PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 427

Taksonomia 27
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego.

W trakcie dwóch sesji plenarnych oraz 13 sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów.

Teksty 24 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii Taksonomia nr 27. Teksty 25 recenzowanych artykułów naukowych znajdują się w Taksonomii nr 26.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 427 • 2016

Taksonomia 27 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Maciej Beręsewicz, Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu
e-mails: {maciej.beresewicz; m.szymkowiak}@ue.poznan.pl

ANALIZA SKUPIEŃ WYBRANYCH LOKALNYCH RYNKÓW NIERUCHOMOŚCI W POLSCE Z WYKORZYSTANIEM INTERNETOWYCH ŹRÓDEŁ DANYCH

CLUSTER ANALYSIS OF SELECTED LOCAL REAL ESTATE MARKETS IN POLAND BASED ON INTERNET DATA SOURCES

DOI: 10.15611/pn.2016.427.03

Streszczenie: Głównym celem artykułu jest grupowanie i ocena stopnia podobieństwa wybranych lokalnych rynków nieruchomości (miast) w Polsce ze względu na średnią cenę ofertową mieszkania za m^2 dla danych pochodzących z dwóch źródeł: z portali internetowych oraz publikowanych przez Narodowy Bank Polski. Uwzględnienie dwóch różnych źródeł danych umożliwiło wyodrębnienie grup miast podobnych ze względu na średnią cenę ofertową mieszkania za m^2. Pozwoliło również ocenić, na ile uzyskane wyniki grupowania z wykorzystaniem danych pochodzących z Internetu są zbieżne z wynikami grupowania dla danych publikowanych przez statystykę publiczną. Ze względu na charakter posiadanych danych zastosowano wybrane miary odległości służące do badania podobieństwa kilku szeregów czasowych oraz metodę pełnego wiązania, dzięki której możliwe było znalezienie grup miast w Polsce podobnych ze względu na kształtowanie się średniej ceny ofertowej mieszkań za m^2. W obliczeniach skorzystano z pakietu TSclust programu R, w którym zaimplementowane zostały najważniejsze miary odległości umożliwiające osiągnięcie tak postawionego celu.

Słowa kluczowe: analiza skupień, internetowe źródła danych, pakiet TSclust, rynek nieruchomości.

Summary: The main purpose of the article is to group and assess the degree of similarity between selected local real estate markets (cities) in Poland in terms of the average offer price per 1 square metre of apartment space. To achieve this goal the author of the study used data from real estate portals and those published by the National Bank of Poland and the Central Statistical Office. By taking into account two different types of data sources the author managed to identify a group of similar cities with respect to the average offer price per 1 m^2 of apartment space. It was also possible to assess to what extent the grouping results based on Internet data are consistent with data published by official statistics. Given the character of available data, the author chose functions of distance used to measure

similarity between time series and the method of complete linkage, which proved useful in identifying clusters of cities of interest that were similar in terms of the average offer price per 1 m^2 of apartment space. The functions used in calculations were implemented in the TSclust Package in the R programme.

Keywords: cluster analysis of time series data, Internet data sources, TSclust package, real estate market.

1. Wstęp

Analiza skupień szeregów czasowych jest obszarem badawczym, który w ostatnich latach jest intensywnie rozwijany w literaturze przedmiotu [Montero, Vilar 2015]. Ma ona również coraz więcej różnorodnych praktycznych zastosowań obejmujących m.in. zagadnienia finansowe czy medyczne. Jest także wykorzystywana w problemie rozpoznawania mowy. Jednym z kluczowych elementów w analizie skupień jest obliczenie odległości pomiędzy obiektami, które podlegają następnie procesowi przypisywania do klas z wykorzystaniem odpowiedniej metody grupowania. Na potrzeby wyznaczania macierzy odległości można przy tym wykorzystać jedną z wielu miar.

