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:
tutaj podajemy najważniejsze słowa kluczowe (łącznie nie powinny być dłuższe niż 150 znaków).

krótkie streszczenie:
Praca dotyczy aktualnego problemu z matematyki stosowanej, a mianowicie stworzenia symulatora ligi NBA przy użyciu znanych technik matematycznych/statystycznych. Celem pracy jest stworzenie symulatora rozgrywek w lidze NBA przy użyciu metod bootstrapowych. Wynikiem pracy będzie oprogramowanie do symulacji rozgrywek w lidze NBA 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 2018

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:
tutaj podajemy najważniejsze słowa kluczowe w języku angielskim (łącznie nie powinny być dłuższe niż 150 znaków)

short summary:
Tutaj piszemy krótkie streszczenie pracy w języku angielskim (nie powinno być dłuższe niż 530 znaków).

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*

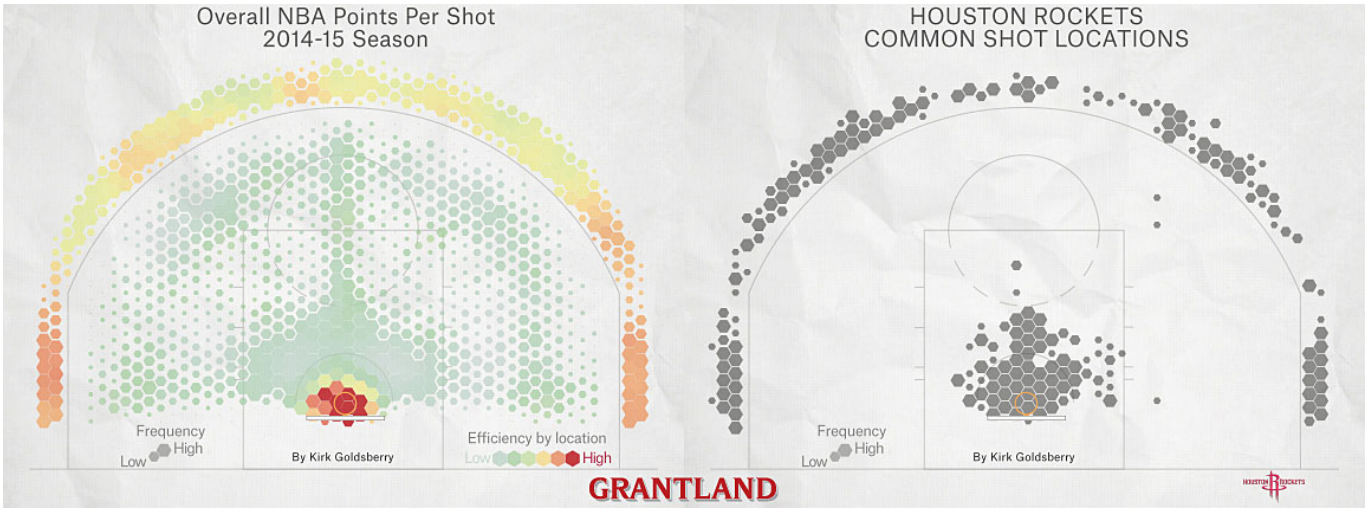
stamp of the faculty

Wrocław, 2018

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 masy 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 w postaci Los Angeles Lakers). Trudność w przewidzeniu wyników wynika bezpośrednio z zasad ligi — dozwolone są w niej wymiany zawodników między klubami, liczne i często katastrofalne kontuzje, czy nabory do ligi, w których najgorsze drużyny mają pierwszeństwo wyboru nowych zawodników chcących dołączyć do NBA (z tego powodu wiele organizacji celowo przegrywa mecze w celu wylosowania najwyższych miejsc w naborze). W obecnych czasach każda drużyna zatrudnia sztab analityków, którzy badają wpływ każdego czynnika obecnego na parkiecie na losy meczu. Idealnym przykładem wpływu statystyki na styl gry zespołu są Houston Rockets, którzy od sezonu 2014/2015 porzucili rzuty 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%. Wizualizacja tego systemu znajduje się na Rysunku 1. Poza analizowa-



