Politechnika Wrocławska

Wydział Matematyki
Kierunek studiów: Matematyka Stosowana
Specjalność: –

Praca dyplomowa – inżynierska

Zastosowanie metod bootstrapowych do prognozowania wyników w zawodach sportowych

Michał Cerazy

słowa kluczowe:
bootstrap, symulacje Monte Carlo, estymacja rozkładu, symulacja wyników koszykarskich

krótkie streszczenie:
Praca dotyczy aktualnego problemu z matematyki stosowanej, a mianowicie zaprojektowanie symulatora ligi NBA przy użyciu znanych technik matematycznych/statystycznych. Celem pracy było stworzenie symulatora rozgrywek w lidze NBA przy użyciu metod bootstrapowych. Zaproponowano dwa modele symulacji sezonu zasadniczego, bazujące na rezultatach starć z poprzednich lat, oraz jeden do symulowania rozgrywek pucharowych. Wynikiem pracy jest oprogramowanie do symulacji rozgrywek w lidze NBA, wykonane przy pomocy języka Python, umożliwiające prognozowanie wyników w następnym roku.

Opiekun pracy dyplomowej	dr hab. inż. Krzysztof Burnecki	.....	.....
	Tytuł/stopień naukowy/imię i nazwisko	ocena	podpis

*Do celów archiwalnych pracę dyplomową zakwalifikowano do:**

a) kategorii A (akta wieczyste)

b) kategorii BE 50 (po 50 latach podlegające ekspertyzie)

** niepotrzebne skreślić*

pieczątka wydziałowa

Wrocław, rok 2019

Wrocław University
of Science and Technology

Faculty of Pure and Applied Mathematics
Field of study: Applied Mathematics
Specialty: –

Engineering Thesis

Application of bootstrap methods to the forecasting of sporting event results

Michał Cerazy

keywords:
bootstrap, Monte Carlo simulations, fitting of distribution, simulation of basketball game outcomes

short summary:
The main subject of the thesis concerns the recent problem from applied mathematics field — creation of NBA league simulator using known mathematical/statistical techniques. The goal was to design a simulator based on bootstrap methods. For this thesis purposes two models of regular season simulation, based on past years games outcomes, and one one model for playoff phase were designed. The result of this thesis is a Python language software capable of predicting outcome of an NBA season in the next year.

Supervisor	dr hab. inż. Krzysztof Burnecki	.....	.....
	Title/degree/name and surname	grade	signature

*For the purposes of archival thesis qualified to:**

a) category A (perpetual files)

b) category BE 50 (subject to expertise after 50 years)

** delete as appropriate*

stamp of the faculty

Wrocław, 2019

Spis treści

Wstęp

Dzięki powszechnemu dostępowi do internetu i rozpowszechnieniu kultury masowej amerykańska liga koszykarska NBA zyskała popularność na całym świecie, przyciągając do siebie najlepszych graczy i masę fanów. Z powodu nieprzewidywalności i złożoności tego sportu podejmowano wiele prób przewidywania wyników rozgrywek, które często toczyły się inaczej, niż by zakładano (najlepszym tego przykładem może być sezon 2003/2004, kiedy to nisko notowani Detroit Pistons pokonali faworytów, czyli Los Angeles Lakers). Trudność w przewidzeniu wyników wynika z powodu licznych i często katastrofalnych kontuzji. Problematyczne w tym są również zasady ligi — dozwolone są w niej wymiany zawodników między klubami, podpisywanie umów z nowymi koszykarzami, czy nabory do ligi, w których najsłabsze drużyny mają najwyższe szanse na pierwszeństwo wyboru nowych zawodników chcących dołączyć do NBA (z tego powodu wiele organizacji celowo przegrywa swoje mecze).

Celem pracy inżynierskiej jest próba przewidzenia rezultatów wybranego sezonu ligi NBA przy pomocy niecodziennego modelu, używając wyłącznie informacji o wynikach poszczególnych drużyn z poprzednich sezonów — w przeciwieństwie do większości istniejących już modeli indywidualny wpływ graczy na przebieg spotkań nie jest rozpatrywany. Na potrzeby tego zadania zaprojektowano kilka algorytmów opierających się na metodzie bootstrap, sprawdzono ich skuteczność dla kilku wybranych sezonów w zależności od okresu używanych danych i nadanych wag, a następnie przy pomocy najskuteczniejszych modeli dokonano predykcji wyników trwającego sezonu 2018/2019. Pierwsza część pracy przybliża zasady ligi, system rozgrywek i tendencje panujące w niej. Drugi rozdział objaśnia teorię wykorzystaną w pracy, a mianowicie generatory liczb losowych, symulacje Monte Carlo i próbę bootstrap. Następny z rozdziałów to opis przygotowanych danych, wyprowadzonych własnych oznaczeń, jak i zaproponowanych modeli predykcji rozgrywek. Kolejna część pracy zawiera porównanie wyników przeprowadzonych symulacji, wybór najlepszego modelu, oraz prognozę na trwający obecnie sezon 2018/2019.

Do przeprowadzania symulacji rozgrywek korzystano z języka programistycznego Python. Pomocne okazały się pakiety takie jak *Matplotlib*, *Pandas*, *Numpy*, *Scipy*, *Random* oraz *Seaborn*. Dane przygotowano przy pomocy pakietu MS Excel oraz dodatku Power Query. Po obliczeniu skumulowanej liczby wygranych dla każdej drużyny dokonano przekształcenia jej do prawdopodobieństwa przy pomocy makra napisanego w języku Visual Basic for Applications.

Rozdział 1

Opis ligi NBA

NBA (National Basketball Association) została założona 6 czerwca 1946 roku. Pierwotnie była znana jako Basketball Association of America i składała się z 11 zespołów, a swoją obecną nazwę zyskała w roku 1949, kiedy to wchłonęła konkurencyjną National Basketball League. Kolejna fuzja z inną ligą miała miejsce w 1976 roku, a mianowicie połączenie z bardziej widowiskową American Basketball Association. Po tym wydarzeniu NBA znacznie zyskała na atrakcyjności — z ABA zaczerpnięto pomysł rzutów za trzy punkty oraz organizację konkursu wsadów, jak i przyjęto cztery dodatkowe kluby [16]. Od 2004 roku w lidze gra 30 zespołów, 29 ze Stanów Zjednoczonych i 1 z Kanady. Liga podzielona jest na dwie konferencje po 15 drużyn, te natomiast składają się z dywizji po 5 organizacji. Szczegółowy podział na konferencje i dywizje, oraz nazwy wszystkich drużyn zawarte zostały w tabelach 1.1 i 1.2, natomiast dokładne rozmieszczenie na mapie kontynentu znajduje się na rysunku 1.1 [14].

Dodatkowo od 2004 roku niektóre kluby zmieniły swoje nazwy lub lokalizacje. W niektórych historycznych zestawieniach lub zbiorach danych mogą widnieć jako (podane w formacie nazwa obecna — poprzednia):

- Charlotte Hornets — Charlotte Bobcats
- Brooklyn Nets — New Jersey Nets
- Oklahoma City Thunder — Seattle SuperSonics
- New Orleans Pelicans — New Orleans Hornets

Sezon w NBA składa się z dwóch części: zasadniczej i następującej po niej pucharowej (playoffs). W sezonie zasadniczym każda drużyna rozgrywa 82 mecze, grając z każdym innym zespołem od 2 do 4 gier. Przykładowa tabela z wynikami na zakończenie sezonu

Tabela 1.1: Drużyny konferencji Wschodniej

konferencja Wschodnia		
Atlantic Division	**Southeast Division**	**Central Division**
Boston Celtics (BOS)	Atlanta Hawks (ATL)	Chicago Bulls (CHI)
Brooklyn Nets (BRK)	Charlotte Hornets (CHO)	Cleveland Cavaliers (CLE)
New York Knicks (NYK)	Miami Heat (MIA)	Detroit Pistons (DET)
Philadelphia 76ers (PHI)	Orlando Magic (ORL)	Indiana Pacers (IND)
Toronto Raptors (TOR)	Washington Wizards (WAS)	Milwaukee Bucks (MIL)

Rysunek 1.1: Rozmieszczenie drużyn NBA [14]

Tabela 1.2: Drużyny konferencji Zachodniej

konferencja Zachodnia		
Northwest Division	**Southwest Division**	**Pacific Division**
Denver Nuggets (DEN)	Dallas Mavericks (DAL)	Golden State Warriors (GSW)
Minnesota Timberwolves (MIN)	Houston Rockets (HOU)	Los Angeles Clippers (LAC)
Oklahoma City Thunder (OKC)	Memphis Grizzlies (MEM)	Los Angeles Lakers (LAL)
Portland Trail Blazers (POR)	New Orleans Pelicans (NOP)	Phoenix Suns (PHX)
Utah Jazz (UTA)	San Antonio Spurs (SAS)	Sacramento Kings (SAC)

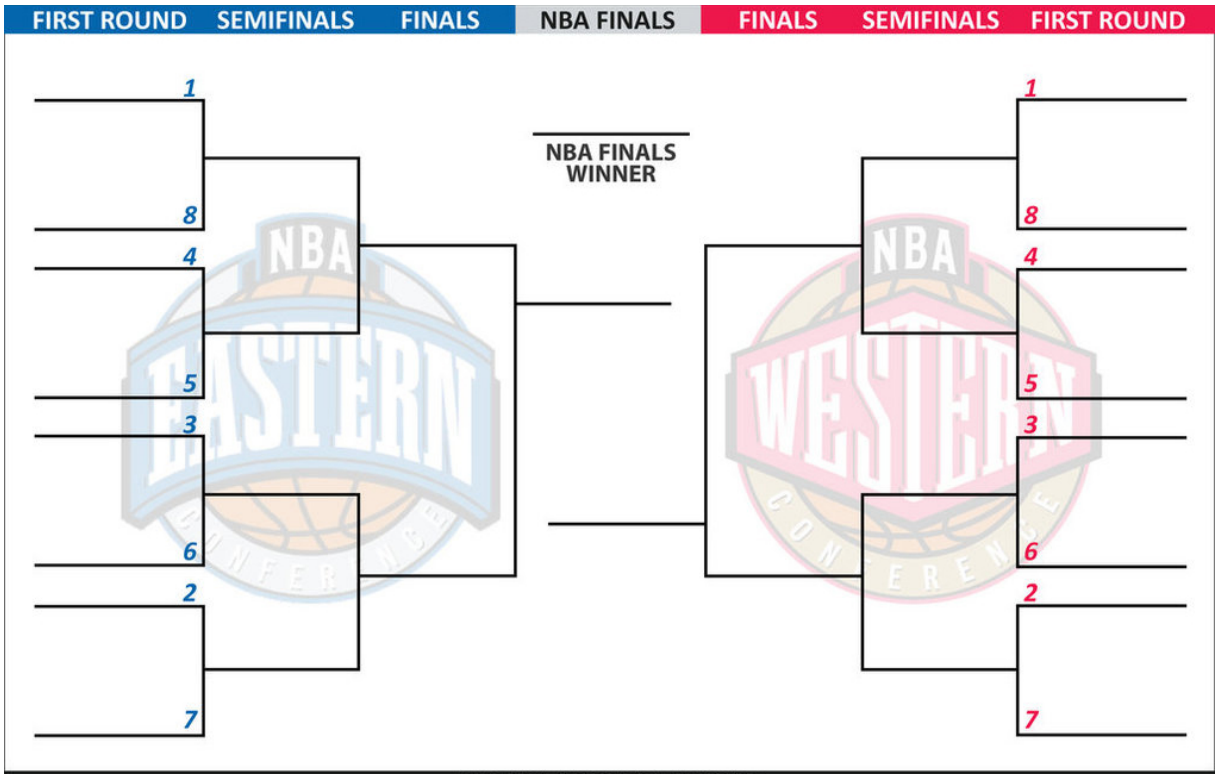
została zawarta w tabeli 1.3. Terminarz rozgrywek wyznaczany jest wedle następujących reguł:

1. drużyny z różnych konferencji grają ze sobą 2 spotkania (1 na wyjeździe i 1 na własnym boisku),

2. drużyny z tej samej dywizji grają ze sobą 4 spotkania (2 na wyjeździe i 2 na własnym boisku),

3. drużyny z tej samej konferencji oraz różnych dywizji grają ze sobą 3 albo 4 spotkania (przynajmniej po jednym na wyjeździe i własnym boisku).

