ПРИКЛАДНА МАТЕМАТИКА

DOI: 10.20535/kpisn.2022.1-2.287916
УДК 004.852

О.Р. Чертов[1], І.С. Жук[1]
[1]КПІ ім. Ігоря Сікорського, Київ, Україна
*Відповідальний автор: ivan.sergeyevich.zhuk@gmail.com

ІМІТАЦІЙНА МОДЕЛЬ ФУТБОЛЬНОГО СЕЗОНУ З МАТЧАМИ З ФІКСОВАНИМ РЕЗУЛЬТАТОМ

Проблематика. Футбол – це величезна індустрія, порівнянна із традиційними економічними галузями. У числовому виразі це десятки мільярдів доларів. Однією з найважливіших проблем, з якою стикається ця галузь, є договірні матчі (матчі з фіксованим результатом).
Найбільш ефективним способом запобігання цій загрозі є багатосторонній підхід у поєднанні із заходами, спрямованими на розширення потенціалу правоохоронних органів та спортивних організацій. Однією зі складових цього підходу є використання математичних методів виявлення підозрілих щодо фіксованого результату матчів.
Мета дослідження. Метою роботи є розроблення алгоритму моделювання договірних матчів, пов'язаних із заробітком на ставках, з використанням якого формуються матчі, результати яких відмінні від очікуваних і можуть розглядатись як аномальні.
Методика реалізації. Розрахунок ймовірностей забиття голів командами під час гри на основі реальних даних сезону; розроблення імітаційної моделі футбольного сезону без договірних матчів і її аналіз за допомогою статистичного моделювання; розроблення алгоритму моделювання договірних матчів, пов'язаних із заробітком на ставках, і його аналіз.
Результати дослідження. Розроблено імітаційну модель футбольного сезону, яка дозволяє з використанням отриманих на реальних даних ймовірнісних розподілів кількості голів, забитих командами протягом домашніх або виїзних ігор, змоделювати результати матчів, враховуючи силу команд і тип гри, а також змоделювати ситуації «договірного» матчу, замінивши поточні результати. За загальними розподілами типів результатів матчів та різницями голів усіх матчів змодельований сезон є подібним до реального сезону. За критерієм Колмогорова–Смирнова різниця наведених розподілів на рівні значущості 0,001 є статистично незначущою.
Висновки. Розроблену імітаційну модель футбольного сезону можна використовувати для дослідження ефективності методів виявлення підозрілих щодо фіксованого результату матчів і їх порівняльного аналізу.
Ключові слова: футбольний сезон; групування команд; розподіл Пуассона; гістограма різниць голів; тип результату матчу.

Вступ

Футбол – це величезна індустрія, порівнянна з традиційними економічними галузями. У числовому виразі це десятки мільярдів доларів. Однією з найважливіших проблем, з якою стикається ця галузь, є договірні матчі (матчі з фіксованим результатом). Корупція і причетність до транснаціональної організованої злочинності підривають довіру до результатів та дискредитують репутацію спортивних змагань. Крім того, величезні доходи, отримані внаслідок проведення договірних матчів, можуть бути використані для фінансування інших незаконних видів діяльності [1].

Найбільш ефективним способом запобігання цій загрозі є багатосторонній підхід у поєднанні із заходами, спрямованими на розширення потенціалу правоохоронних органів та спортивних організацій [1]. Однією із складових цього підходу є використання математичних методів виявлення підозрілих щодо фіксованого результату матчів.

Відомі методи прогнозування результату матчу за рахунок аналізу ставок на матч чи аналізу дій футболістів на полі використовують велику кількість даних, які не завжди доступні [2–5]. Для подолання цього обмеження може бути застосовано метод виявлення підозрілих

Пропозиція для цитування цієї статті: О.Р. Чертов, І.С. Жук, "Імітаційна модель футбольного сезону з матчами з фіксованим результатом", *Наукові вісті КПІ*, № 1–2, с. 82–94, 2022. doi: 10.20535/kpisn.2022.1-2.287916

Offer a citation for this article: O. R. Chertov, I. S. Zhuk, "Soccer season simulation with fixed matches", *KPI Science News*, no. 1–2, pp. 82–94, 2022. doi: 10.20535/kpisn.2022.1-2.287916

на фіксований результат матчів на основі конформних предикторів і степеневих мартингалів, який використовує загальнодоступні публічні дані. Апарат конформних предикторів та степеневих мартингалів був використаний у роботах [6-7] для розробки методів виявлення як точкових, так і групових аномалій, які не вимагають знання про розподіл даних.

Для дослідження ефективності методів виявлення підозрілих щодо фіксованого результату матчів і їх порівняльного аналізу необхідно розробити імітаційну модель футбольного сезону, що враховує наявність матчів з фіксованим результатом.

Розрізняють два типи матчів з фіксованим результатом: пов'язані з підкупом команд з метою заробітку на ставках у букмекерських конторах і що ставлять за мету турнірні цілі [8, 9]. Важливо зазначити, що саме перший тип договірних матчів пов'язаний із криміналом, незаконним збагаченням й викликає максимальну тривогу ФІФА та ООН [1].

Алгоритми формування договірних матчів різних типів різняться між собою логікою. Для руху по турніру достатньо домовлятися про мінімальні й неаномальні результати, що сприяють підвищенню місця команди в загальнокомандному рейтингу. Це не має бути публічно помітним і зазвичай не пов'язано зі ставками.

Договірні матчі, пов'язані з метою заробітку на ставках, мають завдання отримати результат матчу, відмінний від очікуваного, щоб максимально заробити на цьому. Чим неймовірнішим буде результат, тим більше можна заробити на ставці.

У роботі розроблено алгоритм моделювання договірних матчів, пов'язаних із заробітком на ставках, з використанням якого формуються матчі, результати яких відмінні від очікуваних і можуть розглядатись як аномальні.

Постановка задачі

Метою роботи є розроблення алгоритму моделювання договірних матчів, пов'язаних із заробітком на ставках, з використанням якого формуються матчі, результати яких відмінні від очікуваних і можуть розглядатись як аномальні.

Початкові дані та припущення (обмеження) задачі

Початковими даними є реальний сезон. Він складається із двох таблиць: таблиці матчів і таблиці команд. Параметри цих таблиць описано у табл. 1 і табл. 2 відповідно. У таблиці матчів дані упорядковано за датою проведення матчу D_k, а в таблиці команд — за кількістю набраних очок $p(t_i)$. Параметри табл. 2 — загальні очки команд, кількість виграних, нічийних та програних матчів — можна порахувати із таблиці матчів (табл. 1).

Таблиця 1. Опис параметрів таблиці матчів

k	Номер матчу
D_k	Дата проведення матчу
h_k	Назва команди-господаря матчу
o_k	Назва команди-гості матчу
α_k	Кількість голів, забитих командою h_k
β_k	Кількість голів, забитих командою o_k

Таблиця 2. Опис параметрів турнірної таблиці

i	Номер команди
t_i	Назва команди
$p(t)_i$	Загальні очки команди
$w(t)_i$	Кількість ігор команди, які завершились перемогою цієї команди
$d(t)_i$	Кількість ігор команди, які завершились нічиєю
$l(t)_i$	Кількість ігор команди, які завершились поразкою цієї команди
$group(t)_i$	Ранг команди (група, до якої належить команда)

Як реальний сезон було обрано сезон 2013–2014 років Ліги ІІ Франції, для якого в судовому порядку було доведено наявність договірних матчів [10, 11].