W analizie skupień szeregów czasowych zaproponowano wiele miar służących do wyznaczenia odległości między dwoma szeregami czasowymi. Przykłady stanowić mogą: odległość Minkowskiego, Frècheta, DTW (*Dynamic Time Warping*), odległość wykorzystująca autokorelacje czy nieparametryczne estymatory jądrowe [Montero, Vilar 2015]. Wszystkie omówione miary odległości między szeregami czasowymi opisane zostały w literaturze przedmiotu [Fréchet 1906; Berndt, Cliford 1994; Galeano, Peña 2000]. Zostały one ponadto oprogramowane w pakiecie TSclust [Montero, Vilar 2014].

Wspomniane miary odległości można również wykorzystać w ocenie stopnia podobieństwa szeregów czasowych zawierających dane pochodzące z Internetu. Dotyczy to również danych z portali nieruchomości, które są jednym z najważniejszych źródeł informacji o cenach ofertowych mieszkań. W związku z dostępnością nowych źródeł danych zaistniała potrzeba statystycznej oceny znajdujących się tam informacji [Beręsewicz 2015]. Z jednej strony umożliwiają one zmniejszenie kosztów wybranych badań (automatyczne pobieranie danych z Internetu). Z drugiej zaś należy mieć na uwadze obciążenie wyników, które jest konsekwencją selektywności oraz braku reprezentatywności danych zgromadzonych w Internecie – na przykład na portalach poświęconych nieruchomościom. Z tego punktu widzenia szczególnego znaczenia nabiera ocena reprezentatywności danych pochodzących z Internetu oraz możliwość porównania oszacowań z tego typu źródeł z tym, co jest dostępne w obszarze statystyki publicznej.

Głównym celem artykułu jest grupowanie i ocena stopnia podobieństwa wybranych lokalnych rynków nieruchomości (miast) w Polsce ze względu na średnią

cenę ofertową mieszkania za m^2 dla danych z dwóch źródeł: portali internetowych oraz publikowanych przez Narodowy Bank Polski. Uwzględnienie dwóch różnych źródeł danych umożliwiło wyodrębnienie grup miast podobnych ze względu na średnią cenę ofertową mieszkania za m^2. Pozwoliło również ocenić, na ile uzyskane wyniki grupowania z wykorzystaniem danych pochodzących z Internetu są zbieżne z wynikami grupowania dla danych publikowanych przez statystykę publiczną.

Na potrzeby egzemplifikacji omawianych metod wykorzystany został program R i pakiet TSclust [Montero, Vilar 2015]. Analiza uwzględnia rzeczywiste dane odnoszące się do średniej ceny ofertowej mieszkań za m^2 na wtórnym rynku nieruchomości zgromadzone w postaci odpowiednich szeregów czasowych dla wybranych miast w Polsce. Dane pochodziły z jednej strony z publikacji NBP i GUS, a z drugiej z Internetu, tj. z portali internetowych dedykowanych wtórnemu rynkowi nieruchomości (OtoDom.pl, Dom.Gratka.pl, Szybko.pl, Nieruchomości-online.pl oraz Morizon.pl).

2. Źródła danych o rynku nieruchomości

W Polsce głównymi źródłami danych o rynku nieruchomości wykorzystywanymi przez Główny Urząd Statystyczny są rejestry administracyjne oraz badania częściowe. Kluczowe informacje dotyczące transakcji nieruchomościami pozyskiwane są z Rejestru Cen i Wartości Nieruchomości (RCiWN) będącego w gestii starostw powiatowych oraz urzędów na prawach powiatu. Dane te są podstawą rocznego opracowania pt. *Obrót Nieruchomościami* przygotowywanego przez GUS [GUS 2015a]. Raport ten zawiera zagregowane na poziomie województw oraz powiatów dane dotyczące m.in. przeciętnych cen i powierzchni oraz liczby transakcji dla lokali, budynków, nieruchomości zabudowanych i gruntowych. Niestety, poziom agregacji przestrzennej i czasowej nie uwzględnia podziału na rynek pierwotny oraz wtórny.

