний вигляд якого представлений у наступному виразі :

$$y^* = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \sum_{i=0}^{n} \beta_i x_i \,, \qquad (4.1)$$

де $x_0 = 1$, β_i - коефіцієнти регресії.

Процедура побудови лінійної регресійної моделі була автоматизована в табличному редакторі Exel корпорації Microsoft. Зокрема для побудови лінійної регресії була застосована функція «ЛИНІЙН», яка потрібна для розрахунків статистик для ряду із застосуванням методу найменших квадратів.

Дані розрахунків з використанням даної функції приведені на діаграмі з графіком регресії і її рівнянням (рис. 4.1).

На рис. 4.1 наведено побудовану регресійну модель впливу кількості нових виявлених шкідливих технологій (технік атак) на кількість виявлених атак в цілому. На рис. 4.2 і 4.3 наведені побудовані регресійні моделі впливу географічних факторів на кількість виявлених атак.

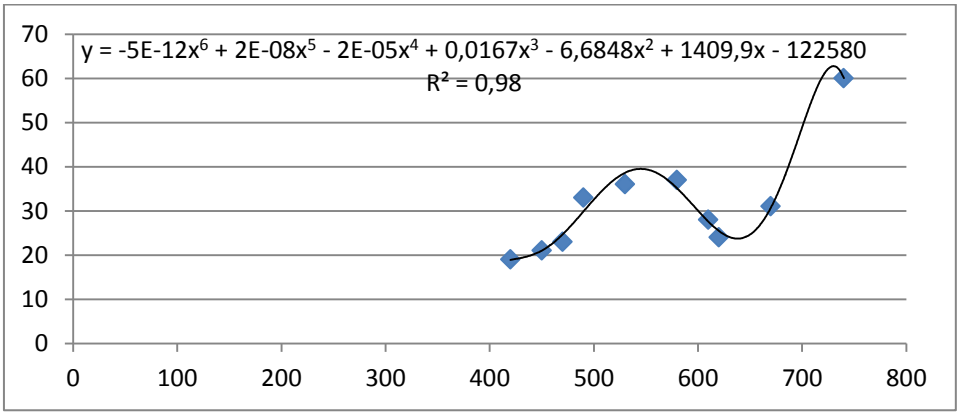

Рис. 4.1 Регресійна модель кількості виявлених атак на веб-ресурси, млн та кількості нових виявлених шкідливих технологій, млн.

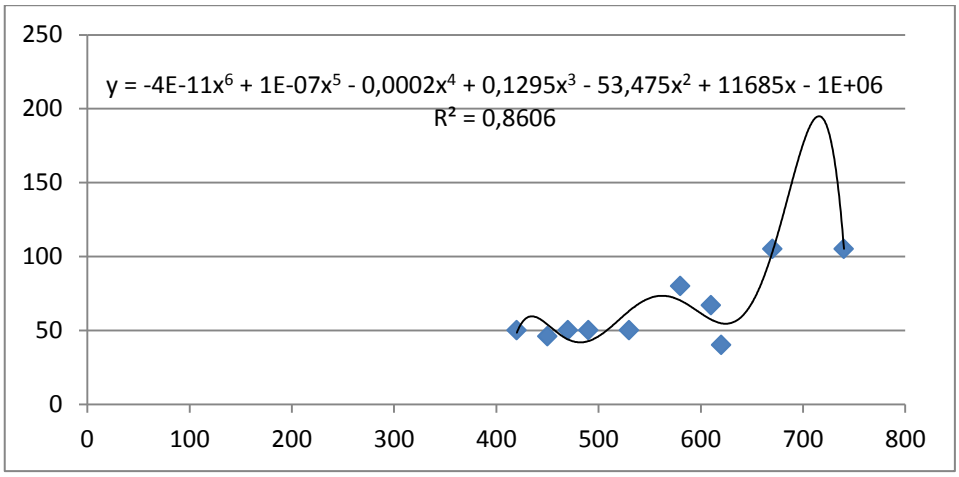

Рис. 4.2 Регресійна модель кількості виявлених атак на веб-ресурси, млн та країни, що найбільше потерпала від атак.

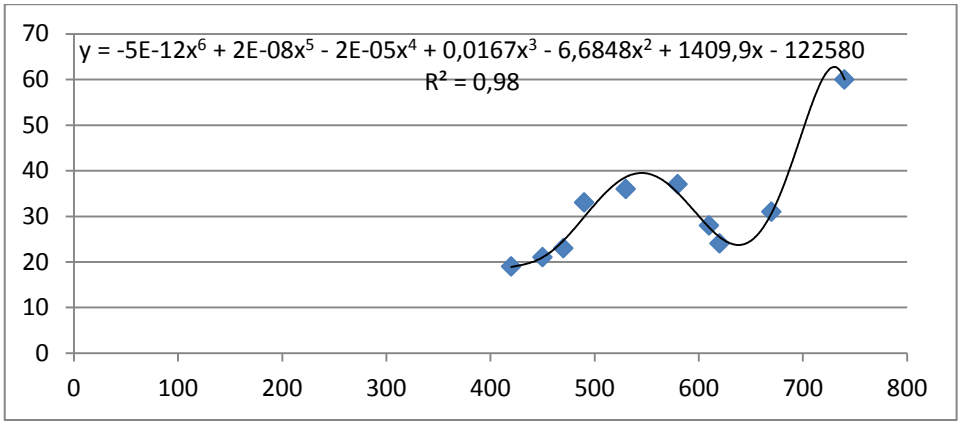

Рис. 4.3 Регресійна модель кількості виявлених атак на веб-ресурси, млн та країни, звідки розповсюджувалася найбільша кількість атак.