Rysunek 1: Decyzje rzutowe Houston i reszty ligi

niem przebiegu gry, analitycy oraz statystycy podejmują próby przewidywania wyników sezonu, co staje się bardzo istotne w razie kontuzji lub wymiany gracza. 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. Celem tej pracy inżynierskiej jest próba przewidzenia rezultatów wybranego sezonu ligi NBA przy pomocy bardzo prostego modelu, używając wyłącznie informacji o

wynikach poszczególnych drużyn w poprzednich sezonach — indywidualny wpływ graczy na przebieg spotkań nie jest rozpatrywany. Na potrzeby tego zadania zaprojektowano kilka algorytmów opierających się na metodzie **bootstrapu parametrycznego**, sprawdzono ich skuteczność dla kilku wybranych sezonów w zależności od okresu używanych danych i nadawania wag, a następnie przy pomocy najskuteczniejszych modeli dokonano predykcji wyników trwającego sezonu 2018/2019.

Rozdział 1

Opis ligi NBA, jak wyglada sezon

National Basketball Association (NBA) 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 rywalizującą National Basketball League. Kolejna fuzja miała miejsce w 1976 roku, a mianowicie wchłonięcie połączenie z bardziej widowiskową American Basketball Association ocaliło ją od bankructwa i stagnacji. Po tym wydarzeniu NBA zyskała na atrakcyjności — z ABA zaczerpnięto pomysły rzutów za trzy punkty i organizacji konkursu wsadów, jak i przyjęto cztery dodatkowe organizacje. 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 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.

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. Terminarz wyznaczany jest wedle następujących reguł:

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)

Tabela 1.1: Drużyny Konferencji Wschodniej

Rysunek 1.1: Rozmieszczenie drużyn NBA

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)

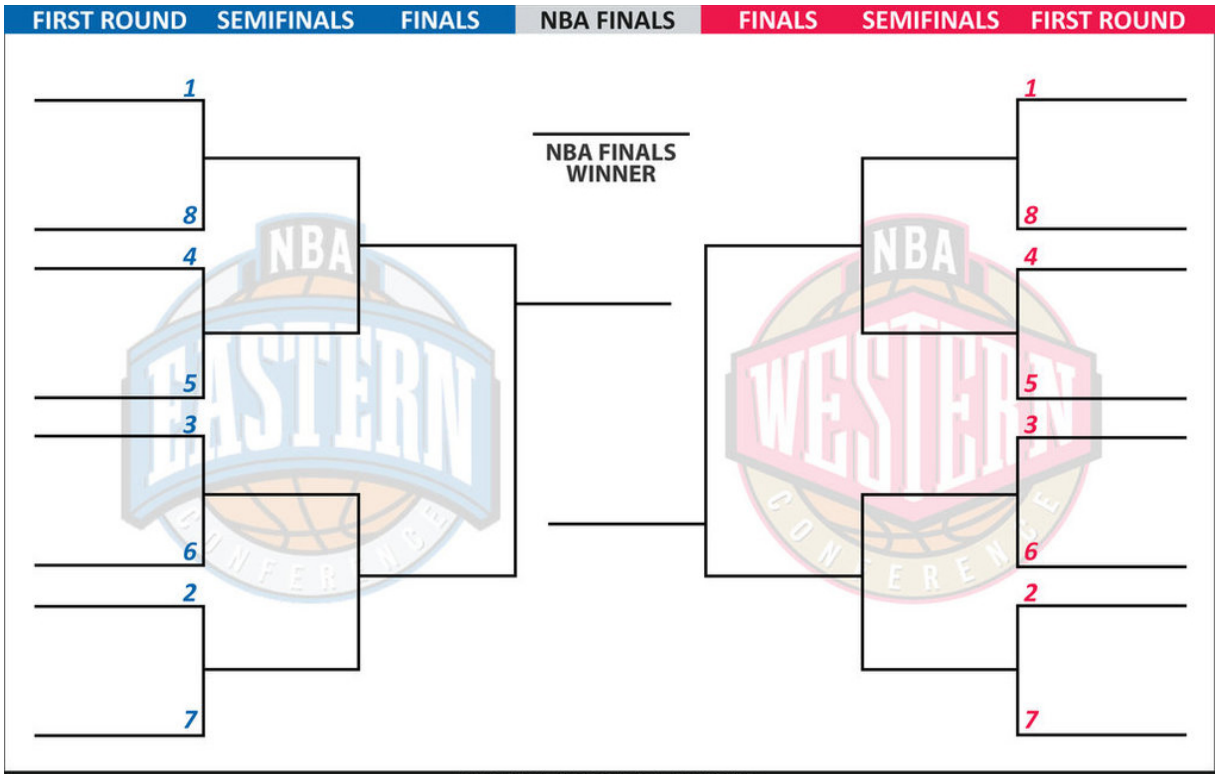
Tabela 1.2: Drużyny Konferencji Zachodniej