Mecze koszykówki nie mogą zakończyć się remisem (w razie remisu po regulaminowym czasie gry rozgrywa się dogrywki aż do wyłonienia zwycięzcy). Po zakończeniu sezonu następuje wspomniana wyżej faza pucharowa; wchodzi do niej po 8 najlepszych zespołów



Rysunek 1.2: Grafika obrazująca system rozgrywek pucharowych [13]

z każdej konferencji (w razie takiej samej ilości zwycięstw dla obu zespołów decydują wyniki ich bezpośrednich spotkań). W tej fazie drużyny grają ze sobą maksymalnie 7 meczów, czyli zespół, który pierwszy wygra 4 mecze, przechodzi do następnego etapu. W fazie playoff jasno zdefiniowane są lokalizacje odgrywania spotkań — lepszy bilans zwycięstw w sezonie zasadniczym skutkuje przewagą parkietu. Seria spotkań grana jest w formacie 2–2–1–1–1, czyli mecze numer 1, 2, 5 i 7 rozgrywane są u lepszej z drużyn. Przy doborze przeciwników bierze się pod uwagę pozycję w tabeli konferencji: drużyna z miejsca pierwszego gra z zespołem na ósmym miejscu, druga z siódmą, i tak dalej. Zwycięzca serii przechodzi do następnego etapu z czterema drużynami, po którym następują finały konferencji — najlepsze drużyny ze swoich grup spotykają się w finałach NBA. Dla lepszego zrozumienia systemu rozgrywek pucharowych na rysunku 1.2 zamieszczono tzw. „drzewko playoff" [13], czyli grafikę oddającą przebieg tej fazy.

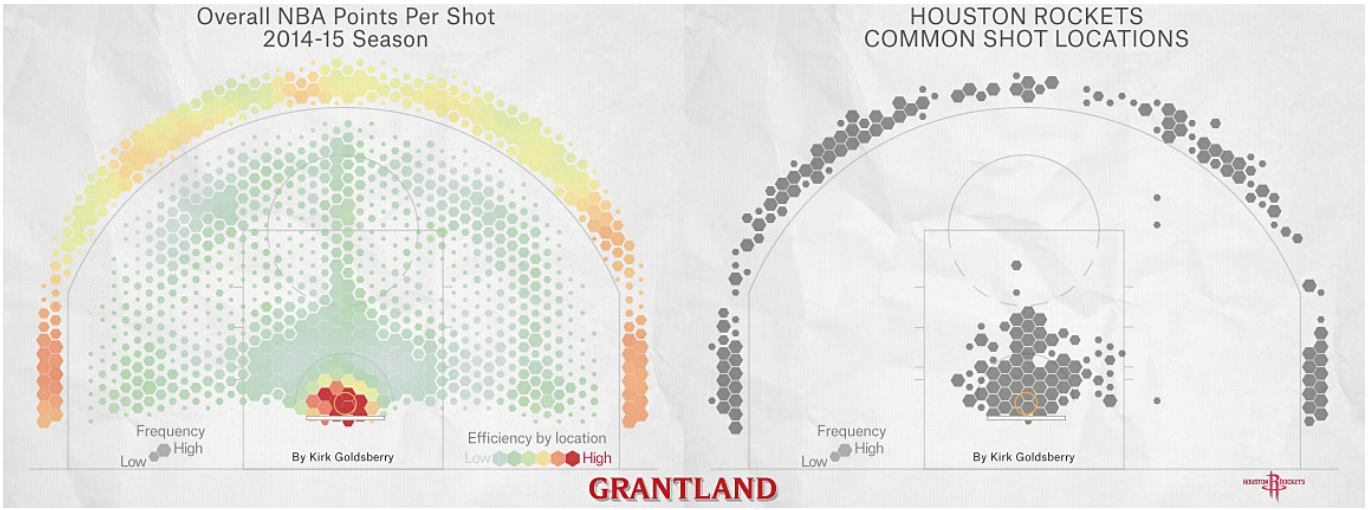
Każdego lata, po zakończeniu fazy playoff, ma miejsce nabór młodych zawodników do NBA (tzw. Draft). Jest to okazja dla graczy uniwersyteckich, którzy chcieliby rozpocząć karierę zawodową, lub zawodników grających w innych ligach. Zasady Draftu są proste: każdy sportowiec może się do niego zgłosić tylko raz oraz aby móc zagrać w lidze trzeba do niego obowiązkowo przystąpić — jest to świetny sposób na utrzymanie balansu w NBA, ponieważ najlepsze zespoły mają najmniejsze szanse na otrzymanie praw do utalentowanych, młodych zawodników. Nabór nagradza najsłabsze kluby, ponieważ zespół z najgorszym wynikiem w sezonie zasadniczym ma największe szanse na otrzymanie pierwszego wyboru, druga najgorsza drużyna ma drugie co do wielkości szanse, i tak dalej. Kolejność wyborów jest ustalana poprzez losowanie, dlatego często w historii zdarzało się, że najsłabszy zespół nie uzyskał pierwszeństwa. Niestety, coraz częstszym zjawiskiem w NBA staje się tak zwane „tankowanie", czyli celowe przegrywanie w celu uzyskania najwyższego możliwie wyboru w Drafcie. Zespoły na poziomie średniej ligowej, mając świadomość, że nie są w

Tabela 1.3: Liczby zwycięstw drużyn w wybranych sezonach

Drużyna	**Zwycięstwa w sezonie 14/15**	**Zwycięstwa w sezonie 17/18**
Atlanta Hawks	60	24
Boston Celtics	40	55
Brooklyn Nets	38	28
Charlotte Hornets	33	36
Chicago Bulls	50	27
Cleveland Cavaliers	53	50
Dallas Mavericks	50	24
Denver Nuggets	30	46
Detroit Pistons	32	39
Golden State Warriors	67	58
Houston Rockets	56	65
Indiana Pacers	38	48
Los Angeles Clippers	56	42
Los Angeles Lakers	21	35
Memphis Grizzlies	55	22
Miami Heat	37	44
Milwaukee Bucks	41	44
Minnesota Timberwolves	16	47
New Orleans Pelicans	45	48
New York Knicks	17	29
Oklahoma City Thunder	45	48
Orlando Magic	25	17
Philadelphia 76ers	18	52
Phoenix Suns	39	21
Portland Trail Blazers	51	49
Sacramento Kings	29	27
San Antonio Spurs	55	47
Toronto Raptors	49	59
Utah Jazz	38	48
Washington Wizards	46	43

stanie walczyć o wysokie cele, decydują się na wejście w przebudowę i kilkuletnie poświęcanie zwycięstw na rzecz rozwoju młodzieży. Podczas porównywania wyników symulacji z rzeczywistymi w tej pracy, wielokrotnie można było zauważyć, które zespoły skupiły się na walce o młodych zawodników, a które o tytuły — najlepszym wskaźnikiem tego jest liczba wygranych.

Po sukcesie rewolucji statystycznej w baseballu pod koniec lat dziewięćdziesiątych, statystyka i analityka sportowa wywiera coraz większy wpływ na sport Stanów Zjednoczonych. Doprowadziło to do tego, że obecnie każda drużyna zatrudnia sztab analityków, którzy badają wpływ czynników obecnych na parkiecie na losy meczu. Idealnym przykładem wpływu statystyki na styl gry zespołu są Houston Rockets, którzy od sezonu 2014/2015 zrezygnowali z rzutów z półdystansu na rzecz rzutów za 3 punkty i tych spod obręczy. Doprowadziło to do sytuacji, w której 82% ich rzutów było oddawanych z tych pozycji, podczas gdy druga najlepsza drużyna w tym aspekcie osiągała poziom 71% [15]. Wizualizacja tego systemu znajduje się na rysunku 1.3. Poza analizowaniem przebiegu gry, analitycy



Rysunek 1.3: Decyzje rzutowe Houston i reszty ligi [15]

oraz statystycy podejmują próby przewidywania wyników sezonu, co staje się bardzo istotne w razie kontuzji lub wymiany gracza [9] [8]. Niemal wszystkie zaproponowane dotąd modele symulacji rozgrywek opierają się na statystykach zawodników, ich wpływie na atak i obronę, udziale w zwycięstwach, oraz ulubionych pozycjach rzutowych.

Rozdział 2

Podstawowe pojęcia wykorzystane w pracy

Wykorzystana w tej pracy wiedza teoretyczna opiera się na rozkładzie jednostajnym, generowaniu zmiennej losowej z tego rozkładu, a także estymacji parametru przy wykorzystaniu symulacji Monte Carlo i bootstrap. W przypadku zasady bootstrap, w celu wybrania lepszego podejścia do tego zadania, dokonano porównania dwóch jej modeli: parametrycznego i nieparametrycznego.

2.1 Rozkład jednostajny

Definicja 2.1. Ciągła zmienna losowa X ma rozkład jednostajny o parametrach a i b, takich że $a, b \in \mathbb{R}$ oraz $a < b$, oznaczany jako $\mathcal{U}(a, b)$ wtedy, gdy jej gęstość jest postaci

$$f(x) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(x)\,[5]. \tag{2.1}$$

Wartość oczekiwana tego rozkładu wynosi

$$E(X) = \frac{a+b}{2}, \tag{2.2}$$

natomiast wariancja jest równa

$$Var(X) = \frac{(b-a)^2}{12}. \tag{2.3}$$

Funkcja charakterystyczna tego rozkładu wyrażona jest wzorem

$$\phi(t) = \frac{e^{itb} - e^{ita}}{(b-a)it}. \tag{2.4}$$

W przypadku, gdy zmienna losowa X posiada ciągłą dystrybuantę $F(x)$, to $U = F(x)$ ma rozkład $\mathcal{U}(0, 1)$.

