

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

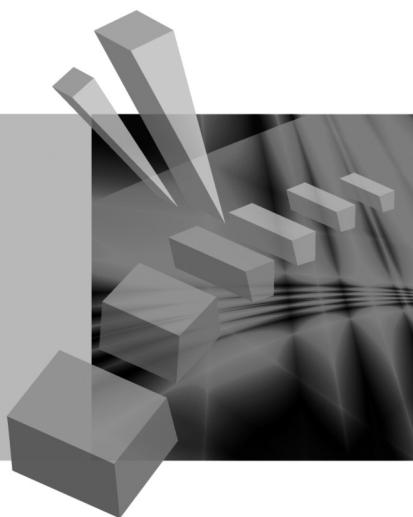
RESEARCH PAPERS

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 wspomaganie 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
Sylwia 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 Dębkowska, 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 Pawelczyk, 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

Joanna Banaś, Małgorzata Machowska-Szewczyk

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

OCENA INTENSYWNOŚCI WYKORZYSTANIA SKRZYNEK POCZTY ELEKTRONICZNEJ ZA POMOCĄ UPORZĄDKOWANEGO MODELU PROBITOWEGO

Streszczenie: Celem artykułu jest analiza opinii studiujących użytkowników kont pocztowych na temat posiadanych skrzynek poczty. Źródło danych stanowiły kwestionariusze ankiety wypełniane przez studentów wybranych wydziałów szczecińskich uczelni, które zawierały pytania dotyczące zarówno charakterystyki skrzynki poczty elektronicznej, jak i intensywności jej używania. Do oceny obsługi poczty elektronicznej wykorzystano uporządkowany model probitowy, w którym występuje jakościowa zmienna zależna, opisująca kategorie uporządkowane przy założeniu rozkładu normalnego składnika losowego.

Słowa kluczowe: uporządkowany model probitowy, badanie ankietowe, metoda regresji krokowej.

1. Wstęp

Poczta elektroniczna jest obecnie jedną z najczęściej używanych form komunikacji między ludźmi i instytucjami. Większość portali społecznościowych oferuje bezpłatne założenie konta i dostęp do poczty elektronicznej przez stronę internetową WWW. Wiele szkół, wyższych uczelni czy firm również umożliwia upoważnionej grupie zakładanie własnych skrzynek poczty elektronicznej. Podstawowym celem artykułu jest analiza opinii studiujących użytkowników kont pocztowych na temat posiadanych skrzynek poczty. Badanie zostało przeprowadzone na podstawie kwestionariuszy ankiety wypełnianych przez studentów wybranych wydziałów szczecińskich uczelni. Kwestionariusz zawierał pytania dotyczące zarówno charakterystyki skrzynki poczty elektronicznej, jak i intensywności jej używania. Każdy respondent mógł także ocenić sposób prezentacji danych w serwisie internetowym i częstość występowania problemów z wykorzystaniem niektórych usług.

Ponieważ większość odpowiedzi na pytania występujące w kwestionariuszu ankiety dostarcza informacji o charakterze jakościowym, wyrażonej w skali nominalnej lub porządkowej, to do oceny obsługi poczty elektronicznej wykorzystano uporządkowany model probitowy, w którym występuje jakościowa zmienna zależna, opisu-

jąca kategorie uporządkowane, przy założeniu normalnego rozkładu składnika losowego. Model taki umożliwiłby zbadanie wpływu poszczególnych zmiennych na poziom oceny posiadanej skrzynki poczty elektronicznej.

2. Opis metody

Uporządkowany model probitowy służy do opisu jakościowej zmiennej zależnej, której warianty można uporządkować. Przykładem takiej zmiennej jest subiektywna ocena sytuacji finansowej rodziny z wartościami: 1 – zła, 2 – średnia, 3 – dobra, 4 – bardzo dobra. Mogą to być również zmienne ilościowe, których wartości nie są dokładnie znane, a dostępna jest informacja, z jakiego przedziału pochodzą. Model ten po raz pierwszy został przedstawiony w pracy Zavoina i McKelveya [1975]. Aby dokonać formalnego zapisu modelu probitowego, zakłada się istnienie nieobserwowalnej zmiennej Y^* [Kostrzewska 2010; Dudek 2007]. Na podstawie jej wartości otrzymanych dla i -tej obserwacji n -elementowej próbki przypisuje się kategorii zmiennej zależnej Y następująco:

