

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

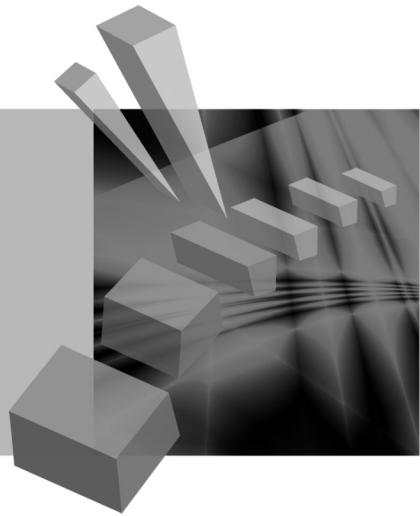
RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon [http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php](http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php)

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczak , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Tomasz Szubert

Uniwersytet Ekonomiczny w Poznaniu

W CO GRAĆ, ABY JAK NAJMNIJ PRZEGRĄĆ? PRÓBA KLASYFIKACJI SYSTEMÓW GRY W ZAKŁADACH BUKMACHERSKICH

Streszczenie: W opracowaniu próbowano udowodnić, że jakkolwiek strategia gry w zakładach bukmacherskich jest nieopłacalna dla graczy, a zyski z tej rozrywki czerpią jedynie firmy organizujące te zakłady. Dla realizacji postawionego zadania wykorzystano informacje o kursach meczów piłkarskich z 6 ostatnich lat w wybranych krajach europejskich. W badaniu posłużono się m.in. analizą korelacji i analizą skupień oraz regresją logistyczną. W toku przeprowadzonych analiz okazało się, że jedynie tzw. progresja z nieograniczoną, maksymalną stawką zakładu jest w stanie przynieść dodatnią stopę zwrotu z gry, ale wymaga ona zaangażowania bardzo dużego kapitału (niekiedy rzędu kilku milionów złotych).

Słowa kluczowe: sportowe zakłady bukmacherskie, analiza skupień, regresja logistyczna.

1. Wstęp

Zakłady bukmacherskie oparte na typowaniu wyników sportowych zyskują w ostatnim czasie coraz większą popularność w pewnych kręgach społeczeństwa. Wpływ na to miało wiele elementów: upowszechnienie Internetu, bogata oferta, a przede wszystkim chęć szybkiego i łatwego bogacenia się przez graczy. Działa tu bowiem pewna iluzja polegająca na zbytnej pewności swojej wiedzy na ten temat i wierze w przewidywalność tych zdarzeń [Dzik 2004, s. 578]. Choć w przypadku wydarzeń sportowych nie działa czysto losowy przypadek, to rywalizacja ta polega mimo wszystko na konfrontacji ludzkich zachowań, a te, jak wiadomo, są często nieprzewidywalne.

2. Cel i zakres pracy

Głównym celem artykułu jest próba porównania ze sobą kilku różnych rodzajów systemów gry w bukmacherskich zakładach sportowych. Do analizy wybrano następujące strategie: a) na gospodarza (oznaczane przez bukmacherów jako 1), czyli typowanie zwycięstwa gospodarza spotkania, b) na remis (X) – typowanie remisu w zdarzeniu, c) na gości (2) – typowanie zwycięstwa gości spotkania, d) na faworyta

– typowanie zwycięstwa tej drużyny, która ma największe szanse wygranej (co wynika z ustalonych przez bukmachera kursów), e) na średniego – typowanie takiego zdarzenia, którego prawdopodobieństwo realizacji nie jest ani najwyższe, ani najniższe, f) na słabego – typowanie zwycięstwa tej drużyny, która ma najmniejsze szanse wygranej, g) 1X – typowanie zwycięstwa gospodarzy lub remisu (wygrana gracza następuje, gdy zajdzie jedno z tych zdarzeń), h) X2 – typowanie zwycięstwa gości lub remisu (jak wyżej), i) 12 – typowanie zwycięstwa gospodarzy lub gości (jak wyżej), j) progresja na remisy – typowanie remisu w zdarzeniu, ale w taki sposób, że jeżeli nie wygra się w danym przypadku, w kolejnym podwaja się stawkę, jaką zagrało się ostatnio¹, k) zakład akumulacyjny „na dwóch faworytów” – typowanie zwycięstwa drużyn, które są uznawane za faworytów w dwóch oddzielnych spotkaniach. Wówczas stawka, jaką można wygrać, jest iloczynem pojedynczych kursów, ale wygrana następuje tylko wtedy, gdy oba te zdarzenia zajdą jednocześnie.

