

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

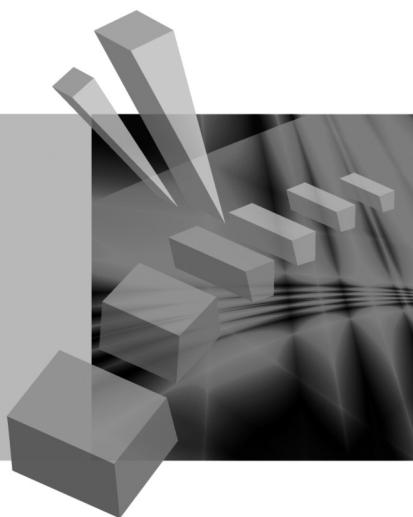
RESEARCH PAPERS

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,
Miroslaw 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 Błaszczyk , 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

Mariusz Kubus

Politechnika Opolska

ZASTOSOWANIE WSTĘPNEGO UWARUNKOWANIA ZMIENNEJ OBJAŚNIANEJ DO SELEKCJI ZMIENNYCH

Streszczenie: Paul i in. [2008] zaproponowali metodę wstępnego uwarunkowania zmiennej objaśnianej, którą realizuje się w dwóch krokach. W pierwszym przewidywane są wartości zmiennej objaśnianej za pomocą metody głównych składowych z nauczycielem. W drugim budowany jest jeszcze raz model regresji, ale na zbiorze uczącym, w którym rolę zmiennej objaśnianej odgrywają jej wartości teoretyczne estymowane w kroku pierwszym. Tu wykorzystuje się LASSO, które pozwala wyeliminować zmienne nieistotne. W artykule zaproponowano modyfikację oryginalnej metody oraz za pomocą symulacji oceniono ich przydatność w zadaniu selekcji zmiennych w przypadku modelu liniowego z interakcjami.

Słowa kluczowe: selekcja zmiennych, metoda wstępnego uwarunkowania, interakcje.

1. Wstęp

Zagadnienie selekcji zmiennych zajmuje obecnie znaczące miejsce w literaturze z zakresu statystycznych metod uczenia czy *data mining*. Zdolności predykcyjne modeli zależą w dużym stopniu od jakości danych, te natomiast zawierają często wiele zmiennych nieistotnych. Modelowanie w tych warunkach prowadzi do nadmiernego dopasowania do danych, a co za tym idzie, mniej dokładnych predykcji dla nowych obiektów, które nie brały udziału w etapie uczenia. Identyfikacja zmiennych istotnie wpływających na badane zjawisko ma walory interpretacyjne oraz może się przyczynić do zmniejszenia kosztów gromadzenia danych w przyszłości. W ostatnich latach pojawił się nowy obszar zastosowań dla selekcji zmiennych, który stanowi duże wyzwanie dla statystyków. Mowa o biologii obliczeniowej, a dokładniej o analizie macierzy ekspresji genów pozyskiwanych z DNA. Jednym z podstawowych zadań jest tu znajdowanie genów reagujących zmianą ekspresji na zmiany środowiskowe (np. podanie leku). Analizowane zbiory danych liczą nawet dziesiątki tysięcy zmiennych (genów) i jednocześnie nieliczne próby uczące (ze względu na kosztowność badań), zwykle nie więcej niż 200, a czasem tylko kilkadziesiąt obserwacji. Modelowanie predykcyjne jest wtedy szczególnie trudne ze względu na: brak stopni swobody, problem nadmiernego dopasowania do danych, niestabilność esty-

mowanych współczynników oraz problem ze współliniowością. Takim zbiorom danych szczególnie dedykowana jest dyskutowana w tym artykule metoda wstępnego uwarunkowania zmiennej objaśnianej (*pre-conditioning*) zaproponowana przez Paula i in. [2008].

Prezentowane w literaturze metody redukcji wymiaru przestrzeni cech można podzielić na dwa główne podejścia. Pierwsze polega na selekcji oryginalnych zmiennych wejściowych na podstawie wybranego kryterium. Istotną rolę odgrywa tu technika przeszukiwania przestrzeni wszystkich podzbiorów zmiennych. Wszystkie metody w tym nurcie klasyfikowane są obecnie do trzech grup (zob. np. [Guyon i in. 2006]): dobór zmiennych przed zastosowaniem algorytmu uczącego (*filters*), wyszukiwanie optymalnego podzbioru związane z oceną jakości modelu (*wrappers*) lub selekcja wewnątrz algorytmu uczącego (*embedded methods*). Drugie podejście do problemu redukcji wymiaru polega na transformacji oryginalnych zmiennych do przestrzeni o mniejszym wymiarze. Obok popularnej metody głównych składowych zaproponowano też jej modyfikację: metodę głównych składowych z nauczycielem (*supervised principal components*) [Bair i in. 2006]. Innym praktykowanym sposobem jest taksonomia cech (zob. np. [Li, Hong 2001; Hastie i in. 2001]). Dyskutowana w artykule metoda *pre-conditioning* wykracza jednak poza ramy podanej klasyfikacji.

