18 kwietnia 2016 r.

Prof. dr hab. Mirosław Szreder
Kierownik Katedry Statystyki
Uniwersytetu Gdańskiego
ul. Armii Krajowej 101
81-824 Sopot

Recenzja rozprawy doktorskiej mgr. Macieja Beręsewicza pt. *Internet data sources for real estate market statistics* napisanej pod kierunkiem prof. UEP dr hab. Elżbiety Gołaty, z udziałem promotora pomocniczego dr. Marcina Szymkowiaka

Powierzona mi do oceny rozprawa doktorska dotyczy ważnego i bardzo aktualnego zagadnienia, jakim jest ocena możliwości wykorzystania w badaniach statystycznych internetowych zbiorów danych oraz kombinacji tych zbiorów z wybranymi rejestrami administracyjnymi. Podjęty w dysertacji mgr. Macieja Beręsewicza temat otwartości statystyki na nowe zasoby informacji, w tym przede wszystkim na zwiększające się szybko ilości danych dostępnych w Internecie, uznaję za wart badań naukowych oraz szerokiej dyskusji w gronie daleko wykraczającym poza środowisko statystyków. Problematyka dużych zbiorów danych i ogromnych, nieznanych nigdy wcześniej możliwości ich gromadzenia i przetwarzania, a więc tego wszystkiego co powszechnie rozumie się pod pojęciem Big Data, zajmuje uwagę nie tylko statystyków, ale również specjalistów z zakresu: sztucznej inteligencji, informatyki i socjologii. Jednym z kluczowych wyzwań – jak słusznie zauważa Autor rozprawy – jest zdolność użytkowników tych zbiorów danych do oceny ich jakości, w tym reprezentatywności, przesądzającej o możliwości dokonywania na ich podstawie uogólnień na większe populacje.

Celem rozprawy doktorskiej mgr. Macieja Beręsewicza jest „ocena internetowych źródeł danych dla potrzeb analiz statystycznych rynku nieruchomości" (s. 5). Uważam, że jest to poprawnie sformułowany cel badawczy. Sądzę ponadto, że w tytule rozprawy lepiej było dodać słowo „ocena", jako że obecne brzmienie tytułu: *Internet data sources for real estate market statistics* nie sugeruje w żaden sposób rodzaju zagadnień z zakresu internetowych zbiorów danych, którym praca jest poświęcona. Pozytywnie oceniam sformułowane na s. 5 rozprawy trzy cele szczegółowe, podporządkowane głównemu celowi badawczemu. Zastrzeżeń nie zgłaszam do hipotez badawczych, z których naczelną hipotezę, brzmiącą następująco: *Internetowe źródła danych umożliwiają estymację charakterystyk rynku nieruchomości z akceptowaną wielkością błędu*, uważam za ambitną i odważną. Weryfikacja tej hipotezy wymaga bowiem rozwiązania kilku problemów natury ogólniejszej, wykraczającej poza specyfikę rynku nieruchomości. W szczególności chodzi o określenie wielkości i

rodzajów błędów, jakimi mają prawo być obarczone dane internetowe, a także znalezienie mechanizmu oceny reprezentatywności tego typu zbiorów danych. Z problemami tymi Autor rozprawy poradził sobie w stopniu wyższym niż satysfakcjonujący. W rzeczywistości więc, mgr M. Beręsewicz zrealizował nie tylko postawiony na wstępie cel badawczy, ale wniósł pewien oryginalny wkład w rozwój badań nad nowymi zbiorami informacji wykorzystywanymi w badaniach statystycznych, nie tylko na rynku nieruchomości.

Recenzowana rozprawa napisana jest w języku angielskim, składa się z pięciu rozdziałów i liczy łącznie 208 stron (bez załączników). Układ treści poszczególnych rozdziałów jest logicznie poprawny i wewnętrznie spójny. Wysoko oceniam stronę techniczną rozprawy, zwłaszcza redakcję tekstu wraz z konsekwentnym używaniem zbioru symboli, precyzyjnie wyjaśnionych na wstępie pracy, a także profesjonalnie sporządzony zestaw bibliografii z podanymi numerami stron w rozprawie, na których przywołano poszczególne pozycje literatury.