5 Результати проведення аналізу

Перевірка адекватності моделей, побудованих на основі рівнянь регресії, починається з перевірки значимості кожного коефіцієнта регресії.

Значимість коефіцієнтів регресії здійснюється за допомогою відомого t-критерію Стьюдента [17]:

$$t_p = \frac{|a_i|}{\sqrt{\sigma^2_{a_i}}}, \qquad (5.1)$$

де $\sigma^2_{a_i}$ - дисперсія коефіцієнта регресії;

Параметр моделі визнається статистично значущим, якщо $t_p > t_{kp}$.

Проведемо оцінку коефіцієнтів регресії. Результати оцінки для першого рівняння наведені в табл. 5.1. Перше рівняння регресії має вигляд:

$$y = -5E-12x^6 + 2E-08x^5 - 2E-05x^4 + 0{,}0167x^3 - 6{,}6848x^2 + 1409{,}9x - 122580$$

Таблиця 5.1 Результати оцінки коефіцієнтів регресії першого рівняння

Номер коефіцієнту регресії	Значення коефіцієнту регресії	Дисперсія	Середнє квадратичне відхилення	Значення $t_{розр}$
1	5,00E-12	2138565539	46244,6271	1,08E-16
2	2,00E-08			4,32E-13
3	2,00E-05			4,32E-10
4	0,0167			3,61E-07
5	6,6848			1,45E-04
6	1409,9			3,05E-02
7	122580			2,65E+00

Результати оцінки для першого рівняння наведені в табл. 5.2. Друге рівняння регресії має вигляд:

$$y = -4E-11x^6 + 1E-07x^5 - 0{,}0002x^4 + 0{,}1295x^3 - 53{,}475x^2 + 11685x - 1E+06$$

Таблиця 5.2 Результати оцінки коефіцієнтів регресії другого рівняння

Номер коефіцієнту регресії	Значення коефіцієнту регресії	Дисперсія	Середнє квадратичне відхилення	Значення $t_{розр}$
1	4E-11	1,4232E+11	377250,1	1,06E-16
2	1E-07			2,65E-13
3	0,0002			5,30E-10
4	0,1295			3,43E-07
5	53,475			1,42E-04
6	11685			3,10E-02
7	1E+06			2,65E+00

Результати оцінки для третього рівняння наведені в табл. 5.3. Третє рівняння регресії має вигляд:

$$y = -1E-11x^6 + 4E-08x^5 - 6E-05x^4 + 0{,}0473x^3 - 20{,}53x^2 + 4708{,}7x - 445935$$

Таблиця 5.3 Результати оцінки коефіцієнтів регресії третього рівняння

Номер коефіцієнту регресії	Значення коефіцієнту регресії	Дисперсія	Середнє квадратичне відхилення	Значення $t_{розр}$
1	1,00E-11	2,8311E+10	168258,806	5,94E-17
2	4,00E-08			2,38E-13
3	6,00E-05			3,57E-10
4	0,0473			2,81E-07
5	20,53			1,22E-04
6	4708,7			2,80E-02
7	445935			2,65E+00

Критичні значення t-розподілу при 22 ступенях свободи приведені в табл. 5.4.

Таблиця 5.4. Критичні значення t-розподілу

Ймовірність	t-значення
0,05	2,074
0,01	2,819
0,001	3,792

В наших моделях всі коефіцієнти підтверджують правило $|t_{розр}| > t_{критич}$. Це означає, що моделі є достатньо адекватними, статистично значимими і можуть використовуватися для прогнозування. На основі зроблених висновків можна робити прогнози щодо тенденцій кібератак, а також імовірно вказувати, які чинники могли вплинути на ріст чи зниження кількості атак. Завдяки цьому отримані дані є підґрунтям для покращення ефективності систем безпеки web-ресурсів, посилаючись на той чи інший фактор безпеки. Наприклад, якщо значимим фактором є кількість нових виявлених технік, то це свідчить про те, що безпеку web-ресурсів необхідно забезпечувати з оглядом на тенденцію появи нових видів атак. Значимість географічного фактору може сказати про те, чи є безпечним використання домену тієї чи іншої країни для своїх web-ресурсів. Інтерпретацію отриманих результатів і зроблених висновків викладено у наступному розділі.

