Internet data sources for real estate market statistics

Streszczenie rozprawy doktorskiej

mgr Maciej Beręsewicz

Katedra Statystyki

Uniwersytet Ekonomiczny w Poznaniu

Poznań 2016

Informacja w systemie statystyki publicznej pozyskiwana jest drogą badań statystycznych, zarówno próbkowych, jak i pełnych, których etapy przeprowadzania są ustalone i dobrze opisane w literaturze (por. Särndal et al., 2003; Bethlehem, 2009; Lohr, 2009; Groves et al., 2011). Przykładem klasycznego źródła danych statystycznych są spisy powszechne, które są kosztowne i przeprowadzane zwykle co 10 lat, a pozyskane z nich informacje publikowane są często z dużym opóźnieniem w stosunku do badania. Natomiast badania próbkowe mają na celu dostarczenie oszacowań wybranych charakterystyk badanej zbiorowości na podstawie informacji pozyskanej jedynie od części jednostek z niej wylosowanych. Zakres badań częściowych jest szerszy niż spisów powszechnych i dotyczy wybranych tematów m.in. badania aktywności ekonomicznej ludności, jakości życia czy wydatków gospodarstw domowych.

Jednakże, w związku z ograniczeniami budżetowymi badań oraz rosnącą frakcją odmów udziału w badaniach częściowych statystycy coraz częściej zwracają uwagę na potrzebę wykorzystania dostępnych źródeł danych pośród, których wymienia się m.in. rejestry administracyjne, Internetowe źródła danych (IZD) czy szerzej „big data". W literaturze takie podejście określa się zmianą paradygmantu statystyki publicznej – odejścia od klasycznych źródeł danych statystycznych do wykorzystania oraz łączenia istniejących źródeł.

Rejestry administracyjne, powstały w wyniku regulacji prawnych bez konsultacji statystyków (Wallgren and Wallgren, 2014). Powszechne wykorzystanie rejestrów od lat 80 XX w. w krajach skandynawskich (m.in. Finlandia, Norwegia, Dania) oraz Holandii wskazało na ogromne możliwości tych źródeł dla potrzeb statystyki. Przy okazji ostatniego spisu powszechnego (NSP 2011) oprócz badania częściowego (próba 20% populacji) wykorzystano również szereg rejestrów administracyjnych. Należy zaznaczyć, że rejestry administracyjne w znacznym stopniu różnią się od spisów czy badań częściowych głównie dlatego, że zostały stworzone do innych celów

niż statystyczne. Dodatkowo, rejestry zawierają zdefiniowany na potrzeby administracji publicznej zakres informacji. Na przykład, Rejestr Cen i Wartości Nieruchomości zawiera informacje o transakcjach, natomiast nie zawiera informacji o mieszkaniach oferowanych do sprzedaży. Dlatego statystycy skierowali swoje zainteresowania również na Internet oraz „big data", które w połączeniu z istniejącymi źródłami danych mogą dostarczyć pełniejszej informacji.

Internetowe źródła danych oraz „big data" opisane są literaturze poświęconej informatyce, technologiom informacyjnym, e-commerce czy socjologii (Abramowicz et al., 2002; Miller, 2011; Lazer et al., 2014). Nieliczne opracowania statystyczne wskazują na wybrane problemy metodologiczne związane z wykorzystaniem IZD. Jednakże nie rozpatrują ich kompleksowo jako źródeł danych statystycznych, jako szczególnego źródła informacji. Informacja statystyczna powinna charakteryzować się między innymi: przydatnością (istotnością informacji), dokładnością (estymacji), terminowością i punktualnością, dostępnością i przejrzystością, porównywalnością (w czasie i przestrzeni) oraz spójnością. Szczególnie ważnym kryterium źródeł danych statystycznych jest ich reprezentatywność umożliwiająca estymację charakterystyk populacji generalnej. W związku z tym, aby IZD (m.in. portale internetowe) podobnie, jak rejestry administracyjne mogły zostać wykorzystane na potrzeby statystyki, istotna jest ich ocena oraz wskazanie szans i zagrożeń związanych z ich wykorzystaniem, ze szczególnym naciskiem na odniesienie do teorii estymacji.