Pierwszy rozdział poświęcony jest charakterystyce nowych, w szczególności internetowych źródeł danych wykorzystywanych w badaniach statystycznych, ich specyfice i wyzwaniom, jakie rodzą we wnioskowaniu statystycznym oraz w zastosowaniach przez instytucje statystyki publicznej. Autor zawarł w tym fragmencie rozprawy krótką dyskusję na temat definicji Big Data, a ponadto zaproponował własną definicję i klasyfikację internetowych źródeł danych (por. Diagram 1.1 na s. 28). Niewątpliwym wkładem Autora do nauki jest w tym fragmencie określenie i przedyskutowanie, w oparciu o aktualną literaturę światową, następujących ważnych aspektów danych pochodzących z rejestrów administracyjnych i zbiorów internetowych: zgodności pojęciowej z wymogami statystyki publicznej, reprezentatywności, obciążeń i błędów, wewnętrznej spójności danych oraz kontroli ich jakości. Pewien niedosyt może natomiast budzić mało szczegółowe omówienie odrębności pomiędzy rejestrami administracyjnymi i internetowymi zbiorami danych, na przykład faktu, iż urzędowe rejestry administracyjne gromadzą najczęściej informacje o całej populacji, podczas gdy internetowe źródła danych – co słusznie ujmuje w definicji Autor (s. 26) – są traktowane jako dane próbkowe. Pochwalić za to wypada Doktoranta za skorzystanie przy omawianiu tej problematyki z jednej z najlepszych pozycji poświęconych rejestrom administracyjnym w statystyce, autorstwa Anders i Britt Wallgren z 2014 r.

Problematyka jakości danych internetowych oraz ich przydatności w analizach statystycznych jest kontynuowana przez Doktoranta w rozdziale drugim. Jest ona tutaj ściślej powiązana, w przeciwieństwie do rozdziału pierwszego, z rynkiem nieruchomości. Treść tego rozdziału świadczy o dobrej znajomości przez Autora wymogów informacyjnych rynku

nieruchomości, a także o umiejętności krytycznej analizy zastanego stanu wiedzy na temat źródeł danych statystycznych o ofertach i transakcjach na rynku nieruchomości. Oryginalną propozycję Autora, dotyczącą integracji danych statystycznych z uwzględnieniem internetowych źródeł danych, oceniam jako interesującą i wartościową. Podobnie wysoko oceniam dyskusję, jaką Autor przeprowadził w podrozdziale 2.5 na temat najważniejszych aspektów oceny jakości danych pochodzących ze źródeł internetowych.

Teoretycznym podstawom badania reprezentatywności danych internetowych poświęcony jest rozdział 3. Autor zwraca na wstępie uwagę na możliwe źródła błędów systematycznych (np. będących wynikiem autoselekcji jednostek do próby, ang. *self-selection mechanism*) oraz omawia sposoby niwelowania tych błędów. Cała dalsza część tego rozdziału zawiera oryginalne rozważania i dyskusje na temat pojęcia i pomiaru reprezentatywności danych ze źródeł internetowych. Jest to ważny, z punktu widzenia celu podjętych badań, fragment rozprawy. Autor zawarł w nim własną propozycję dwustopniowej procedury pomiaru stopnia reprezentatywności danych, uwzględniającą poziom agregacji dostępnych danych oraz rodzaj źródła danych (statystyczne i niestatystyczne). Zaletą tej propozycji jest to, że może być wykorzystana w odniesieniu do innych poza internetowymi zbiorów danych, a ponadto zastosowanie tej procedury nie jest ograniczone jedynie do rynku nieruchomości. Pozytywnie oceniam wartość naukową i aplikacyjną tej propozycji, a także właściwe zrozumienie przez Autora znaczenia referencyjnych zbiorów danych w zaproponowanej procedurze.

Ostanie dwa rozdziały rozprawy mają charakter empiryczny i dotyczą ściśle polskiego rynku nieruchomości. W rozdziale 4 zilustrowana i zweryfikowana została zaproponowana wcześniej przez Autora procedura oceny reprezentatywności danych próbkowych (dostępnych w internecie) w odniesieniu do rynku nieruchomości w Polsce. Omawiając kolejne etapy tej procedury i prezentując odpowiednie zbiory danych internetowych oraz zbiorów danych referencyjnych (Narodowego Banku Polskiego i Głównego Urzędu Statystycznego), Doktorant zwraca uwagę zarówno na aspekty statystyczne tej analizy, jak i wartości poznawcze uzyskanych wyników. Wnioski z tej analizy są cennym uzupełnieniem zaproponowanej metody oceny reprezentatywności zbiorów danych. Niezależnie od tego, szereg prawidłowości ujawnionych w tej analizie stanowi wzbogacenie wiedzy o jakości zasobów informacyjnych charakteryzujących rynek nieruchomości w Polsce. Konsekwencją wykrycia przez Autora obciążeń w oszacowaniach niektórych wielkości charakteryzujących nieruchomości będące przedmiotem transakcji, m.in. średniej ceny ofertowej metra kwadratowego powierzchni, są poszukiwania szczegółowej wiedzy na temat tych obciążeń, zawarte w rozdziale 5. Mgr M. Beręsewicz wykorzystał w tym fragmencie rozprawy koncepcję statystyki małych obszarów, a