6 Висновки

Таким чином, результати дослідження свідчать про наявність кореляції вибраних факторів з основними дослідженими тенденціями розвитку кібератак. Побудовані регресійні моделі для найбільш поширених корелюючих факторів є адекватними і статистично значимими. Кореляція між факторами N (кількість виявлених атак) і N_1 (кількість нових виявлених шкідливих технологій) досить логічна, адже з появою нових технологій кібератак, для яких ще не знайдені методи протидії, кількість атак в цілому збільшиться. Достатньої кореляції для факторів популярності різновидів атак та платформ веб-розробки не було виявлено, але географічна кореляція виявилася більш цікавою.

Географічні кореляції можуть свідчити про різні речі. У випадку кореляції з першим географічним фактором маємо негативну кореляцію з країнами, що розповсюджують атаки. Це означає, що в більш технологічно розвинених країнах спостерігається більша агресивність проведення кібератак. У випадку з урахуванням впливу другого географічного фактору маємо позитивну кореляцію з країнами, що потерпають від атак, тобто країни з гіршим рівнем розвитку потерпають від кібератак в більшій мірі, ніж розвинені країни.

Це може означати, що розвинені країни, на відміну від країн, що розвиваються, забезпечили себе достатнім захистом, або ж це може свідчити про зміну політичних і соціальних причин кібератак. Це, нажаль, говорить про збільшення небезпеки для країн, що розвиваються. Однак ці висновки неостаточні, адже вони залежать від коректності вхідних даних для проведення аналізу.

В цілому результати проведення регресійного аналізу факторів, що характеризують розвиток кібератак, можуть бути уточнені за рахунок розширення і забезпечення адекватності вхідних статистичних даних. Також вони можуть бути використаними з метою прогнозування розвитку кібератак та розробки методів і засобів протидії їм.

ЛІТЕРАТУРА

1. Закон України Про основні засади забезпечення кібербезпеки України - Відомості Верховної Ради (ВВР), 2017, № 45, ст.403.
2. Center for Internet Security - Cybersecurity Threats [Електронний ресурс] / Center for Internet Security – Режим доступу: https://www.cisecurity.org/cybersecurity-threats/ , вільний.
3. Бурячок, В. Л. Інформаційна та кібербезпека: соціотехнічний аспект: підручник / [В. Л. Бурячок, В. Б. Толубко, В. О. Хорошко, С. В. Толюпа]; За заг. ред. В. Б. Толубка. - К.: ДУТ, 2015. - 288 с.
4. Елисеева И. И. Общая теория статистики: Учебник / Елисеева И. И., Юзбашев М. М. Под ред. И. И. Елисеевой. — 4-е издание, переработанное и дополненное. — М: Финансы и Статистика, 2002. — 480 с.
5. Айвазян С.А. Прикладная статистика и основы эконометрики. / Айвазян С.А. – М.: Юнити. 2001. – 432с.

6. Positive Research 2015-2017 [Електронний ресурс] / Positive Technologies // 2015-2017. – Режим доступу: https://www.ptsecurity.com/ru-ru/research/analytics/ , вільний.
7. Развитие информационных угроз в 2015-2017 годах. Статистика. [Електронний ресурс] / АО KasperskyLab // 2015-2017. – Режим доступу: https://securelist.ru/all/?category=736 , вільний.
8. McAfee Labs Threat Report [Електронний ресурс] / McAfee Labs // September 2017. – Режим доступу: https://www.mcafee.com/us/security-awareness/articles/mcafee-labs-threats-report-sep-2017.aspx , вільний.
9. Cisco 2016 Annual Security Report [Електронний ресурс] / Cisco // 2015-2017. – Режим доступу: https://www.cisco.com/c/m/en_us/offers/sc04/2016-annual-security-report/index.html , вільний.
10. A Rising Tide: New Hacks Threaten Public Technologies [Електронний ресурс] / TrendLabs // 2015-2017. – Режим доступу: https://www.trendmicro.com/vinfo/us/security/research-and-analysis/threat-reports/roundup/a-rising-tide-new-hacks-threaten-public-technologies , вільний.
11. Stack Overflow Annual Developer Survey. [Електронний ресус] / Stack Exchange Inc // 2015-2017. – Режим доступу: https://insights.stackoverflow.com/survey, вільний.
12. Legatum Prosperity Index. [Електронний ресурс] / LEGATUM INSTITUTE // 2015-2017. – Режим доступу: http://www.prosperity.com/rankings , вільний
13. Дрейпер, Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – М.: Вильямс, 2007. – 912 с.
14. Соколов, Г.А. Введение в регрессионный анализ и планирование регрессионных экспериментов в экономике: Учебное пособие / Г.А. Соколов, Р.В. Сагитов. – М.: ИНФРА-М, 2012. – 202 с.
15. Фёрстер Э. Методы корреляционного и регрессионного анализа = Methoden der Korrelation - und Regressiolynsanalyse. / Фёрстер Э., Рёнц Б. — М.: Финансы и статистика, 1981. — 302 с.
16. Стрижов В.В. Методы выбора регрессионных моделей. / Стрижов В.В., Крымова Е.А. – М.: ВЦ РАН, 2010. – 60 с.
17. Ромакін В. В. Комп'ютерний аналіз даних: Навчальний посібник. – Миколаїв: Вид-во МДГУ ім. Петра Могили, 2006. — 144 с.

Надійшла – 15.12.2017.