W artykule proponowana jest modyfikacja metody wstępnego uwarunkowania zmiennej objaśnianej. Za pomocą symulacji dokonano oceny przydatności w zadaniu selekcji zmiennych zarówno metody oryginalnej, jak i proponowanej modyfikacji. Badania przeprowadzono dla modelu liniowego z interakcjami, a między predyktorami wprowadzano zależności liniowe. Wzięto również pod uwagę przypadek, gdy liczba zmiennych jest znacznie większa od liczby obiektów.

2. Metoda wstępnego uwarunkowania zmiennej objaśnianej

Metoda wstępnego uwarunkowania zmiennej objaśnianej (*pre-conditioning*) [Paul i in. 2008] realizowana jest w dwóch krokach. W pierwszym przewidywane są wartości zmiennej objaśnianej za pomocą metody głównych składowych z nauczycielem. Następnie empiryczne wartości zmiennej objaśnianej zamieniane są wartościami teoretycznymi estymowanymi przez metodę głównych składowych z nauczycielem. W drugim kroku budowany jest jeszcze raz model regresji, ale na zbiorze uczącym, w którym rolę zmiennej objaśnianej odgrywają jej wartości teoretyczne estymowane w kroku pierwszym. Tu proponuje się metodę LASSO [Tibshirani 1996], która pozwala wyeliminować zmienne nieistotne. Główną ideą całej tej procedury jest wyeliminowanie szumu w kroku pierwszym, by uzyskać lepsze rezultaty selekcji zmiennych w kroku drugim.

2.1. Metoda głównych składowych z nauczycielem

Zaproponowana przez autorów pracy [Bair i in. 2006] metoda głównych składowych z nauczycielem (*supervised principal components*) jest odpowiedzią na problem

klasycznej metody głównych składowych stosowanej w regresji, gdzie kierunki maksymalizujące wariancję nie muszą być współliniowe ze zmienną objaśnianą. Główną ideą metody jest wyznaczenie głównych składowych tylko dla podzbioru zmiennych – najbardziej skorelowanych ze zmienną objaśnianą. Szczegółowo algorytm wygląda następująco. W pierwszym kroku oceniana jest ważność predyktorów, a więc budowany jest ranking. Oryginalnie funkcję kryterium pełni moduł współczynnika regresji prostej. Następnie dla różnych wartości progowych wyznaczanych jest m głównych składowych, ale tylko dla zmiennych, które w rankingu przekroczyły ustalony próg. Wykonywana jest regresja względem nich oraz wybierana optymalna wartość progowa i liczba głównych składowych za pomocą sprawdzania krzyżowego.

2.2. Propozycja modyfikacji

Oryginalnie zaproponowana metoda *pre-conditioning* pozostawia możliwości modyfikacji, polegające na zastosowaniu innych metod regresji w poszczególnych krokach. W artykule proponuje się wykorzystanie metody *elastic net* [Zou, Hastie 2005] do selekcji zmiennych. Z kolei do wstępnego uwarunkowania proponowana jest metoda POLYMARS [Kooperberg, Bose, Stone 1997], ze względu na jej odporność na zmienne nieistotne oraz walory predykcyjne. W przeprowadzonych wstępnie badaniach symulacyjnych dla modelu liniowego z interakcjami (2) metoda ta wykazywała zdecydowaną przewagę nad innymi nieparametrycznymi metodami regresji ze względu na dokładność predykcji oraz czas obliczeń.