Команди реального сезону поділено на чотири групи (табл. 3). Принцип групування є таким: перший кластер формується з $M+1$ перших команд сезону, останній кластер формується з $L+1$ останніх команд сезону. Числа M і L обираються із правил підвищення або пониження у класі. Для Ліги ІІ Франції $M=3$ та $L=3$. Інші команди розбиваються на 2 групи: перша група складається з команд, близьких до команд першого кластера, друга група — з команд, близьких до команд останнього кластера. Для отримання цих двох груп у межах поточного дослідження було використано алгоритм 2-середніх [12]. Результати групування команд цього сезону показано в табл. 3. Певні групи команд табл. 3 оформлені окремими кольорами: світло-червоним кольором оформлено команди кластера № 1, зеленим кольором — команди кластера № 2, жовтим — команди кластера № 3, білим — команди кластера № 4.

Таблиця 3. Групування команд сезону 2013–2014 років Ліги II Франції

№ команди в загальному заліку	Назва команди	Кількість набраних очок	Кількість перемог	Кількість нічиїх	Номер кластера, ранг
1	'Metz'	76	22	10	1
2	'Lens'	65	17	14	1
3	'Caen'	64	18	10	1
4	'Nancy'	61	16	13	1
5	'Niort'	58	15	13	2
6	'Dijon'	57	14	15	2
7	'Brest'	56	15	11	2
8	'Angers'	55	14	13	2
9	'Tours'	55	15	10	2
10	'Troyes'	52	15	7	3
11	'Creteil'	50	12	14	3
12	'Le Havre'	48	11	15	3
13	'Arles-Avignon'	46	10	16	3
14	'Clermont'	45	10	15	3
15	'Nimes'	44	10	14	3
16	'Auxerre'	43	10	13	3
17	'Laval'	42	10	12	4
18	'Chateauroux'	40	10	10	4
19	'Istres'	36	9	9	4
20	'CA Bastia'	24	4	12	4

На основі таблиці матчів і таблиці команд реального сезону обчислюють частоти голів, що їх забивають команди у матчі. Для кожної команди, обчислюють два типи частот. Перший тип частот (домашній) $\nu\left(X_i | group(t_j) = g\right)$ – це частоти того, що команда t_i заб'є $X_i = x_i$, $x_i = \overline{0, K_i}$ голів на власному полі за умови, що команда супротивника t_j належить до групи g. Другий тип частот (виїзний) $\nu\left(Y_j | group(t_i) = d\right)$ – це частоти того, що команда t_j заб'є $Y_j = y_j$, $y_j = \overline{0, K_j}$ голів на виїзді за умови, що команда супротивника t_i належить до групи d. У цих формулах під K_i і K_j розуміємо максимальну кількість голів в аналізованому реальному сезоні, які, відповідно, забила команда t_i в домашніх іграх та команда t_i у виїзних іграх. Таким чином, для кожної команди отримуємо чотири множини частот першого типу і чотири множини частот другого типу. На основі отриманих частот розраховують параметри теоретичних законів розподілу ймовірностей забиття голів командами під час їх гри з іншими командами відповідного рангу.

Вхідними даними алгоритму моделювання сезону є:

– N команд реального сезону з атрибутами, описаними в табл. 2;

– $P_g\left(X_i\right) = P\left(X_i | group(t_j) = g\right)$, $g = \overline{1,4}$, i, $j = \overline{1, N}$ – ймовірність того, що команда t_i заб'є $X_i = x_i$, $x_i = \overline{0, K_i}$ голів на власному полі за умови, що команда супротивника t_j належить до групи g;

– $\breve{P}_d\left(Y_j\right) = P\left(Y_j | group(t_i) = d\right)$, $d = \overline{1,4}$, i, $j = \overline{1, N}$ – ймовірність того, що команда t_j заб'є $Y_j = y_j$, $y_j = \overline{0, K_j}$ голів на виїзді за умови, що команда супротивника t_i належить до групи d.

Результатом роботи алгоритму є модельний сезон, сформований на основі заданого реального сезону футбольного турніру. Модельний сезон складається з таблиці матчів (табл. 1) та турнірної таблиці сезону (табл. 2).

Припущення алгоритму є такими:

1) кількість голів, які команда-господарка заб'є під час гри, має розподіл Пуассона;

2) номер і групу команди розглядають як попередні оцінки її сили.

Розв'язання сформульованої задачі розглянемо в такій послідовності:

1. Розрахунок ймовірностей забиття голів командами під час гри на основі реальних даних сезону.

2. Розроблення імітаційної моделі футбольного сезону без договірних матчів і її аналіз за допомогою статистичного моделювання.

3. Розроблення алгоритму моделювання договірних матчів, пов'язаних із заробітком на ставках, і його аналіз.

Розрахунок ймовірностей забиття голів командами під час гри на основі реальних даних сезону

Однією з найбільш поширених моделей для прогнозування результату футбольних матчів є розподіл Пуассона [13] – його використовують для розрахунку ймовірностей кількості голів, які команда забиває у матчі. При цьому враховують припущення, що гол є незалежною подією, оскільки він не впливає на ймовірність того, скільки голів буде забито далі.

У розрахунку імовірностей голів, забитих командою, враховують тип гри – домашня або виїзна, а також силу команди супротивника, тобто до якої групи її віднесено. Алгоритм побудови розподілів Пуассона кількості голів команд сезону складається з таких етапів:

1. Формуємо множину значень голів, забитих командою t_i протягом ігор сезону. Отримуємо множину $G_i^{(g)} = \left\{k_i^{(g)}, k_i^{(g)}+1, \ldots, K_i^{(g)}\right\}$ де $k_i^{(g)}$ та $K_i^{(g)}$ – відповідно мінімальна та максимальна кількість голів, забитих командою t_i протягом усіх її домашніх ігор сезону, в яких команда-суперник t_j належала до групи g.

2. $\forall k \in G_i^{(g)}$ для кожного зі значень кількості голів k з множини $G_i^{(g)}$ обчислюємо частоту появи кількості голів k як голів, забитих командою t_i протягом ігор сезону, в яких команда t_i була домашньою, а команда-суперник t_j належала до групи g:

$$\nu_g\left(X_i = k\right) = \nu\left(X_i = k \middle| group\left(t_j\right) = g\right) = \frac{\left|\left\{m_{ij} \middle| X_i = k, group\left(t_j\right) = g\right\}\right|}{\left|\left\{m_{ij} \middle| group\left(t_j\right) = g\right\}\right|},$$

де m_{ij} – матч між командами (t_i, t_j), де t_i є домашньою командою матчу, а t_j є виїзною командою матчу; $\left\{m_{ij} \middle| X_i = k, group\left(t_i\right) = g\right\}$ – множина, що містить лише ті домашні матчі команди t_i, в яких вона забила $X_i = k$ голів і суперник t_j належав до групи g; $\left\{m_{ij} \middle| group\left(t_j\right) = g\right\}$ – множина, що містить усі домашні матчі команди t_i, у яких суперник t_j належав до групи g, а знак модуля над множиною підраховує кількість елементів у цій множині.

