

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

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

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregów czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Maciej Beręsewicz

Uniwersytet Ekonomiczny w Poznaniu

PRÓBA ZASTOSOWANIA RÓŻNYCH MIAR ODLEGŁOŚCI W UOGÓLNIONYM ESTYMATORZE PETERSENA

Streszczenie: Celem artykułu jest próba rozszerzenia zaproponowanego przez Lavallée'a i Rivesta [2012] estymatora CReG o identyfikację jednostek na podstawie miar odległości oraz zastosowanie do oszacowania wielkości wtórnego rynku nieruchomości w Poznaniu z wykorzystaniem internetowych źródeł danych. W artykule zostanie szczegółowo omówiony uogólniony estymator Petersena (CReG) oraz będą zastosowane 4 miary odległości w celu identyfikacji powtarzających się ofert sprzedaży mieszkań. Przeprowadzone zostanie badanie symulacyjne z wykorzystaniem informacji z portali OtoDom oraz Gratka. Obliczenia wykonano w pakiecie statystycznym R.

Słowa kluczowe: uogólniony estymator Petersena, losowanie pośrednie, internetowe źródła danych, parowanie statystyczne, wtórny rynek nieruchomości.

1. Wstęp

Internetowe źródła informacji zaczynają odgrywać coraz większą rolę w codziennym życiu. Z pomocą specjalistycznych portali internetowych można wyszukiwać informacje o wycieczkach, hotelach, nieruchomościach czy produktach. Portale te coraz częściej odgrywają znaczącą rolę przy wyborze usługi bądź artykułów. W Polsce, według badania GUS – Społeczeństwo informacyjne [GUS 2013], 70,5% gospodarstw domowych w 2012 r. miało dostęp do Internetu (66,6% w 2011 r.). Co ważne, w 2011 r. 44% Polaków wyszukiwało informacje o towarach lub usługach, natomiast 30,3% dokonywało zakupów przez Internet.

Internet jest również ważnym źródłem informacji o wtórnym rynku nieruchomości. Nieliczne badania poświęcone potencjalnym klientom nieruchomości [Strączkowski 2008] wskazują, że Internet jest też głównym źródłem pozyskiwania informacji o rynku nieruchomości. Narodowy Bank Polski wykorzystuje portale internetowe w celu tworzenia raportów rocznych oraz kwartalnych dotyczących wtórnego rynku nieruchomości w zakresie struktury oraz dynamiki cen.

Jednak wykorzystanie internetowych źródeł danych z punktu widzenia teorii statystyki jest dopiero odkrywane. Niektóre urzędy statystyczne, takie jak Statistics Netherlands, próbują wykorzystać portale internetowe do estymacji parametrów w wybranych populacjach (m.in. rynek mieszkaniowy, komunikacja lotnicza) [Dass i Arends-Tóth 2012; Dass i in. 2011]. Wiąże się z tym wiele metodologicznych problemów, począwszy od określenia populacji, jej wielkości, metod próbkowania poprzez jakość danych, a na estymacji i wnioskowaniu skończywszy.

Celem artykułu jest próba estymacji wielkości wtórnego rynku nieruchomości w Poznaniu na 1 września 2013 r. z wykorzystaniem internetowych źródeł danych. Jest to istotny problem, ponieważ nieznaną jest wielkość rynku wtórnego, a oszacowanie może być podstawą tworzenia nowych wskaźników opisujących sytuację w gospodarce.

Na potrzeby badania został zaadaptowany uogólniony estymator Petersena, który wykorzystuje metodę losowania pośredniego wprowadzoną przez Lavallée'a [1995], a szczegółowo omówioną w [Deville i Levellée 2006]. Podjęta zostanie próba rozszerzenia zaproponowanego przez Lavallée'a i Rivesta [2012] estymatora CReG o wykorzystanie różnych funkcji odległości w celu detekcji jednostek podobnych na dwóch portalach internetowych. Wykorzystane zostaną też dane pochodzące z dwóch głównych portali internetowych Gratka oraz OtoDom [por. MegaPanel 2013].