$$y_i = \begin{cases} 0, & \text{gdy } y_i^* \leq \mu_1 \\ 1, & \text{gdy } \mu_1 < y_i^* \leq \mu_2 \\ 2, & \text{gdy } \mu_2 < y_i^* \leq \mu_3, \\ \vdots & \\ J, & \text{gdy } y_i^* > \mu_J \end{cases} \quad (1)$$

gdzie $\mu_1, \mu_2, \dots, \mu_J$ to wartości progowe, umożliwiające nadawanie kodów zmiennej Y , Y^* to nieobserwowalna zmienna, o której zakłada się, że $Y^* = \beta_0 + \mathbf{X}\boldsymbol{\beta} + U$, gdzie $\mathbf{X} = [X_1, \dots, X_K]$ jest wektorem losowym obserwowanych zmiennych, $\boldsymbol{\beta} = [\beta_1, \dots, \beta_K]^T$ – wektorem parametrów modelu, β_0 to wyraz wolny, U – składnik losowy o rozkładzie normalnym standaryzowanym.

Do modelowania prawdopodobieństwa zajścia zdarzenia, że zmienna Y przyjmie ustaloną wartość $y_i \in \{0, 1, 2, \dots, J\}$ pod warunkiem, że zmienne $X_1 = x_{i1}, \dots, X_K = x_{iK}$ można wykorzystać założenie, że składnik losowy U ma rozkład normalny standaryzowany i wyznaczyć je za pomocą dystrybuanty rozkładu normalnego standaryzowanego Φ :

$$\begin{aligned} P(Y = y_i / X_1 = x_{i1}, \dots, X_K = x_{iK}) &= P(y_i = j / X_1 = x_{i1}, \dots, X_K = x_{iK}) = \\ &= P(\mu_j < y_i^* \leq \mu_{j+1}) = P(\mu_j < \mathbf{x}_i \mathbf{b} + b_0 + u_i \leq \mu_{j+1}) = \\ &= P(\mu_j - \mathbf{x}_i \mathbf{b} - b_0 < u_i \leq \mu_{j+1} - \mathbf{x}_i \mathbf{b} - b_0) = \\ &= \Phi(\mu_{j+1} - \mathbf{x}_i \mathbf{b} - b_0) - \Phi(\mu_j - \mathbf{x}_i \mathbf{b} - b_0). \end{aligned} \quad (2)$$

Zatem funkcja prawdopodobieństwa przyjęcia wariantu i przez zmienną Y dla konkretnej obserwacji jest następująca:

$$P(Y = y_i / \mathbf{X} = \mathbf{x}_i) = \begin{cases} \Phi(\mu_1 - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = 0 \\ \Phi(\mu_2 - \mathbf{x}_i \mathbf{b} - b_0) - \Phi(\mu_1 - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = 1 \\ \Phi(\mu_3 - \mathbf{x}_i \mathbf{b} - b_0) - \Phi(\mu_2 - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = 2 \\ \vdots & \\ \Phi(\mu_J - \mathbf{x}_i \mathbf{b} - b_0) - \Phi(\mu_{J-1} - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = J-1 \\ 1 - \Phi(\mu_J - \mathbf{x}_i \mathbf{b} - b_0), & , \text{ gdy } y_i = J \end{cases} \quad (3)$$

gdzie $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iK}]$ – wektor wartości zmiennych X_1, \dots, X_K zaobserwowanych dla i -tej jednostki, $\mathbf{b} = [b_1, \dots, b_K]^T$ oraz b_0 – wartości estymatorów parametrów $\boldsymbol{\beta}$ oraz β_0 odpowiednio uzyskane z próbki.

Po oszacowaniu za pomocą metody największej wiarygodności parametrów modelu \mathbf{b} oraz b_0 (przy czym wartości estymatorów nie interpretuje się bezpośrednio) warto dokonać predykcji efektów zmian wartości zmiennych objaśniających na wartość prawdopodobieństwa przynależności do każdej z tych grup. Owe efekty krańcowe wpływu zmian wartości zmiennej X_k , $k \in \{1, 2, \dots, K\}$ na prawdopodobieństwo, że zmienna zależna Y przyjmie dla i -tej obserwacji ustaloną wartość porządkową $y_i \in \{0, 1, 2, \dots, J\}$, wyznacza się, wykorzystując pochodną cząstkową [Kostrzewska 2010]:

$$\frac{\partial P(Y = y_i / \mathbf{X} = \mathbf{x}_i)}{\partial x_{ik}} = \begin{cases} b_k \varphi(\mu_1 - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = 0 \\ b_k (\varphi(\mu_2 - \mathbf{x}_i \mathbf{b} - b_0) - \varphi(\mu_1 - \mathbf{x}_i \mathbf{b} - b_0)) & , \text{ gdy } y_i = 1 \\ b_k (\varphi(\mu_3 - \mathbf{x}_i \mathbf{b} - b_0) - \varphi(\mu_2 - \mathbf{x}_i \mathbf{b} - b_0)) & , \text{ gdy } y_i = 2 \\ \vdots & \\ b_k (\varphi(\mu_J - \mathbf{x}_i \mathbf{b} - b_0) - \varphi(\mu_{J-1} - \mathbf{x}_i \mathbf{b} - b_0)) & , \text{ gdy } y_i = J-1 \\ -b_k \varphi(\mu_J - \mathbf{x}_i \mathbf{b} - b_0) & , \text{ gdy } y_i = J \end{cases} \quad (4)$$

gdzie φ oznacza funkcję gęstości rozkładu normalnego standaryzowanego.

Jeżeli zmienna niezależna X_k , $k \in \{1, 2, \dots, K\}$ jest dychotomiczna, to jej wpływ na zmianę prawdopodobieństwa, że zmienna objaśniana Y przyjmie dla i -tej obserwacji ustaloną wartość porządkową $y_i \in \{0, 1, 2, \dots, M\}$, wyznaczany jest jako różnica odpowiednich prawdopodobieństw według wzoru (5):

$$\begin{aligned} & P(Y = y_i / X_1 = x_{i1}, \dots, X_k = 1, \dots, X_K = x_{iK}) - \\ & P(Y = y_i / X_1 = x_{i1}, \dots, X_k = 0, \dots, X_K = x_{iK}). \end{aligned} \quad (5)$$

Wpływy cząstkowe oblicza się zazwyczaj w punkcie wyznaczonym jako wektor wartości średnich poszczególnych zmiennych niezależnych. Oszacowane parametry modelu wskazują na taki sam wpływ konkretnej zmiennej objaśniającej na prawdopodobieństwo każdego ze stanów osiąganych przez zmienną zależną. Wpływy cząstkowe umożliwiają ocenę oddziaływania każdej ze zmiennych indywidualnie na zmiany prawdopodobieństwa każdego ze stanów zmiennej objaśnianej.

3. Materiał badawczy

Aby przeprowadzić ocenę intensywności wykorzystania skrzynek poczty elektronicznej przez studentów, zgromadzono materiał badawczy techniką CAWI [Mazurek-Łopacińska 2005]. Na podstawie wypełnionych przez studentów kilku sześciennych uczelni kwestionariuszy ankiety [Górski 2011] od marca do maja 2011 r. zbudowano uporządkowany model probitowy. Spośród dostępnych informacji wyodrębniono grupę 252 studentów, którzy zadeklarowali posiadanie adresu poczty elektronicznej. Zmienną zależną zdefiniowano na podstawie odpowiedzi na pytanie 4. w kwestionariuszu ankiety:

Ile wiadomości e-mail wysyłasz w ciągu dnia?

Zmienna zależna Y może dla i -tej obserwacji w próbie przyjmować wartości y_i określone następująco:

$$y_i = \begin{cases} 0, & \text{gdy } y_i^* \leq 2 & \text{(nie więcej niż 2)} \\ 1, & \text{gdy } 2 < y_i^* \leq 5 & \text{(od 3 do 5)} \\ 2, & \text{gdy } 5 < y_i^* \leq 9 & \text{(od 6 do 9)} \\ 3, & \text{gdy } y_i^* > 9 & \text{(10 lub więcej)} \end{cases}$$