Відносні частоти $\breve{\nu}_d(Y_j = k)$ дискретної випадкової величини кількості голів Y_j, забитих командою t_j протягом ігор сезону, в яких команда t_j була виїзною, обчислюють аналогічно:

$$\breve{\nu}_d\left(Y_j = k\right) = \nu\left(Y_j = k \middle| group\left(t_i\right) = d\right) = \frac{\left|\left\{m_{ij} \middle| Y_j = k, group\left(t_i\right) = d\right\}\right|}{\left|\left\{m_{ij} \middle| group\left(t_i\right) = d\right\}\right|},$$

де $\left\{m_{ij} \middle| Y_j = k, group\left(t_i\right) = d\right\}$ – множина, що містить лише ті виїзні матчі команди t_j, в яких вона забила $Y_j = k$ голів, і суперник t_i належав до групи d; $\left\{m_{ij} \middle| group\left(t_i\right) = d\right\}$ – множина, що містить усі виїзні матчі команди t_j, в яких суперник t_i належав до групи d.

3. Згідно з оцінкою за методом максимальної правдоподібності, параметр $\lambda_i^{(g)}$ розподілу Пуассона для домашніх матчів команди t_i, в яких суперник належав до групи g, визначають за формулою

$$\lambda_i^{(g)} = \sum_{k=0}^{K_i} k \cdot \nu_g\left(X_i = k\right).$$

Параметр $\breve{\lambda}_j^{(d)}$ розподілу Пуассона для виїзних матчів команди t_j, в яких суперник належав до групи d, знаходять за формулою

$$\breve{\lambda}_j^{(d)} = \sum_{k=0}^{K_j} k \cdot \breve{\nu}_{gd}\left(Y_j = k\right).$$

Розподіли Пуассона кількостей забитих голів командою t_i в її домашніх матчах і командою t_j в її виїзних матчах обчислюють за формулами

$$P\left(X_i = x_i\right) = \frac{e^{-\lambda_i^{(g)}} \lambda_i^{(g)x_i}}{x_i!};$$

$$\breve{P}_d\left(Y_j = y_j\right) = \frac{e^{-\breve{\lambda}_j^{(d)}} \breve{\lambda}_j^{(d)y_j}}{y_j!}.$$

4. Для оцінювання, наскільки гарною є відповідність між фактичною кількістю матчів з тією чи іншою кількістю голів та моделлю Пуассона, використовуємо критерій Хі-квадрат. Для перевірки розподілу ймовірності Пуассона кількості голів у домашніх матчах команди t_i, суперник якої належить до групи g, статистику Хі-квадрат визначаємо за формулою [14]

$$\chi^2_{i,g} = \sum_{i=0}^{n} \frac{(\nu_g(X_i = k) M_i^{(g)} - P_g(X_i = k) M_i^{(g)})^2}{P_g(X_i = k) M_i^{(g)}},$$

де $M_i^{(g)} = \left|\left\{m_{ij} | group(t_j) = g\right\}\right|$ – кількість усіх домашніх матчів команди t_i, в яких суперник t_j належав до групи g.

Статистику Хі-квадрат для виїзних матчів команди t_j, в яких суперник t_i належав до групи d, знаходять аналогічно.

Критичне значення $\chi^2_\alpha(n)$ визначають на рівні значущості $\alpha = 0{,}05$. Кількість степенів вільності $n = K_i^{(g)} - k_i^{(g)}$.

Якщо виконується умова $\chi^2_{i,g} \le \chi^2_\alpha(n)$, вважають, що отриманий розподіл досить точно описує випадкову величину.

Якщо розподіл Пуассона неточно описує випадкову величину, можуть бути використані такі розподіли [13]:

– розподіл Пуассона з розширеною кількістю нулів;
– від'ємний біноміальний розподіл;
– геометричний розподіл;
– рівномірний розподіл.

Параметри вказаних розподілів можуть бути знайдені за методом найменших квадратів, методом максимальної правдоподібності або методом моментів. Адекватність отриманих розподілів також перевіряють за критерієм Хі-квадрат.

Розробка імітаційної моделі футбольного сезону

I етап: отримання початкової реалізації модельного сезону

1. Використовуючи функції ймовірності $P_g(X_i)$ та $\breve{P}_d(Y_j)$, моделюємо результат матчу (t_i, t_j) між командами t_i та t_j.

1.1. Генеруємо початкову кількість голів x_i, забитих командою t_i, як значення випадкової величини з функцією ймовірності $P_g(X_i)$:

$$S_{ix_i} \le r_i < S_{i(x_i+1)},$$

де

$$S_{in+1} = \sum_{k=0}^{n} P_g(X_i = k), S_{i0} = 0, S_{i(K_i+1)} = 1,$$

де S_{in} – ймовірність того, що випадкова величина R набуде значення $r_i < S_{in}$; початкове значення кількості голів x_i визначають за нижньою границею S_{ix_i} інтервалу, в який потрапляє випадкове число r_i.

1.2. Враховуючи частоту нічийних матчів домашньої команди в реальному сезоні визначаємо, як завершиться цей матч, – внічию чи ні. Для цього генеруємо бінарну випадкову величину, де «1» відповідає події «матч завершиться внічию», а «0» – «матч завершиться перемогою якоїсь команди», де ймовірність події «1» для команди t_i дорівнює $d(t_i)/(w(t_i) + d(t_i) + l(t_i))$.

1.3. Якщо отримуємо «1», то матч завершується нічиєю і кількість голів, забитих виїзною командою, $y_j = x_i$.

1.4. Якщо отримуємо «0», то генеруємо початкову кількість голів y_j, забиту командою t_j як значення випадкової величини з функцією ймовірності $\breve{P}_d(Y_j)$. Кількості голів y_j визначають через визначення інтервалу, в який потрапляє значення r_j рівномірно розподіленої на інтервалі [0; 1] випадкової величини R:

$$\breve{S}_{jy_j} \le r_j < \breve{S}_{j(y_j+1)},$$

де

$$\breve{S}_{jn+1} = \sum_{k=0}^{n} \breve{P}_d(Y_j = k), \breve{S}_{j0} = 0, \breve{S}_{j(K_j+1)} = 1,$$

де $\breve{S}_{jn}$ – ймовірність того, що випадкова величина R набуде значення $r_j < \breve{S}_{jn}$. Початкове значення кількості голів y_j визначають за нижньою границею $\breve{S}_{jy_j}$ інтервалу, в який потрапляє випадкове число r_j. Якщо при цьому $y_j = x_i$, то збільшуємо результат у сильнішої команди на 1 м'яч, тобто якщо $g \le d$, то збільшуємо x_i на 1, інакше збільшуємо y_i на 1.

2. Повторюємо крок 1 для всіх $N(N-1)$ матчів, де N – це кількість команд, які брали участь у цьому сезоні.

У підсумку маємо модельний сезон футбольного турніру:

$$M = \left\{m_{ij} = (x_i, y_j), i, j = 1, N, i \ne j\right\},$$

де x_i є кількістю голів, забитих командою t_i у матчі m_{ij}, y_j є кількістю голів, забитих командою t_j у матчі m_{ij}.