Hipoteza, jaka przyświecała podjętemu opracowaniu, zakłada, że żaden z wybranych systemów gry na dłuższą metę nie przynosi oczekiwanych rezultatów w postaci zysków dla graczy (stąd tak przewrotny tytuł artykułu), a działalność taka jest opłacalna jedynie dla firm prowadzących zakłady sportowe (stąd tak duża liczba takich placówek).

W badaniu wykorzystano kilka metod statystycznych. Po pierwsze, przeprowadzono symulacje gry w zakładach bukmacherskich na rzeczywistych danych pochodzących ze strony internetowej betexplorer.com, prowadzącej niezwykle obszerne archiwum kursów bukmacherskich. Po drugie, przy wykorzystaniu analizy korelacji próbowano znaleźć czynniki mogące podpowiadać graczom, na co zwrócić uwagę przy wyborze zdarzenia, które chce się obstawiać. Po trzecie, dokonano klasyfikacji wybranych strategii gry z wykorzystaniem hierarchicznej analizy skupień. Na koniec zastosowano regresję logistyczną, aby odkryć, jakie zdarzenia mają największe szanse wystąpienia. Wszystkie te metody miały służyć jako narzędzie do zrealizowania celu badawczego i zweryfikowania postawionej hipotezy.

W analizie oparto się na danych dotyczących kursów meczów piłkarskich odbywających się w 6 krajach: Polsce, Anglii, Francji, Hiszpanii, Niemczech i Włoszech w latach 2005–2011. Dyscyplina ta i wybrane państwa zapewniły największą wiarygodność przeprowadzonej analizy, chodziło bowiem o jak najmniejsze ryzyko manipulowania wynikami zdarzeń, co niekiedy ma miejsce w mniej popularnych dyscyplinach i bardziej egzotycznych zakątkach świata, kończąc się nierzadko aresztowaniem nawet samych zawodników podejrzanych o tzw. ustawianie wyników [Hughes, Pfanner 2009].

¹ Wybór remisu nie jest konieczny, jednak to remisy mają najczęściej kursy powyżej 2,00, a tylko taki kurs zapewni, że odzyskamy postawiony wcześniej kapitał i dodatkowo osiągniemy zysk.

3. Rezultaty badawcze

Analizę skuteczności poszczególnych strategii gry rozpoczęto od sprawdzenia, jaką stopę zwrotu można uzyskać, grając wybranym systemem podczas wszystkich meczów w danym roku² i w danej lidze (tab. 1). Dodatkowo sprawdzono, jaki zysk można by osiągnąć, grając na wszystkie mecze z 6 sezonów i 6 lig (wiersz „ogółem”).

Jak widać z przytoczonych danych, jest tylko jedna strategia (progresja na remisy), która zapewnia zysk dla gracza, za to jest on bardzo wysoki (wynosi średnio 58,3%, gdyby typować wszystkie mecze z analizowanych lat i lig). Z czego wynika tak duża jej skuteczność, zostanie opisane w dalszej części opracowania, gdzie jednocześnie zostaną pokazane wady tej strategii, bo nie jest ona mimo wszystko złotym środkiem na zarabianie w nieskończoność.