2.2 Generatory liczb pseudolosowych rozkładu jednostajnego

Załóżmy, że F to dystrybuanta rozkładu jednostajnego $\mathcal{U}(0, 1)$. Podczas generowania ciągu zbliżonego do realizacji rozkładu jednostajnego zazwyczaj stosuje się następującą

metodę: wybierzmy funkcję G określoną i mającą wartości na odcinku $[0,1]$, o wartości początkowej $u_0 \in [0,1]$, i zdefiniujmy

$$u_1 = G(u_0),\ u_2 = G(u_1),\ \ldots,\ u_i = G(u_{i-1}),\ i = 1,\, 2,\, \ldots \tag{2.5}$$

Znając postać funkcji G i początkową wartość u_0 możliwe jest ponowne wygenerowanie każdego elementu zdefiniowanego powyżej ciągu. Prowadzi to do następującej definicji:

Definicja 2.2. Generator liczb pseudolosowych z rozkładu jednostajnego $\mathcal{U}(0,1)$ to taki algorytm dla funkcji G o wartości startowej $u_0 \in [0,1]$, który wyznacza wartości u_n, mając przy okazji podaną własność: dla każdego n, wygenerowane $u_1, u_2, \ldots, u_n$ oddają zachowanie próby losowej $U_1, U_2, \ldots, U_n \sim \mathcal{U}(0,1)$ [4].

Popularne testy zgodności, takie jak test Craméra-von Misesa [1], nie powinny odrzucać hipotezy, wedle której $u_1, u_2, \ldots, u_n$ to realizacja próby $U_1, U_2, \ldots, U_n \sim \mathcal{U}(0,1))$ [4]. Na potrzeby pracy korzystano z generatora liczb pseudolosowych o rozkładzie jednostajnym pochodzącym z biblioteki *Random* zawartej w języku Python.

2.3 Szacowanie parametrów modelu przy użyciu metody Monte Carlo

Załóżmy, że θ to parametr rozkładu zmiennej losowej X oraz można go przedstawić jako $\theta = Eh(X)$, przy czym h to pewna znana funkcja. Dodatkowym założeniem jest to, że można wygenerować próbę o rozkładzie X.

Definicja 2.3. Niech $X_1, X_2, \ldots, X_m$ będzie próbą losową pewnego rozkładu X dla pewnego m. Średnia $\bar{h} = m^{-1}(h(X_1) + h(X_2) + \cdots + h(X_m))$ nazywana jest **estymatorem** $Eh(X) = \theta$ **wyznaczonym metodą Monte Carlo** [7].

Można zauważyć, że $\bar{h}$ to średnia próbkowa próby $h(X_1), h(X_2), \ldots, h(X_m)$, użycie jej do oszacowania parametru θ jest możliwe dzięki prawu wielkich liczb. Głównym problemem tego zagadnienia jest estymacja parametru θ. W jednym z najprostszych przypadków funkcja $h(x) = x$, a zatem parametr θ jest wartością oczekiwaną EX. Całość tej metody opiera się na generowaniu próby losowej lub pseudolosowej z rozkładu jednostajnego, zastosowaniu funkcji h do przekształcenia elementów tej próby, a następnie wyznaczeniu estymatora $\bar{h}$ parametru θ. W celu dokładnego opisania metody posłużmy się przykładem: obliczmy

$$\theta := \int_0^1 g(x)\, dx. \tag{2.6}$$

W przypadku, gdy nie jesteśmy w stanie policzyć tej całki analitycznie, pozostaje nam zastosowanie metod numerycznych lub symulacyjnych. Algorytm szacowania parametru θ przy użyciu metody Monte Carlo:

1. wygeneruj $U_1, U_2, \ldots, U_n \sim \text{IID}\ \ \mathcal{U}(0,1)$,
2. oszacuj θ korzystając z

$$\hat{\theta}_n := \frac{g(U_1) + g(U_2) + \cdots + g(U_n)}{n}. \tag{2.7}$$

Modelowanie przy pomocy metody Monte Carlo jest przydatne przy badaniu skomplikowanych procesów losowych, które można rozbić na dwie kategorie. Pierwsza skupia w sobie systemy, w których proces jest sprecyzowany, ale poprzez jego złożoność trudno obliczyć jego parametry teoretyczne. Druga kategoria to eksperymenty losowe o modelu matematycznym trudnym do skonstruowania — dzięki metodzie Monte Carlo można dokonać ich klasyfikacji, generując wyniki modelowe, po czym przyrównać je z danymi eksperymentalnymi. Omawiana w następnym rozdziale metoda bootstrap w swoich założeniach wywodzi się z symulacji Monte Carlo.

2.4 Bootstrap nieparametryczny

Próba bootstrap to metoda służąca do oceny podstawowych charakterystyk pewnego estymatora (na przykład średniej lub wariancji), używając wielokrotnych symulacji opartych na znanej realizacji jego rozkładu. Narzędzie to zostało stworzone przez Bradleya Efrona i opublikowane w artykule „Bootstrap methods: another look at the jackknife" z 1979 roku [2]. Załóżmy, że $x_1, x_2 \dots x_n$ to realizacja pewnej próby losowej, a $\hat{F}$ jest dystrybuantą empiryczną tej próby. Dystrybuanta $\hat{F}$ to znane przybliżenie pewnego nieznanego rozkładu F, dlatego też rozkład $\hat{\Theta}$ estymować będziemy przy pomocy $\hat{F}$, czyli dokonamy oceny rozkładu estymatora $\hat{\Theta}$ w oparciu o generowanie prób z rozkładu $\hat{F}$.

Definicja 2.4. Próba losowa $X^* = (X_1^*, X_2^*, \dots, X_n^*)$ o rozkładzie $\hat{F}$ dla ustalonej realizacji $x = (x_1, x_2, \dots, x_n)$ nazywana jest nieparametryczną próbą bootstrap, lub prościej **próbą bootstrap** [4].

Otrzymywanie realizacji próby bootstrap bazuje na wykonaniu n-krotnego losowania ze zwracaniem elementów próby pierwotnej, tak więc losowość w próbie X^* polega na losowym wyborze elementu $x_1, x_2, \dots, x_n$. W ten sposób powstaje populacja, w której każda zmienna X_i^* jest niezależna od pozostałych, oraz z jednakowym prawdopodobieństwem przyjmuje dowolną wartość próby.

Zasada bootstrap mówi, że rozkład $T(X^*)$ dla ustalonych $x_1, x_2, \dots, x_n$ ma kształt zbliżony do rozkładu $T(X)$. Wynika to z faktu, że rozkład statystyki $(T(X^*) - \hat{\Theta})$ jest bliski rozkładowi statystyki $(T(X) - \Theta)$ [11]. Dodatkowo fakt, iż rozkład statystyki $T(X^*)$ ma przesunięte położenie względem rozkładu statystyki $T(X)$ o $\hat{\Theta} - \Theta$, jest nieistotny z punktu porównywania kształtu [4]. Dzięki temu można dokonać oceny rozkładu $\Theta = T(X)$ wykonując poniższe kroki:

1. dokonaj losowania niezależnych prób bootsrap $X_1^*, X_2^*, \dots, X_k^*$ korzystając z realizacji $x_1, x_2, \dots, x_n$,
2. wyznacz $\Theta_1^* = T(X_1^*)$, $\Theta_2^* = T(X_2^*)$, $\dots$, $\Theta_k^* = T(X_k^*)$.

Otrzymany w ten sposób rozkład $(\Theta_1^*, \Theta_2^*, \dots, \Theta_k^*)$ nazywany jest estymatorem rozkładu $\hat{\Theta}$ otrzymanym metodą bootstrap. Dla zadowalającego przybliżenia tego rozkładu wymagane jest co najmniej $k = 1000$ prób bootstrap, a im większa ich liczba, tym dokładniejsze oszacowanie.

2.4.1 Bootstrap parametryczny

Założenia bootstrapu parametrycznego są podobne do bootstrapu klasycznego — jedyną różnicą jest fakt, że zamiast symulowania niezależnych prób bootstrapowych z dystrybuanty

empirycznej, generowane są niezależne próby z rozkładu pewnego parametrycznego modelu [10]. W tym przypadku do danych dopasowany jest pewien model teoretyczny (często przy pomocy metody największej wiarygodności), a próby liczb losowych generowane są z owego dopasowanego modelu. Proces symulowania tak zdefiniowanej próby przebiega podobnie do innych procesów bootstrapowych, przy czym wielkość takiej próby zazwyczaj odpowiada rozmiarowi oryginalnego zbioru danych. Prowadzi to do następującej definicji.

Definicja 2.5. Próbę bootstrap, której nieznany rozkład F można przybliżyć pewnym znanym rozkładem parametrycznym $\hat{F}$ nazywamy **parametryczną próbą bootstrap** [3].

Rozdział 3

Metodologia, algorytmy

W tej części pracy zawarto opis danych i procesu ich przygotowania, wprowadzono używane dalej oznaczenia oraz współczynniki, a także zaproponowano modele do przewidywania wyników sezonu. Wszystkie z zaproponowanych schematów przedstawiono w postaci algorytmów i pseudokodów, co znacznie ułatwia ich zrozumienie i implementację w dowolnym języku programowania.

3.1 Opis danych

Dane, które będą wykorzystywane do symulacji sezonu, tak jak i terminarze, zostały zebrane z najpopularniejszej statystycznej stronie internetowej poświęconej koszykówce. [12]. Na potrzeby tej pracy zebrano wyniki starć pomiędzy drużynami, począwszy od sezonu 2004/2005 aż do 2017/2018. Początkowo symulowano rozgrywki sezonów 2014/2015 oraz 2017/2018 w celu wybrania najlepszego modelu, a następnie skorzystano z niego, aby przewidzieć wyniki rozgrywek we wciąż trwającym sezonie 2018/2019. Posiadając liczbę wygranych jednej drużyny z drugą na przestrzeni lat, dokonano następujących transformacji danych:

W zależności od przedziału czasowego, jaki będziemy rozpatrywać, zebrano wyniki w określonych rozgrywkach (na przykład, przy wyznaczaniu wyników sezonu 2014/2015 i przedziale 5 lat, używać będziemy danych z lat 2009 do 2014). Dzięki uzyskanej w ten sposób liczbie wygranych możemy wyznaczyć stosunek zwycięstw do porażek dla wybranych zespołów (przykład: Boston Celtics i Atlanta Hawks grały ze sobą 10 razy, druga z drużyn wygrała zaledwie 4 razy, dlatego też w starciu z Celtami ich szansa na wygraną wynosi 0.4). Po zastosowaniu tej metody dla wszystkich zespołów uzyskano macierz o liczbie wierszy i kolumn odpowiadającej ilości drużyn w NBA, zawierającą prawdopodobieństwa na wygraną z każdym zespołem w lidze. Przykładowa postać danych zawarta została w tabeli 3.1.