2. Wtórny rynek nieruchomości

Fundamentalnym założeniem, będącym podstawą badania, jest ujęcie wtórnego rynku nieruchomości jako populacji trudnej do zbadania (*hard-to-reach population*). Podyktowane jest to między innymi problemami natury metodologicznej, tj. brakiem dostępnego pełnego operatu losowania, jak również trudnościami pozyskania informacji od oferujących nieruchomości, co ogranicza możliwości stosowania klasycznych metod wnioskowania statystycznego. W przypadku wtórnego rynku nieruchomości w Polsce przemawiają za tym następujące przesłanki:

- a) brak dostępnego spisu wszystkich dostępnych w danym okresie nieruchomości mieszkaniowych oferowanych do sprzedaży,
- b) oferowanie, jak i sprzedaż możliwe są zarówno przez pośredników jak i bezpośrednio,
- c) brak informacji o tym, gdzie oraz kiedy dane mieszkanie jest oferowane do sprzedaży.

Pod względem badawczym można podjąć próbę podziału według kryterium podmiotowego – rynku pierwotnego i wtórnego. W przypadku rynku pierwotnego istnieje spis wszystkich planowanych bądź nowo oddanych nieruchomości, natomiast w przypadku rynku wtórnego sytuacja nie jest tak klarowna. Organizacja wtórnego rynku w Polsce pozwala na sprzedaż mieszkań przez pośredników, jak również pozarynkowo (bezpośrednio). W efekcie nie możemy określić operatu

losowania ani wskazać rejestru administracyjnego, który zawiera wszystkie oferowane w danym czasie nieruchomości. Problem ten zaznaczony został m.in. w: [Widłak 2010; Łaszek i Widłak 2008].

Możliwości badania wtórnego rynku nieruchomości pojawiają się w Internecie. Coraz częściej, aby móc sprzedać nieruchomość, należy umieścić informacje na specjalistycznych portalach internetowych (dalej oznaczone jako SPI), które pośredniczą między oferentami a klientami. Otwiera to ogromne możliwości przed badaczami, jednak niesie za sobą również wiele niebezpieczeństw (m.in. jakość danych).

W Polsce istnieje wiele SPI świadczących usługi pośrednictwa w sprzedaży nieruchomości. Różnią się one zarówno pod względem zawartości (poświęcone rynkowi pierwotnemu lub wtórnemu), jak również jakością informacji na nich zawartych (por. otodom.pl, nieruchomista.pl). Kluczowe z punktu widzenia dalszej analizy jest określenie zagrożeń, które wiążą się z wykorzystaniem internetowych źródeł danych:

- **Reprezentatywność** – dotyczy określenia, jaka część ogłoszeń pojawia się na portalach internetowych, ilu z działających w danym rejonie pośredników umieszcza swoje ogłoszenia na portalach internetowych.
- **Selektywność** – niektóre portale internetowe mogą być częściej wybierane lub odwiedzane.
- **Jakość danych** – dotyczy w głównej mierze weryfikacji prawdziwości ogłoszenia, identyfikacja ogłoszeń powtarzających się, poprawność umieszczonych informacji (błędy pomiaru, błędy losowe, nielosowe w tym systematyczne).
- **Wielość źródeł informacji i ich zróżnicowanie** – istnienie wielu różnych, pokrywających się źródeł informacji, trudność w określeniu wielkości populacji.

Firmy tworzące portale internetowe mają inny cel niż instytucje państwowe. Konkurują między sobą liczbą ogłoszeń oraz skutecznością dotarcia do klientów. Często sytuacją na polskich portalach internetowych jest umieszczanie tego samego ogłoszenia przez jednego bądź wielu pośredników.

W związku z tym należy wziąć pod uwagę, że relacja między ogłoszeniodawcą i ofertą może być „wiele do wielu”. W związku z tym można zastosować znane w literaturze podejście określane jako losowanie pośrednie (*Indirect Sampling*) oraz jego rozszerzoną postać – uogólniony estymator Petersena (*Generalized Capture Recapture estimator*). Wychodząc z założenia, że połączeń oferta–pośrednik jest wiele do wielu, podjęto próbę zastosowania uogólnionego estymatora Petersena, w którym wykorzystano różne miary odległości do identyfikacji podobnych ogłoszeń.

3. Losowanie pośrednie

Losowanie Pośrednie zostało zaproponowane przez Lavallée’a [1995] w celu oszacowania wag dla osób w panelowym badaniu gospodarstw domowych. Lavallée zaproponował uogólnioną metodę podziału wag (*Generalized Share Weight Met-*

hod), która szacuje wejściowe wagi dla gospodarstw domowych. Podejście to stosowane jest w przypadku, gdy nie dysponujemy operatem losowania dla populacji, którą chcemy zbadać, lecz jedynie pomocniczym operatem dla innej populacji, która jest połączona z interesującą nas populacją. Głównym założeniem jest możliwość zbadania wszystkich jednostek z populacji z wykorzystaniem pomocniczego operatu losowania.