Z innych strategii najgorzej prezentuje się metoda obstawiania najsłabszej drużyny (średnia strata na poziomie 16,0%). Nie opłaca się zatem wybierać najwyższych kursów (typowych dla słabych zespołów), licząc, że drużyna skazana na porażkę przypadkiem zwycięży, a my dzięki niej osiągniemy duży zysk. Podobnie rzecz ma się w przypadku typowania gości danego meczu (średnia strata 15,0%). Drużyna, która rozgrywa mecz u siebie, zazwyczaj ma bowiem większe szanse zwycięstwa, gdyż przemawia za nią znajomość własnego boiska, doping kibiców, większa presja na osiągnięcie dobrego wyniku itp. I właśnie taka strategia typowania drużyny mającej osiągnąć zwycięstwo sprawdza się częściej, aczkolwiek stopa zwrotu też jest ujemna. Ciekawe jest także porównanie strategii „na faworyta” oraz „na dwóch faworytów”. W tym drugim przypadku, choć mnożąc oba kursy, można oczekiwać większych zysków, to jednak często zdarza się tak, że spośród dwóch drużyn, które obstawiliśmy, zwycięża tylko jedna, co nie zapewnia nam nawet zwrotu zainwestowanego kapitału.

Wracając do wspomnianej już strategii „progresja na remisy”, dokładne stopy zwrotu z gry na poszczególne ligi oraz sezony zaprezentowano na rys. 1. Jak widać, rzeczywiście strategia ta jest bardzo opłacalna, ale z kolei w 2010 r., typując mecze w lidze angielskiej, na koniec sezonu osiągnęlibyśmy stratę – bagatela – 92,5% zainwestowanego kapitału. Dzieje się tak, ponieważ grając tą strategią, za każdym razem, gdy w danym spotkaniu nie wystąpi remis, w kolejnym meczu musimy podwajać stawkę, co czasami prowadzi do tak absurdalnych sytuacji, że niekiedy gracz, chcąc odzyskać wcześniej zainwestowane pieniądze, musi wyłożyć nawet 10 mln złotych! W dłuższej perspektywie, obstawiając następnie 20, 40, 80 mln złotych, za którymś razem osiągnęlibyśmy zysk, ale firmy bukmacherskie, wiedząc, jak działa ta strategia, nakładają limity w postaci ograniczenia maksymalnej stawki, jaką gracz może obstawić na dane zdarzenie. Limit ten jest różny u różnych bukmacherów, zależy też od rodzaju typowanego zdarzenia, ale najczęściej oscyluje wokół 10. potęgi minimalnej stawki [Pope, Peel 1989, s. 323-341]. Limity te służą ponadto jako

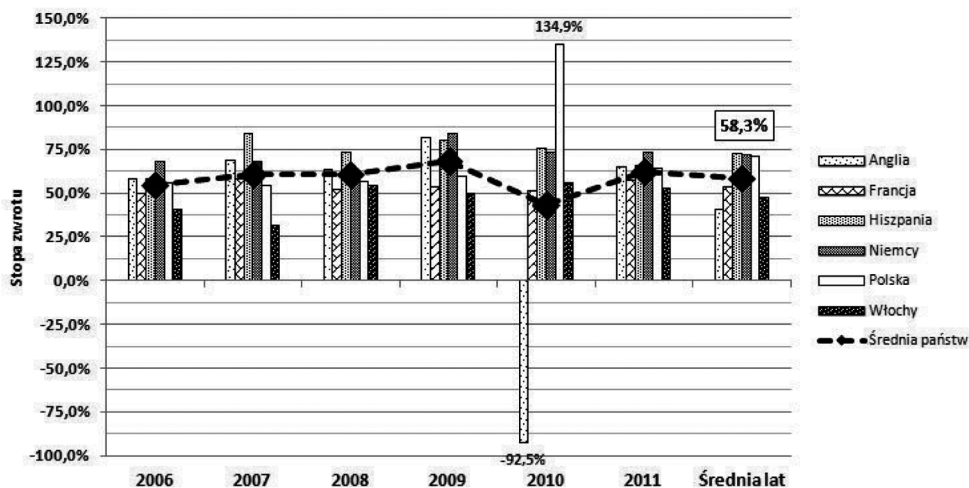
² Podany w tabelach, na rysunkach oraz w opisie wyników rok oznacza końcowy rok sezonu.

