DOI: 10.18276/sip.2018.54/3-23

Marek Ogryzek*

Uniwersytet Warmińsko-Mazurski w Olsztynie

PARAMETRYCZNA OCENA JAKOŚCI ESTYMACJI MAP OPRACOWYWANYCH METODAMI GEOSTATYSTYCZNYMI

Streszczenie

Geostatystyczne metody opracowania map to jedna z kartograficznych metod prezentacji wyników badań. Określenie jakości estymacji lub porównanie wyników pomiędzy estymacjami jest możliwe na podstawie analizy statystyk opisowych, tak zwanej statystyki jakości estymacji. Dokładne dopasowanie modelu do danych może nie generować najniższych błędów, dlatego przy wyborze optymalnego modelu ważne jest stosowanie walidacji wyników. W praktyce ze względu na dużą ilość statystyk jakości estymacji oraz brak jasnych kryteriów wyboru wybór jednej, najlepszej metody jest problematyczny. W ramach artykułu zostanie przedstawiona autorska metoda oceny jakości estymacji wykorzystująca algorytm optymalizacyjny dla map tworzonych metodami geostatystycznymi na przykładzie średniej ceny transakcyjnej gruntów rolnych niezabudowanych z zasobu Agencji Nieruchomości Rolnych OT Olsztyn. Mapy średnich cen gruntów jako narzędzie do gospodarowania nieruchomościami rolnymi Skarbu Państwa przedstawiają zróżnicowanie średnich cen transakcyjnych gruntów przeznaczonych pod zabudowę oraz gruntów rolnych.

Słowa kluczowe: zasób ANR, geostatystyka, estymacja, walidacja, interpolacja, algorytmy optymalizacyjne, MPOJE (metoda parametrycznej oceny estymacji)

* Adres e-mail: marek.ogryzek@uwm.edu.pl.

Wstęp

Pierwsze prawo geografii ekonomicznej sformułowane w 1970 roku przez Toblera (1970), mówiące o tym, że obiekty położone bliżej siebie są bardziej do siebie podobne niż obiekty od siebie oddalone, nazywane jest również pierwszym prawem analiz przestrzennych. Prekursorem wykorzystywania metod ilościowych do analiz przestrzennych był Schaeffer (1953), a Isard (1956, 1960) zwrócił szczególną uwagę na efekt sąsiedztwa. Badanie i analizowanie układów oraz powiązań przestrzennych doprowadziło do powstania ekonometrii przestrzennej (Florax, Van der Vlist, 2003).

Dane pochodzące z rynku nieruchomości mają charakter danych geoprzestrzennych, stąd naturalna wydaje się konieczność ich analizy metodami uwzględniającymi położenie nieruchomości i tak zwany wpływ sąsiedztwa (Kulczycki, Ligas, 2009). Jest to jeden z typów danych we wprowadzonej przez Cressiego (1993) klasyfikacji danych przestrzennych. Dane mogą być obserwowane/mierzone w sposób ciągły (w dowolnym punkcie) niezależnie od charakteru mierzonego/obserwowanego atrybutu, który może mieć zarówno charakter ciągły, jak i skokowy. Średnie ceny transakcyjne gruntów są takimi danymi, a mapy średnich cen transakcyjnych gruntów przedstawiają zgodnie z Rozporządzeniem (2011), na podstawie danych gromadzonych przez starostów w rejestrze cen i wartości nieruchomości, badań statystycznych oraz analiz i zestawień charakteryzujących rynek nieruchomości w skali całego kraju w powiązaniu z zasadniczym trójstopniowym podziałem terytorialnym państwa, różnicowanie średnich cen transakcyjnych gruntów przeznaczonych pod zabudowę oraz gruntów rolnych. Uśrednienie tych danych dla jednostek statystycznych powoduje bardzo duże błędy statystyczne, a w praktyce brak uzyskania wiarygodnej informacji. Ważna zatem staje się metoda wyznaczania średniej ceny transakcyjnej (Ogryzek, Kurowska, 2008).

W niniejszym artykule została przedstawiona procedura opracowania map metodami geostatystycznymi na przykładzie cen transakcyjnych gruntów rolnych niezabudowanych z zasobu Agencji Nieruchomości Rolnych Oddział Terenowy Olsztyn (ANR OT Olsztyn).