Przy analizie ocen skrzynek poczty elektronicznej na podstawie kwestionariusza ankiety przyjęto 31 potencjalnych zmiennych objaśniających X_k . Na podstawie każdej zmiennej o charakterze jakościowym, której liczba możliwych kategorii była większa niż 2, utworzono odpowiednią liczbę nowych zmiennych binarnych, odzwierciedlających poszczególne kategorie. Wprowadzono oznaczenie $X_{k,j}$ dla zmiennych binarnych utworzonych na podstawie zmiennej X_k , gdzie $k \in \{1, \dots, 31\}$ jest numerem zmiennej, j zaś oznacza numer wariantu wyjściowej zmiennej. Zmiennej $X_{k,j}$ przypisuje się wartość jeden, jeżeli wystąpił wybrany wariant (zmienna X_k przyjęła j -ty wariant cechy), natomiast w przeciwnym przypadku wartość zero. Tabela 1 przedstawia zbiór zmiennych potencjalnie objaśniających (zmiennie $X_{18,3}$, $X_{26,1}$, $X_{31,3}$, $X_{31,4}$ usunięto, gdyż żaden z badanych nie wybrał takiej kategorii, co zaznaczono przez przekreślenie).

Tabela 1. Zmienne objaśniające (wytluszczone zmienne bazowe).

Zmienna	Opis zmiennych binarnych
1	2
X_1	Czas posiadania adresu poczty elektronicznej. Zmienne binarne $X_{1,1}, X_{1,2}, X_{1,3}$ odpowiadające kategoriom: krócej niż 5 lat , 5-9 lat, 10 lat lub więcej
X_2	Liczba adresów poczty elektronicznej. Zmienne binarne $X_{2,1}, X_{2,2}, X_{2,3}$ odpowiadające kategoriom: 1, 2-3, 4 lub więcej
X_3	Częstość sprawdzania poczty elektronicznej w ciągu dnia. Zmienne binarne $X_{3,1}, X_{3,2}, X_{3,3}$ odpowiadające kategoriom: rzadko, średnio , często
X_4	Główni adresaci wysyłanych wiadomości e-mail. Zmienne binarne $X_{4,1}, X_{4,2}, X_{4,3}, X_{4,4}, X_{4,5}$ odpowiadające kategoriom: rodzina, osoby związane z pracą, znajomi, osoby związane ze szkołą, inni
X_5	Posiadanie <i>uczelnianej</i> skrzynki pocztowej (mail-box) (1 – tak, 0 – nie)
X_6	Posiadanie skrzynki pocztowej (mail-box) na portalu <i>wp</i> (1 – tak, 0 – nie)
X_7	Posiadanie skrzynki pocztowej (mail-box) na portalu <i>onet</i> (1 – tak, 0 – nie)
X_8	Posiadanie skrzynki pocztowej (mail-box) na portalu <i>interia</i> (1 – tak, 0 – nie)
X_9	Posiadanie skrzynki pocztowej (mail-box) na portalu <i>o2 (tlen)</i> (1 – tak, 0 – nie)
X_{10}	Posiadanie skrzynki pocztowej (mail-box) na portalu <i>gmail</i> (1 – tak, 0 – nie)
X_{11}	Posiadanie <i>innej</i> skrzynki pocztowej (mail-box) (1 – tak, 0 – nie)
X_{12}	Serwery, na których znajduje się najczęściej używana skrzynka pocztowa. Zmienne binarne $X_{12,1}, X_{12,2}, X_{12,3}, X_{12,4}, X_{12,5}, X_{12,6}, X_{12,7}$ odpowiadające kategoriom: <i>uczelniany</i> , <i>wp</i> , <i>onet</i> , <i>interia</i> , <i>o2 (tlen)</i> , <i>gmail</i> , inny
X_{13}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>uczelnianej</i> poczty elektronicznej. Zmienne binarne $X_{13,1}, X_{13,2}, X_{13,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{14}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>wp</i> posiadanej poczty elektronicznej. Zmienne binarne $X_{14,1}, X_{14,2}, X_{14,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{15}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>onet</i> posiadanej poczty elektronicznej. Zmienne binarne $X_{15,1}, X_{15,2}, X_{15,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{16}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>interia</i> posiadanej poczty elektronicznej. Zmienne binarne $X_{16,1}, X_{16,2}, X_{16,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{17}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>o2</i> posiadanej poczty elektronicznej. Zmienne binarne $X_{17,1}, X_{17,2}, X_{17,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{18}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w serwisie <i>gmail</i> posiadanej poczty elektronicznej. Zmienne binarne $X_{18,1}, X_{18,2}, X_{18,3}$ odpowiadające kategoriom: nigdy, rzadko , często
X_{19}	Ocena częstotliwości pojawiania się problemów dotyczących wykorzystania niektórych usług w <i>innym</i> serwisie posiadanej poczty elektronicznej. Zmienne binarne $X_{19,1}, X_{19,2}, X_{19,3}$ odpowiadające kategoriom: nigdy, rzadko, często
X_{20}	Ocena sposobu prezentacji danych (interfejs) w serwisie <i>uczelnianej</i> poczty elektronicznej. Zmienne binarne $X_{20,1}, X_{20,2}, X_{20,3}$ odpowiadające kategoriom: nieprzyjemny , zadowolający, przyjemny