Tabela 1. Porównanie stóp zwrotu z gry w zakładach bukmacherskich według różnych strategii

Wyszczególnienie	Na faworyta	Na średniego	Na słabego	Na gospodarza	Na remisy	Na gości	Na 1X	Na X2	Na X12	Progresja	Na 2 faworytów
Ogółem	-3,8%	-11,8%	-16,0%	-5,4%	-11,5%	-15,0%	-7,2%	-12,6%	-7,2%	+58,3%	-7,6%
Lata											
2006	-6,6%	-13,2%	-17,6%	-10,8%	-11,0%	-15,9%	-9,4%	-12,9%	-9,5%	54,6%	-12,8%
2007	-8,6%	-8,9%	-14,4%	-6,1%	-9,8%	-16,7%	-7,8%	-12,4%	-9,3%	60,9%	-16,6%
2008	-1,8%	-12,7%	-22,0%	-6,9%	-13,2%	-15,7%	-8,4%	-13,4%	-6,8%	60,4%	-3,8%
2009	-0,4%	-14,6%	-17,1%	-3,8%	-14,9%	-13,5%	-8,4%	-12,5%	-4,7%	68,1%	-1,1%
2010	-1,4%	-10,7%	-16,8%	-2,9%	-9,5%	-17,6%	-4,5%	-13,4%	-6,8%	43,2%	-3,0%
2011	-4,2%	-10,4%	-8,1%	-2,2%	-10,3%	-10,3%	-4,8%	-10,8%	-6,0%	62,6%	-8,1%
Kraje											
Anglia	-3,4%	-12,1%	-18,9%	-0,9%	-10,4%	-23,2%	-4,7%	-14,8%	-6,5%	40,8%	-6,8%
Francja	-6,8%	-5,3%	-15,7%	-8,3%	-5,1%	-14,5%	-7,5%	-8,6%	-10,0%	53,5%	-13,3%
Hiszpania	-0,6%	-20,5%	-11,9%	-4,8%	-18,0%	-9,7%	-8,3%	-14,5%	-4,9%	72,7%	-0,9%
Niemcy	-7,5%	-9,6%	-5,3%	-8,4%	-9,9%	-4,9%	-8,7%	-7,7%	-7,2%	71,6%	-15,0%
Polska	-4,5%	-13,1%	-17,3%	-5,4%	-14,2%	-15,8%	-7,9%	-15,3%	-7,8%	70,9%	-9,0%
Włochy	-1,2%	-10,0%	-25,4%	-5,3%	-12,0%	-19,9%	-6,7%	-14,7%	-7,0%	47,5%	-2,4%

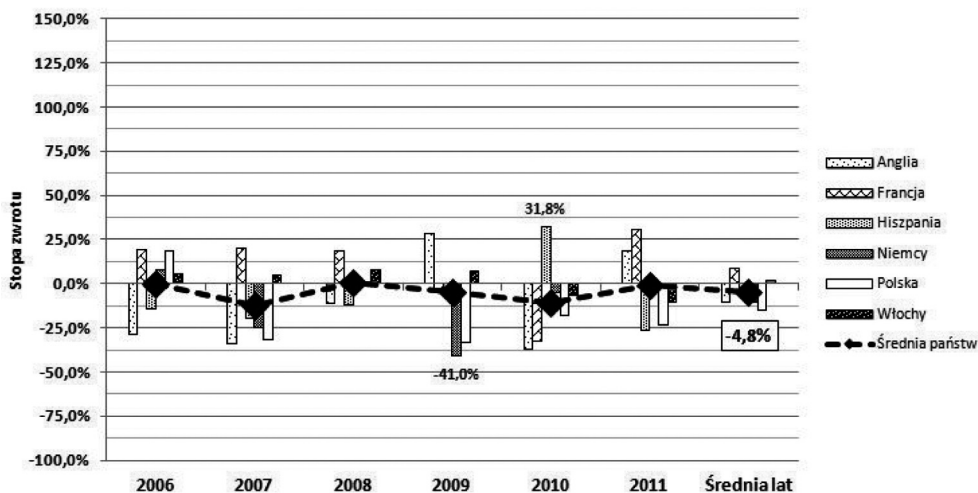
Źródło: opracowanie własne na podstawie kursów zdarzeń sportowych podawanych przez serwis Betexplorer.com.

zabezpieczenie przed niebezpieczeństwem poniesienia olbrzymich strat, które mogą spotkać nałogowych graczy, nieumiejących samemu stawiać sobie takich ograniczeń [Goodie 2005, s. 481-502].