Ze względu na formę użytych danych nie ma znaczenia, czy podczas symulacji wykorzystywany będzie bootstrap klasyczny lub parametryczny. Używanie pierwszego polegałoby

Tabela 3.1: Przykładowa tabela z prawdopodobieństwami zwycięstw

	ATL	BOS	CHO
ATL		0.61	0.54
BOS	0.39		0.71
CHO	0.46	0.29	

na losowaniu ze zwracaniem wyników (zwycięstwo lub porażka) z danych historycznych — jest to jednoznaczne z wyliczeniem prawdopodobieństwa wygranej w oparciu o historię i symulowaniu rozkładu dwupunktowego.

3.2 Własne oznaczenia

Podczas prób symulacji dokonano intuicyjnego założenia, wedle którego największy wpływ na postawę sezonu mają rozgrywki bezpośrednio go poprzedzające. W tym celu dobrano system wag — z powodu dynamicznych zmian w lidze, najstarsze sezony otrzymują najmniejszą rangę, która stopniowo zwiększa się, im bliżej do zawodów rozpatrywanych w symulacji. Ważona ilość zwycięstw Z_{ij} $i-$tej drużyny D_i z $j-$tą drużyną D_j wynosi

$$Z_{ij} = \sum_{k=1}^{n} (1 + x \cdot k) \cdot R_{ijk}, \tag{3.1}$$

gdzie n to ilość sezonów, z których pobrano dane, x to ustalona waga kolejnych rozgrywek, a R_{ijk} to wynik starć drużyny D_i z drużyną D_j w $k-$tym sezonie. Podczas testowania skuteczności modeli dobierano różne wagi w celu znalezienia tego zwracającego najlepsze predykcje. W trakcie symulacji program przechodzi przez dokładnie określony terminarz rozgrywek — drużyna gra z przeciwnikiem tyle razy, ile spotkań wyznaczono w rozkładzie. Algorytmy losowania opisano szczegółowo w rozdziale 1. Na potrzeby testów mających na celu wyłonienie najlepszego modelu zdefiniowano współczynnik G równy

$$G = \sum_{i=1}^{n} |\hat{\Theta}_i - \Theta_i|, \tag{3.2}$$

gdzie n to ilość drużyn (obecnie 30), $\hat{\Theta}_i$ to wartość estymowana przy pomocy próby bootstrap dla i-tej drużyny, a Θ_i to jej rzeczywisty wynik w porównywanym sezonie. Wartość G to bezwzględna suma błędów modelu, a więc im jest ona mniejsza, tym lepsze jest dopasowanie.

3.3 Modele symulacji rozgrywek

Zaproponowane w tym rozdziale modele opierają się na zasadzie bootstrapu nieparametrycznego. Przygotowane dane zawierają prawdopodobieństwa poszczególnych drużyn na wygraną, co można potraktować jako znane realizacje rozkładów ich zwycięstw.

3.3.1 Model I — uśredniony

Pierwszy ze stworzonych modeli polega na obliczeniu ogólnego stosunku zwycięstw do porażek dla każdej drużyny w wybranym okresie — wszystkie wygrane zespołu zostają podzielone przez łączną liczbę rozegranych spotkań, wynikiem czego jest liczba z przedziału $[0, 1]$ określana jako P_i, gdzie i to $i-$ta drużyna. Schemat symulowania wyników spotkań między drużynami został szczegółowo opisany w algorytmie 3.1 i pseudokodzie 3.2.

Algorytm 3.1. Model uśredniony

1. wstaw liczniki zwycięstw drużyn $B_1 = 0, B_2 = 0, \ldots, B_{30} = 0$

2. dla $i = 1$, $i = 2$, ..., $i = 30$:
   (a) dla $j = i$, $j = i + 1$, ..., $j = 30$:
      i. znajdź drużyny D_i i D_j
      ii. odczytaj średnie ilości zwycięstw W_i i W_j dla drużyn D_i i D_j
      iii. wyznacz prawdopodobieństwo zwycięstwa W_{ij} przez drużynę D_i równe $W_{ij} = \frac{W_i}{W_i+W_j}$
      iv. w terminarzu znajdź liczbę spotkań S_{ij} pomiędzy drużynami D_i i D_j
      v. wykonaj S_{ij} razy:
         A. symuluj liczbę U z rozkładu jednostajnego $\mathcal{U} \sim U[0, 1]$
         B. jeżeli $W_{ij} \leq U$, to zwiększ licznik zwycięstw $B_i = B_i + 1$, w przeciwnym razie zwiększ licznik zwycięstw $B_j = B_j + 1$

Pseudokod 3.2. Model uśredniony

```
1: B_1 ← 0, B_2 ← 0, ..., B_30 ← 0
2: i ← 1
3: while i ≤ 30 do
4:     for j ← i to 30 do
5:         znajdź drużyny D_i i D_j
6:         znajdź W_i i W_j
7:         W_ij ← W_i / (W_i + W_j)
8:         znajdź liczbę spotkań S_ij pomiędzy D_i i D_j
9:         for p ← 1 to S_ij do
10:            U ~ U[0, 1]
11:            if W_ij ≤ U then
12:                B_i ← B_i + 1
13:            else
14:                B_j ← B_j + 1
15:            end if
16:        end for
17:    end for
18:    i ← i + 1
19: end while
```

3.3.2 Model II — rywalizacji

Drugi z zaproponowanych modeli zakłada zwracanie uwagi na historyczne wyniki przeciwko konkretnej drużynie. W zawodowym sporcie niejednokrotnie można trafić na zażarte rywalizacje między dwoma klubami lub tendencję do wygrywania z pewnym przeciwnikiem. Schemat symulowania wyników przy wykorzystaniu tego modelu został szczegółowo opisany w algorytmie 3.3 i pseudokodzie 3.4.

Algorytm 3.3. Model rywalizacji

1. wstaw liczniki zwycięstw drużyn $B_1 = 0, B_2 = 0, \ldots, B_{30} = 0$
2. dla $i = 1$, $i = 2$, ..., $i = 30$:
   (a) dla $j = i$, $j = i + 1$, ..., $j = 30$:

i. znajdź drużyny D_i i D_j

ii. odczytaj z macierzy wyników stosunek zwycięstw $W_{ij} = \frac{Z_{ij}}{Z_{ij}+Z_{ji}}$ drużyny D_i przeciwko drużynie D_j

iii. w terminarzu znajdź liczbę spotkań S_{ij} pomiędzy drużynami D_i i D_j

iv. wykonaj S_{ij} razy:

A. symuluj liczbę U z rozkładu jednostajnego $\mathcal{U} \sim U[0,1]$

B. jeżeli $W_{ij} \leq U$, to zwiększ licznik zwycięstw $B_i = B_i + 1$, w przeciwnym razie zwiększ licznik zwycięstw $B_j = B_j + 1$

Pseudokod 3.4. Model rywalizacji

```
1: B_i ← 0, B_j ← 0, ..., B_30 ← 0
2: i ← 1
3: while i ≤ 30 do
4:     for j ← i to 30 do
5:         znajdź drużyny D_i i D_j
6:         znajdź Z_ij i Z_ji
7:         W_ij = Z_ij / (Z_ij + Z_ji)
8:         znajdź liczbę spotkań S_ij pomiędzy D_i i D_j
9:         for p ← 1 to S_ij do
10:            U ~ U[0,1]
11:            if W_ij ≤ U then
12:                B_i ← B_i + 1
13:            else
14:                B_j ← B_j + 1
15:            end if
16:        end for
17:    end for
18:    i ← i + 1
19: end while
```

3.3.3 Model III — fazy pucharowej

Po symulacji całego sezonu, czyli 1230 spotkań, 8 najlepszych drużyn z każdej konferencji przechodzi do fazy playoff, gdzie toczy rozgrywki zgodnie z systemem opisanym w rozdziale 1. Na tym etapie rozgrywek symulacja spotkań różni się od części zasadniczej: zamiast jednego z zasugerowanych powyżej modeli korzysta się ze wcześniejszej symulacji fazy zasadniczej. W celu oddania trendów panujących w wygenerowanych rozgrywkach (potencjalnych kontuzjach, spadkach lub zwyżkach formy), użyta zostaje jedynie informacja o ilości wygranych przed rozpoczęciem playoffów. Szczegółowy opis symulacji tej fazy został opisany w algorytmie 3.5 i pseudokodzie 3.6.

Algorytm 3.5. Model playoff

1. wybierz drużyny D_i i D_j
2. wstaw liczniki zwycięstw $ZW_i = 0$ i $ZW_j = 0$
3. odczytaj symulowane ilości zwycięstw z sezonu zasadniczego B_i i B_j dla wybranych drużyn D_i i D_j

4. wyznacz prawdopodobieństwo zwycięstwa drużyny D_i nad drużyną D_j równe $W_{ij} = \frac{B_i}{B_i+B_j}$
5. powtarzaj dopóki $ZW_i = 4$ albo $ZW_j = 4$:
   (a) symuluj liczbę U z rozkładu jednostajnego $U \sim U[0,1]$
   (b) jeżeli $W_{ij} \leq U$, wstaw $ZW_i = ZW_i + 1$, w przeciwnym razie wstaw $ZW_j = ZW_j + 1$
6. jeżeli $ZW_i = 4$, to przenieś drużynę D_i do następnego etapu, w przeciwnym razie przenieś drużynę D_j

Pseudokod 3.6. Model playoff

znajdź drużyny D_i i D_j
$ZW_i \leftarrow 0, ZW_j \leftarrow 0$
znajdź B_i i B_j
$W_{ij} = \frac{B_i}{B_i+B_j}$
while $ZW_i < 4$ and $ZW_j < 4$ **do**
    $U \sim \mathcal{U}[0,1]$
    if $W_{ij} \leq U$ **then**
        $B_i \leftarrow B_i + 1$
    else
        $B_j \leftarrow B_j + 1$
    end if
end while
if $ZW_i = 4$ **then**
    D_i przechodzi do następnej rundy
else
    D_j przechodzi do następnej rundy
end if

Ilości zwycięstw drużyn w kolejnych symulowanych rozgrywkach są zapisywane i zapamiętywane, podobnie jak informacje o przejściach do kolejnych faz rozgrywek pucharowych.

3.4 Predykcja wyników na podstawie symulacji

Korzystając z metod bootstrapowych, zaproponowano kilka modeli pozwalających na oszacowanie przebiegu rozgrywek w analizowanym sezonie. Przed korzystaniem z opisanych poniżej schematów przeprowadzono symulację opartą na algorytmach zdefiniowanych w rozdziale 3, dzięki czemu do symulacji podchodzono z przygotowanymi wcześniej danymi.