II етап: обчислення очок команд

1. Обчислюємо попередні очки кожної команди $p(t_i)$, $i = 1, N$:

$$p(t_i) = 3w(t_i) + d(t_i);$$

$$w(t_i) = \left|\left\{m_{ij} | x_i > y_j\right\}\right|;$$

$$d\left(t_i\right) = \left|\left\{m_{ij} \middle| x_i = y_j\right\}\right|.$$

Тут $w(t_i)$ та $d(t_i)$ є, відповідно, кількістю перемог та ігор у нічию для команди ti.

III етап: групування команд за отриманими очками

Для групування команд модельного сезону використовують ту саму методику, яку було використано для групування команд реального сезону.

Аналіз результатів моделювання футбольного сезону

Спочатку було перевірено якість моделювання сезонів на основі етапів I та II розробленої моделі без спотворення результатів матчів. Для дослідження адекватності запропонованого методу імітаційного моделювання було змодельовано 100 сезонів.

На рис. 1 показано абсолютні частоти типів результатів за класами матчів для реального сезону. Також на рис. 1 подано середні значення абсолютних частот типів результатів за 100 змодельованими сезонами, згруповані за класами матчів. У кожному модельному сезоні використовували групування команд, яке було отримано для команд з реального сезону. Як випливає з рис. 1, у змодельованих сезонах загалом зберігається якісний характер залежностей між кількістю домашніх перемог, нічиїх та домашніх поразок за класами матчів реального сезону.

Також було перевірено достовірність різниці між типами результатів змодельованого і реального сезонів для кожного класу матчів за критерієм Колмогорова–Смирнова [15] на рівні значущості $\alpha = 0{,}001$. Критичне значення статистики $\lambda_\alpha = 1{,}95$. Результати застосування критерію для кожного класу матчів наведено в табл. 4. Критерій полягає в перевірці такої нульової гіпотези $H_0 : F_r\left(x\right) = F_m\left(x\right)$, де $F_r(x)$ є емпіричною функцією розподілу вибірки типів результату матчів реального сезону, а $F_m(x)$, відповідно, змодельованого сезону. Гіпотеза H_0 виконується за умови $\lambda \le \lambda_\alpha$. За обраного рівня значущості для усіх класів матчів різницю між типами результатів змодельованого і реального сезонів було визнано статистично незначущою.

Таблиця 4. Результати перевірки критерію Колмогорова–Смирнова на рівні значущості $\alpha = 0{,}001$ для встановлення значущості різниці між типами результатів змодельованого і реального сезонів за класами матчів

Клас матчів	Значення статистики λ
[1,1]	0,339
[1,2]	0,598
[1,3]	0,876
[1,4]	0,102
[2,1]	1,319
[2,2]	0,732
[2,3]	0,077
[2,4]	0,495
[3,1]	1,910
[3,2]	0,311
[3,3]	0,442
[3,4]	0,749
[4,1]	1,824
[4,2]	0,777
[4,3]	0,317
[4,4]	0,073

На рис. 2 показано абсолютні частоти типів результатів реального сезону, а також середні значення абсолютних частот типів результатів за 100 змодельованими сезонами. Сумарне відхилення від реального сезону за типами результату матчу у змодельованих сезонах становить 13 %. Перевіримо достовірність різниці двох наведених вибірок різниць голів за критерієм Колмогорова–Смирнова [15] з тим же рівнем значущості $\alpha = 0{,}001$, що й раніше. Для розглянутих вибірок статистика критерію $\lambda = 0{,}32$, критичне значення статистики $\lambda_\alpha = 1{,}95$. Отже, різниця між розглянутими загальними вибірками типів результату матчів є статистично незначущою на рівні значущості $\alpha = 0{,}001$. Таким чином, за розподілом типів результатів за сезоном змодельований сезон є подібним до реального.

На рис. 3 зображено гістограми різниць голів у матчах реального й одного змодельованого сезонів. Перевіримо достовірність різниці двох наведених вибірок різниць голів за критерієм Колмогорова–Смирнова [15]. Введемо рівень значущості $\alpha = 0{,}001$. Критерій полягає у перевірці такої нульової гіпотези $H_0 : F_r\left(x\right) = F_m\left(x\right)$, де $F_r(x)$ є емпіричною функцією розподілу вибірки різниць голів реального сезону, а $F_m(x)$, відповідно, змодельованого сезону. Для розглянутих вибірок статистика критерію $\lambda = 0{,}58$, критичне значення статистики $\lambda_\alpha = 1{,}95$. Отже, різниця між розглянутими вибірками різниць голів є статистично незначущою на рівні значущості $\alpha = 0{,}001$. Таким чином, за розподілом різниць голів усіх матчів змодельований сезон є подібним до реального сезону.