1. Metodyka

Na potrzeby artykułu przeprowadzono analizę możliwości opracowania map metodami geostatystycznymi, wykorzystując ceny transakcyjne sprzedaży nieruchomości

ści niezabudowanych z zasobu ANR OT Olsztyn w latach 2004–2014. W pierwszym etapie prac przygotowano procedurę opracowania map metodami geostatystycznymi. Następnie sporządzono procedurę walidacji otrzymanych wyników, której zastosowanie ma ułatwić wybór najlepszej dla tego typu danych metody. Narzędzie parametrycznej oceny ma być uniwersalne i niezależne od typu danych. Na potrzeby analizy ustalono, że jednostką administracyjną będzie obszar gmin z zasobu ANR OT Olsztyn. Analizę wykonano dla lat 2004–2014 dla każdego roku oddzielnie, co oznacza, że procedurę powtórzono 10 razy dla różnych danych tego samego typu na tym samym obszarze. Mapy i parametry oceny opracowano z wykorzystaniem oprogramowania ArcGIS z uwzględnieniem podziału terytorialnego na gminy. Procedura powinna przebiegać według następujących etapów:

Etap 1 –wstępna analiza danych.

Etap 2 – opracowanie map metodami geostatystycznymi.

Etap 3 – wykonanie walidacji.

Etap 4 – porównanie parametrów oceny estymacji.

Etap 5 – wybór optymalnej metody geostatystycznej.

W etapie 1 należy zebrać odpowiednio dużą ilość danych, żeby w późniejszym etapie wykonać walidację podzbiorem, czyli wyznaczyć zbiór do estymacji oraz testowy zbiór pomiarów. Liczba punktów minimalna to 30, ale najlepiej, gdy liczba punktów pomiarowych waha się między 100 a 150m a dodatkowe dane są reprezentatywne, niezależne, uzyskane z tego samego źródła danych i opracowane tą samą metodą oraz wystarczająco dokładne. Część z danych pomiarowych posłuży do stworzenia podzbioru walidacyjnego. Szacuje się, że liczba danych do walidacji nie powinna być mniejsza niż 5%, a większa niż 15%. W niniejszym opracowaniu będzie to 10% danych.

W tym etapie wykorzystano zestaw narzędzi Explore Data, który zawiera narzędzia do analizy danych przestrzennych, tak zwaną Exploratory Spatial Data Analysis (ESDA). Narzędzia te umożliwiają uzyskanie informacji o danych i wybraniu najbardziej odpowiedniej metody i parametrów modelu interpolacji. Na przykład przy użyciu krigingu rozłącznego metoda wymaga normalnego rozkładu danych. Jeśli dane nie są tak rozproszone, należy uwzględnić transformację danych w ramach modelu interpolacji lub w przypadku niektórych metod nie wykonywać interpolacji tą metodą. Skośność i kurtoza wskazują na rozkład normalny na histogramie, a NormalQQPlot/General QQPlot są narzędziami do sprawdzenia, czy zbiór danych posiada rozkład normalny, i zbadania, czy dwa zestawy danych są do siebie podobne. Na tym etapie wykorzystamy miary symetrii rozkładu(skośność i kurtozę)

w celu uzyskania informacji, czy rozkład danych ma rozkład normalny (rozkład gaussowski).

W etapie 2 w oprogramowaniu ArcGIS wykorzystano instrument Geostatistical Wizard, w którym możliwe jest generowanie modeli powierzchni według podziału na metody statystyczne i deterministyczne oraz z barierami. Szczegółowy opis metod geostatystycznych użytych w badaniach zaprezentowano w literaturze zagranicznej (Hengel, 2011; Krivoruchko, 2011; Sarna, 2009), a także polskiej (Urbański, 2008; Zawadzki, 2011) oraz w pomocy programu ArcGIS. Geostatystyka to zbiór narzędzi statystycznych uwzględniających w analizie danych ich przestrzenną i czasową lokalizację, a opartych na teorii funkcji losowych. Należy pamiętać, że celem geostatystyki nie musi być tylko interpolacja (estymacja) przestrzenna, ale również zrozumienie zmienności przestrzennej lub czasowej zjawiska, symulowanie wartości oraz optymalizacja sieci pomiarowej. W niniejszej analizie opracowane i walidowane będą mapy estymacji.

Według Waters (1989) „interpolacja przestrzenna to procedura szacowania wartości cechy w niepróbowanych punktach na obszarze objętym istniejącymi pomiarami”. W procesie interpolacji na podstawie istniejących danych oblicza się najczęściej wartości w centrach poszczególnych komórek rastra. Związek między wartościami tej samej cechy (atrybutu) w różnym czasie lub w różnych punktach przestrzeni nazywa się autokorelacją i powinien on być zachowany zgodnie z pierwszym prawem analiz przestrzennych. Etap ten należy wykonywać według schematu:

1. Wybranie metody analizy statystycznej.
2. Wybranie typu metody analizy statystycznej.
3. Badanie trendu danych.
4. Wyznaczenie semiwariogramu empirycznego.
5. Wybór sektorów i sąsiedztwa.
6. Wyznaczenie parametrów oceny estymacji.