Należy również wskazać problemy związane z określeniem definicji Internetowych źródeł danych czy big data w kontekście istniejących źródeł informacji statystycznej. Dlatego na potrzeby pracy przyjęto następującą definicję: Internetowe źródła danych są nieprobabilistyczną próbą, która została utworzona przy wykorzystaniu Internetu oraz jest utrzymywana przez prywatne podmioty (ang. *Internet data source is a self-selected (non-probabilistic) sample that is created through the Internet and maintained by entities external to NSIs and administrative regulations.*) Dodatkowo autor na potrzeby pracy wprowadził ograniczenie definicji Internetu jedynie do sieci WWW.

Praca ma charakter metodologiczno-empiryczny. Proponowane podejścia zostały przedstawione na przykładzie wtórnego rynku nieruchomości, który jest w niewielkim stopniu opisany przez oficjalne badania. Dodatkowo w badaniu założono, że wtórny rynek nieruchomości jest populacją trudną do zbadania, która charakteryzuje się brakiem dostępnego operatu losowania oraz trudnością w określeniu jednostek badania.

Głównym celem rozprawy jest ocena Internetowych źródeł danych na potrzeby opisu rynku nieruchomości. Aby osiągnąć cel zasadniczy, sformułowano następujące cele szczegółowe:

G1 Wskazanie źródeł błędów nielosowych w Internetowych źródeł danych o rynku wtórnym

nieruchomości.

G2 Ocenę reprezentatywności internetowych źródeł danych o rynku wtórnym nieruchomości.

G3 Oszacowanie obciążenia internetowych źródeł danych o wtórnym rynku nieruchomości.

Aby osiągnąć powyższe cele, sformułowano następujące hipotezy badawcze. Główna hipoteza badawcza weryfikowana w pracy brzmi: Internetowe źródła danych umożliwiają estymację charakterystyk rynku nieruchomości z akceptowalnym błędem (ang. *Internet data sources enable acceptable estimation of real estate market characteristics.*). Szczegółowe hipotezy badawcze zostały określone następująco

H1 Zaproponowane autorskie podejście pozwala ocenić reprezentatywność Internetowych źródeł danych o wtórnym rynku nieruchomości.

H2 Internetowe źródła danych charakteryzują się systematycznym obciążeniem, które różni się między źródłami i miastami (domenami).

H3 Błąd doboru (ang. *self-selection error*) w Internetowych źródłach danych ma charakter informatywny (tj. zależy od badanej cechy y).

H4 Internetowe źródła danych umożliwiają estymację ceny ofertowej mieszkania na wtórnym rynku nieruchomości z akceptowalnym błędem mierzonym względnym obciążeniem (ang. *absolute relative bias*).

Wybór źródeł danych, domen oraz okresu badania został podyktowany dostępnością danych. Badanie empiryczne ograniczono do 12 miast (Białystok, Gdańsk, Katowice, Kraków, Łódź, Lublin, Olsztyn, Opole, Poznań, Szczecin, Warszawa, Wrocław) oraz okresu I kw. 2012 do IV kw. 2014. Przyjęto, że domeną, dla której przeprowadzono estymację obciążenia średniej ceny ofertowej, jest miasto. Na potrzeby badania wybrane zostały trzy portale internetowe - Dom.Gratka.pl, OtoDom.pl i Nieruchomości-Online.pl. Narodowy Bank Polski oraz Główny Urząd Statystyczny (NBP/GUS) prowadzą badanie cen nieruchomości mieszkaniowych i komercyjnych, które dotyczy m.in. wtórnego rynku nieruchomości. Wyniki badania NBP/GUS zostały przyjęte, jako dane referencyjne do których odnoszone są wyniki otrzymane na podstawie IZD. Do oceny reprezentatywności IZD wykorzystano również badanie Społeczeństwo Informacyjne prowadzone przez GUS. Dodatkowo, w przypadku oceny reprezentatywności IZD dla Poznania uwzględniono dodatkowe źródło w postaci Rejestru Cen i Wartości Nieruchomości.