3. Selekcja zmiennych za pomocą regularyzowanej regresji liniowej

Selekcja zmiennych za pomocą regularyzowanej regresji liniowej jest szczególnie rekomendowana w przypadku dużych wymiarów przestrzeni cech. Ogromną popularnością cieszy się zaproponowana przez Tibshiraniego [1996] metoda LASSO oraz algorytm LARS [Efron i in. 2004] oferujący również rozwiązanie przybliżone LASSO. LARS jest szybkim algorytmem, który wprowadza iteracyjnie do modelu zmienne niosące najwięcej informacji o niewyjaśnionej części zmienności zmiennej objaśnianej (zob. także [Kubus 2011]). Maksymalna liczba iteracji jest równa liczbie zmiennych. Ostatecznie decyzję o liczbie zmiennych wprowadzanych do modelu podejmuje się za pomocą kryteriów jakości modelu. W regularyzowanej regresji liniowej parametry modelu β estymowane są przez minimalizację sumy funkcji straty (często kwadratowej) oraz komponentu kary P , który zależy od wartości współczynników oraz parametrów regularyzacji. Zou i Hastie [2005] zaproponowali (*elastic net*):

$$P(\lambda, \alpha, \beta) = \lambda \cdot \sum_{j=1}^p \left(\alpha \beta_j^2 + (1 - \alpha) |\beta_j| \right), \quad (1)$$

gdzie λ decyduje o rozmiarze kary za wielkość współczynników β_j , p jest liczbą zmiennych objaśniających, natomiast parametr α nadaje wagi wyrażeniom zaczerpniętym z regresji grzbietowej oraz LASSO. Metoda ta cechuje się tym, że jeśli pewna zmienna objaśniająca jest istotna, to zmienne współliniowe z nią są włączane do modelu. Tej cechy nie ma LASSO. Wbrew powszechnemu postulatowi o niewprowadzaniu do modelu zmiennych redundantnych, własność ta może być w pewnych sytuacjach korzystna, np. można ją wykorzystać do modelowania interakcji, co zilustrowane będzie w przeprowadzonym dalej eksperymencie.

Zmodyfikowaną wersję LASSO zaproponował Meinshausen [2007]. W metodzie *relaxed* LASSO estymatory uzyskuje się w dwukrokowej procedurze. Najpierw algorytm LARS dokonuje selekcji zmiennych, a następnie za pomocą drugiego parametru regularyzacji wartości bezwzględne współczynników jeszcze raz są zmniejszane (*shrinkage*). Metoda ta jest szczególnie rekomendowana dla wysokich wymiarów przestrzeni cech i dużej liczby zmiennych nieistotnych. Wyniki, jakie uzyskał Meinshausen [2007], dotyczą jednak przypadku zmiennych niezależnych.

4. Eksperyment

Ponieważ w wielu zastosowaniach praktycznych można się spodziewać zależności między zmiennymi objaśniającymi oraz zachodzących między nimi interakcji, badania przeprowadzono właśnie w takich warunkach. Rozważany będzie model:

$$y = \beta_0 + \sum_{j=1}^{10} \beta_j x_j + \beta_{11} x_1 x_2 + \beta_{12} x_2 x_3 + \beta_{13} x_4 x_5 + \beta_{14} x_6 x_7 x_8 + \sigma \varepsilon, \quad (2)$$

gdzie $\varepsilon \sim N(0;1)$. Motywacją wyboru takiego modelu był też fakt, że symulacje prezentowane w artykułach źródłowych ([Bair i in. 2006; Meinshausen 2007; Paul i in. 2008; Zou, Hastie 2005]) przeprowadzane są na ogół dla klasycznego modelu liniowego względem zmiennych, czasem tylko dla przypadku zmiennych ortogonalnych. Liczebności generowanych zbiorów uczących oraz testowych ustalono odpowiednio na 200 oraz 500. Szum gaussowski wprowadzano tylko w zbiorach uczących, a jego poziom losowano według formuły: $\sigma = m \cdot sd(y)$, gdzie $m \in \{0,5; 0,6; 0,7; 0,8; 0,9; 1\}$. Rozważono dwa przypadki ze względu na liczbę zmiennych p . Po wprowadzeniu dodatkowych zmiennych nieistotnych do zbiorów danych uzyskano: $p = 100$ lub $p = 1000$. Wprowadzano też zależności liniowe między zmiennymi w ten sposób, by co piąty predyktor był kombinacją liniową poprzedzających go czterech:

$$x_{5+k*5} = \alpha_{k1} x_{1+k*5} + \alpha_{k2} x_{2+k*5} + \alpha_{k3} x_{3+k*5} + \alpha_{k4} x_{4+k*5} + \sigma_k \varepsilon, \quad (3)$$

dla $k \in \{0,1, \dots, p/5 - 1\}$. Tym razem dla uzyskania silnej zależności poziom szumu losowano nieco mniejszy: $\sigma_k = m_k \cdot sd(x_{5+k*5})$, gdzie $m_k \in \{0,1; 0,2; 0,3; 0,4\}$.