W etapie 3 należy ocenić jakość estymacji. W tym celu trzeba wyznaczyć statystyki jakości estymacji oraz przeprowadzić walidację wyników estymacji. Jest to szczególnie ważne w przypadku modelowania takich danych, jak ceny nieruchomości, które obarczone są znacznym szumem (zawierają wyraźny błąd) lub też posiadają kilka wartości odstających. W związku z tym w niniejszych badaniach walidację zostaną wykonane dwiema metodami, to znaczy walidacją podzbiorem (*jackknifing*) oraz kroswalidacją (*crossvalidation*).

Etap 4 jest autorską metodą oceny jakości estymacji na podstawie analizy parametrów estymacji. Do podstawowych statystyk oceny jakości estymacji należą:

- pierwiastek średniego błędu kwadratowego (RMSE – *root mean square error*);
- średni błąd metody (ME – *mean error*);
- odchylenie standardowe średniego błędu metody (SD,s).

Do parametrycznej oceny jakości estymacji wykorzystamy podstawowe wyżej opisane statystyki opisowe, to znaczy miary tendencji centralnej (ME i RMSE), miarę rozproszenia (SD.s), uzyskane przy walidacji oraz walidacji podzbiorem z etapu 3.

W etapie 5 na podstawie przeprowadzonej analizy z wykorzystaniem algorytmu optymalizacyjnego (Ogryzek, 2012) warunki (warunek) oceny jakości estymacji można zapisać w postaci formuły matematycznej. Załóżmy, iż:

$C_1, C_2, \dots, C_m$ – ocena jakości estymacji metody geostatystycznej,

 $\Delta_1, \Delta_2, \dots, \Delta_n$ – waga parametru,

$a_1, a_2, \dots, a_n$ – różnica w procentach do najlepszego parametru a ,

$b_1, b_2, \dots, b_n$ – różnica w procentach do najlepszego parametru b ,

$n_1, n_2, \dots, n_n$ – różnica w procentach do najlepszego parametru n ,

to warunek:

$$\begin{cases} \Delta_1 a_1 + \Delta_2 b_1 + ... + \Delta_n n_n = C_1 \\ \Delta_{21} a_2 + \Delta_{22} b_2 + ... + \Delta_{2n} n_n = C_2 \\ \\ \Delta_{n1} a_n + \Delta_{n2} b_n + ... + \Delta_m n_m = C_n \end{cases}$$

Ta metoda geostatystyczna, której parametr C_n jest najbliżej bądź równy 0, jest metodą optymalną, $C_n \leq 0 = \max$.

2. Analiza cen transakcyjnych gruntów niezabudowanych w latach 2004–2014

Badania przeprowadzono według procedury opisanej w metodyce. Na początku opracowano wstępną analizę danych wejściowych. Należy zaznaczyć, iż w opracowaniu pracowano na danych źródłowych z ANR OT Olsztyn i żadna z transakcji (nawet te odstające) nie została odrzucona. Do analizy przygotowane dane z lat 2004–2014 oddzielnie dla każdego roku. Przeprowadzono procedurę dla danych zawierających 100-procentowy udział oraz dla danych zawierających losowo wybrane 90% udziału. Przyjęto, że 10% danych będzie stanowiło zbiór walidacyjny. W trakcie analizy ustalono, że dane (zarówno te 100-procentowe, jak i te 90-procentowe) nie mają rozkładu normalnego. Za pomocą narzędzia Histogram w oprogramowaniu ArcGIS spraw-

dzono kształtu rozkładu danych. Przeglądając statystyki średnie i mediany, można określić centralne położenie dystrybuanty. Zauważyć można, że na histogramach rozkład nie jest w kształcie charakterystycznego „dzwonu”, a co z tym jest związane – średnie i mediany nie są sobie bliskie. Oznacza to, że rozkład ten nie jest zbliżony do normalnego. W związku z tym wykluczono metody, których użycie wymaga rozkładu normalnego, na przykład kriging rozłączny. Do analizy wybrano metody:

- **IDW** – Inverse Distance Weighting,
- **RBF** – Radial Basis Function,
- **LPI** – Local Polynomial Interpolation,
- **KO** – Kriging Ordinary,
- **KS** – Kriging Simple,
- **KU** – Universal kriging.