W rozprawie przedstawiono rozszerzenie estymatora statystyki małych obszarów opartego na liniowym modelu mieszanym zaproponowanym przez Fosen and Zhang (2011); Zhang (2012a).

Proponowany model został zaadaptowany na potrzeby oszacowania oraz dekompozycji obciążenia średniej ofertowej ceny metra kwadratowego oszacowanej na podstawie IZD w 12 miastach w Polsce. Zaproponowany model uwzględnia cztery efekty losowe: efekt skorelowanych źródeł danych, efekt domeny (miasta) oraz efekt interakcji między źródłem danych a domeną (miastem) i efekt autokorelacji obciążenia w czasie.

Pierwszy rozdział poświęcony jest przeglądowi literatury oraz podsumowaniu prac nad nowymi źródłami danych w statystyce publicznej. Przedstawiono rys historyczny wykorzystania statystycznych i niestatystycznych źródeł danych. Zaproponowana została definicja Internetowych źródeł danych. Rozdział kończy podsumowanie wyzwań i zagrożeń związanych z użyciem IZD na potrzeby statystyki publicznej.

Rozdział drugi poświęcony jest IZD w świetle danych statystycznych. Na początku wskazano niejednoznaczności definicji populacji i jednostki w IZD oraz na rynku nieruchomości. Następnie porównano cechy IZD ze spisami, badaniami częściowymi oraz rejestrami administracyjnymi. Ponadto, omówiono wybrane źródła danych o rynku nieruchomości w Polsce. Przedstawiono również autorską propozycję integracji źródeł danych statystycznych uwzględniającą IZD. Rozważania kończy identyfikacja błędów w IZD wraz z przykładami rynku nieruchomości opracowana na podstawie Zhang (2012b).

Trzeci rozdział poświęcony jest teoretycznym podstawom badania reprezentatywności IZD. Dodatkowo, wskazano możliwe przyczyny obciążeń wynikające z nielosowego charakteru IZD Przedstawiono pojęcie reprezentatywności na podstawie Kruskal and Mosteller (1979a,b,c); Bethlehem (2009), które zostało omówione w kontekście IZD. Zaproponowano także autorską procedurę pomiaru reprezentatywności IZD uwzględniającą poziom agregacji dostępnych danych (dane indywidualne, dane zagregowane) oraz ich źródło: statystyczne (spisy, badania częściowe, sprawozdawczość) i niestatystyczne (głównie rejestry administracyjne). Możliwości pomiaru reprezentatywności przedstawiono w zależności od dostępności danych oraz omówiono możliwe do zastosowania miary.

Rozdziały czwarty i piąty zawierają wyniki badań empirycznych. W rozdziale czwartym za-stosowano zaproponowaną w rozdziale trzecim dwu stopniową procedurę badania reprezentatywności. W pierwszym kroku oceniono zasadność wykorzystania Internetu na potrzeby opisu rynku nieruchomości w Polsce. Wykorzystano do tego celu dane z prowadzonego przez GUS badania Społeczeństwo Informacyjne. W drugim kroku oceniono rozkłady powierzchni oraz liczby pokoi w mieszkaniach oferowanych do sprzedaży na wtórnym rynku nieruchomości w 12 miastach w Polsce. Jako dane referencyjne wykorzystano wyniki badania prowadzonego przez NBP i GUS. Porównano również rozkłady cen mieszkań oferowanych do sprzedaży na wtórnym

rynku nieruchomości z cenami transakcyjnymi pochodzącymi z Rejestru Cen i Wartości Nieruchomości.