Drugim oficjalnym źródłem informacji o wtórnym rynku nieruchomości jest badanie pt. *Badanie cen nieruchomości mieszkaniowych i komercyjnych* przeprowadzane przez Narodowy Bank Polski przy współpracy GUS, które jest wyszczególnione w *Programie badań statystyki publicznej na rok 2015* [GUS 2015b]. Celem badania jest m.in. analiza cen ofertowych i transakcyjnych lokali wraz z atrybutami cenotwórczymi na rynku pierwotnym i wtórnym. Badanie to ma zakres przestrzenny ograniczony jedynie do miast wojewódzkich oraz ich aglomeracji [NBP 2015]. Głównym źródłem pozyskiwanych informacji są ankietowani pośrednicy oraz inne podmioty zajmujące się obrotem nieruchomościami, a także RCiWN. Wynikiem badania jest kwartalna informacja m.in. o przeciętnych cenach ofertowych i transakcyjnych na rynku pierwotnym i wtórnym.

Z kolei internetowe źródła danych są w niewielkim stopniu wykorzystywane na potrzeby opisu pierwotnego i wtórnego rynku nieruchomości przez statystykę

publiczną. Taki stan rzeczy wynika głównie z braku informacji o jakości oraz reprezentatywności informacji zawartych na serwisach ogłoszeniowych [Beręsewicz 2015]. Jest również konsekwencją tego, że na portalach internetowych można pozyskać informacje jedynie o cenach ofertowych a nie transakcyjnych. Jednakże w literaturze możemy znaleźć przykłady, w których takie serwisy są wykorzystywane przez urzędy statystyczne. Na przykład Urząd Statystyczny w Holandii wykorzystuje dane zawarte na portalu Funda.nl do opisu rynku nieruchomości oraz łączenia ich z rejestrem transakcji [Hoekstra, ten Bosch, Harteveld 2012]. Dodatkowo serwisy ogłoszeniowe publikują oficjalne raporty oraz umożliwiają analizę rynku nieruchomości na niedostępnym dla statystyki publicznej poziomie[1]. Ponadto umożliwiają uzyskanie informacji np. o średnich cenach ofertowych nieruchomości w dowolnych przekrojach i w dowolnym momencie[2]. Elektroniczny charakter portali pozwala także na automatyczne pozyskiwanie danych bez dodatkowych obciążeń dla respondentów.

W artykule autorzy, wykorzystując dane NBP i pochodzące z portali internetowych, podjęli próbę oceny stopnia podobieństwa wybranych lokalnych rynków nieruchomości (miast) w Polsce ze względu na średnią cenę ofertową mieszkania za m^2. W tym celu zastosowali analizę skupień szeregów czasowych dla różnych miar odległości.

3. Przegląd miar odległości w analizie skupień szeregów czasowych

W zagadnieniu związanym z analizą skupień szeregów czasowych zaproponowano wiele miar służących do wyznaczenia odległości między dwoma szeregami czasowymi. Poniżej zostaną omówione te miary, które wykorzystano w badaniu empirycznym. Pełen przegląd wielu innych miar można znaleźć w pracy [Montero, Vilar 2014; Fréchet 1906; Berndt, Cliford 1994; Galeano, Peña 2000]. Zakładać przy tym będziemy, że dysponujemy dwoma szeregami czasowymi $X_T = (X_1, X_2, \ldots, X_T)^T$ i $Y_T = (Y_1, Y_2, \ldots, Y_T)^T$, dla których będziemy chcieli wyznaczyć odległość mierzącą stopień ich niepodobieństwa.