Następnie wykorzystując wybrane metody geostatystyczne, opracowano mapy w oprogramowaniu ArcGIS. Dla każdego roku opracowano 12 map – sześć metod po dwa rodzaje danych (dane 90% i 100%), co dało w sumie 132 mapy. W następnym etapie przeprowadzono walidację dla każdej z map dwiema metodami, to znaczy metodą kroswalidacji i walidacji podzbiorem. Każdą z macierzy wynikowych zapisano w formie tabeli liczbowej. Uproszczony schemat pracy został przedstawiony na rysunku 1 na przykładzie danych z roku 2014.

W kolejnym etapie wyznaczono parametry oceny estymacji zgodnie z przyjętą procedurą (sześć parametrów dla każdej metody w każdym badanym roku) i przy wykorzystaniu programowania liniowego wyznaczono dla każdego roku optymalną metodę geostatystyczną. W badaniu wprowadzono takie same wagi dla każdego parametru:

$C_1, C_2, \dots, C_m$ – ocena jakości estymacji metody geostatystycznej,

$\Delta_1, \Delta_2, \dots, \Delta_n$ – waga parametru,

$a_1, a_2, \dots, a_n$ – różnica w procentach do najlepszego parametru RMSE w grupie danych 100%,

$b_1, b_2, \dots, b_n$ – różnica w procentach do najlepszego parametru ME w grupie danych 100%,

$c_1, c_2, \dots, c_n$ – różnica w procentach do najlepszego parametru SD,s w grupie danych 100%,

$d_1, d_2, \dots, d_n$ – różnica w procentach do najlepszego parametru RMSE dla danych 90%,

$e_1, e_2, \dots, e_n$ – różnica w procentach do najlepszego parametru ME dla danych 90%,

$f_1, f_2, \dots, f_n$ – różnica w procentach do najlepszego parametru SD,s dla danych 90%,

$n_1, n_2, \dots, n_n$ – różnica w procentach do najlepszego parametru,

Rysunek 1. Schemat wykonania procedury wyboru optymalnej metody estymacji

Źródło: opracowanie własne.

Tabela 2. Zestawienie wyników POJE w analizie cen nieruchomości rolnych niezabudowanych w latach 2004–2014

Metoda	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Suma
IDW	1	6	3	4	5	5	3	1	2	1	5	36
LPI	3	2	5	3	4	6	6	5	6	5	6	51
RBF	2	3	1	5	3	2	4	2	1	4	4	31
KO	4	5	4	2	2	1	1	3	5	2	2	31
KS	6	1	6	6	6	3	5	6	4	6	1	50
KU	5	4	2	1	1	4	2	4	3	3	3	32

Źródło: opracowanie własne.

W tabeli 2 każdej z metod przyporządkowano miejsce od 1 do 6, które metoda uzyskała zgodnie z przyjętą procedurą wyboru optymalnej metody estymacji (rys. 1). Na 11 analizowanych przypadków tylko w dwóch potwierdziło się, że tam, gdzie błąd RMS jest najniższy, metoda ta jest optymalna. W 2007 roku odnotowano specjalną sytuację, że dwie metody uzyskały tę samą liczbę punktów – był to KO i KU. O kolejności decydowała precyzja na poziomie trzeciego miejsca po przecinku. Jest to jednak sytuacja, gdy kriging uniwersalny jest odmianą krigingu zwykłego. Najczęściej jako metodę optymalną wybierano IDW (trzy razy), a pozostałe metody – po dwa razy. Wyjątek stanowi tu LPI, która nie została w żadnym z przypadków przyporządkowana jako optymalna. Wynik ten uzasadnia celowość stosowania metody oceny jakości estymacji, a nie tak jak zwyczajowo jest przyjmowane, założenia, że „najlepsza” jest metoda, gdzie błąd RMS jest najniższy. Dodatkowo na potrzeby niniejszego artykułu wykonano analizę syntetyczną cen za lata 2004–2014. Optymalną metodą przy analizie syntetycznej metodą parametrycznej oceny jakości Estymacji (MPOJE) został KS (rys. 2), jednak metoda ta miała również najniższy wynik błędu RMS.

W badaniu przeprowadzonym na podstawie analizowanych danych i wyników uzyskano informację, że nie ma jednej optymalnej metody estymacji przy modelowaniu rynku nieruchomości rolnych niezabudowanych. W związku tym należy przeprowadzać ocenę jakości estymacji dla każdego zbioru danych zawierającego ceny transakcyjne.

Rysunek 2. Mapa średniej ceny transakcyjnej gruntów niezabudowanych gmin województwa warmińsko-mazurskiego w latach 2004–2014 metodą krugingu zwykłego

Źródło: opracowanie własne na podstawie danych z ANR OT Olsztyn z lat 2004–2013.