W rozdziale 4 zauważono różnice między IZD, a badaniem przeprowadzonym przez NBP i GUS oraz Rejestrem Cen i Wartości Nieruchomości. Wyniki analizy wskazują, że istnieją dwie grupy nieruchomości mieszkalnych, które nie są reprezentowane on-line: (1) tanie, o małej powierzchni i (2) drogie, o dużej powierzchni. Zidentyfikowano następujące potencjalne źródła tej sytuacji: (1) nieruchomości sprzedawane przez komornika sądowego po cenie niższej od ceny rynkowej w formie licytacji; (2) sprzedaż w rodzinie bądź wśród osób bliskich; (3) umowa między stronami transakcji bez publikacji w Internecie; (4) brokerzy nie publikują wszystkich ofert na portalach ogłoszeniowych; (5) nieruchomości, które są najcenniejsze są przedstawione określonej, wąskiej grupie klientów. W przypadku porównań z badaniem NBP i GUS, uzyskane wyniki wskazują znaczące różnice w strukturze mieszkań jednopokojowych o małej powierzchni (mniej niż 40 m2) oraz dużych (ponad 80m2) oferowanych do sprzedaży w Poznaniu.

W rozdziale piątym przedstawiono estymator statystyki małych obszarów oparty na liniowym modelu mieszanym zaproponowanym przez Fosen and Zhang (2011); Zhang (2012a), który został następnie rozszerzony na potrzeby oszacowania i dekompozycji obciążenia średniej ceny ofertowej metra kw. Zaproponowany model zakłada, że istnieje "złoty standard" czyli nieobciążone oszacowanie badanej charakterystyki pochodzące najczęściej z badań reprezentacyjnych lub pełnych. W pracy przyjęto jako „złoty standard" badanie prowadzone przez NBP i GUS. Proponowany model uwzględniał cztery efekty losowe i umożliwił detekcję systematycznego obciążenia szacunków średniej ofertowej ceny m2 mieszkania oszacowanej na podstawie IZD.

Uwzględnienie efektu losowego dla domeny, źródła danych i interakcji między domeną a źródłem, znacząco wpływa na poprawę modelu. Ponadto, otrzymane wyniki wskazały bardzo silną autokorelację obciążenia w czasie. Największą różnicę w cenie ofertowej m2 zauważono w źródle charakteryzującym się najmniejszą popularnością (w tym przypadku Nieruchomosci-Online.pl), natomiast najmniejsze różnice obserwowane są dla najpopularniejszego źródła danych. Wyraz wolny modelu był nieistotny, co może wskazywać na brak istotnych statystycznie różnic między IZD a badaniem przeprowadzonym przez NBP i GUS. Względny błąd średniej ofertowej ceny m2 oszacowany na podstawie IZD waha się między 0,2% do ponad 4%.

Dodatkowo, na podstawie oszacowanego modelu wykazano, że obciążenie różni się między badanymi miastami oraz jest skorelowane ze średnią ceną ofertową w danym mieście. Sugeruje to, że mechanizm doboru ma charakter informatywny (ang. *informative self-selection mechanism*). Obciążenie jest wyższe dla miast, w których wartości nieruchomości są wysokie (m.in.

Kraków, Warszawa i Gdańsk), natomiast niższe w tych miastach, w których oferowane są tańsze mieszkania (m.in. Białystok, Katowice i Szczecin).

Rozprawę doktorską kończy podsumowanie wraz z dyskusją nad dalszymi możliwymi kierunkami prac.

Bibliografia (wybrana)

Abramowicz, W., Kalczynski, P. J., and Wecel, K. (2002). *Filtering the Web to feed data warehouses*. Springer Science & Business Media.

Bethlehem, J. (2009). *Applied survey methods: A statistical perspective*. John Wiley & Sons.

Fosen, J. and Zhang, L.-C. (2011). The approach to quality evaluation of the micro-integrated employment statistics. ESSnet Data Integration.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*, volume 561. John Wiley & Sons.

Kruskal, W. and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review*, 47(1):13–24.

Kruskal, W. and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, 47(2):111–123.

Kruskal, W. and Mosteller, F. (1979c). Representative sampling III: The current statistical literature. *International Statistical Review*, 47(3):245–265.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(14).

Lohr, S. (2009). *Sampling: design and analysis*. Cengage Learning.

Miller, G. (2011). Social scientists wade into the tweet stream. *Science*, 333(6051):1814–1815.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Wallgren, A. and Wallgren, B. (2014). *Register-based Statistics*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc., second edition.

Zhang, L.-C. (2012a). On the accuracy of register-based census employment statistics. European Conference on Quality in Official Statistics.

Zhang, L.-C. (2012b). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63.