3.4.1 Predykcja wyników sezonu zasadniczego

Sezon zasadniczy odgrywa bardzo ważną rolę w rozgrywkach NBA — w trakcie jego trwania drużyny toczą walkę o najlepsze miejsca w fazie pucharowej, sprawdzają swoje siły w trakcie gier z potencjalnymi rywalami w playoff, a także wzmacniają swoje drużyny poprzez rotację zawodników. Podczas przejść symulacji zapisywano łączną liczbę wygranych każdego zespołu, dzięki czemu otrzymano rozkłady zwycięstw wszystkich składów. Po przeanalizowaniu rozkładów zauważono, że mediany i wartości średnie

Tabela 3.2: Porównanie średnich i median wygranych spotkań w symulacji

Drużyna	**Mediana**	**Średnia**
ATL	47	46.95
BOS	44	44.13
CHO	39	39.07
CHI	44	43.97
CLE	49	49.13
DAL	41	41.53
DEN	36	36.18
DET	36	35.97
GSW	65	65.09
HOU	51	50.96
IND	43	43.57
LAC	54	53.97
LAL	24	23.98
MEM	47	47.20
MIA	46	45.59
MIL	36	28.79
MIN	29	35.97
BRK	29	29.23
NOP	35	34.73
NYK	31	30.99
ORL	28	28.33
PHI	21	20.83
POR	45	45.09
SAC	31	31.06
SAS	61	61.38
OKC	51	51.16
TOR	50	50.36
UTA	41	41.46
WAS	44	43.80

nie różnią się znacznie od siebie, dlatego w dalszych rozważaniach jako przyszłą liczbę wygranych zespołu w sezonie zasadniczym przyjmuje się medianę rozkładu zwycięstw w symulacji. Wyniki porównania zawarto w tabeli 3.2. Dodatkowo zbadano normalność rozkładów zwycięstw wszystkich drużyn. Niestety, pomimo kształtu gęstości zbliżonego do normalnego, wszystkie testy statystyczne odrzuciły hipotezę o rozkładzie normalnym.

3.4.2 Modele predykcji wyników fazy pucharowej

Rozgrywki playoff są niezwykle trudne do przewidzenia — niejednokrotnie zdarzyło się, że faworyci zostali pokonani przez znacznie niżej notowanego przeciwnika (najlepszy przykład to seria DAL-GSW w 2007 roku, kiedy Dallas Mavericks kończąc sezon z najlepszym bilansem zwycięstw w lidze, przegrali pierwszą rundę w 6 meczach z ósmą drużną konferencji). Do symulacji tego etapu zaproponowano 2 odmienne od siebie modele, bazujące na innych założeniach.

Model IV — Prawdopodobieństwo przejść

Model ten polega na oszacowaniu szans poszczególnych drużyn na przejście do wybranych etapów rozgrywek pucharowych. Podczas każdej z symulacji sezonu zbierano informacje o przejściach zespołów do kolejnych faz playoff. W ten sposób estymowano prawdopodobieństwa przejść do odpowiednio pierwszej rundy, drugiej rundy, finałów konferencji oraz finałów. Według zaproponowanego schematu do kolejnych etapów awansują zespoły z największą ilością przejść, a więc największym prawdopodobieństwem awansu. Do rundy pierwszej każdej konferencji dostanie się 8 drużyn najczęściej zakwalifikowanych w symulacjach, do drugiej 4 zespoły o największym prawdopodobieństwie przejścia do drugiego etapu, kolejne etapy wyznaczane są analogicznie.

Model V — Najczęstsze kombinacje

Drugi z badanych modeli porównuje najczęściej pojawiające się kombinacje zespołów w poszczególnych etapach. Podczas symulacji rozgrywek pucharowych zapisywano listy z drużynami przechodzącymi do kolejnych etapów playoff, dzięki czemu zachowane zostały również układy, w jakich toczyły się rozgrywki. Innymi słowy, dla rundy pierwszej wyszukiwana jest najczęściej pojawiająca się kombinacja zespołów w pierwszej rundzie, czynność ta jest powtarzana w kolejnych etapach. Mechanizm ten jest w stanie zwrócić bardzo precyzyjną prognozę, wymaga jednak dużej liczby powtórzeń symulacji do poprawnego działania.

Rozdział 4

Symulacje

Poniższa część pracy skupia się na wyborze najlepszego z zaproponowanych modeli — analizowano je w oparciu o kilka czynników i badano zgodność z rzeczywistymi rezultatami. Po wyznaczeniu najdokładniejszego z nich dokonano predykcji przebiegu sezonu i rozgrywek pucharowych 2018/2019.

4.1 Dobór optymalnego modelu

Podczas testowania modeli badano zmiany wag poszczególnych sezonów oraz okres zbierania danych. Podjęto decyzję o przeprowadzaniu symulacji dla sezonów 2014/2015 oraz 2017/2018 o następujących parametrach:

- wagi wynoszące odpowiednio $x = 0$, $x = 0.5$, $x = 1$,
- okresy zbierania danych wynoszące odpowiednio 5 i 10 lat, a zatem
  - dla sezonu 2014/2015 wykorzystywano dane z lat 2009-2014 oraz 2004-2014
  - dla sezonu 2017/2018 wykorzystywano dane z lat 2012-2017 oraz 2007-2017

Wszystkie przeprowadzone w ten sposób symulacje zostały wykonane dla 10000 powtórzeń próby bootstrap.

4.1.1 Wyniki symulacji dla sezonów zasadniczych

Podczas symulowania podstawowej części sezonu, jako przewidzianą liczbę wygranych przyjęto medianę rozkładu symulowanych zwycięstw, co zostało opisane w Podrozdziale 3.4.1. Przy pomocy czynnika G zdefiniowanego w rozdziale 3.2 dokonano oceny modeli dla określonych wcześniej parametrów. Dodatkowo obliczono sumę błędów dla obu badanych sezonów, co znacznie uprości wybór lepszego modelu. Wyniki symulacji zawarto w tabeli 4.1. Jak można zauważyć, najlepsze oceny otrzymał Model I dla danych z 5 lat i wagą $x = 1$. Do dalszych rozważań zakwalifikowany został również Model 2 z zebranymi 10 sezonami o wadze $x = 0.5$ (okazał być się najlepszy w prognozie drugiego symulowanego sezonu). Pomimo gorszej oceny, korzystne będzie przetestowanie dwóch różnych modeli w następnym etapie. Po przeanalizowaniu rezultatów symulacji okazało się, że wyniki dla sezonu 2017/2018 były znacznie lepsze niż dla 2014/2015. Na rysunkach 4.1–4.4 zawarto wykresy pudełkowe wygenerowanych rozkładów zwycięstw dla wszystkich drużyn w sezonie 2017/2018 przy użyciu wybranych wcześniej modeli i parametrów.

Analizując wykresy pudełkowe, można zauważyć, że:

Tabela 4.1: Współczynnik G jakości dopasowania modelu

Parametry	**Sezon 14/15**	**Sezon 17/18**	**Suma**
M I, 5 lat, waga 0	306	280	586
M I, 5 lat, waga $0,5$	295	275	570
M I, 5 lat, waga 1	293	270	563
M I, 10 lat, waga 0	318	290	608
M I, 10 lat, waga $0,5$	304	275	579
M I, 10 lat, waga 1	303	283	586
M II, 5 lat, waga 0	304	308	612
M II, 5 lat, waga $0,5$	305	292	597
M II, 5 lat, waga 0	305	289	594
M II, 10 lat, waga 0	331	299	630
M II, 10 lat, waga $0,5$	314	275	589
M II, 10 lat, waga 1	331	298	629

- liczba dopasowań bardzo dobrych, czyli leżących w przedziale $[Q_1, Q_3]$ jest lepsza dla modelu II i wynosi 6, podczas gdy dla modelu I wartość ta jest równa 5. Większość drużyn osiąga wyniki na podobnym poziomie.
- liczba wartości odstających, leżących poza przedziałem $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ jest większa dla modelu II, jest równa 8, gdzie dla modelu I równa się ona 7.

Patrząc na te wyniki, nie można jednoznacznie odrzucić jednego modelu na korzyść drugiego, dlatego do wybrania najlepszego z nich dokonano również analizy rozgrywek pucharowych. Ponownie analizując tabelę z wartościami G, można zauważyć, że w dla rozgrywek 2017/2018 symulacje były bardziej trafne. Wyraźnie większa dokładność w przypadku późniejszego sezonu wynika z faktu, że od kilku lat obserwuje się te same drużyny w czołówce ligi. Przed 2014 rokiem równowaga została zachwiana przez przejście jednego z najlepszych graczy ligi, LeBrona Jamesa z Miami Heat do Cleveland Cavaliers, jak i rozpowszechnienie przez Golden State Warriors systemu szybkiej gry opartej na rzutach z dystansu. W trakcie lata 2018 roku LeBron ponownie zmienił klub, tym razem na Los Angeles Lakers, co może mieć znaczący wpływ na jakość predykcji trwającego sezonu (drużyny, w których gra pojawiają się w finałach NBA nieprzerwanie od 2010 roku).

4.1.2 Wyniki symulacji dla fazy playoff

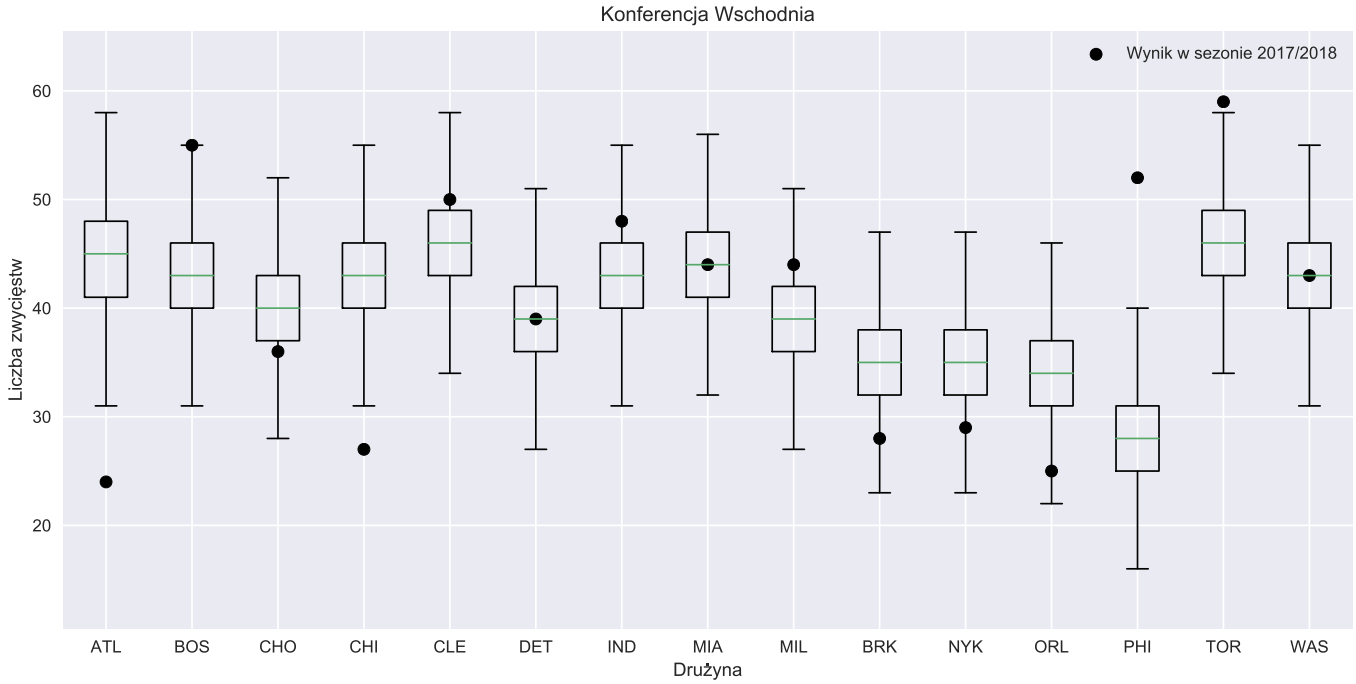
Dla wybranych w poprzedniej części modeli i parametrów przeprowadzona została symulacja rozgrywek pucharowych, korzystając z algorytmów zdefiniowanych w rozdziale 3.4.2. W tabeli 4.2 porównano przewidziane na ich podstawie serie wraz z rzeczywistym przebiegiem rozgrywek w 2018 roku. Podczas symulowania tej fazy trzymano się zasady, wedle której drużyna z lepszym bilansem wypisana jest jako pierwsza.