ściślej model Fosena i Zhanga z 2011 roku i adaptował go do potrzeb dekompozycji obciążenia i estymacji wielkości tego błędu. Za oryginalne należy uznać wnioski sformułowane przez Autora, dotyczące mechanizmów i współzależności w zakresie generowania obciążenia ocen średniej ceny ofertowej mieszkań na podstawie danych ze źródeł internetowych.

Odnosząc się do osiągnięć badawczych zawartych w recenzowanej rozprawie doktorskiej, pragnę najpierw zwrócić uwagę na dobre przygotowanie Doktoranta do samodzielnego prowadzenia badań z zakresu analiz statystycznych. W rozprawie swojej mgr M. Beręsewicz udowodnił, że potrafi poprawnie formułować oryginalne cele badawcze i odpowiadające im hipotezy, a także proponować adekwatne metody i techniki badań. Dysponując szeroką wiedzą z zakresu badań statystycznych oraz wiedzą o rynku nieruchomości, Doktorant zaprojektował i zrealizował naukowe badanie, opisane w rozprawie, które stanowi wzbogacenie dotychczasowego stanu wiedzy na temat wartości poznawczej internetowych zbiorów danych. W szczególności Autor zwrócił uwagę na użyteczność tego typu danych w badaniach statystyki publicznej.

Do najważniejszych osiągnięć zawartych w rozprawie, stanowiących oryginalne osiągnięcie badawcze Doktoranta, zaliczam:

1. wskazanie i przedyskutowanie konsekwencji zakwalifikowania internetowych baz danych do zbiorów danych próbkowych generowanych przez nieprobabilistyczny mechanizm autoselekcji;
2. sformułowanie własnej definicji reprezentatywności próby i zaproponowanie odpowiadającej jej procedury pomiaru stopnia reprezentatywności danych (internetowych lub innych);
3. rozwinięcie koncepcji statystyki małych domen w kierunku jej zastosowań w szacowaniu obciążeń rezultatów wnioskowania statystycznego na podstawie danych próbkowych (głównie internetowych);
4. zilustrowanie i zweryfikowanie własnych propozycji metodycznych empirycznym przykładem rynku nieruchomości w Polsce.

Z pełnym przekonaniem stwierdzam, że wymienione wyżej osiągnięcia uznać należy za wartościowy wkład Doktoranta do dyscypliny ekonomia.

Rozprawa ta ma także pewne słabości, ale waga ich jest niewielka. Zaliczam do nich:

a) zbyt słabo podkreśloną wewnętrzną niejednorodność tzw. nowych zbiorów danych, w tym licznych odrębności pomiędzy rejestrami administracyjnymi a internetowymi zbiorami danych;

b) zbyt mało uwagi poświęconej możliwości zastosowania mechanizmów ważenia i kalibracji w odniesieniu do danych ze zbiorów internetowych (w procesie dążenia do uzyskania nieobciążonych ocen z wnioskowania);

c) nieskromne podkreślenie przez Doktoranta aż pięciokrotnie na dwóch stronach zakończenia rozprawy (s. 185-186) pionierskiego charakteru badań i jego całościowego charakteru: „*is the first attempt of this kind at a comprehensive assessment*", „*are the first comprehensive attempts*", „*have not been previously discussed in such detail in statistical literature*", „*has not been previously underlined in the statistical literature*", "*This is the first comprehensive approach*".

W rozprawie znalazłem bardzo niewiele błędów językowych i redakcyjnych.

Całość rozprawy doktorskiej mgr. Macieja Beręsewicza stanowi oryginalne i bogate merytorycznie opracowanie na temat możliwości wykorzystania internetowych zbiorów danych dla celów analiz i badań statystycznych na rynku nieruchomości. Doktorant posiada solidną i usystematyzowaną wiedzę z zakresu statystyki, a ponadto dobrze porusza się w problematyce rynku nieruchomości. Rekomenduję zatem dopuszczenie mgr. Macieja Beręsewicza do publicznej obrony rozprawy doktorskiej przygotowanej pod kierunkiem dr hab. Elżbiety Gołaty, prof. nadzw. UEP.

Mirosław Szreder

prof. dr hab. Mirosław Szreder