Niech BB oznacza populację, którą chcemy zbadać, a AA populację dla której dysponujemy operatem losowania, która jest połączona z populacją BB . Niech $N^A N^A$ oznacza liczebność populacji AA , $s^A s^A$ próbę pobraną z populacji AA , $m^A m^A$ liczebność próby $s^A s^A$, $\pi_j^A \pi_j^A$ prawdopodobieństwo inkluzji pierwszego rzędu dla jednostki j z $U^A U^A$, $m^B m^B$ – liczebność jednostek z populacji BB (klastrow, zespołów), kk – jednostki z populacji BB , $s^B s^B$ – próbę z populacji BB . Dla każdej jednostki k z populacji BB i próby $s^A s^A$. Mamy:

$$w_k = \frac{1}{L_k^B} \sum_{j=1}^{N^A} l_{j,k} \frac{t_j}{\pi_j^A},$$

gdzie $t_j = 1$ jeżeli $j \in s^A$ i 0 w przeciwnym wypadku, $l_{j,k} = 1$, $l_{j,k} = 0$, gdy występuje połączenie między j oraz k , w przeciwnym wypadku 0 , $L_k^B = \sum_{j=1}^{N^A} l_{j,k}$, $L_k^B = \sum_{j=1}^{N^A} l_{j,k}$. Wielkość populacji $N^B N^B$ określona jest sumą wag $N^B = \sum_{k=1}^{m^B} w_k N^B = \sum_{k=1}^{m^B} w_k$. Więcej o losowaniu pośrednim można znaleźć w książce [Lavallée 2007].

4. Uogólniony estymator Petersena

Uogólniona metoda podziału wag sprawdza się w przypadku, gdy pomocniczy operat losowania umożliwia zbadanie wszystkich jednostek z interesującej nas populacji. Innymi słowy, istnieje przynajmniej jedno połączenie każdej jednostki z populacji B z populacją A . W przypadku, gdy pomocniczy operat losowania nie pozwala na zbadanie wszystkich jednostek z populacji B , należy zastosować inne podejście. W związku z tym Lavallée i Rivest [2013] zaproponowali uogólnienie tej metody, wykorzystując rozwiązanie znane z badań populacyjnych, które pozwala na wykorzystanie dwóch źródeł informacji do oszacowania wielkości badanej populacji w przypadku, gdy powiązania między jednostkami nie są jeden do jednego.

Uogólniony estymator Petersena (*Generalized Capture Recapture estimator*) wykorzystuje wagi oszacowane uogólnioną metodą podziału wag. Pierwotnie estymator Petersena (*Capture Recapture estimator*) nieznannej liczebności populacji ma następującą postać

$$\hat{N}_{PET} = \frac{n_1 n_2}{n_{12}},$$

gdzie n_1 – liczebność jednostek złapanych za pierwszym razem, n_2 – liczebność jednostek złapanych za drugim razem, n_{12} – liczebność złapanych za pierwszym i drugim razem.

Uogólniony estymator Petersena ma więc następującą postać:

$$\hat{N}_{CReG}^B = \frac{\hat{N}_{N1}^B \hat{N}_{N2}^B}{\hat{N}_{A1,A2}^B},$$

gdzie $\hat{N}_{N1}^B = \sum_{k=1}^{N_{N1}^B} w_k^{A1}$, $\hat{N}_{N2}^B = \sum_{k=1}^{N_{N2}^B} w_k^{A2}$, $\hat{N}_{A1,A2}^B = \sum_{k=1}^{N_{A1,A2}^B} w_k^{A1} w_k^{A2}$, gdzie $w_k^{A1}, w_k^{A2}, w_k^{A1,A2}$ określone są zgodnie z wzorem wykorzystującym uogólnioną metodę dzielenia wag.

Estymator ten charakteryzuje się takimi samymi własnościami jak klasyczny estymator Petersena – jest nieobciążony, ale charakteryzuje się dużą wariancją. Zastosowanie tego estymatora pozwala na uwzględnienie dwóch źródeł informacji, z których wiemy, że zawierają informacje o tych samych jednostkach.