Po przeanalizowaniu wyników okazało się, że prognozy wynikające z modelu V są całkowicie nietrafne — kombinacje zespołów w następnych rundach nie pokrywają się z poprzednimi. Podczas predykcji rundy pierwszej wskazał on poprawnie 12 z 16 drużyn (wyniki dla sezonu symulowanego modelem I), z czego dla konferencji Wschodniej błędnie wyznaczone były aż 3 z 8 organizacji. Dla porównania, podczas szacowania przebiegu modelu II pomylił się w przypadku 5 drużyn. Przy dalszych rozważaniach model ten został odrzucony — pomimo porównywalnej dokładności z modelem IV (takie same liczby

Tabela 4.2: Przebieg serii playoff w symulacjach dla 2018 roku

Prognozy dla wybranego Modelu I		
Model IV	**Model V**	**Rzeczywistość**
Eastern Conference First Round		
CLE-WAS	TOR-CLE	TOR-WAS
MIA-BOS	WAS-CHI	CLE-IND
ATL-CHI	ATL-BOS	PHI-MIA
TOR-IND	BRK-MIA	BOS-MIL
Western Conference First Round		
GSW-DAL	NOP-OKC	HOU-MIN
HOU-OKC	UTA-LAC	OKC-UTA
LAC-MEM	POR-SAS	POR-NOP
SAS-NOP	GSW-HOU	GSW-SAS
Eastern Conference Semifinals		
CLE-MIA	IND-TOR	TOR-CLE
TOR-ATL	CHO-ATL	BOS-PHI
Western Conference Semifinals		
GSW-HOU	SAS-GSW	HOU-UTA
SAS-LAC	UTA-MEM	GSW-NOP
Eastern Conference Finals		
CLE-TOR	CLE-CHI	BOS-CLE
Western Conference Finals		
GSW-SAS	GSW-SAS	HOU-GSW
Finals		
GSW-TOR	GSW-TOR	GSW-CLE
Zwycięzca		
GSW	GSW	GSW

Prognozy dla wybranego Modelu II		
Model IV	**Model V**	**Rzeczywistość**
Eastern Conference First Round		
MIA-WAS	CHI-CLE	TOR-WAS
BOS-CLE	BOS-IND	CLE-IND
CHI-TOR	MIA-TOR	PHI-MIA
ATL-IND	ATL-BRK	BOS-MIL
Western Conference First Round		
SAS-POR	SAS-DAL	HOU-MIN
LAC-HOU	GSW-POR	OKC-UTA
OKC-MEM	HOU-NOP	POR-NOP
GSW-NOP	LAC-OKC	GSW-SAS
Eastern Conference Semifinals		
MIA-BOS	MIA-CLE	TOR-CLE
ATL-CHI	ATL-BOS	BOS-PHI
Western Conference Semifinals		
SAS-LAC	SAS-LAC	HOU-UTA
OKC-GSW	OKC-GSW	GSW-NOP
Eastern Conference Finals		
MIA-CHI	MIA-CHI	BOS-CLE
Western Conference Finals		
SAS-GSW	SAS-GSW	HOU-GSW
Finals		
MIA-SAS	MIA-SAS	GSW-CLE
Zwycięzca		
SAS	SAS	GSW



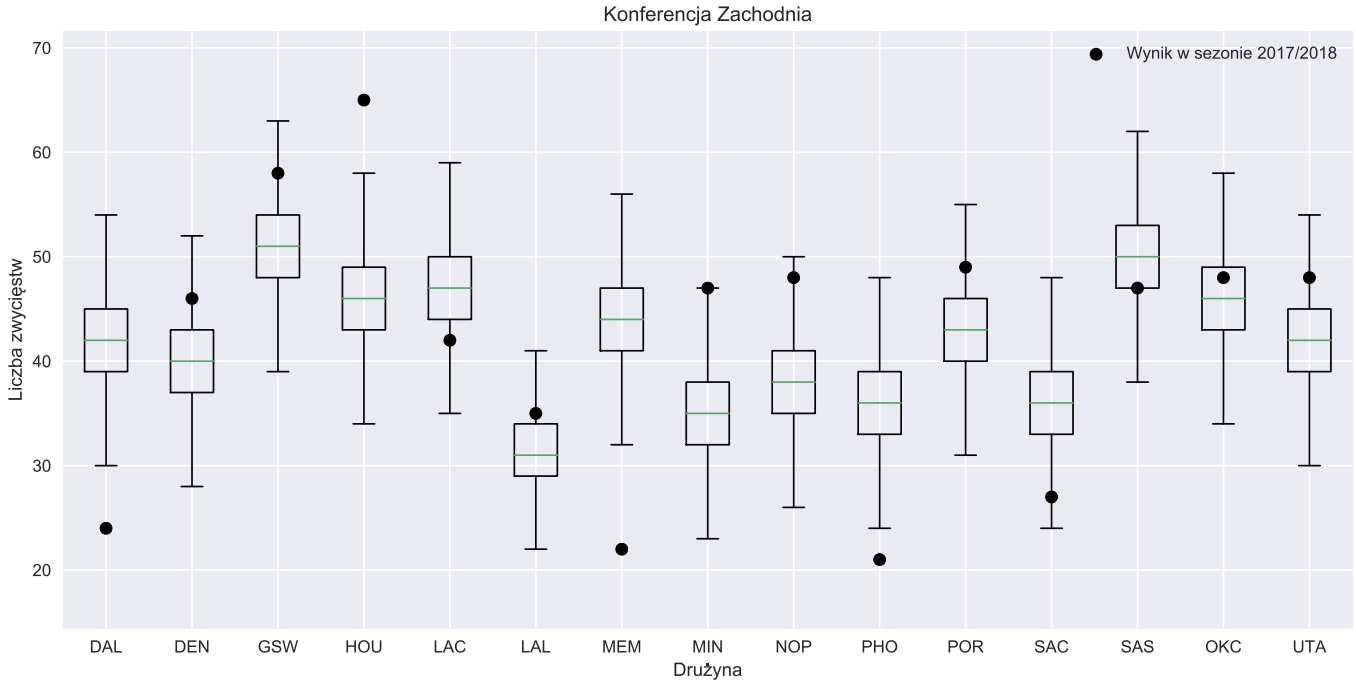
Rysunek 4.1: Wykres pudełkowy wyników symulacji sezonu przy użyciu modelu I, na podstawie danych z 5 lat i wadze $x = 1$, konferencja Wschodnia

poprawnie wytypowanych drużyn) jego ogromną wadą jest brak konsekwentności, czego najlepszym przykładem jest przewidziana seria Charlotte-Atlanta w finałach konferencji Wschodniej (Charlotte nie zostało wytypowane do gry w pierwszej rundzie).

Po odrzuceniu jednej z metod predykcji fazy pucharowej zbadano przebieg przewidywanych rozgrywek w oparciu o model IV wykorzystujący dane o sezonach zasadniczych generowanych przy wykorzystaniu modeli dobranych we wcześniejszych rozważaniach. Obydwa modele sprawdzały się podobnie w przypadku rundy pierwszej, jednak od rundy drugiej szacunki zaczęły być znacznie dokładniejsze:

- w przypadku modelu I do Półfinałów konferencji Wschodniej zakwalifikowały się 2 z 4 drużyn, dla modelu II natomiast nie przeszła żadna z rzeczywistych drużyn,
- dla Półfinałów konferencji Zachodniej model I wygrał ponownie, przewidując słusznie 2 drużyny, przy 1 dla konkurenta,
- w finałach konferencji Model I jeszcze raz okazał się lepszy, za każdym razem dobrze przewidując przynajmniej jednego finalistę, podczas gdy drugi z systemów był skuteczny tylko dla 1 z 4 zespołów,
- model I słusznie przewidział zwycięzcę całej ligi.

Pomimo faktu, że model IV nie przewidział pojawienia się w tej części sezonu zespołów z Milwaukee, Filadelfii, Minnesoty i Utah, był w stanie stosunkowo dobrze wytypować dalsze etapy rozgrywek, dlatego też zostanie wykorzystany w dalszych pracach.