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

2. drużyny z tej samej dywizji grają ze sobą 4 spotkania (2 na wyjedzie 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 wyjedzie 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 z każdej konferencji (w razie takiej samej ilości zwycięstw dla obu zespołów decydują mecze bezpośrednie pomiędzy nimi). W tej fazie drużyny grają ze sobą maksymalnie 7 meczów, czyli zespół, który pierwszy wygra 4 mecze, przechodzi do następnego etapu.



Rysunek 1.2: Drzewko Playoff

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 grane są u lepszej z drużyn. Przy doborze przeciwników w tej fazie bierze się pod uwagę pozycję w tabeli konferencji: drużyna z miejsca pierwszego gra z zespołem o ósmym bilansie w danej konferencji, druga z siódmą, i tak dalej. Zwycięzca serii przechodzi do następnego etapu z czterema drużynami, po którym następują finały konferencji — najlepsze drużyny ze swoich konferencji spotykają się w finałach NBA. Dla lepszego zrozumienia systemu rozgrywek pucharowych na Rysunku 1.2 zamieszczono tzw. „drzewko Playoff".

Wyniki z sezonu regularnego 2014/2015 prezentują się następująco:

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

Tabela 1.3: Liczby zwycięstw drużyn w wybranych sezonach

Zwycięzca	**Przegrany**	**Wynik**
Eastern Conference First Round		
Atlanta Hawks	Brooklyn Nets	4-2
Chicago Bulls	Milwaukee Bucks	4-2
Cleveland Cavaliers	Boston Celtics	4-0
Washington Wizards	Toronto Raptors	4-0
Western Conference First Round		
Golden State Warriors	New Orleans Pelicans	4-0
Houston Rockets	Dallas Mavericks	4-1
Los Angeles Clippers	San Antonio Spurs	4-3
Memphis Grizzlies	Portland Trail Blazers	4-1
Eastern Conference Semifinals		
Atlanta Hawks	Washington Wizards	4-2
Cleveland Cavaliers	Chicago Bulls	4-2
Western Conference Semifinals		
Golden State Warriors	Memphis Grizzlies	4-2
Houston Rockets	Los Angeles Clippers	4-3
Eastern Conference Finals		
Cleveland Cavaliers	Atlanta Hawks	4-0
Western Conference Finals		
Golden State Warriors	Houston Rockets	4-1
Finals		
Golden State Warriors	Cleveland Cavaliers	4-2

Tabela 1.4: Rozgrywki pucharowe w 2015 roku

Rozdział 2

Teoria, matematyka

2.1 Rozkład jednostajny

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

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

Wartość oczekiwana 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 wyrażana 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 normalnego $\mathcal{U}(0, 1)$. Podczas generowania ciągu zbliżonego do realizacji rozkładu jednostajnego zazwyczaj stosuje się następującą metodę: wybierzmy funkcję G określoną i mająca 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 2.1. 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 $U_1, U_2, \ldots, U_n \sim \mathcal{U}(0,1)$ (popularne testy nie odrzucają hipotezy, wedle której $u_1, u_2, \ldots, u_n$ to realizacja próby $U_1, U_2, \ldots, U_n \sim \mathcal{U}(0,1)$).