5. Przygotowanie danych

Na potrzeby badania napisany został pająk internetowy (w języku R, pakiety *httr*, *RCurl*, *XML*), który 1 września 2013 r. pobrał wszystkie ogłoszenia znajdujące się na portalu OtoDom i Gratka, zostały one określone jako aktualne z ostatnich 7 dni¹. Są to ogłoszenia, które zostały zarówno zaktualizowane (ogłoszenie mogło być nawet z poprzedniego roku), jak i dodane w ostatnim tygodniu od daty pobrania ogłoszeń.

Algorytm pobrania danych (pająk internetowy) miał następujące kroki²:

1. Zadać zapytanie do bazy danych o nieruchomości w Poznaniu.
2. Wejść na 1 stronę wyników wyszukiwania.
3. Znajdź ostatnią stronę wyszukiwania i przypisz wynik do zmiennej n .
4. Dla każdej strony wyszukiwania ($i = 1, \dots, n$):
 - 4.1. Pobierz informacje z danej strony z wynikami wyszukiwania (informacje ogólne).
 - 4.2. Wejść na każdą stronę z ogłoszeniami, które znajdują się na stronie wyszukiwania ($j = 1, \dots, 25$ dla OtoDom lub $j = 1, \dots, 20$ dla Gratka).

¹ Zostało to stwierdzone na podstawie wyszukiwania zaawansowanego, które było dostępne na obu portalach.

² Jest to uproszczony algorytm (pseudokod), który został zastosowany do obydwu portali internetowych.

- i. Dla każdej strony pobierz wszystkie informacje na niej zawarte.
- ii. Jeżeli przejdziesz do ostatniej strony z ogłoszeniami, przejdź do kolejnej strony wyszukiwania (*next i*).

Kolejnym etapem było ujednoczenie nazw zmiennych, które nie zostały predefiniowane w kodzie algorytmu. W kolejnym kroku należało ujednoczyć wartości zmiennych, które pojawiały się na obydwu stronach internetowych, ale miały różne opisy (np. jednopokojowe, 1 pokój). Następnie zaimputowano braki danych w istniejących zmiennych, wykorzystując opisy, które towarzyszyły każdemu ogłoszeniu. Zdarzały się sytuacje, że ogłoszenie nie mało podanych podstawowych danych w formie tabelarycznej³, a znajdowały się one w formie tekstowej (opisie) dołączonej do ogłoszenia.

Po procesie ujednoczenia zmiennych, wyłączono z analizy ogłoszenia, które nie spełniały założeń badania, np. oferty sprzedaży mieszkań na rynku pierwotnym, w domach jednorodzinnych, sprzedaży odstępnego lub wynajmu.

Istotne z punktu badania były dwie zmienne – OFERENT oraz położenie mieszkania. Zmienna OFFERENT określała podmiot, który umieścił ogłoszenie na portalu. W przypadku portalu OtoDom początkowa liczba podmiotów oferujących ogłoszenia wynosiła 124, na Gratce 166, część wspólna 49. Po oczyszczeniu nazw podmiotów na OtoDom było ich 118, na Gratce 161, a część wspólna wynosiła 111. Podmioty występujące na obydwu portalach różniły się nawet liczbą umieszczonych ogłoszeń. W tabeli 1 przedstawiony jest rozkład wartości bezwzględnej z różnicy ogłoszeń tych samych oferentów między dwoma badanymi portalami. Można zauważyć, że niektóre podmioty (pośrednicy) preferowały portal OtoDom, a inne portal Gratka. Dodatnia różnica oznacza, że pomimo obecności tych samych oferentów na obydwu badanych portalach, występują różnice w liczbie publikowanych ofert. Jest to istotne z punktu widzenia analizy, ponieważ wpływa na selektywność oraz reprezentatywność danych zawartych na badanych portalach.

Tabela 1. Rozkład bezwzględnej różnicy ogłoszeń oferowanych przez tych samych oferentów na dwóch portalach

Minimum	Kwartył 1	Mediana	Średnia	Kwartył 3	Max
0	0	0	5,518	3	183

Źródło: opracowanie własne na podstawie portali OtoDom i Gratka.