Rysunek 4.2: Wykres pudełkowy wyników symulacji sezonu przy użyciu modelu I, na podstawie danych z 5 lat i wadze $x = 1$, konferencja Zachodnia

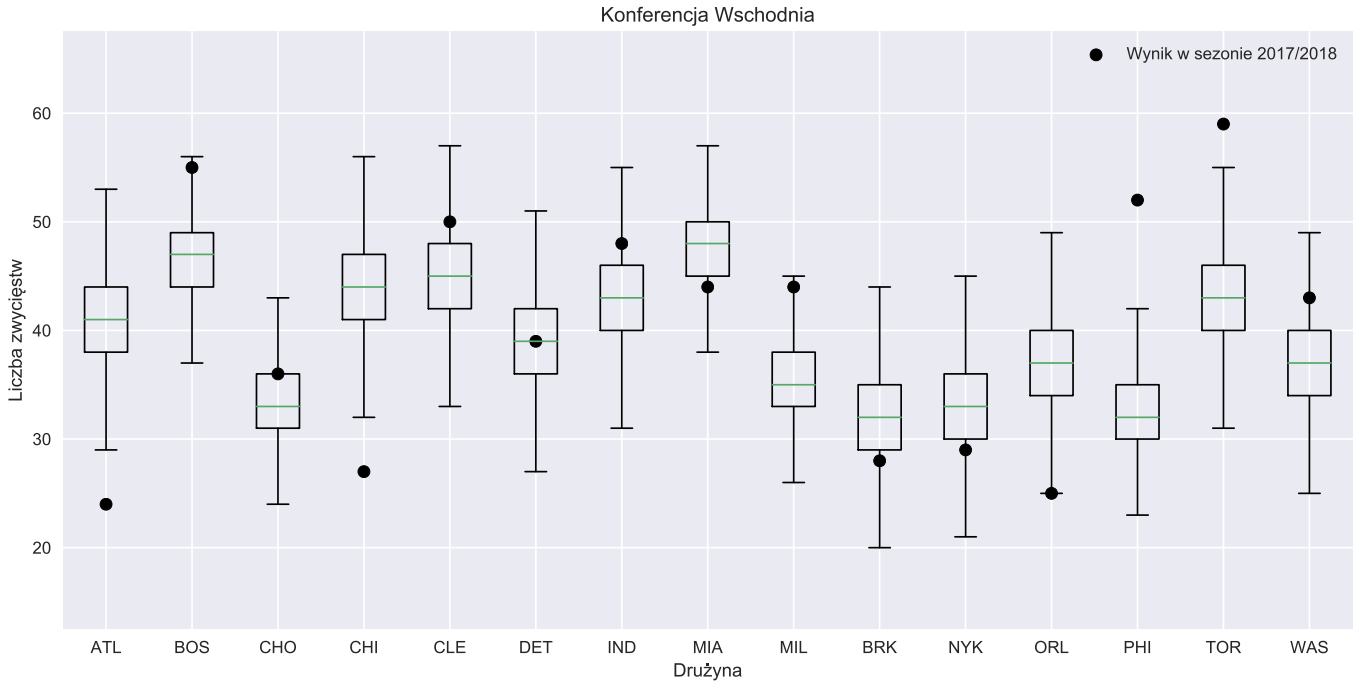
4.1.3 Symulacja sezonu 2018/2019

W poprzedniej części pracy udowodniono, że przewidywania najtrafniejsze będą, jeżeli sezon zasadniczy symulowany będzie przy pomocy modelu I, danych z ostatnich 5 lat, i wagi $x = 1$. W przypadku fazy pucharowej do predykcji rozgrywek playoff wykorzystany zostanie model IV. Wyniki symulacji można odczytać z rysunków 4.5 i 4.6, jak i tabeli 4.3. Porównując rezultaty przeprowadzonych prognoz z poprzednimi symulacjami, można zauważyć pewną tendencję — zaproponowane w tej pracy modele nie zakładają sytuacji, w której zespół wchodzi w stan przebudowy i zaczyna przegrywać. Z tego powodu oszacowane liczby wygranych można uznać za bardzo optymistyczne, ponieważ wiadome jest, że zespoły takie jak Brooklyn Nets, Phoenix Suns, Atlanta Hawks, czy New York Knicks najprawdopodobniej nie osiągną przewidywanych wyników. Podobnie sytuacja ma się w fazie pucharowej, gdzie z powodu nagłego osłabienia Cleveland zespoły, które do tej pory w cieniu wzmacniały swoje składy, mają realną szansę na walkę o najwyższe cele. Po zestawieniu wyników modelu I z przewidywaniami ESPN [6] można zauważyć, że rezultaty zaproponowane w obu modelach pokrywają się z nielicznymi wyjątkami — są to kluby, które znacząco wzmocniły się tego lata lub weszły w przebudowę.

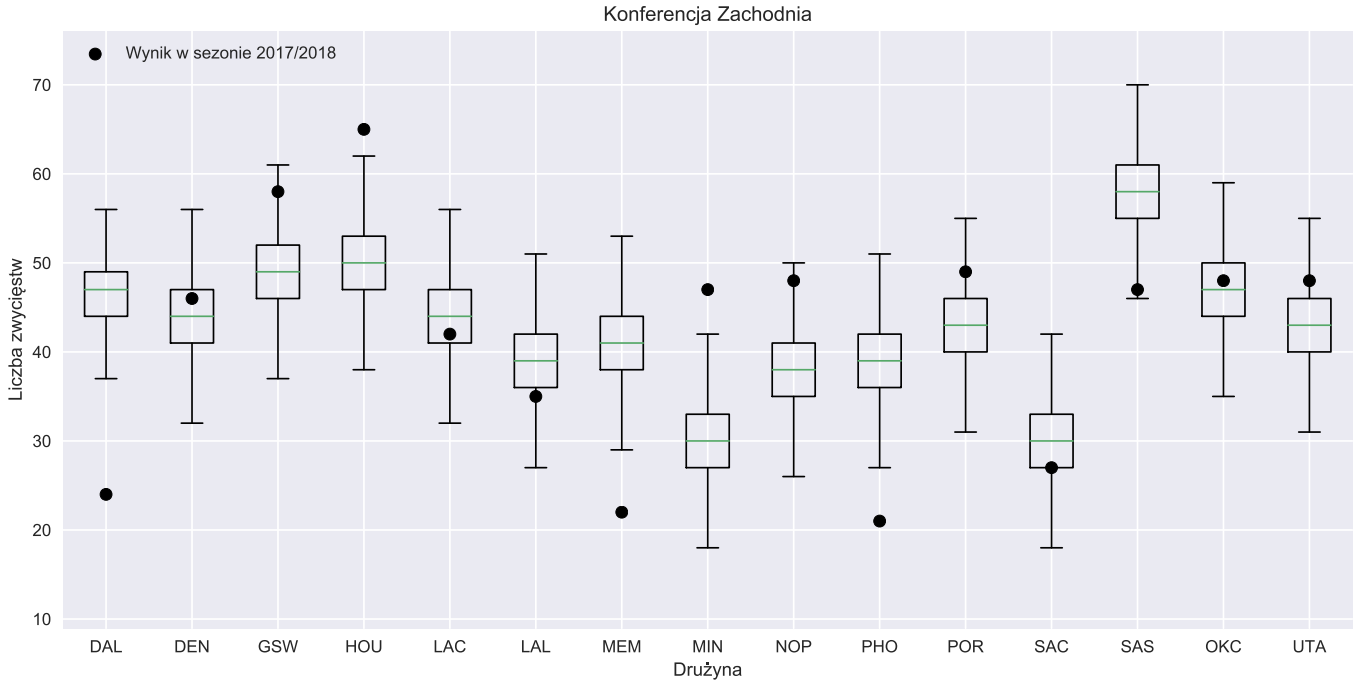
Tabela 4.3: Wyniki symulacji sezonu 2018/2019

Prognozy sezonu zasadniczego		
	Liczba zwycięstw	
Drużyna	**Model I**	**ESPN**
Atlanta Hawks	41	22
Boston Celtics	45	58
Brooklyn Nets	33	32
Charlotte Hornets	40	35
Chicago Bulls	40	28
Cleveland Cavaliers	46	31
Dallas Mavericks	38	33
Denver Nuggets	41	47
Detroit Pistons	40	44
Golden State Warriors	51	58
Houston Rockets	48	57
Indiana Pacers	43	47
Los Angeles Clippers	46	35
Los Angeles Lakers	33	46
Memphis Grizzlies	39	33
Miami Heat	43	47
Milwaukee Bucks	40	47
Minnesota Timberwolves	38	45
New Orleans Pelicans	41	45
New York Knicks	34	28
Oklahoma City Thunder	46	49
Orlando Magic	35	30
Philadelphia 76ers	36	53
Phoenix Suns	36	27
Portland Trail Blazers	44	43
Sacramento Kings	35	24
San Antonio Spurs	49	44
Toronto Raptors	47	55
Utah Jazz	43	49
Washington Wizards	43	44

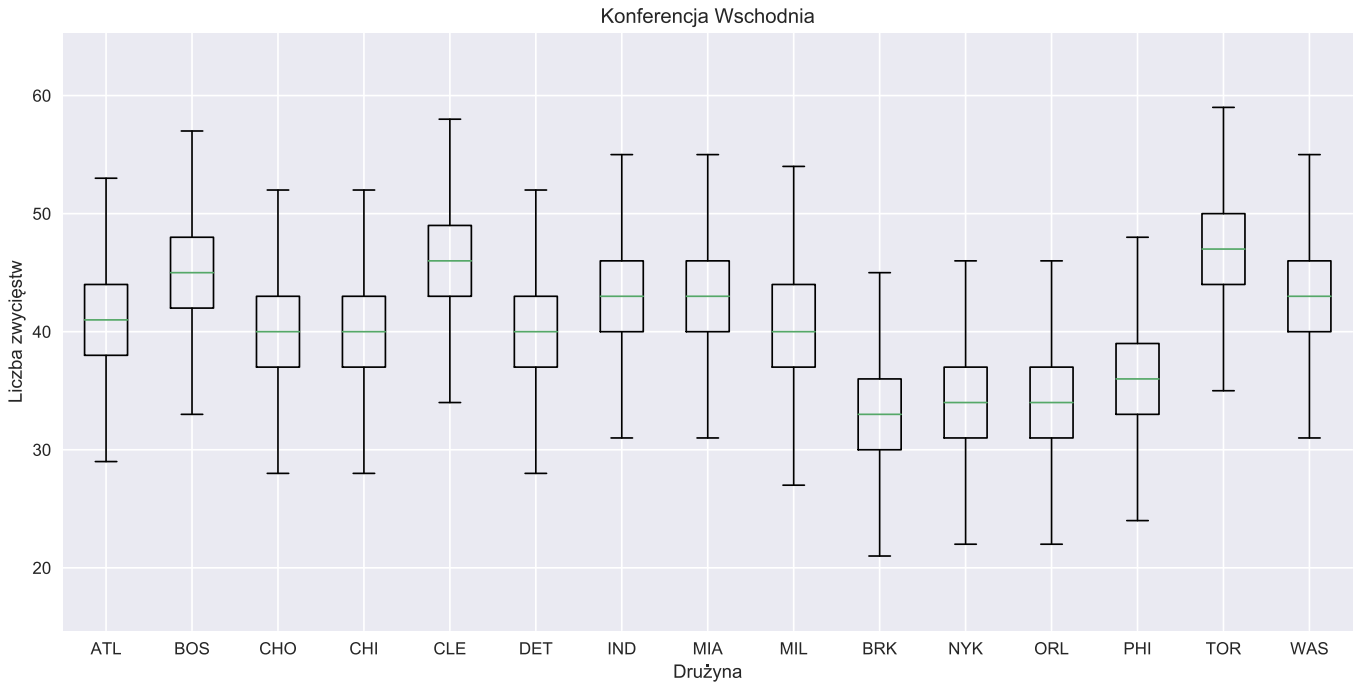
Prognozy fazy pucharowej	
Wschód	**Zachód**
First Round	
TOR-CHO	GSW-NOP
IND-WAS	LAC-OKC
BOS-MIA	HOU-POR
CLE-ATL	SAS-UTA
Conference Semifinals	
TOR-IND	GSW-LAC
CLE-BOS	SAS-HOU
Conference Finals	
TOR-CLE	GSW-SAS
Finals	
GSW-TOR	
Zwycięzca	
GSW	