1	2
X_{21}	Ocena sposobu prezentacji danych (interfejs) w serwisie posiadanej poczty elektronicznej na <i>wp</i> . Zmienne binarne $X_{21,1}, X_{21,2}, X_{21,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{22}	Ocena sposobu prezentacji danych (interfejs) w serwisie posiadanej poczty elektronicznej na <i>onet</i> . Zmienne binarne $X_{22,1}, X_{22,2}, X_{22,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{23}	Ocena sposobu prezentacji danych (interfejs) w serwisie posiadanej poczty elektronicznej <i>interia</i> . Zmienne binarne $X_{23,1}, X_{23,2}, X_{23,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{24}	Ocena sposobu prezentacji danych (interfejs) w serwisie posiadanej poczty elektronicznej na <i>o2</i> . Zmienne binarne $X_{24,1}, X_{24,2}, X_{24,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{25}	Ocena sposobu prezentacji danych (interfejs) w serwisie posiadanej poczty elektronicznej na <i>gmail</i> . Zmienne binarne $X_{25,1}, X_{25,2}, X_{25,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{26}	Ocena sposobu prezentacji danych (interfejs) w serwisie innej posiadanej poczty elektronicznej. Zmienne binarne $X_{26,1}, X_{26,2}, X_{26,3}$ odpowiadające kategoriom: nieprzyjazy, zadowolający, przyjazny
X_{27}	Korzystanie z poczty elektronicznej za pomocą urządzenia przenośnego (1 – tak, 0 – nie)
X_{28}	Płeć (1 – kobieta, 0 – mężczyzna)
X_{29}	Wiek. Zmienne binarne $X_{29,1}, X_{29,2}, X_{29,3}$ odpowiadające kategoriom: 18-20 lat, 21-23, 24 lub więcej lat
X_{30}	Kierunek studiów. Zmienne binarne $X_{30,1}, X_{30,2}, X_{30,3}, X_{30,4}$ odpowiadające kategoriom: Informatyka (ZUT w Szczecinie), Lekarski (PUM) , Zarządzanie i inżynieria produkcji (ZUT w Szczecinie), Ekonomia (US)
X_{31}	Rodzaj studiów. Zmienne binarne $X_{31,1}, X_{31,2}, X_{31,3}, X_{31,4}$ odpowiadające kategoriom: stacjonarne I stopnia, stacjonarne II stopnia , niestacjonarne I stopnia, niestacjonarne II stopnia.

Źródło: opracowanie własne na podstawie kwestionariusza ankiety [Górski 2011].