Wykorzystanie zmiennej położenie mieszkania wymagało ujednoczenia nazw ulic na podstawie informacji zarówno z ogłoszenia, jak i z opisów. Następnie każde ogłoszenie zgeokodowano, używając funkcji *geocode* z pakietu *ggmap* w programie **R**. Kolejnym etapem przetwarzania danych było usunięcie obserwacji niereal-

³ Dotyczy portali zawierających oddzielny tabelaryczny opis nieruchomości, w których zapisane były informacje o powierzchni, liczbie pokoi itd. Dodatkowo prawie każde ogłoszenie zawierało opis tekstowy, który jest najczęściej bardzo rozbudowany i zawiera więcej szczegółów.

nych (np. mieszkanie za 1 zł bądź za 100 mln zł). Wykorzystano w tym celu test zgodności ceny mieszkania z rozkładem lognormalnym zawartym w funkcji *get-Outliers* (pakiet *extremevalues*)⁴.

Ostatecznie do badania zostały wybrane 2532 ogłoszenia z Gratki (pierwotnie 2780) oraz 2187 z OtoDom (pierwotnie 2425). Pobieranie, przetwarzanie i analiza danych została wykonana z wykorzystaniem pakietu **R**.

6. Badanie symulacyjne

Wykorzystano 4 miary odległości – *Euklidesową*, *Maksimum*, *Manhattan* oraz *Canberra*. Wybrano również 6 zmiennych służących do opisu danego ogłoszenia – *cena*, *powierzchnia*, *piętro*, *liczba pokoi*, *liczba poziomów* oraz *współrzędne geograficzne* mieszkania. Przeprowadzono następujące badanie symulacyjne.

Dla każdej miary odległości:

1. Losowano 1000 razy pośredników, wykorzystując losowanie wprost proporcjonalne do liczby ogłoszeń, które dany pośrednik miał w ofercie.

2. Następnie w celu identyfikacji ofert podobnych przeprowadzono analizę skupień metodą Warda, wykorzystując daną miarę odległości.

Identyfikacja ofert podobnych opierała się na wskazaniu liczby skupień poprzez odcięcie na wartości 0,15. Ustalenie punktu odcięcia wyznaczono na podstawie merytorycznej oceny dendrogramu oraz składu skupień.

7. Uzyskane wyniki

W tabeli 2 znajduje się zestawienie liczebności otrzymanych skupień oraz średniej liczby ich wystąpień. Wynika z niej, że najczęściej pojawiały się skupienia o liczebności od 1 do 4. Pojawiały się również ogłoszenia, które występowały więcej niż 10 razy. Oznacza to, że niektóre oferowane nieruchomości mieszkaniowe były umieszczane kilka razy.

Tabela 2. Liczebność skupienia oraz średnia liczba wystąpień skupienia o danej liczebności

Liczebność skupienia	1	2	3	4	5	6	7	8	9	10	12	13	11	14
Średnia liczba wystąpień	922	713	106	121	31	30	15	14	3	4	1	1	2	1

Źródło: opracowanie własne na podstawie portali OtoDom i Gratka.

W tabeli 3 znajduje się rozkład liczby pośredników przypadających na jedno ogłoszenie. Okazuje się, że przeciętnie 1,27 przypadają na jedno ogłoszenie, a najwięcej mieszkań było wystawionych przez jednego oferenta. Prawie 14% ogłoszeń miało dwóch, a 3% – 3 oferentów. Pojawiały się również ogłoszenia wystawione przez 8 podmiotów.

⁴ Opis procedury można znaleźć w winiecie pakietu pod adresem <http://www.markvanderloo.eu/>.

Tabela 3. Liczba oferentów przypadających na jedno ogłoszenie oraz średnia liczba ich wystąpień

Liczba oferentów	1	2	3	4	5	6	7	8
Średnia liczba wystąpień	1592	272	60	26	8	2	1	1

Źródło: opracowanie własne na podstawie portali OtoDom i Gratka.

Tabela 4. Wyniki badania symulacyjnego

Miara	Średnia	Mediana	CV	Min	Max	Q5	Q95
Bez wykorzystania miary odległości	4829,24	4886,05	217,08	1174,74	8101,39	2937,24	6439,03
Euklidesowa	2654,31	2664,71	103,89	974,31	4312,01	1829,12	3513,95
Maximum	2605,17	2618,29	106,36	652,67	4022,80	1701,46	3482,38
Manhattan	2714,63	2707,06	114,09	810,08	5132,76	1750,05	3644,45
Canberra	2728,47	2764,95	110,27	866,79	4539,76	1737,36	3635,23

Źródło: opracowanie własne na podstawie portali OtoDom i Gratka.