Zwróćmy uwagę, że zależności (3) wprowadzono zarówno dla zmiennych nieistotnych, jak i dla zmiennych z modelu. Wszystkie współczynniki (modelu (2) oraz w zależności (3)), a także realizacje zmiennych objaśniających (z wyjątkiem x_{5+k*5}) generowano losowo z jednowymiarowego, standaryzowanego rozkładu normalnego.

W badaniu wzięto pod uwagę następujące metody:

- metodę głównych składowych z nauczycielem (spc),
- oryginalną metodę wstępnego uwarunkowania zmiennej objaśnianej (spc+lars),
- metodę głównych składowych z nauczycielem do wstępnego uwarunkowania oraz *elastic net* do selekcji zmiennych (spc+en),
- metodę POLYMARS do wstępnego uwarunkowania oraz *elastic net* do selekcji zmiennych (polymars+en).

Liczby wprowadzanych do modelu zmiennych obrazuje tab. 1.

Tabela 1. Średnie liczby wybieranych zmiennych nieistotnych/zmiennych z modelu (z odchyleniami standardowymi). W nawiasach podano mediany. Wyniki dla 100 symulacji

	spc	spc+lars	spc+en	polymars+en
$p = 100$	10,6±18,8 / 5,8±1,9 (2) / (6)	13,7±22,2 / 6,55±2,1 (3) / (6,5)	7,8±9,2 / 6,7±1,8 (5) / (7)	6,7±6,65 / 6,7±1,8 (4) / (7)
$p = 1000$	47,0±138,7 / 4,9±1,5 (6) / (5)	42,7±69,8 / 5,45±1,5 (7) / (6)	5,65±6,1 / 4,0±1,7 (5) / (4)	5,9±4,8 / 4,65±1,8 (4) / (4)

Źródło: obliczenia własne.

W przypadku $p = 100$ zaproponowane modyfikacje identyfikują więcej zmiennych z modelu. Mediany wprowadzanych zmiennych nieistotnych dla spc i spc+lars są wprawdzie mniejsze, ale średnie i odchylenia standardowe wysokie, co świadczy o niestabilności tych metod. W przypadku $p = 1000$ zaproponowane modyfikacje wprowadzają mniej zmiennych nieistotnych, a ponadto wyniki są bardziej stabilne. Na przykład metoda spc wprowadzała czasem ponad 600 zmiennych nieistotnych do modelu, natomiast oryginalna metoda *pre-conditioning* (spc+lars) dość często ok. 200 zmiennych nieistotnych. Zaproponowane modyfikacje identyfikowały jednak mniej zmiennych z modelu. Uzyskane wyniki należałoby więc zweryfikować błędami predykcji.

W tym miejscu należy podkreślić, że diskutowane metody stosowano wyłącznie w celu wstępnej selekcji zmiennych. Celem jest zbudowanie modelu liniowego z interakcjami, więc w drugim kroku analizy wprowadzone będą wyrażenia interakcyjne dla uzyskanego podzbioru zmiennych. Takie postępowanie wiąże się z szybkim wzrostem wymiaru przestrzeni cech i nie zawsze jest praktyczne. W przypadku zaproponowanych w artykule modyfikacji wzrost ten nie jest znaczny. Dla $p = 1000$

wymiar nowej przestrzeni w żadnej ze 100 symulacji nie przekraczał 200, a więc był przynajmniej pięciokrotnie mniejszy od wymiaru przestrzeni pierwotnej. Do selekcji zmiennych w drugim etapie analizy zastosowano metodę *relaxed* LASSO. Błędy predykcji estymowane na zbiorach testowych dla tak uzyskanych modeli liniowych z interakcjami ilustrują tab. 2-3. Dla porównania przedstawiono też błędy dla metody POLYMARS, którą stosowano na pełnym zestawie zmiennych. Dla zbadania istotności różnic między średnimi zastosowano test rangowanych znaków Wilcozona dla obserwacji zestawionych w pary (zob. np. [Aczel 2005]). W przypadku $p = 100$ średnie w zaproponowanych modyfikacjach okazały się istotnie mniejsze od średniej uzyskanej w oryginalnej metodzie *pre-conditioning* (p -wartości w porównaniach z *spc+en* oraz *polymars+en* wynosiły odpowiednio 0,0022 oraz 0,0013). Z kolei nie ma istotnych różnic w błędach predykcji pomiędzy proponowanymi modyfikacjami a metodą POLYMARS stosowaną dla pełnego zestawu zmiennych wejściowych.