4. Wyniki estymacji uporządkowanego modelu probitowego

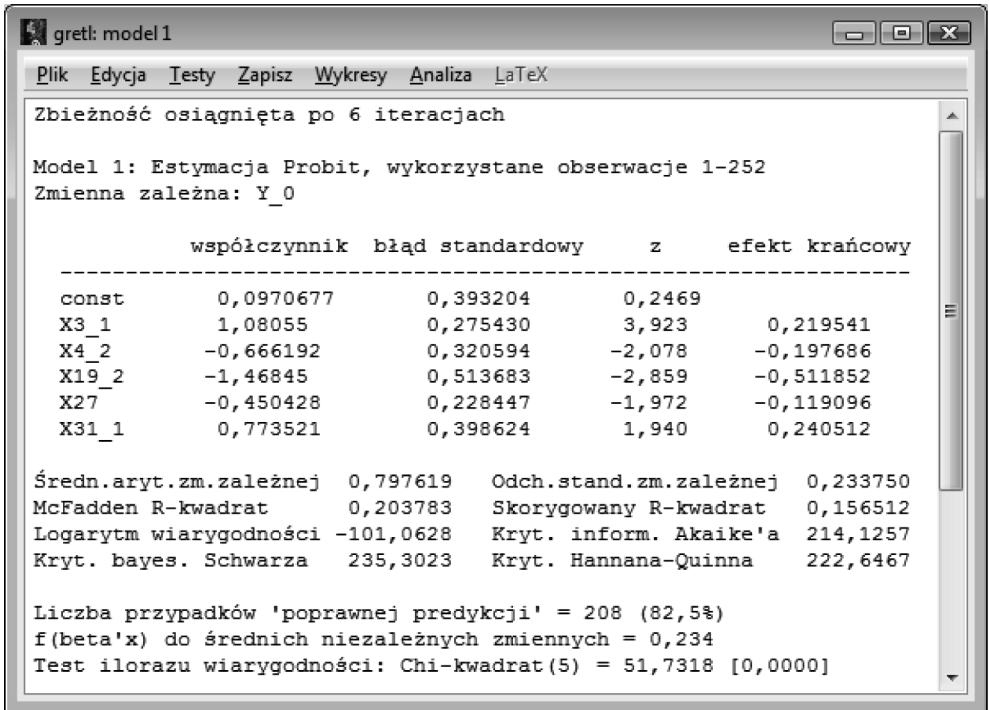
Liczności występowania kolejnych wartości zmiennej zależnej oraz wskaźniki struktury przedstawiono w tab. 2. Widać, że najczęściej (w niemal 80% przypadków) badani wysyłają przeciętnie nie więcej niż 2 wiadomości e-mail w ciągu dnia. Prawie 17% respondentów wysłało dziennie średnio od 3 do 5 wiadomości. Niecałe 3% ankietowanych – od 7 do 9, a powyżej 10 – mniej niż 1% badanych.

Współczynniki uporządkowanego modelu probitowego zostały wyznaczone w niekomercyjnym programie gretl 1.9.5cvs [<http://www.kufel.torun.pl/>] metodą regresji krokowej wstecz (por. rys. 1). Statystyczną łączną istotność (na poziomie 0,05) oszacowanych parametrów modelu sprawdzono testem ilorazu wiarygodności.

Tabela 2. Rozkład wartości zmiennej zależnej

Wartości	$Y=0$	$Y=1$	$Y=2$	$Y=3$
Liczność	201	42	7	2
Odsetek	0,798	0,167	0,028	0,008

Źródło: opracowanie własne na podstawie danych [Górski 2011].



```

gretl: model 1
Plik  Edycja  Testy  Zapisz  Wykresy  Analiza  LaTeX
-----
Zbieżność osiągnięta po 6 iteracjach

Model 1: Estymacja Probit, wykorzystane obserwacje 1-252
Zmienna zależna: Y_0

-----
                współczynnik   błąd standardowy       z       efekt krańcowy
-----
const           0,0970677           0,393204           0,2469
X3_1            1,08055                0,275430           3,923           0,219541
X4_2           -0,666192              0,320594           -2,078          -0,197686
X19_2          -1,46845              0,513683           -2,859          -0,511852
X27            -0,450428             0,228447           -1,972          -0,119096
X31_1          0,773521              0,398624           1,940           0,240512

Średn. aryt. zm. zależnej  0,797619   Odch. stand. zm. zależnej  0,233750
McFadden R-kwadrat      0,203783   Skorygowany R-kwadrat      0,156512
Logarytm wiarygodności -101,0628   Kryt. inform. Akaike'a     214,1257
Kryt. bayes. Schwarz    235,3023   Kryt. Hannana-Quinna      222,6467

Liczba przypadków 'poprawnej predykcji' = 208 (82,5%)
f(beta'x) do średnich niezależnych zmiennych = 0,234
Test ilorazu wiarygodności: Chi-kwadrat(5) = 51,7318 [0,0000]
  
```

Rys. 1. Okno wynikowe dla modelu probitowego

Źródło: obliczenia własne w programie gretl.

Z otrzymanych rezultatów wynika, że zmiany w przeciętnej liczbie wysyłanych dziennie wiadomości e-mail można wyjaśnić częstością sprawdzania poczty elektronicznej, kategorią odbiorcy, częstością występowania problemów w serwisach mniej popularnych, korzystaniem z poczty na urządzeniu przenośnym i rodzajem studiów.