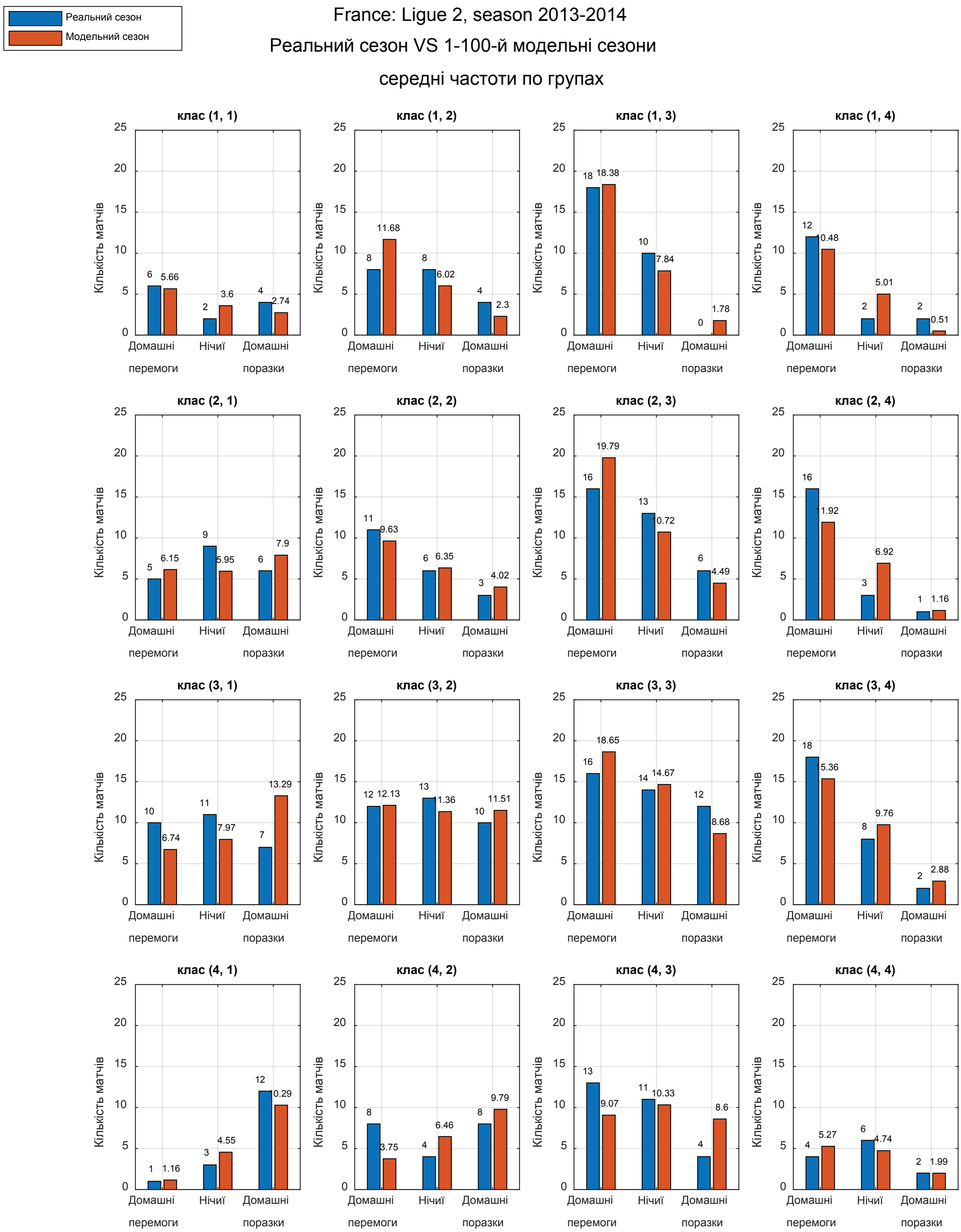

Рис. 1. Розподіл типів результатів матчів за класами матчів у змодельованих сезонах

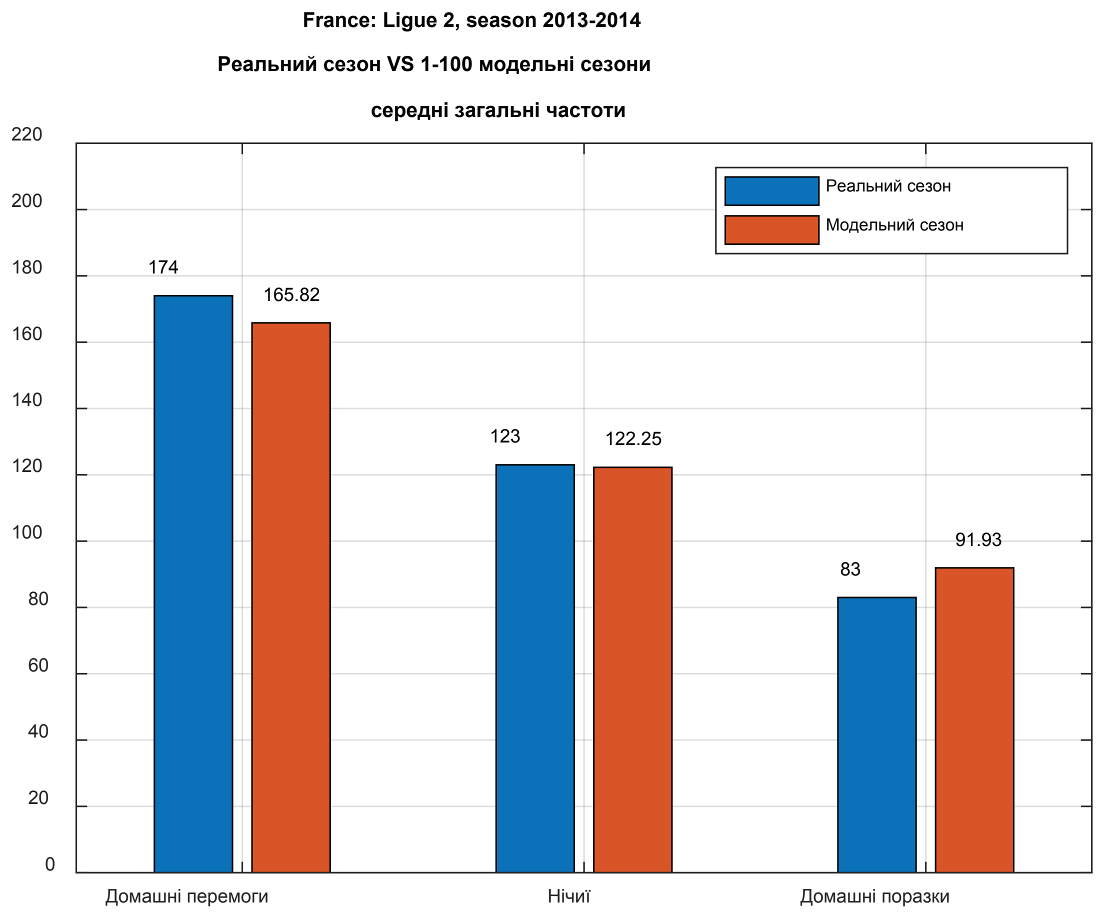

Рис. 2. Розподіл типів результатів матчів загалом у змодельованих сезонах

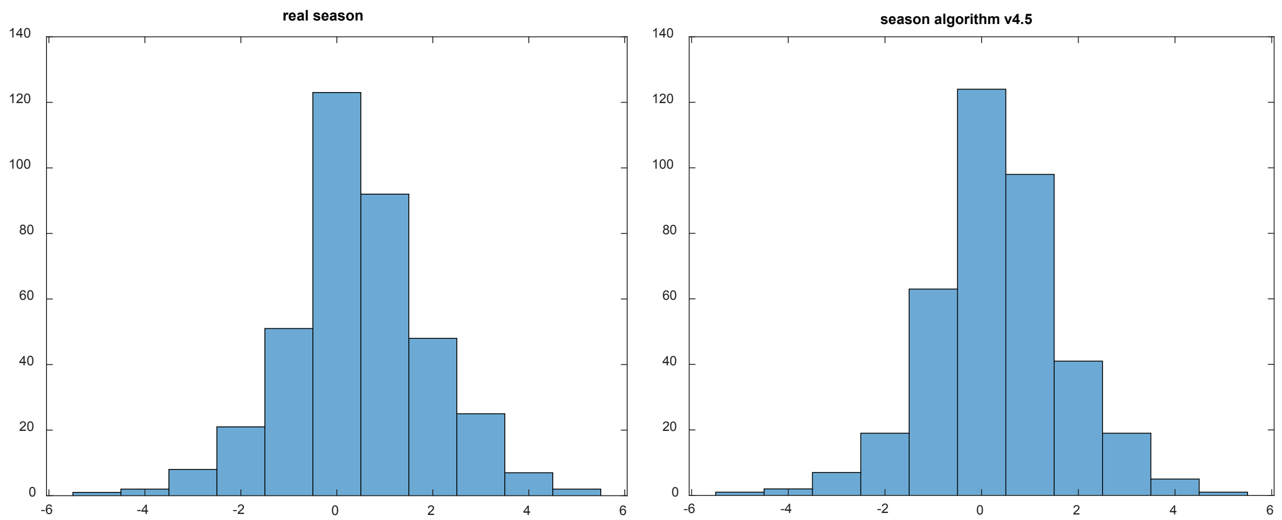

Рис. 3. Розподіл різниць голів реального і змодельованого сезонів

Алгоритм моделювання договірних матчів, пов'язаних із заробітком на ставках

Договірний матч, щоб на ставках на ньому могли заробити зловмисники, має різнитися від очікуваного результату, тобто мати аномальний характер. Це його найважливіша властивість. Тому його значення не повинно перебувати в області очікуваних значень результату матча, пов'язаних із силою команд. Запропоновано ввести порогове значення ймовірності p_A, яку не мають перевищувати аномальні різниці голів матчів класу і, таким чином, серед яких можуть міститись результати договірних матчів. Значення ймовірності p_A запропоновано обирати в діапазоні $0 < p_A < 0{,}4$.