Dynamic Time Wraping **(DTW)** – odległość ta została dogłębnie przeanalizowana przez D. Sankoffa i J.B. Kruskala w ich publikacji z 1983 r. [Sankoff, Kruskal 1983]. Jako miara odległości służąca do badania stopnia niepodobieństwa dwóch szeregów czasowych została zaproponowana w pracy [Berndt, Cliford 1994]. Miara odległości DTW umożliwia znalezienie najmniejszej odległości między dwoma szeregami czasowymi przy dopuszczeniu nieliniowej transformacji czasu dla obu szeregów. Wyraża się ona wzorem:

[1] M.in. http://ceny.szybko.pl, http://www.morizon.pl/statystyki/.

[2] W tym przypadku konieczna jest ocena obciążenia uzyskanych wyników. Można tego dokonać, wykorzystując modele mieszane [Beręsewicz, Szymkowiak 2015].

$$d_{\mathrm{DTW}}(X_T, Y_T) = \min_{r \in M} \left(\sum_{i=1,\ldots,m} |X_{a_i} - Y_{b_i}| \right), \tag{1}$$

gdzie M jest zbiorem wszystkich możliwych sekwencji m par zachowujących kolejność obserwacji postaci $r = \left(\left(X_{a_1}, Y_{b_1}\right), \ldots, \left(X_{a_m}, Y_{b_m}\right) \right)$, $a_i, b_j \in \{1, \ldots, T\}$ oraz $a_1 = b_1 = 1$, $a_m = b_m = T$, $a_{i+1} = a_i$ lub $a_i + 1$, $b_{i+1} = b_i$ lub $b_i + 1$ dla $i \in \{1, \ldots, m-1\}$. DTW jako miara odległości jest bardziej odpowiednia niż odległość euklidesowa, zwłaszcza w sytuacjach gdy porównujemy szeregi o podobnej strukturze, ale przesunięte w czasie. Odległość ta może być wykorzystywana również w przypadku braku części danych lub pewnych ich niedokładności. Najważniejsza jest bowiem tutaj kolejność występowania poszczególnych faz szeregu czasowego. Do wad tej miary odległości należy zaliczyć jest złożoność, zwłaszcza dla dużych zbiorów danych. Sposób wyznaczania odległości dla dwóch szeregów czasowych z wykorzystaniem miary DTW został opisany w pracy [Wołyński, Górecki 2013].

Odległość wykorzystująca współczynnik korelacji (CORT) – to miara zaproponowana w pracy [Douzal, Nagabhushan 2007]. Podobieństwo pomiędzy dwoma szeregami czasowymi jest szacowane z wykorzystaniem współczynnika korelacji, który wyraża się wzorem:

$$\mathrm{CORT}(X_T, Y_T) = \frac{\sum_{t=1}^{T-1} (X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1} (X_{t+1} - X_t)^2} \sqrt{\sum_{t=1}^{T-1} (Y_{t+1} - Y_t)^2}}. \tag{2}$$

Wartości tej miary bliskie –1 lub 1 oznaczają bliskie podobieństwo między dwoma szeregami czasowymi. Wartość równa 0 oznacza stochastyczną niezależność pomiędzy rozważanymi szeregami czasowymi. Odległość wykorzystująca współczynnik korelacji wyraża się następującym wzorem:

$$d_{\mathrm{CORT}}(X_T, Y_T) = \phi[\mathrm{CORT}(X_T, Y_T)] \cdot d(X_T, Y_T), \tag{3}$$

gdzie $d(X_T, Y_T)$ może oznaczać przykładowo miarę odległości DTW, a funkcja ϕ wyraża się wzorem:

$$\phi_k(u) = \frac{2}{1 + \exp(ku)}, \quad k \geq 0, \tag{4}$$

gdzie u to argument funkcji (w tym przypadku współczynnik korelacji), a $k \geq 0$ to parameter kontrolujący wagę przypisaną do odległości $d(X_T, Y_T)$.