2.3 Metoda Monte Carlo

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

Definicja 2.2. Niech $X_1, X_2, \ldots, X_m$ będzie próbą pseudolosową 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**.

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, który nie musi posiadać interpretacji probabilistycznej. Całość tej metody opiera się na generowaniu próby losowej lub pseudolosowej 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 Zasada bootstrap

Próba bootstrap to metoda służąca do oceny rozkładu pewnego estymatora (na przykład 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. 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 2.3. 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 **próbą bootstrap**.

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. Efron wykazał, że rozkład $T(X^*)$ dla ustalonych $x_1, x_2, \dots, x_n$ ma kształt zbliżony do rozkładu $T(X)$, ponadto rozkład statystyki $(T(X^*) - \hat{\Theta})$ jest bliski rozkładowi statystyki $(T(X) - \Theta)$. 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 $\Theta_1^* = T(X_1^*) - \hat{\Theta}$, $\Theta_2^* = T(X_2^*) - \hat{\Theta}$, $\dots$, $\Theta_k^* = T(X_k^*) - \hat{\Theta}$.

Otrzymany w ten sposób rozkład $(\Theta_1^*, \Theta_2^*, \dots, \Theta_k^*)$ to przybliżenie rozkładu błędów estymacji $\hat{\Theta} - \Theta$ przy pomocy statystyki T. Histogram tego rozkładu nazywany jest estymatorem rozkładu $\hat{\Theta}$ otrzymanym metodą bootstrap. Dla zadowalającego przybliżenia rozkładu $(T(X^*) - \hat{\Theta})$ wymagane jest przynajmniej $k = 1000$ prób bootstrap, a im większa ich ilość, 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. 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 proby przebiega podobnie do innych procesów bootstrapowych, przy czym wielkość takiej próby zazwyczaj odpowiada rozmiarowi oryginalnego zbioru danych. Prowadzi to do następującej definicji:

Definicja 2.4. Próbę bootstrap, której nieznany rozkład F można przybliżyć pewnym znanym rozkładem parametrycznym $\hat{F}$ nazywamy **parametryczną próbą bootstrap**.

Rozdział 3

Metodologia, algorytmy

3.1 Opis danych

Dane, które będą wykorzystywane do symulacji sezonu zostały zebrane ze strony sportowej basketballreference.com.
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 ilość wygranych jednej drużyny z drugą na przestrzeni lat dokonano następujących transformacji danych:
W zależności od interwału czasowego, jaki będziemy rozpatrywać, zebrano wyniki w określonych rozgrywkach (na przykład, przy wyznaczaniu wyników sezony 2014/2015 i interwale 5 lat, używać będziemy danych z lat 2009 do 2014). Dzięki uzyskanej w ten sposób liczbie wygranych w możemy stosunek zwycięstw do porażek dla wybranych zespołów (przykład: Boston Celtics i Atlanta Hawks grały ze sobą 10 razy, Jastrzębie wygrały zaledwie 4 razy, dlatego też w starciu z Celtami ich stosunek wygranych do przegranych wynosi 0.4). Po zastosowaniu tej metody dla wszystkich zespołów uzyskano macierz o rozmiarze 30 wierszy i kolumn zawierającą prawdopodobieństwa na wygraną z każdym zespołem w lidze.

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_i $i-$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 zaciągamy dane, x 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. Podczas 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. Podczas 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 lepszego dokonano dopasowania.

3.3 Modele symulacji rozgrywek

Zaproponowane w tym rozdziale modele opierają się na zasadzie bootstrapu parametrycznego. Podchodząc do symulacji znane są prawdopodobieństwa poszczególnych drużyn na wygraną, co pozwala na przybliżenie rozkładu teoretycznego danego sezonu.