Алгоритм моделювання договірних матчів, пов'язаних із заробітком на ставках, складається з таких етапів:

1. З використанням отриманих модельних сезонів турніру будують гістограми різниць голів усіх матчів окремо для кожного класу. Приклади гістограм, отриманих за 100 модельними сезонами, показано на рис. 4. Як випливає з рис. 4, гістограми мають очікувані закономірності в результатах матчів (сильніші команди мають кращі результати у грі зі слабшими командами; вдома команди грають краще, ніж на виїзді).

2. Використовуючи гістограму різниць голів, збудовану за класами матчів за 100 сезонами, визначають множину нормальних різниць голів $D_{ij}^{(N)}$ за допомогою такого ітераційного алгоритму: спочатку $D_{ij}^{(N)}$ є пустою множиною, тобто $D_{ij}^{(N)} = \varnothing$.

2.1. Обирають значення різниці голів $\tilde{d}$, яке за гістограмою класу матчів (i, j) має найбільшу частоту появи h_d серед тих значень d, які $d \notin D_{ij}^{(N)}$.

2.2. Додають значення $\tilde{d}$ до множини $D_{ij}^{(N)}$.

2.3. Обчислюють сумарну частоту появи усіх значень із множини $D_{ij}^{(N)}$:

$$p_{ij}^{(N)} = \sum_{d \in D_{ij}^{(N)}} h_d .$$

2.4. Якщо $p_{ij}^{(N)} \geq 1 - p_A$, переходять на крок 2.5. Інакше переходять на крок 2.1.

2.5. Значення можливих різниць голів $d^* \notin D_{ij}^{(N)}$ утворюють множину $D_{ij}^{(A)}$ аномальних різниць класу матчів (i, j).

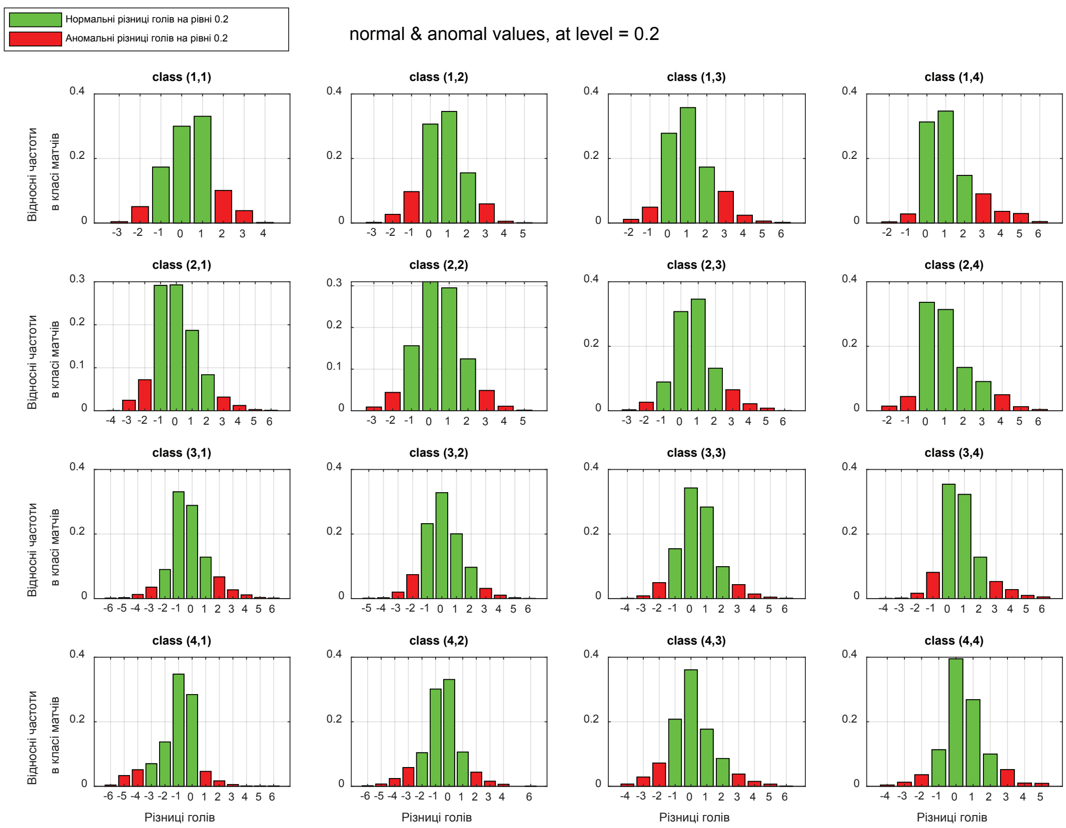

Рис. 4. Гістограми різниць голів за кожним класом матчів

Значення множини $D_{ij}^{(A)}$ використовують для моделювання результату договірного матчу.

Для прикладу на рис. 4 показано нормальні (зелений колір) й аномальні (червоний колір) різниці голів за кожним класом матчів за рівня аномальності $p_A = 0{,}2$. Результати вибору значень за гістограмою з рис. 4 за рівня аномальності $p_A = 0{,}2$ відображено в табл. 5.

Таблиця 5. Нормальні й аномальні різниці м'ячів

Клас матчів	Аномальні різниці м'ячів	Нормальні різниці м'ячів
[1,1]	{–3, –2, 2, 3, 4}	{–1, 0, 1}
[1,2]	{–3, –2, –1, 3, 4, 5}	{0, 1, 2}
[1,3]	{–2, –1, 3, 4, 5, 6}	{0, 1, 2}
[1,4]	{–2, –1, 3, 4, 5, 6}	{0, 1, 2}
[2,1]	{–4, –3, –2, 3, 4, 5, 6}	{–1, 0, 1, 2}
[2,2]	{–3, –2, 3, 4, 5}	{–1, 0, 1, 2}
[2,3]	{–3, –2, 3, 4, 5, 6}	{–1, 0, 1, 2}
[2,4]	{–2, –1, 4, 5, 6}	{0, 1, 2, 3}
[3,1]	{–6, –5, –4, –3, 2, 3, 4, 5, 6}	{–2, –1, 0, 1}
[3,2]	{–5, –4, –3, –2, 3, 4, 5, 6}	{–1, 0, 1, 2}
[3,3]	{–4, –3, –2, 3, 4, 5, 6}	{–1, 0, 1, 2}
[3,4]	{–4, –3, –2, –1, 3, 4, 5, 6}	{0, 1, 2}
[4,1]	{–6, –5, –4, 1, 2, 3, 4, 5, 6}	{–3, –2, –1, 0}
[4,2]	{–6, –5, –4, –3, 2, 3, 4, 6}	{–2, –1, 0, 1}
[4,3]	{–4, –3, –2, 3, 4, 5, 6}	{–1, 0, 1, 2}
[4,4]	{–4, –3, –2, 3, 4, 5}	{–1, 0, 1, 2}

3. Вибирають N матчів сезону турніру випадковим чином, які будуть перетворюватись на договірні. Далі розглядають усі дії на прикладі одного такого матчу.

4. Визначають клас матчів, до якого належить аналізований матч.