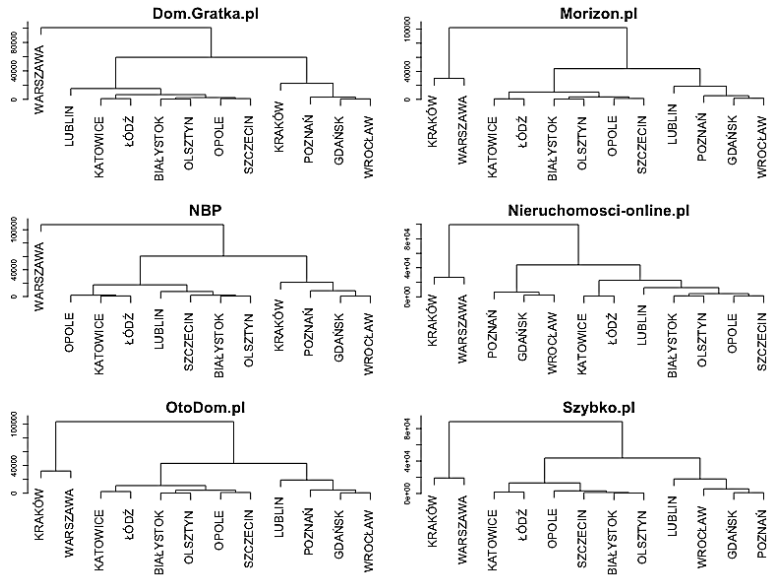
4. Analiza porównawcza wybranych lokalnych rynków nieruchomości dla danych z portali internetowych i NBP – opis procedury badawczej

Z pięciu największych portali internetowych dedykowanych wtórnemu rynkowi nieruchomości (OtoDom.pl, Dom.Gratka.pl, Szybko.pl, Nieruchomości-online.pl oraz Morizon.pl) zebrano dane na temat cen ofertowych mieszkań za m^2 będących w sprzedaży dla 12 miast w Polsce (Białystok, Gdańsk, Katowice, Kraków, Lublin, Łódź, Olsztyn, Opole, Poznań, Szczecin, Warszawa i Wrocław)[3]. Dane obejmowały lata 2012–2014 i zostały pobrane z Internetu z wykorzystaniem specjalnie przygotowanego zgodnie z techniką Web-scrapingu oprogramowania. Dla każdego miasta i portalu obliczono w ujęciu kwartalnym średnią cenę ofertową mieszkań za m^2. W ten sposób utworzono odpowiednie szeregi czasowe dla 12 miast i 5 portali.

Z badania prowadzonego przez NBP wzięto informacje na temat średnich cen ofertowych mieszkań za m^2 w rozważanych miastach w ujęciu kwartalnym za lata 2012–2014. Następnie, uwzględniając odpowiednie miary odległości (DTW oraz wykorzystującą współczynnik korelacji), wyznaczono macierze odległości pomiędzy szeregami czasowymi w ramach każdego portalu i dla danych pochodzących z NBP dla rozpatrywanych miast. W ostatnim kroku procedury badawczej, z wykorzystaniem wyznaczonych macierzy odległości i metody pełnego wiązania (najdalszego sąsiada), stworzono odpowiednie dendrogramy oraz dokonano analizy stopnia podobieństwa kształtowania się cen mieszkań za m^2 oferowanych do sprzedaży w wybranych miastach. Dokonano ponadto oceny zgodności uzyskanych skupień dla danych pochodzących z portali internetowych z danymi pochodzącymi z badania przeprowadzonego przez NBP. W tym celu wykorzystano współczynnik zgodności wspólnych węzłów pomiędzy dwoma drzewami [Galili 2016]. Ponadto wykorzystując współczynnik korelacji kofenetycznej dokonano oceny dopasowania wyznaczonych dendrogramów do macierzy odległości [Sokal, Rohlf 1962].