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 wygląda następująco:

Algorytm 3.1. Model uśredniony

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

znajdź drużyny D_i i D_j
znajdź W_i i W_j
$W_{ij} \leftarrow \frac{W_i}{W_i+W_j}$
znajdź liczbę spotkań S_{ij} pomiędzy D_i i D_j
for $p \leftarrow 1$ to S_{ij} **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 for
end for
$i \leftarrow i + 1$
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 zwykłą łatwość w pokonaniu szczególnego przeciwnika. Algorytm symulowania wyników spotkań między drużynami wygląda następująco:

Algorytm 3.3. Model rywalizacji

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

$Z_i \leftarrow 0, Z_j \leftarrow 0, \ldots, B_{30} \leftarrow 0$
$i \leftarrow 1$
while $i \leq 30$ **do**
for $j \leftarrow i$ to 30 **do**
znajdź drużyny D_i i D_j
znajdź Z_{ij} i Z_{ji}
$W_{ij} = \frac{Z_{ij}}{Z_{ij}+Z_{ji}}$
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 wcześniej modeli korzysta się ze wcześniejszej symulacji fazy zasadniczej. W celu oddania trendów panujących w wygenerowanych rozgrywkach (a mianowicie potencjalnych kontuzjach, spadkach lub zwyżkach formy), użyta zostaje jedynie informacja o ilości wygranych przed rozpoczęciem Playoffów. Algorytm symulowania tej fazy jest postaci:

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

```
1: znajdź drużyny D_i i D_j
2: ZW_i ← 0, ZW_j ← 0
3: znajdź B_i i B_j
4: W_ij = B_i / (B_i + B_j)
5: while ZW_i < 4 and ZW_j < 4 do
6:     U ∼ U[0, 1]
```

```
    if W_ij ≤ U then
        B_i ← B_i + 1
    else
        B_j ← 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 modeli 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 zawodnikami. 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 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 3.1?

3.4.2 Modele predykcji wyników fazy pucharowej

Rozgrywki Playoff są niezwykle trudne do przewidzenia — niejednokrotnie zdarzyło się już, 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

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

Tabela 3.1: Porównanie średnich i median wygranych spotkań w symulacji

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 do Rundy Pierwszej w symulacjach, do Drugiej 4 zespoły o największym prawdopodobieństwie przejścia do drugiego etapu, kolejne etapy wyznaczane są analogicznie. PRZYKĄŁDOWA TABELA

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, jednak wymaga dużej liczby powtórzeń symulacji do poprawnego działania.

Rozdział 4

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 symulacji bootstrap.

4.1 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 symulowanych sezonów, dzięki czemu wybór lepszego modelu będzie ułatwiony. 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 wyników 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:

- 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, OPISZ BOXPLOT!!!!!!!!!!!

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

Tabela 4.1: Współczynnik G jakości dopasowania modelu

- 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.

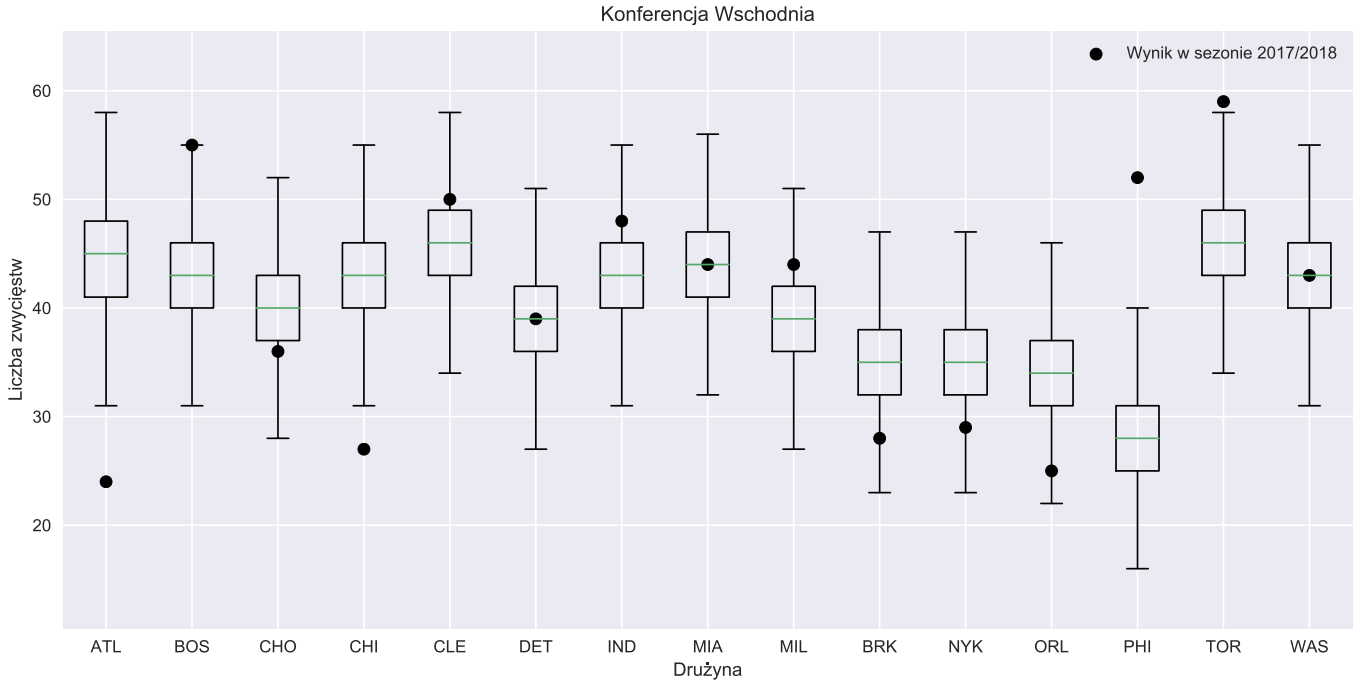
Patrząc na te wyniki nie można jednoznacznie odrzucić jednego modelu na korzyść drugiego, dlatego do wybrania najlepszego z nich dokonamy również analizy rozgrywek pucharowych. Ponownie analizując tabelę z wartościami G można zauważyć, że w dla rozgrywek 2017/2018 symulacje były bardziej trafne. Wyraźnie 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).