5. Якщо результат обраного матчу не міститься в області очікуваних значень результату матчів цього класу матчів (тобто цей результат і так схожий на договірний матч), то випадковим чином обирають інший матч. Інакше переходять до п. 6.

6. Різницю голів обирають випадково із множини аномальних значень різниць голів для відповідного класу матчів (табл. 5). Під час цього випадкового вибору використовують ймовірності, пропорційні частотам аномальних значень з гістограми різниць голів для відповідного класу матчів.

7. Рахунок договірного матчу обирають випадково з відповідного стовпця таблиці рахунків матчів (табл. 6), де номер стовпця таблиці дорівнює обраній різниці голів договірного матчу. Під час цього випадкового вибору використовують ймовірності на основі арифметичної прогресії з $a_1 = 1$ та $a_n = 0{,}1, n \ge 2$ (рис. 5):

$$P(j) = \frac{a_j}{S} = \frac{a_1 + \frac{(a_n - a_1)}{(n-1)}(j-1)}{\frac{(a_1 + a_n)n}{2}}$$

$$= \frac{1 + \frac{(0{,}1-1)}{(n-1)}(j-1)}{\frac{(1+0{,}1)n}{2}} = \frac{1 - \frac{0{,}9(j-1)}{(n-1)}}{0{,}55n}$$

$$= \frac{(n-1-0{,}9j+0{,}9)}{0{,}55n(n-1)} = \frac{(n-0{,}1-0{,}9j)}{0{,}55n(n-1)}$$

$$= \frac{(n-1+1-0{,}1-0{,}9j)}{0{,}55n(n-1)} = \frac{1}{0{,}55n} + \frac{(0{,}9-0{,}9j)}{0{,}55n(n-1)},$$

$$1 \le j \le n, n \ge 2;$$

$$P(1) = 1, n = 1,$$

де n — кількість варіантів рахунку в обраному стовпці таблиці рахунків матчів, j — порядковий номер рахунку в обраному стовпці таблиці рахунків матчів, S — сума перших n членів арифметичної прогресії.

Якщо $n = 1$, обрана функція ймовірності є сталою функцією зі значенням $P(1) = 1$. Для $n \ge 2$ обрана функція ймовірності є спадною ступінчатою функцією з початковим значенням ймовірності $P(1) = \frac{1}{0{,}55n}$ і кроком. $\frac{0{,}9(1-j)}{0{,}55n(n-1)}$. Обрана функція ймовірності забезпечує поступове зменшення ймовірності зі збільшенням порядкового номера результату матчу. Елементи стовпця таблиці рахунків матчів розміщені у порядку збільшення кількостей голів, забитих командами під час матчу. Отже, чим більше голів кожна команда заб'є під час гри, тим менша ймовірність обрання такого варіанта результату матчу як договірного. Визначення апріорно невідомих ймовірностей як членів арифметичної прогресії використовують у задачах дослідження операцій як один із методів подолання апріорної невизначеності.

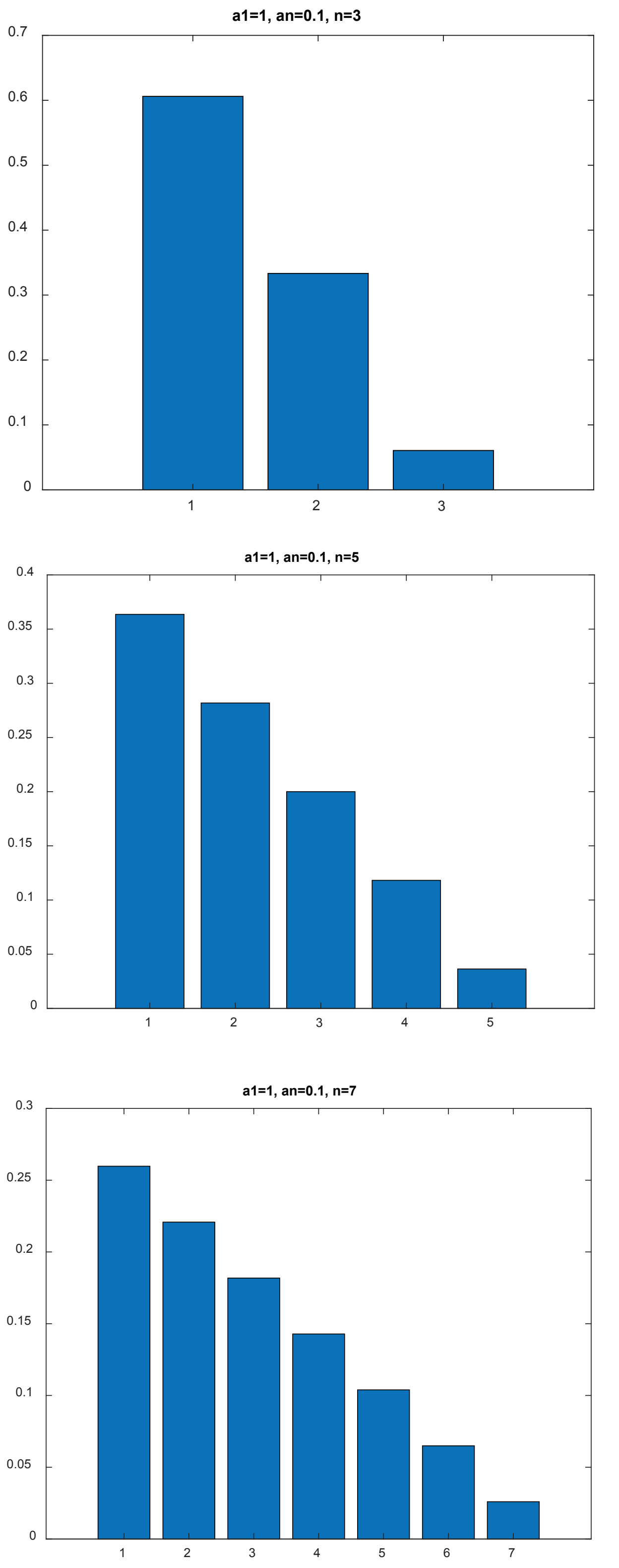

Рис. 5. Ймовірності на основі арифметичної прогресії з $a_1 = 1$ і $a_n = 0{,}1$ для $n \in \{3, 5, 7\}$

Таблиця 6. Рахунки матчів для всіх класів матчів

	Різниця голів матчу						
	0	1	2	3	4	5	6
Рахунок матчу	0:0	1:0	2:0	3:0	4:0	5:0	6:0
	1:1	2:1	3:1	4:1	5:1	6:1	
	2:2	3:2	4:2	5:2	6:2		
	3:3	4:3	5:3	6:3			
	4:4	5:4	6:4				
	5:5	6:5					
	6:6						

Результати моделювання матчів із фіксованим результатом продемонстровано на прикладі одного із змодельованих сезонів (табл. 7). Випадковим чином у змодельованому сезоні було обрано $T_s = 10$ матчів. Було використано рівень аномальності $p_A = 0{,}2$. Для кожного матчу сформовано результат за алгоритмом формування договірних матчів в імітаційній моделі відповідно до аномальних різниць, визначених для кожного класу матчів у табл. 1.