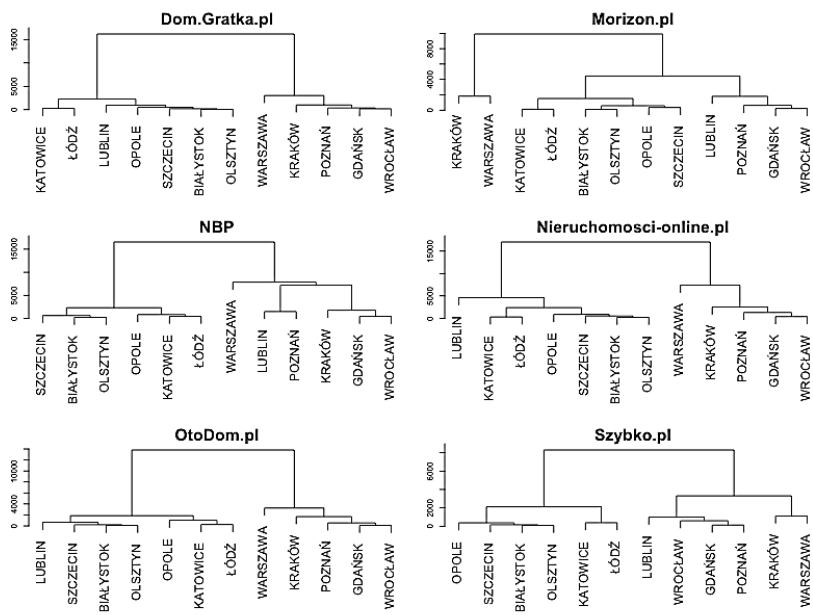
Na rysunkach 1 i 2 przedstawiono wyniki grupowania 12 analizowanych miast ze względu na średnią cenę ofertową m^2 z wykorzystaniem odpowiedniej miary odległości i metody pełnego wiązania dla danych pochodzących z pięciu wspomnianych wcześniej portali internetowych oraz publikowanych przez NBP [2015].

[3] W raporcie NBP [2015] ze względu na lokalny charakter rynków mieszkaniowych, przedmiotem pogłębionej analizy jest 16 miast – stolic województw oraz Gdynia. Zastosowany przez NBP podział wynika z porównywalnej wielkości oraz stopnia rozwoju rynków nieruchomości. W artykule autorzy musieli ograniczyć się do 12 wskazanych w treści miast, gdyż nie było możliwe pobranie danych jednostkowych ze wszystkich portali internetowych z wykorzystaniem przygotowanego zgodnie z techniką Web-scrapingu oprogramowania (kwestie ograniczeń nałożonych przez administratora portalu). W takiej sytuacji zwrócono się z prośbą o udostępnienie danych do gestora danego portalu. W ten sposób uzyskano dane dla 12 miast za okres 2012–2014. Analiza dla wszystkich portali została zatem przeprowadzona tylko dla tych miast i za ten okres, dla których posiadano odpowiednie dane.



Rys. 1. Grupowanie mieszkań ze względu na średnią cenę ofertową m^2 – miara odległości DTW

Źródło: opracowanie własne.



Rys. 2. Grupowanie mieszkań ze względu na średnią cenę ofertową m^2 – miara odległości wykorzystująca współczynnik korelacji

Źródło: opracowanie własne.

Ponadto w tab. 1 zawarto oszacowania współczynników zgodności wspólnych węzłów i korelacji kofenetycznej. Dzięki tym współczynnikom możliwe było dokonanie oceny, które z otrzymanych dendrogramów powstałych z wykorzystaniem danych pochodzących z badania NBP były najbliższe grupowaniu miast dla danych zebranych z portali internetowych oraz który z dendrogramów najlepiej dopasował się do macierzy odległości.

W przypadku miary odległości DTW najbliższe dendrogramowi NBP wyniki uzyskane zostały dla portalu Dom.Gratka.pl. Natomiast w przypadku miary odległości opartej na współczynniku korelacji w czasie (CORT) najbliższe wyniki uzyskano dla portalu Szybko.pl. Świadczą o tym najwyższe wartości współczynników zgodności wspólnych węzłów i korelacji kofenetycznej (zostały one pogrubione w tab. 1).