4.2 Wyniki symulacji dla fazy Playoff

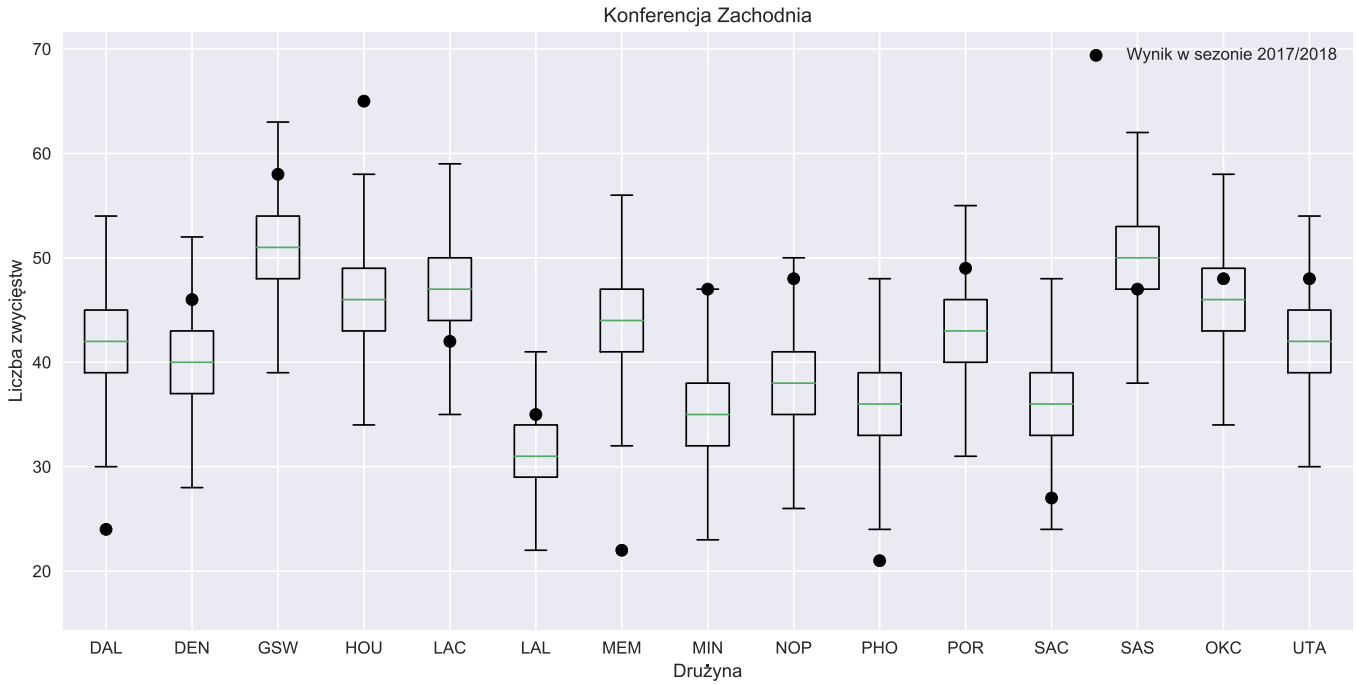
Dla wybranych w poprzedniej części modeli i parametrów przeprowadzona została symulacja rozgrywek pucharowych, korzystając z algorytmów zdefiniowanych w Rozdziale 3.4.2. w TABELACH REFY!!! porównano przewidziane na ich podstawie serie wraz z rzeczywistym przebiegiem rozgrywek w 2018 roku.

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

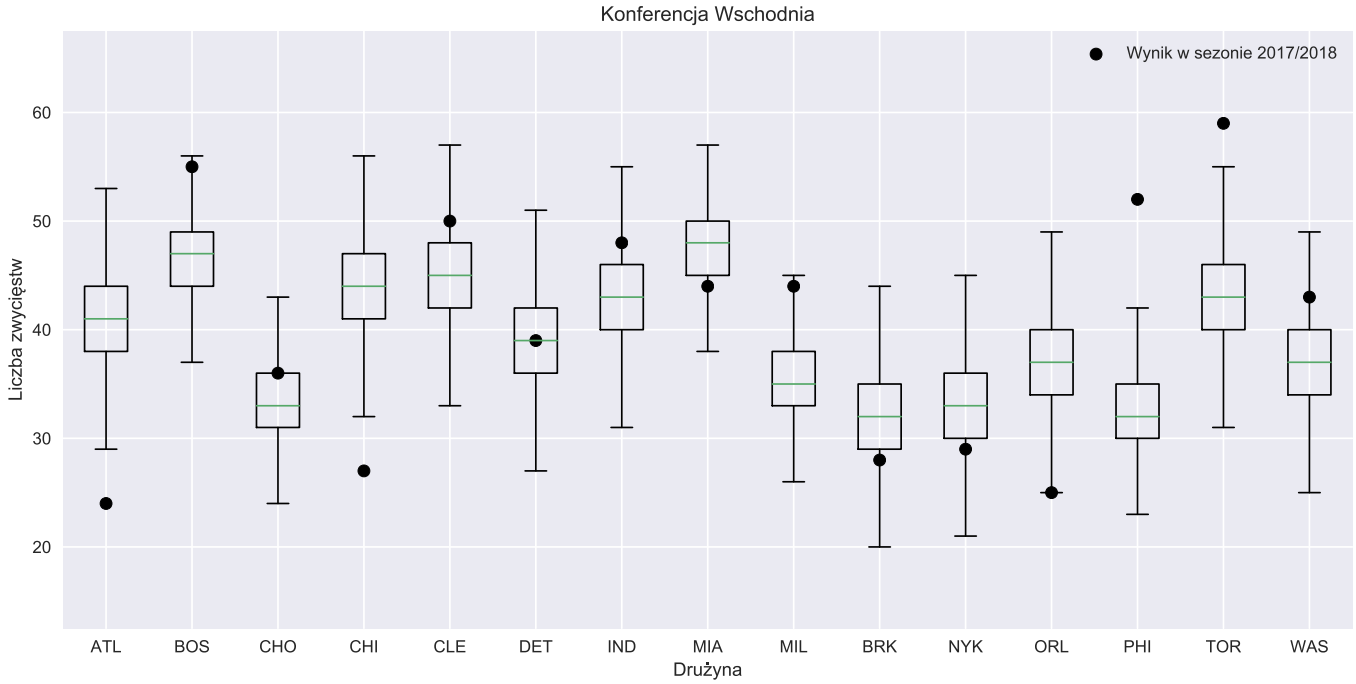
Tabela 4.2: Przebieg serii Playoff w symulacjach dla 2018 roku



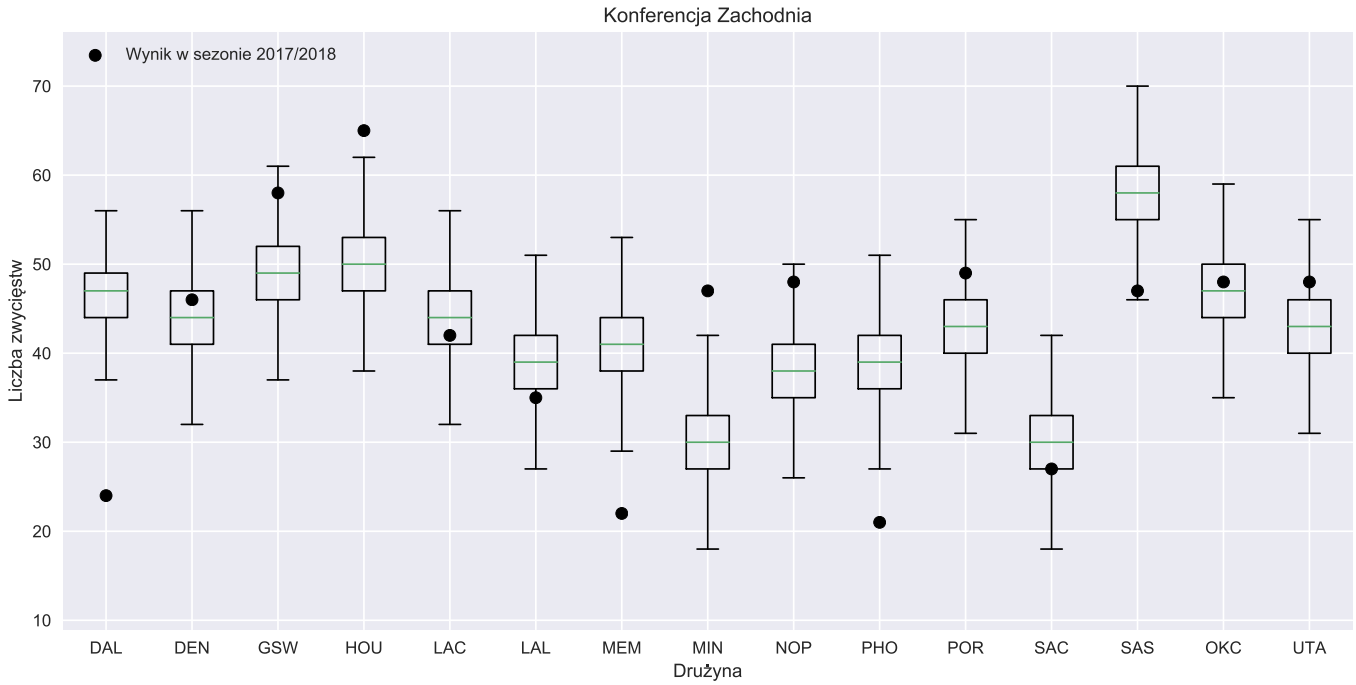
Rysunek 4.1: Model I, 5 lat, waga 1



Rysunek 4.2: Model I, 5 lat, waga 1



Rysunek 4.3: Model II, 10 lat, waga 0,5



Rysunek 4.4: Model II, 10 lat, waga 0,5

Rozdział 5

Wnioski

gęstości symulacji powinny mieć rozkład normalny? niekoniecznie?

Podsumowanie

Dodatek

tabele z prawdopodobieństwami, terminarz sezonu