Tabela 2. Średnie (ze 100 symulacji) błędy predykcji estymowane na zbiorach testowych wraz z odchyleniami standardowymi w przypadku 100 zmiennych objaśniających

	POLYMARS	spc+lars	spc+en	polymars+en
błąd predykcji	2,39 ± 0,93	2,73 ± 0,87	2,39 ± 0,90	2,38 ± 1,04

Źródło: obliczenia własne.

W przypadku $p = 1000$ (tab. 3) średnie błędy predykcji dla modeli liniowych z selekcją zmiennych za pomocą *spc+en* lub *polymars+en* są istotnie mniejsze od średniej uzyskanej metodą POLYMARS (p -wartości wynosiły odpowiednio 0,0016 oraz 0,0435).

Tabela 3. Średnie (ze 100 symulacji) błędy predykcji estymowane na zbiorach testowych wraz z odchyleniami standardowymi w przypadku 1000 zmiennych objaśniających

	POLYMARS	spc+en	polymars+en
błąd predykcji	3,68 ± 1,92	3,18 ± 1,30	3,59 ± 2,01

Źródło: obliczenia własne.

5. Podsumowanie

W artykule zaproponowano modyfikację metody wstępnego uwarunkowania zmiennej objaśnianej (*pre-conditioning*) polegającą na zastosowaniu innych metod regresji w wymaganych krokach. Badanie przeprowadzono dla modelu liniowego z interakcjami, gdzie między predyktorami zachodziły zależności liniowe. Dzięki zaproponowanym modyfikacjom uzyskano lepsze rezultaty w eliminacji zmiennych nieistotnych, co pozwoliło na wprowadzanie wyrażeń interakcyjnych bez nadmiernego

wzrostu wymiaru badanej przestrzeni. Po ponownej selekcji zmiennych metodą *relaxed* LASSO uzyskano modele liniowe o mniejszych błędach predykcji w porównaniu z oryginalną metodą *pre-conditioning*. W przypadku gdy liczba zmiennych była znacznie większa od liczby obserwacji, uzyskano też nieco mniejsze błędy od metody POLYMARS stosowanej na pełnym zestawie zmiennych wejściowych.

Literatura

- Aczel A.D., *Statystyka w zarządzaniu*, PWN, Warszawa 2005.
- Bair E., Hastie T., Paul D., Tibshirani R., *Prediction by supervised principal components*, „J. Amer. Statist. Assoc.” 2006, no 101.
- Efron B., Hastie T., Johnstone I., Tibshirani R., *Least angle regression*, „Annals of Statistics” 2004, no 32 (2).
- Guyon I., Gunn S., Nikravesh M., Zadeh L., *Feature Extraction: Foundations and Applications*, Springer, New York 2006.
- Hastie T., Tibshirani R., Botstein D., Brown P., *Supervised harvesting of expression trees*, „Genome Biol.” 2001, no 2.
- Kooperberg C., Bose S., Stone C.J., *Polychotomous regression*, „Journal of the American Statistical Association” 1997, no 92.
- Kubus M., *Analiza metody LARS w problemie selekcji zmiennych w regresji*, [w:] K. Jajuga, M. Waleśiak (red.), *Taksonomia 18, Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, UE, Wrocław 2011.
- Li H., Hong F., *Cluster-Rasch models for microarray gene expression data*, „Genome Biol.” 2001, no 2.
- Meinshausen N., *Lasso with relaxation*, „Computational Statistics and Data Analysis” 2007, no. 52(1).
- Paul D., Bair E., Hastie T., Tibshirani R., *“Pre-conditioning” for feature selection and regression in high-dimensional problems*, „Annals of Statistics” 2008, no 36(4).
- Tibshirani R., *Regression shrinkage and selection via the lasso*, „J.Royal. Statist. Soc. B.” 1996, no 58.
- Zou H., Hastie T., *Regularization and variable selection via the elastic net*, „Journal of the Royal Statistical Society Series B.” 2005, no 67(2).

THE APPLICATION OF PRE-CONDITIONING OF EXPLANATORY VARIABLE FOR FEATURE SELECTION

Summary: Paul et al. [2008] proposed pre-conditioning of explanatory variable that consists of two steps. Firstly, supervised principal components are applied for prediction and then regression model is constructed again. This time the response is replaced with the previously obtained predictions. The second regression model is LASSO, which eliminates irrelevant variables. The modification of original pre-conditioning is proposed in this paper. The usefulness of original pre-conditioning as well as proposed modifications are assessed using simulations. The experiment was carried out for linear model with interactions.

Keywords: feature selection, pre-conditioning, interactions.