Tabela 1. Oszacowane współczynniki zgodności węzłów pomiędzy dwoma drzewami[4] i korelacji kofenetycznej

Źródło danych	DTW		CORT	
	współczynnik zgodności węzłów wspólnych	współczynnik korelacji kofenetycznej	współczynnik zgodności węzłów wspólnych	współczynnik korelacji kofenetycznej
Dom.Gratka.pl	**0,8695**	**0,9718**	0,7391	0,6925
Morizon.pl	0,7391	0,7304	0,7391	0,6112
Nieruchomosci	0,7826	0,8313	0,7826	0,7868
OtoDom.pl	0,7391	0,7305	0,7826	0,6972
Szybko.pl	0,7391	0,7303	**0,7826**	**0,9417**

Źródło: opracowanie własne.

Dokonując analizy dendrogramu uzyskanego dla miary odległości DTW i danych publikowanych przez NBP oraz pochodzących z portalu Dom.Gratka.pl, można zauważyć, że Warszawa stanowi skupienie jednoelementowe. Może to potwierdzać, że ceny ofertowe mieszkań za m^2 w stolicy znacznie odbiegają od tych w pozostałych miastach. Pozostałe dwa wyraźnie widoczne skupienia miast w obydwu rozpatrywanych przypadkach były identyczne.

Z kolei analiza dendrogramu uzyskanego dla miary odległości CORT i danych NBP oraz portalu Szybko.pl ukazuje pewną różnicę. Polega ona na tym, że Warszawa stanowi odrębne skupienie dla danych NBP, podczas gdy dla danych z portalu Szybko.pl znajduje się ona w jednym skupieniu z Krakowem. Pozostałe dwa skupienia zawierają w obydwu rozpatrywanych przypadkach te same miasta.

[4] Dendrogram uzyskany dla danych publikowanych przez NBP i danych pochodzących z jednego z pięciu rozpatrywanych portali internetowych.

5. Wnioski

Rynek nieruchomości należy do jednego z trudniejszych sektorów gospodarki w zakresie sporządzania różnego rodzaju prognoz i analiz. Wynika to w dużej mierze z jego niskiej transparentności. Dostęp do rzetelnych materiałów, a także wiarygodnych danych jest ograniczony. Szczególną rolę w tym zakresie odgrywają raporty publikowane przez Narodowy Bank Polski i Główny Urząd Statystyczny. Nie zaspokajają one jednak popytu na wszystkie informacje, na które zgłaszane jest zapotrzebowanie przez różnych odbiorców. Może to dotyczyć przykładowo braku informacji o cenach ofertowych mieszkań czy strukturze rynku na różnym poziomie agregacji przestrzennej (mniejsze miasta czy w przekroju powiatów). Remedium na brakujące informacje w takim przypadku stanowić mogą dane pochodzące z Internetu, tj. z portali dedykowanych nieruchomościom. Możemy dzięki nim zwiększyć pokrycie informacyjne w zakresie lokalnych rynków nieruchomości, zwłaszcza w tych obszarach, dla których brakuje danych dostarczanych przez NBP i GUS. Należy jednak podkreślić, że do wyników uzyskanych z wykorzystaniem danych pochodzących z portali internetowych trzeba podchodzić z dużą ostrożnością, co się wiąże przede wszystkim z brakiem reprezentatywności tego typu źródeł informacji. Z tego punktu widzenia niezwykle ważne są prace badawcze pozwalające ocenić, na ile dane z Internetu dają zbieżne wyniki z tymi, których dostarcza statystyka publiczna. Niniejszy artykuł wpisuje się po części w ten nurt, a uzyskane rezultaty potwierdzają, że dane internetowe na temat kształtowania się średnich cen ofertowych mieszkań mogą stanowić alternatywne źródło informacji w stosunku do tych publikowanych przez NBP i GUS. Można również stwierdzić, że dane pochodzące z portali internetowych o wtórnym rynku nieruchomości w Polsce stanowią cenne źródło informacji na temat kształtowania się cen ofertowych mieszkań. Należy jednak mieć na uwadze, że ze względu na to, iż nie są one reprezentatywne, zazwyczaj oszacowania średniej ceny ofertowej mieszkań będą obciążone w stosunku do średniej ceny ofertowej mieszkań publikowanej przez NBP i GUS z wykorzystaniem danych pochodzących z badania reprezentacyjnego pośredników nieruchomości. Ocena wielkości obciążenia jest możliwa z wykorzystaniem metod, jakie oferuje statystyka małych obszarów w postaci modeli mieszanych [Beręsewicz, Szymkowiak 2015]. Dane z portali internetowych umożliwiają jednak uchwycenie głównego obrazu lokalnych rynków nieruchomości ze względu na kształtowanie się ceny ofertowej mieszkań za m^2. W tym zakresie niezwykle przydatna okazuje się analiza skupień szeregów czasowych.