Rys. 1. Stopy zwrotu otrzymywane z gry w zakładach bukmacherskich strategią „progresja na remisje”

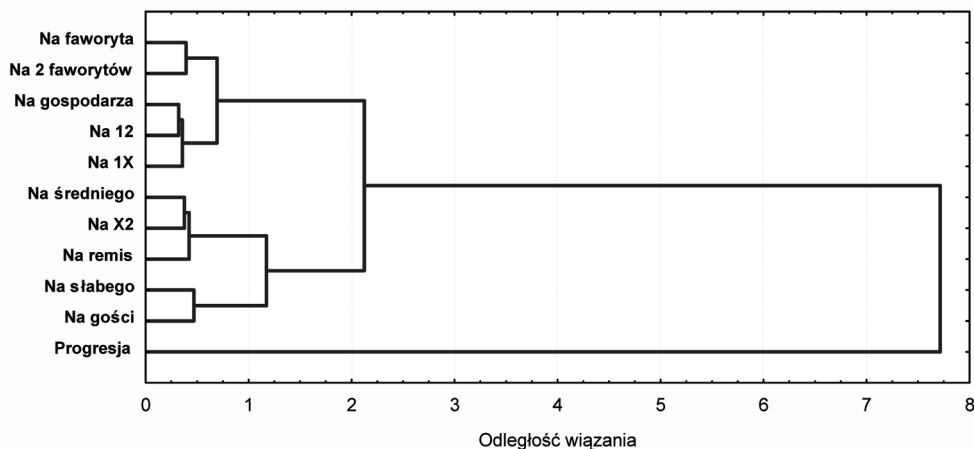
Źródło: jak w tab. 1.



Rys. 2. Stopy zwrotu otrzymywane z gry w zakładach bukmacherskich strategią „progresja na remisje” z ograniczeniem maksymalnej stawki do 1000 zł

Źródło: jak w tab. 1.

Rysunek 2 pokazuje z kolei, jak wyglądałyby stopy zwrotu w przypadku zmodyfikowanej wersji „progresji na remisy”, z maksymalną stawką 1000 zł. W takim przypadku oczekiwany zysk graczy jest już ujemny (wynosi $-4,8\%$), a tylko w 1/3 przypadków można było liczyć na zwrot kapitału. Największy z nich to 31,8% w lidze hiszpańskiej w 2010 r., największa strata wystąpiła w lidze niemieckiej w 2009 r. ($-41,0\%$). Widać więc, że progresja na remisy w takiej bardziej realistycznej formie nie jest już tak opłacalna, ale za to wahania stóp zwrotu nie są aż tak duże i potencjalna strata gracza (gdy przerwie progresję w jakimś niekorzystnym momencie) nie będzie tak olbrzymia jak w wersji bez nałożonego limitu.



Rys. 3. Klasyfikacja systemów gry w zakładach bukmacherskich według uzyskiwanych stóp zwrotu

Źródło: jak w tab. 1.

Po dokonaniu krótkiej charakterystyki wybranych strategii i porównaniu stóp zwrotu otrzymywanych z gry w poszczególnych latach i ligach, dokonano następnie klasyfikacji tych systemów, aby pogrupować je w klasy strategii podobnych do siebie pod kątem uzyskiwanych stóp zwrotu (rys. 3). Grupowanie wykonano metodą Warda z wykorzystaniem odległości euklidesowej. Okazało się, że można wyodrębnić 4 główne grupy strategii: pierwsza to wszystkie systemy związane z grą na faworyta lub na drużynę grającą u siebie (często jest tak, że gdy dwie drużyny o podobnym poziomie rozgrywają mecz, ta, która występuje u siebie, jest właśnie uznawana za faworyta i to jej przypisywane są mniejsze kursy, a tym samym większe prawdopodobieństwo wygranej). Drugą grupę systemów tworzą te związane z grą na remis i średnie kursy (najczęściej właśnie remisy otrzymują takie przeciętne kursy), w trzeciej klasie znajdują się strategie obstawiania gości i słabych drużyn (podobne uzasadnienie jak w przypadku gospodarzy–faworytów), a czwartą, zupełnie odrębną kategorię stanowi progresja na remisy, gdyż tylko ona jako jedyna zapewnia dodatnią stopę zwrotu.