Podsumowanie

W niniejszym artykule przedstawiono procedurę opracowania map metodami geostatystycznymi uwzględniającą walidację wyników oraz autorską metodę oceny jakości estymacji na podstawie parametrów interpolacji opracowaną na podstawie algorytmu optymalizacyjnego. Na podstawie wyników parametrów oceny jakości estymacji w modelach cenowych na obszarze ANR OT Olsztyn w latach 2004–2014 (11 modeli) wnioskujemy, że:

1. Takie same rodzaje danych nie posiadają jednej uniwersalnej metody estymacji.
2. Wybór optymalnej metody estymacji w każdym przypadku powinien być poprzedzony analizą jakości estymacji.
3. Metodą do analizy jakości estymacji może być zaproponowana w artykule metoda parametryczna ocena jakości estymacji (MPOJE).

4. Wybór parametrów oraz klasyfikację ważności (wagi parametrów) w metodzie MPOJE powinien standaryzować analityk danych posiadający wiedzę geostatystyczną.
5. Uniwersalny charakter metody umożliwia jej zastosowania dla różnych rodzajów danych.

Literatura

- Cressie, N.A.C. (1993). *Statistics for Spatial Data*. John Wiley & Sons.
- Florax, R.J., Van der Vlist, A.J. (2003). Spatial Econometric Data Analysis: Moving Beyond Traditional Models. *International Regional Science Review*, 26 (3), 223–243.
- Hengel, T. (2007). *A Practical Guide to Geostatistical Mapping of Environmental Variables*. Office for Official Publications of the European Communities, Italy.
- Isard, W. (1956). *Location and Space Economy: A General Theory Relating to Industrial Location, Market Areas, Land Use, Trade and Urban Structure*. Cambridge: The IMT Press.
- Isard, W. (1960). *Methods of Regional Analysis*. New York: Technology Press of Massachusetts Institute of Technology and Wiley.
- Krivoruchko, K. (2011). *Spatial Statistical Data Analysis for GIS Users*. Esri Press.
- Kulczycki, M., Ligas, M. (2009). Interpolacja danych geoprzestrzennych pochodzących z rynku nieruchomości. *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 17 (2).
- Ogryzek, M. (2012). *Transformacja użytkowania ziemi, jako instrument optymalizacji przestrzeni planistycznej*. Rozprawa doktorska.
- Ogryzek, M., Kurowska, K. (2016). Geostatystyczne metody opracowywania map średnich cen transakcyjnych gruntów rolnych niezabudowanych. *Zeszyty Naukowe Uniwersytetu Szczecińskiego, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania*, 45 (1), 397–408.
- Rozporządzenie Rady Ministrów z 3.10.2011 w sprawie rodzajów kartograficznych opracowań tematycznych i specjalnych. Dz.U. nr 222, poz. 13, 28.
- Sarna, D. (2009). *Geostatistical with Applications in Earth Sciences*. Springer, Italy.
- Schaefer, F.K. (1953). Exceptionalism in Geography: A Methodological Examination. *Annals of the Association of American Geographers*.
- Tobler, W. (1970). A Computer Model Simulating Urban Growth in Detroit Region. *Economic Geography*, 46 (2), 234–240.
- Urbański, J. (2008). *GIS w badaniach przyrodniczych*. Gdańsk: Wyd. UG.
- Zawadzki, J. (2011). *Metody geostatystyczne dla kierunków przyrodniczych i technicznych*. Warszawa: Wyd. OWPW.

PARAMETRIC EVALUATION OF THE QUALITY OF ESTIMATION OF MAPS DEVELOPED BY GEOSTATISTICAL METHODS

Abstract

Geostatistical methods for map development is one of the cartographic methods of presenting research results. Determination of the quality of the estimation or comparison of results between estimations is possible based on the analysis of descriptive statistics, the so-called quality statistics of estimation. Accurate matching of model to data may not generate the lowest errors, that is why it is important to validate the results prior to choosing the optimal model. In practice, choosing the best method is problematic, due to the large amount of quality statistics of estimation and the lack of clear selection criteria. Within the article, the author's method of evaluation of the quality of estimation of maps developed by geostatistical methods on the basis of the average transaction price of the undeveloped land, from The Agricultural Property Agency, Regional Office in Olsztyn, will be presented. Average land price maps, as a tool to manage the agricultural property of the State Treasury, represent diversification of average transaction prices of building land and agricultural land.

Translated by Marek Ogryzek

Keywords: Agricultural Real Estate Agency resources, geostatistics, estimation, validation, interpolation, optimization algorithms, MPOJE (parametric assessment estimation method)

JEL Codes: C1, Q11, Q15, R3