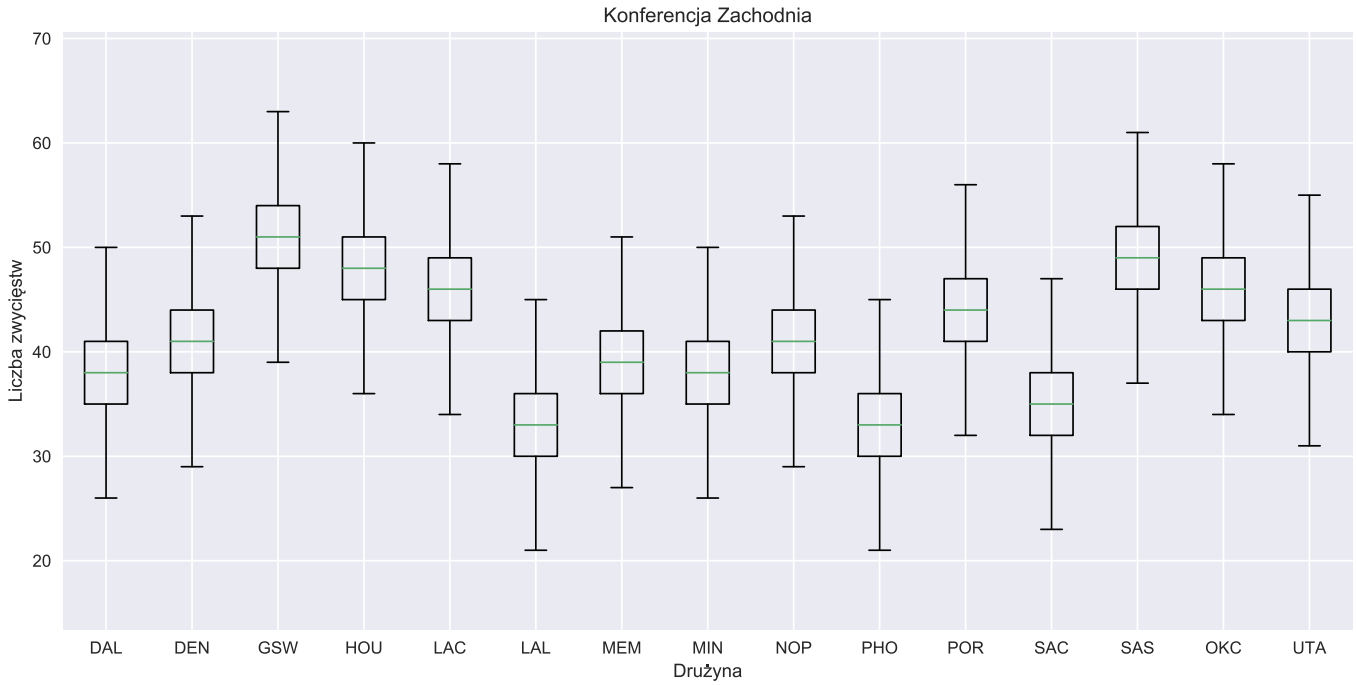
Rysunek 4.3: Wykres pudełkowy wyników symulacji sezonu przy użyciu modelu II, na podstawie danych z 10 lat i wadze $x = 0.5$, konferencja Wschodnia



Rysunek 4.4: Wykres pudełkowy wyników symulacji sezonu przy użyciu modelu II, na podstawie danych z 10 lat i wadze $x = 0.5$, konferencja Wschodnia



Rysunek 4.5: Szacowane wyniki w sezonie 2018/2019



Rysunek 4.6: Szacowane wyniki w sezonie 2018/2019

Rozdział 5

Podsumowanie

Celem tej pracy inżynierskiej było stworzenie oprogramowania zdolnego do przewidywania wyników rozgrywek NBA w oparciu o metody bootstrapowe. Zaproponowano kilka modeli umożliwiających symulowanie sezonu, bazując na danych zawierających wyniki z zeszłych sezonów. Wszystkie przetestowano dla różnych okresów zbierania danych, wag premiujących kolejne lata, jak i sezonu, dla którego dokonano symulacji. Po porównaniu wyników symulacji sezonów zasadniczych, do dalszej analizy wybrano dwa modele — model I o wadze $x = 1$ i okresie zbierania danych równym 5 lat i model II z wagą $x = 0.5$ i okresem 10 lat. Zbadanie przebiegu fazy pucharowej w obu przypadkach pozwoliło wybrać najskuteczniejszy z modeli, którym okazał się wymieniony wcześniej model I. Niestety, rezultaty zaproponowanych metod nie oddają w pełni trendów panujących obecnie w NBA, jak i nie działają najlepiej w przypadku nagłej poprawy gry przez słabe do tej pory zespoły. Dokonano porównania wyników zaproponowanego modelu z prognozami ESPN i zauważono, że w większej części tendencje w nich są podobne.

W trakcie badania skuteczności modeli testowano normalność rozkładu zwycięstw każdej z drużyn w próbie bootstrap, jednak we wszystkich testach hipotezy te zostały odrzucone. Dalsze rozważania na ten temat spowodowały wysnucie hipotezy, wedle której zwycięstwa te pochodzą z mieszanego rozkładu dwumianowego, jednak udowodnienie tego było trudne. Uzyskawszy nie najlepsze wyniki takiego podejścia do szacowania rezultatów sezonu, warte rozważenia jest stworzenie modelu opierającego się na statystykach zawodników grających w poszczególnych klubach i ich wpływie na grę.

Podczas prób symulowania rozgrywek należy pamiętać, że zawodnicy z czasem przestają grać na wysokim poziomie, zmieniają kluby, łapią kontuzje, czy nawet przechodzą na emerytury. Przykładem organizacji, która pomimo regresu w ostatnich sezonach, jest wysoko notowana we wszystkich symulacjach to San Antonio Spurs. Trzon tego zespołu od kilkunastu lat był taki sam i grał na niezwykle wysokim poziomie, zdobywając wielokrotne mistrzostwa od 2000 roku. Jednak ostatnie lata przyniosły spadek formy wielu gwiazd związany z wiekiem, co doprowadziło do zakończenia ich karier. Wszystkie modele klasyfikują tę drużynę bardzo wysoko ze względu na zeszłe dokonania, co prowadzi do mało prawdopodobnych wyników predykcji. W dalszych rozważaniach warto skupić się na możliwości przejścia słabego zespołu w przebudowę, jak i określenie wpływu wieku zawodników na ich grę.

Kolejnym niemożliwym do przewidzenia czynnikiem jest decyzja zarządu klubu o celowym przegrywaniu — skutkiem tego są ich bardzo niskie wyniki i podwyższone wygrane innych zespołów. Wysoce prawdopodobne jest, że zaproponowane w tej pracy modele dałyby bardziej precyzyjne wyniki, gdyby żadna drużyna nie decydowała się na przebudowę

poprzez odpuszczanie meczów. Osobną kwestią jest również etap, w którym zespoły rozwijające młodzież zaczynają nagle wygrywać, ponieważ niejednokrotnie po wielu słabych latach z zaskoczenia osiągały wysokie rezultaty (na przykład Philaledphia 76ers, po kilku sezonach ciągłego przegrywania zajęli niespodziewanie trzecie miejsce na Wschodzie).

Modele symulujące rozgrywki pucharowe również miały te same problemy, jednak w większości przypadków skutecznie udawało im się przewidywać kształt finałów konferencji i wytypować zwycięzców ligi. Okazało się, że najdokładniejsze były dla stabilnego sezonu 2017/2018, można więc założyć, że wszystkie algorytmy działają najlepiej dla stabilnej ligi, w której drużyny grają stale na tym samym poziomie.

Dodatek

Ze względu na dużą ilość tabel z danymi i wygenerowanych dla nich wykresów, wszystkie dodatki zostały zawarte na płycie dołączonej do pracy.

Bibliografia

[1] T. W. Anderson: „On the Distribution of the Two-Sample Cramér-von Mises Criterion", The Annals of Mathematical Statistics 33, #3, 1962, 1148-1159

[2] B. Efron: „Bootstrap methods: another look at the jackknife", The Annals of Statistics 1979, Vol. 7, No. 1, 1–26

[3] B. Efron: „The jackknife, the bootstrap, and other resampling plans", Technical report no. 63, Stanford University, grudzień 1980

[4] J. Koronacki, J. Mielniczuk: „Statystyka dla studentów kierunków technicznych i przyrodniczych", Wydawnictwo Naukowo-Techniczne, Warszawa, 2009

[5] R. Magiera: „Modele i metody statystyki matematycznej. Część II. Wnioskowanie statystyczne", Wydanie drugie rozszerzone, Oficyna Wydawnicza GiS, Wrocław, 2007

[6] ESPN: „Summer Forecast: East standings, West standings for 2018-19", 14 sierpnia 2018, [dostęp: 9 grudnia 2018], *<http://www.espn.com/nba/story/_/id/24365036/nba-standings-predictions-espn-summer-forecast>*

[7] M. Haugh: „Generating Random Variables and Stochastic Processes", Columbia University, 2017, [dostęp: 4 grudnia 2018], *<http://www.columbia.edu/~mh2078/MonteCarlo/MCS_Generate_RVars.pdf>*

[8] M. Oh, S. Keshri, G. Iyengar: „Graphical Model for Baskeball Match Simulation", Sloan Sports Analytics Conference, 2015 [dostęp: 1 grudnia 2018], *<http://www.sloansportsconference.com/wp-content/uploads/2015/02/SSAC15-RP-Finalist-Graphical-model-for-basketball-match-simulation.pdf>*

[9] N. Sandholtz, L. Bornn: „Replaying the NBA", Sloan Sports Analytics Conference, 2018 [dostęp: 5 grudnia 2018], *<http://www.lukebornn.com/papers/sandholtz_ssac_2018.pdf>*

[10] C. Shalizi: „The Bootstrap", 3 luty 2011, [dostęp: 3 grudnia 2018], *<https://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf>*

[11] K. Singh, M. Xie: „Bootstrap: A Statistical Method", Rutgers University, [dostęp: 16 grudnia 2018], *<http://www.stat.rutgers.edu/home/mxie/stat586/handout/Bootstrap1.pdf>*

[12] Basketball-Reference: Basketball Statistics and History, *<https://www.basketball-reference.com/>*

[13] DevianArt, [dostęp: 9 grudnia 2018], *<https://pre00.deviantart.net/d550/th/pre/i/2017/089/a/0/nba_playoff_bracket_by_nbaplayoffs-db410wn.jpg>*

[14] How many teams are in each conference in the NBA?, Maps of World, 9 lipca 2017, [dostęp: 16 grudnia 2018], *<https://www.mapsofworld.com/answers/united-states/many-teams-conference-nba/>*

[15] Moreyball: „The Houston Rockets and Analytics", Harvard Business School, 5 kwietnia 2018, [dostęp: 6 grudnia 2018] *<https://digit.hbs.org/submission/moreyball-the-houston-rockets-and-analytics/>*

[16] NBA History, NBA Hoops Online, [dostęp: 14 listopada 2018], *<https://nbahoopsonline.com/History/>*