Таблиця 7. Приклади утворених договірних матчів у модельному сезоні

Домашня команда	Виїзна команда	Оновлений результат	Початковий результат
‘team 3’	‘team 13’	4:1	0:0
‘team 2’	‘team 15’	6:2	3:2
‘team 4’	‘team 19’	0:1	3:2
‘team 9’	‘team 7’	0:2	2:2
‘team 11’	‘team 7’	5:1	2:3
‘team 16’	‘team 19’	3:0	1:1
‘team 17’	‘team 2’	4:1	0:0
‘team 17’	‘team 7’	5:2	1:2
‘team 17’	‘team 15’	3:0	4:4
‘team 17’	‘team 19’	4:0	4:3

В отриманих результатах є такі, які можуть виглядати як очікувані, наприклад, 0:1 чи 0:2, але ж для відповідних класів матчів такі результати розглядають як аномальні.

Висновки

Для дослідження ефективності методів виявлення підозрілих щодо фіксованого результату матчів і їх порівняльного аналізу актуальним завданням є розроблення імітаційної моделі футбольного сезону з матчами з фіксованим результатом.

Особливістю розробленої імітаційної моделі є те, що команди поділяють на групи, які враховують їх силу за загальними очками в сезоні.

Відповідно, ймовірність забиття голів командою під час матчу розраховують за групами, а не за всім сезоном. Також під час розрахунку цієї ймовірності враховують тип гри — домашня або виїзна. Це дозволяє врахувати особливості гри домашньої та виїзної команди.

Як випливає з розподілів типів результатів матчів, згрупованих за класами матчів, у змодельованих сезонах у середньому зберігається характер залежностей між кількістю домашніх перемог, нічиїх, виїзних перемог, притаманних класам матчів реального сезону.

За загальними розподілами типів результатів матчів та різницями голів усіх матчів змодельований сезон є подібним до реального сезону. За критерієм Колмогорова–Смирнова, різниця наведених розподілів на рівні значущості 0,001 є статистично незначущою.

Гістограми різниць голів по кожному класу матчів мають очікувані закономірності в результатах матчів: сильніші команди мають кращі результати у грі зі слабшими командами; вдома команди грають краще, ніж на виїзді.

Розглянуту імітаційну модель може бути узагальнено для врахування інших параметрів футбольних матчів, на які приймаються ставки — кількість призначених пенальті, кількість попереджень і вилучень тощо.

References

[1] "Resource Guide on Good Practices in the Investigation of Match-Fixing", *United Nations : Office on Drugs and Crime*. //www.unodc.org/unodc/en/safeguardingsport/publications/match-fixing.html (accessed Jan. 05, 2022).

[2] D. Forrest, and I.G. McHale, "Using statistics to detect match fixing in sport", *IMA Journal of Management Mathematics*, Vol. 30, No 4, pp. 431–449, Sep. 2019. doi: 10.1093/imaman/dpz008

[3] S. Anfilets *et al.*, "DEEP MULTILAYER NEURAL NETWORK FOR PREDICTING THE WINNER OF FOOTBALL MATCHES", *International Journal of Computing*, pp. 70–77, Mar. 2020. doi: 10.47839/ijc.19.1.1695

[4] N. Razali, A. Mustapha, F.A. Yatim, and R.A. Aziz, "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL) ", *IOP Conf. Ser.: Mater. Sci. Eng.*, Vol. 226, No 1, p. 012099, Aug. 2017. doi: 10.1088/1757-899X/226/1/012099

[5] T. Narizuka, Y. Yamazaki, and K. Takizawa, "Space evaluation in football games via field weighting based on tracking data", *Sci Rep*, Vol. 11, No 1, p. 5509, Mar. 2021. doi: 10.1038/s41598-021-84939-7

[6] Laxhammar, R., & Falkman, G. (2011). Sequential Conformal Anomaly Detection in trajectories based on Hausdorff distance. 14th International Conference on Information Fusion, 1–8.

[7] Ho, S.-S., Schofield, M., Sun, B., Snouffer, J., & Kirschner, J. (2019). A Martingale-Based Approach for Flight Behavior Anomaly Detection. *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, 43–52. https://doi.org/10.1109/MDM.2019.00-75

[8] focus.ua, "Dirty games. Match-fixing in Ukrainian football", *FOCUS*, Mar. 05, 2013. https://focus.ua/ukraine/263259 (accessed Jan. 05, 2022).

[9] B. Constandt, and E. Manoli, "*Understanding match-fixing in sport: Theory and practice*". 2022.

[10] "Matchs truquйs : coup de filet dans le milieu du football professionnel", *leparisien.fr*, No 18, 2014. https://www.leparisien.fr/faits-divers/corruption-coup-de-filet-dans-le-milieu-du-football-professionnel-18-11-2014-4301229.php (accessed Jan. 05, 2022).

[11] L. Chami, "Matchs truquйs de Ligue 2 : 18 mois ferme pour les anciens dirigeants nomois", *leparisien.fr*, Sep. 13, 2018. https://www.leparisien.fr/sports/football/matchs-truques-de-l2-18-mois-ferme-pour-les-anciens-dirigeants-nimois-13-09-2018-7887090.php (accessed Jan. 05, 2022).

[12] K.P. Murphy, *Probabilistic Machine Learning: Advanced Topics*", MIT Press, 2021.

[13] S.D. Langan, "*Predict Football Matches: Using Spreadsheet Models to Become a Winning Sports Bettor*" (Kindle Edition), 2013, 379 p.

[14] S.M. Ross, "*Introductory Statistics*", Academic Press, 2017.

[15] N.Sh. Kremer, "*Theory of Probability and Mathematical Statistics*", UNITY-DANA, 2001.

O.R. Chertov, I.S. Zhuk

SOCCER SEASON SIMULATION WITH FIXED MATCHES

Background. Football is a huge industry comparable to traditional economic sectors. In numerical terms, it is tens of billions of dollars. But one of the most important problems that this industry faces is match fixing (matches with a fixed result).

The most effective way to prevent this threat is a multilateral approach combined with measures aimed at expanding the potential of law enforcement agencies and sports organizations. One of the components of this approach is the use of mathematical methods for identifying suspicious match-fixing results.

Objective. The purpose of the paper is to develop an algorithm for modelling fixed matches related to earnings on bets, which is used to form matches which results are different from the expected ones and can be considered as anomalous.

Methods. Calculation of probabilities of scoring goals by teams during the game based on real data of the season; development of a simulation model of a football season without fixed matches and its analysis by means of statistical modelling; development of an algorithm for modelling fixed matches related to earnings on bets and its analysis.

Results. A simulation model of a football season has been developed, which allows, using probability distributions of the number of goals scored by teams during home or away games obtained on real data, simulating the results of matches, taking into account the strength of the teams and the game type, as well as to simulate the situations of a “fixed” match, replacing the current results. In terms of the overall distributions of match result types and goal differences of all matches, the simulated season is similar to the real season. According to the Kolmogorov-Smirnov test, the difference between the given distributions at the significance level of 0.001 is statistically insignificant.

Conclusions. The developed simulation model of a football season can be used to study the effectiveness of methods for detecting fixed matches and their comparative analysis.

Keywords: football season; team grouping; Poisson distribution; goal difference histogram; type of match result.

Рекомендована Радою
факультету прикладної математики
КПІ ім. Ігоря Сікорського

Надійшла до редакції
11 січня 2022 року

Прийнята до публікації
27 червня 2022 року