Literatura

Beręsewicz M., 2015, *On the representativeness of Internet data sources for the real estate market in Poland*, Austrian Journal of Statistics, vol. 44, no. 2, s. 45–57.

Beręsewicz M., Szymkowiak M., 2015, *Robust model based approach to assess accuracy of Internet data sources*, First Latin American ISI Satellite Meeting on Small Area Estimation, Santiago, Chile, 3–5.08.2015, http://www.encuestas.uc.cl/sae2015/program_sae.html.

Berndt D.J., Cliford J., 1994, *Using Dynamic Time Warping to Find Patterns in Time Series*, KDD Workshop, https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf, s. 359–370.

Douzal C.A., Nagabhushan P.N., 2007, *Adaptive dissimilarity index for measuring time series proximity*, Advances in Data Analysis and Classification, vol. 1, no. 1, s. 5–21.

Fréchet M.M., 1906, *Sur quelques points du calcul fonctionnel*, Rendiconti del Circolo Matematico di Palermo (1884–1940), vol. 22, no. 1, s. 1–72.

Galeano P., Peña D., 2000, *Multivariate analysis in vector time series*, Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo, vol. 4, no. 4, s. 383–403.

Galili T., 2016, *Package 'dendextend', Extending R's Dendrogram Functionality*, https://cran.r-project.org/web/packages/dendextend/dendextend.pdf (14.09.2015).

GUS, 2015a, *Obrót nieruchomościami w 2014*, Główny Urząd Statystyczny, Departament Handlu i Usług, Warszawa, http://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5492/4/12/1/obrot_nieruchomosciami_w_2014.pdf (14.09.2015).

GUS, 2015b, *Program badań statystyki publicznej na rok 2015*, Główny Urząd Statystyczny, Warszawa, s. 171–172, http://bip.stat.gov.pl/download/gfx/bip/pl/defaultstronaopisowa/526/1/1/pbssp_2015.doc (14.09.2015).

Hoekstra R., ten Bosch O., Harteveld F., 2012, *Automated data collection from web sources for official statistics: First experiences*, Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, vol. 28, no. 3, s. 99–111.

Montero P., Vilar J.A., 2014, *TSclust: An R package for time series clustering*, Journal of Statistical Software, vol. 62, no. 1, s. 1–43.

Montero P., Vilar J.A., 2015, *Package 'TSclust'*, dokumentacja pakietu TSclust, http://cran.r-project.org/web/packages/TSclust/TSclust.pdf (14.09.2015).

NBP, 2015, *Raport o sytuacji na rynku nieruchomości mieszkaniowych i komercyjnych w Polsce w 2014 r.*, Departament Stabilności Finansowej, Warszawa.

Sankoff D., Kruskal J.B., 1983, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison Wesley, Reading, MA.

Sokal R.R., Rohlf F.J., 1962, *The comparison of dendrograms by objective methods*, Taxon, vol. 11, no. 2, s. 33–40.

Wołyński W., Górecki T., 2013, *Analiza skupień*, Wydział Matematyki i Informatyki UAM, Poznań.