W tabeli 4 przedstawiono rozkład wielkości oszacowanej populacji nieruchomości mieszkaniowych w Poznaniu oferowanych do sprzedaży 1 września 2013. Możemy zauważyć, że uwzględnienie niewielkich różnic w prezentowanych ofertach zdecydowanie wpływa na oszacowanie wielkości badanego rynku. Nie uwzględniając podobieństw między zamieszczonymi ogłoszeniami, średnia wielkość rynku nieruchomości oraz współczynnik zmienności oszacowań jest blisko dwa razy wyższy.

8. Wnioski

Wykorzystanie internetowych źródeł informacji, losowania pośredniego oraz analizy skupień pozwoliło na nieobciążone oszacowanie wielkości populacji przy uwzględnieniu pojawiania się ofert, które są wystawiane przez wielu pośredników. Uwzględnienie miar odległości pozwoliło zidentyfikować ogłoszenia dotyczące tych samych nieruchomości i w wyniku poprawić oszacowania Uogólnionego Estymatora Petersena.

Możemy również zauważyć, że wszystkie metody charakteryzują się dużą wariancją, powodującą duże zróżnicowanie w oszacowaniach wielkości populacji. Zastosowanie miar odległości zmniejszyło współczynnik zmienności uogólnionego estymatora Petersena o połowę, co wskazuje na bardziej wiarygodne wyniki.

Literatura

- Daas P., Roos M., de Blois C., Hoekstra R., ten Bosch O., Ma Y. (2011), *New data sources for statistics: experiences at Statistics Netherlands*, Statistics Netherlands, Working Papers.
- Dass P., Arends-Tóth J. (2012), *Secondary data collection*, Statistics Netherlands, Working Papers.
- Deville J.-C., Lavallée P. (2006), *Indirect Sampling: The Foundations of the Generalized Weight Share Method*, 2006, Survey Methodology, 32, 2, 165-176.
- GUS (2013), *Spółczesność informacyjna w Polsce. Wyniki badań statystycznych z lat 2009-2013*, Główny Urząd Statystyczny, Urząd Statystyczny w Szczecinie, Warszawa.
- Łaszek J., Widłak M. (2008), *Badanie cen na rynku mieszkań prywatnych zamieszkałych przez właściciela z perspektywy banku centralnego*, NBP, Bank i Kredyt, 8.
- Lavallée P. (1995), *Crosssectional weighting of longitudinal surveys of individuals and households using the weight share method*, Survey Methodology, 21, 1, 25-32.
- Lavallée P. (2007), *Indirect Sampling*, Series in Statistics, Springer.
- Lavallée P., Rivest L.-P. (2012), *Capture-Recapture Sampling and Indirect Sampling*, Journal of Official Statistics, 28, 1, 1-27.
- MegaPanel (2013), PBI/Gemius, wrzesień.
- Strączkowski Ł. (2008), *Tendencje i determinanty rozwoju lokalnego rynku nieruchomości mieszkaniowych (na przykładzie Miasta Poznania)*, Katedra Inwestycji i Nieruchomości, UEP, rozprawa doktorska.
- Widłak M. (2010), *Dostosowanie indeksów cenowych do zmian jakości. Metoda wyznaczania hedonicznych indeksów cen i możliwości ich zastosowania dla rynku mieszkaniowego*, NBP, Materiały i Studia, 247.
- Widłak M., Tomczyk E. (2010), *Konstrukcja i własności hedonicznego indeksu cen mieszkań dla Warszawy*, NBP, Bank i Kredyt, 41, 1.

AN ATTEMPT TO USE DIFFERENT DISTANCE MEASURES IN THE GENERALIZED PETERSEN ESTIMATOR

Summary: The aim of the article is an attempt to extend *CR_EG* estimator proposed by [Lavallée and Rivest, 2012] by incorporating different distance measures to identify units that refer to the same statistical unit. The estimator will be used to assess the size of secondary real estate market in Poznań. In the article *Generalized Capture Recapture* estimator will be discussed in detail and four different measures will be applied. Simulation study using web portals OtoDom and Gratka will be conducted. All calculations are made using R statistical package.

Keywords: Generalized Petersen estimator, indirect sampling, internet data sources, statistical matching, secondary real estate market.