Wiedząc już, że żadna z przedstawionych dotychczas strategii nie jest zadowalająca dla graczy (progresja w wersji z ograniczeniami również nie zapewniała zysków), spróbowano sprawdzić jeszcze jedną strategię, która nie byłaby aż tak mechaniczna jak poprzednie oraz wymagałaby większej inicjatywy i wiedzy gracza na temat odbywających się spotkań. Chodzi o typowanie zwycięstwa (lub remisu) konkretnego zespołu w każdym jego meczu, nieważne, czy gra u siebie czy na wyjeździe, czy jest uważany za faworyta czy nie. W tym przypadku możliwe byłoby uzależnienie wyboru konkretnej drużyny do typowania od wyników, jakie osiągnęła w poprzednim sezonie (np. od miejsca, jakie zajęła w rozgrywkach rok wcześniej, od liczby punktów, jakie zdobyła, grając u siebie i na wyjeździe, od liczby zwycięstw, remisów itp.) Zmienne te zamieszczono w tab. 2, w której zaprezentowano również wartości współczynników korelacji Pearsona pokazujących, jaki jest związek między daną cechą a uzyskaną stopą zwrotu, czyli czy np. liczba odniesionych w poprzednim sezonie zwycięstw powoduje, że w aktualnym roku drużyna, zachowując wysoką formę, również osiągnie dużo zwycięstw, a tym samym stopa zwrotu z typowania jej wygranych także może być wyższa.

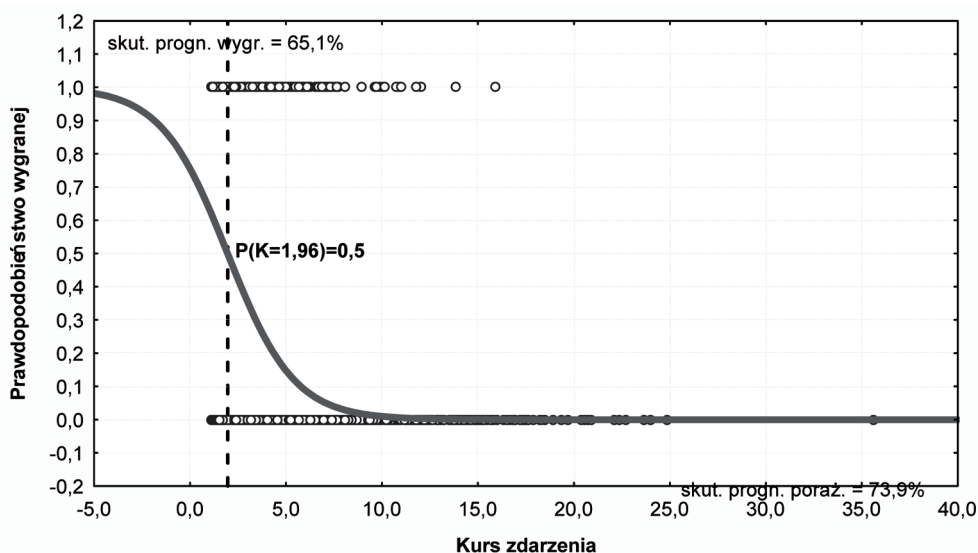
Tabela 2. Związki między wynikami drużyn w sezonie wcześniejszym a stopą zwrotu uzyskiwaną z typowania tego zespołu w aktualnym sezonie