Współczynników modelu probitowego nie można interpretować wprost, dlatego obliczono prawdopodobieństwa warunkowe przyjęcia kolejnych wartości (por. wzory (2)-(3) i tab. 3) jako średnie z teoretycznych prawdopodobieństw warunkowych w próbie oraz wpływy cząstkowe istotnych zmiennych objaśniających na te prawdopodobieństwa, obliczone w punkcie średnich wartości tych zmiennych (por. wzór (4) i tab. 3).

Uzyskany model probitowy jest dobrze dopasowany do danych. Wskazuje na to m.in. brak istotnych różnic między częstościami występowania poszczególnych wartości zmiennej zależnej (por. tab. 2) a średnimi prawdopodobieństw warunkowych (por. tab. 3).

Tabela 3. Prawdopodobieństwa warunkowe $P(Y=j)$ i efekty krańcowe

Średnia liczba wysyłanych wiadomości w ciągu dnia	$Y=0$ (nie więcej niż 2)	$Y=1$ (od 3 do 5)	$Y=2$ (od 6 do 9)	$Y=3$ (10 lub więcej)
$P(Y=j)$	0,795	0,169	0,030	0,006
Wpływ $X_{3,1}$	0,220	-0,191		
Wpływ $X_{4,2}$	-0,198	-0,022	0,210	
Wpływ $X_{19,2}$	-0,512	0,493		
Wpływ X_{27}	-0,119	0,086	0,003	
Wpływ $X_{31,1}$	0,241	-0,058	-0,022	

Źródło: obliczenia własne.

5. Wnioski

Interpretacja wpływów cząstkowych na prawdopodobieństwa warunkowe (por. tab. 3) pozwala sformułować następujące konkluzje:

- przyrostowi prawdopodobieństwa wysyłania do 2 wiadomości dziennie najbardziej sprzyja studiowanie na studiach stacjonarnych I stopnia; rzadkie sprawdzanie poczty sprzyja nieco słabiej;
- na spadek tego prawdopodobieństwa ma wpływ rzadkość pojawiania się problemów z wykorzystaniem usług w serwisach innych niż popularne, pisanie w sprawach zawodowych oraz sprawdzanie poczty na urządzeniach przenośnych;
- rzadkość pojawiania się problemów z wykorzystaniem usług w serwisach innych niż popularne ma wpływ na przyrost prawdopodobieństwa wysyłania od 3 do 5 wiadomości dziennie, natomiast rzadkie sprawdzanie poczty obniża to prawdopodobieństwo;
- pisanie wiadomości związanych z pracą wywołuje przyrost prawdopodobieństwa wysyłania od 6 do 9 wiadomości dziennie.

Interesujące jest również to, że zaproponowany model probitowy nie wykazuje zależności częstości wysyłania wiadomości elektronicznych w ciągu dnia od: okresu posiadania adresu poczty elektronicznej, liczby posiadanych skrzynek pocztowych, rodzaju serwisu pocztowego, problemów z wykorzystaniem usług na portalach popularnych, sposobu prezentacji danych w serwisie, płci i wieku respondenta oraz kierunku studiów.

Literatura

- Dudek H., *An Identification of Farmers' Households in Danger of Poverty on the Ground of Ordered Logit Model*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 1169, UE, Wrocław 2007.
- Górski Ł., *Analiza statystyczna poczty internetowej*, Praca inżynierska, ZUT, Szczecin 2011.
<http://www.kufel.torun.pl/>.
- Kostrzewska J., *Wpływ cech społeczno-demograficznych na wysokość wynagrodzenia zamężnych kobiet*, Zeszyty Naukowe UEK, Seria: Metody Analizy Danych, Kraków 2010.
- Mazurek-Lopacińska K., *Badania marketingowe. Teoria i praktyka*, Wydawnictwo Naukowe PWN, Warszawa 2005.
- Zavoina W., McKelvey R., *A statistical model for the analysis of ordinal level dependent variables*, „Journal of Mathematical Sociology“ 1975.

EVALUATION OF INTENSITY OF MAILBOXES USING WITH THE ORDERED PROBIT MODEL

Summary: This article aims to analyze the opinion of studying users of mailboxes on their mailboxes. The data source survey questionnaires were completed by students of selected departments of Szczecin universities, which included questions about both the characteristics of the mailbox, and the intensity of its use. For the evaluation of e-mail an ordered probit model is used in which there is a qualitative dependent variable, describing structured categories assuming a normal distribution of the random component.

Keywords: ordered probit model, questionnaire survey, stepwise regression.