Wyszczególnienie	Anglia		Francja		Hiszpania		Niemcy		Polska		Włochy	
	r_{xy}	P	r_{xy}	P	r_{xy}	P	r_{xy}	P	r_{xy}	P	r_{xy}	P
M_0 a SZ_1	-0,28	0,00	-0,07	0,50	-0,18	0,07	0,08	0,47	-0,04	0,74	-0,20	0,04
M_0 a SR_1	0,11	0,25	-0,04	0,65	0,34	0,00	0,08	0,45	-0,08	0,50	0,18	0,07
MD_0 a SZD_1	0,07	0,49	0,06	0,54	-0,16	0,11	0,10	0,36	-0,14	0,23	-0,15	0,12
MD_0 a SRD_1	0,01	0,90	-0,02	0,84	0,27	0,00	0,00	0,97	0,06	0,63	0,10	0,33
MW_0 a SZW_1	-0,28	0,00	-0,11	0,29	-0,11	0,25	0,03	0,81	0,01	0,93	-0,27	0,01
MW_0 a SRW_1	0,15	0,13	0,02	0,83	0,12	0,24	-0,08	0,42	-0,10	0,39	-0,06	0,54
Z_0 a SZ_1	0,28	0,00	0,07	0,49	0,17	0,08	-0,08	0,48	-0,03	0,83	0,29	0,00
ZD_0 a SZD_1	-0,04	0,66	-0,08	0,39	0,12	0,22	-0,14	0,17	0,08	0,52	0,17	0,09
ZW_0 a SZW_1	0,30	0,00	0,11	0,27	0,11	0,25	-0,03	0,80	-0,07	0,54	0,23	0,02
P_0 a SZ_1	0,30	0,00	0,05	0,58	0,21	0,04	-0,09	0,41	0,04	0,76	0,28	0,00
PD_0 a SZD_1	-0,04	0,66	-0,10	0,30	0,14	0,17	-0,16	0,14	0,13	0,29	0,17	0,08
PW_0 a SZW_1	0,29	0,00	0,14	0,16	0,15	0,12	-0,02	0,86	-0,01	0,91	0,23	0,02
R_0 a SR_1	0,06	0,53	0,01	0,90	0,16	0,10	-0,11	0,32	-0,11	0,35	-0,13	0,19
RD_0 a SRD_1	0,12	0,23	-0,09	0,35	0,13	0,21	-0,04	0,70	-0,16	0,17	0,05	0,63
RW_0 a SRW_1	0,00	0,98	0,01	0,96	-0,03	0,76	-0,18	0,09	0,01	0,92	-0,07	0,49

Oznaczenia: M – miejsce w tabeli, Z – zwycięstwa, P – punkty, R – remisy, S – stopa zwrotu, D – dom, W – wyjazd, 0 – sezon wcześniejszy, 1 – sezon badany.

Źródło: jak w tab. 1.

Jak się okazało, nie wszystkie związki są statystycznie istotne, ale np. w lidze hiszpańskiej, analizując wszystkie 6 lat i porównując miejsce w tabeli, jakie zajmowała dana drużyna, ze stopą zwrotu podczas obstawiania jej remisów, otrzymano wartość współczynnika korelacji na poziomie $+0,34$. Oznacza to, że im większe (czyt. dalsze) miejsce drużyna zajmowała w poprzednim sezonie w tabeli, czyli im była słabsza (1. miejsce oznacza najlepszą drużynę, ostatnie miejsce jest zarezerwowane dla najsłabszego zespołu), tym wyższą stopę zwrotu można było osiągnąć, wybierając ją jako swój obiekt do typowania remisów. Jest to logiczna zależność, bowiem im dalsze miejsce drużyna zajmowała w tabeli, tym była ona słabsza i najczęściej jedyne, na co ją było stać w nowym sezonie, to walka chociażby o remis. Niestety, z innych zaprezentowanych współczynników korelacji wynika, że ta i inne prawidłowości występują dosyć nieregularnie, zatem znów okazało się, że opieranie się na jakiejś „magicznej wiedzy” w postaci znajomości historii występów konkretnych drużyn raz przynosi efekt, a raz nie, ale na pewno nie jest trwałym sposobem na osiąganie zysków z gry w zakładach sportowych.



Rys. 4. Krzywa logistyczna opisująca prawdopodobieństwo realizacji zdarzenia o danym kursie

Źródło: jak w tab. 1.

W ostatniej części pracy dokonano sprawdzenia, czy niskie kursy, które oznaczają jednocześnie wysokie prawdopodobieństwo realizacji danego zdarzenia, rzeczywiście zapewniają większe szanse sprawdzenia się tego zdarzenia. W tym celu zastosowano regresję logistyczną, gdzie zmienną objaśniającą były poszczególne kursy zdarzeń, a dychotomiczną zmienną objaśnianą – realizacja (wartość 1) bądź

nie (wartość 0) danego zdarzenia. Przykładowy wykres wartości empirycznych oraz krzywej teoretycznej dla ligi angielskiej zaprezentowano na rys. 4. Pokazuje on, że niskie kursy rzeczywiście mają wyższe prawdopodobieństwo wygrania aniżeli kursy wysokie, aczkolwiek nawet kurs na poziomie ok. 16,00 ma szansę realizacji. Pocięszające jest to, że krzywa logistyczna ma spodziewany kształt, tzn. niskie kursy (poniżej pewnej wartości granicznej) zapewniają prawdopodobieństwo wygranej bliższe 1, a te wysokie mają małe szanse realizacji. Taką graniczną wartością jest w tym przypadku 1,96 (w innych ligach był to podobny poziom), a więc gracze, chcąc zmaksymalizować prawdopodobieństwo wygranej, powinni wybierać kursy niższe właśnie od tej wartości. Co ciekawe, wybór takich kursów w 65,1% przypadków rzeczywiście przyniósł wygraną, a wybór kursów powyżej 1,96 w 73,9% przypadków powodował, że zdarzenie się nie zrealizowało. Z danych tych rysuje się zatem pewien optymistyczny obraz, że dobierane przez bukmacherów kursy w pewien sposób oddają jednak rzeczywistość, bowiem niskie kursy oznaczające większe szanse zwycięstwa najczęściej przynoszą takiej drużynie wygraną.

4. Podsumowanie

Mimo popularności i powszechności różnego rodzaju gier hazardowych opartych na wydarzeniach sportowych, analitykom zajmującym się tą tematyką nie udało się do tej pory stworzyć systemu w 100% gwarantującego zysk. Zastosowanie wybranych metod statystycznych miało pozwolić taki model opracować lub przynajmniej odnaleźć system minimalizujący stratę. Okazało się, że z rozpatrywanych strategii gry najlepszym rozwiązaniem jest stosowanie progresji (stopniowe podwajanie stawki), ale tylko, gdyby nie istniały limity na maksymalną kwotę, jaką można zainwestować. Inne rozpatrywane strategie nie były niestety opłacalne.

Literatura

- Dzik B., *Hazard*, [w:] *Psychologia ekonomiczna*, Gdańskie Wydawnictwo Psychologiczne, T. Tyszka (red.), Gdańsk 2004.
- Goodie A., *The role of perceived control and overconfidence in pathological gambling*, „Journal of Gambling Studies” 2005, nr 21.
- Hughes R., Pfanner E., *Arrests in Europe Over Soccer Fixing Investigation*, New York Times, 20.11. 2009, <http://www.nytimes.com/2009/11/21/sports/soccer/21fix.html>.
- Pope P., Peel D., *Information, prices and efficiency in a fixed-odds betting market*, “Economica” 1989, no 56.

HOW TO PLAY TO LOSE THE LEAST? CLASSIFICATION OF SYSTEMS IN SPORTS BETS

Summary: The study undertaken attempts to prove that any strategy of sports betting is not profitable for the players and the profits of this entertainment are derived only by companies organizing those bets. To achieve the task, the information of football matches odds in 6 recent years in selected European countries was used. In the study correlation analysis, cluster analysis and logistic regression were applied. In the course of those analyses, it turned out that only progression with unlimited maximum rate of the bet was able to bring a positive financial return from the game, but it required a very large amount of capital (sometimes of several million zlotys), which for many players was an impossible barrier to overcome. In practice, bookmakers use limits of such maximum rates, blocking this way the chances of profit for players.

Keywords: sports bets, cluster analysis, logistic regression.