

PRÓBA ZASTOSOWANIA METOD GEOSTATYSTYCZNYCH DO TAKSACJI NIERUCHOMOŚCI*

AN ATTEMPT TO APPLY GEOSTATISTICAL METHODS TO REAL ESTATE VALUATION

Piotr Cichociński

Katedra Geomatyki, Akademia Górniczo-Hutnicza w Krakowie

Słowa kluczowe: analiza wariancji, geostatystyka, kriging, taksacja nieruchomości

Keywords: analysis of variance, geostatistics, kriging, real estate valuation

Wstęp

Punktem wyjścia dla proponowanej metody wyceny były obserwacje dokonane w trakcie prac nad przeglądem wykorzystania metod kartograficznych do analizy rynku nieruchomości. Zaprezentowane we wcześniejszej publikacji autora (Cichociński, 2007) przykłady pokazały możliwość zastosowania uproszczonej metodyki, jednocześnie obnażając jednak jej wady. Przy analizie rozkładu przestrzennego wartości nieruchomości na podstawie informacji o cenach jednostkowych popełniany jest bowiem błąd polegający na tym, że zmienność tych cen nie zależy wyłącznie od różnicy w położeniu (lokalizacji) nieruchomości, lecz również od innych cech o charakterze nieprzestrzennym, takich jak: dla lokali – powierzchnia, standard wykończenia, kondygnacja, liczba pomieszczeń; dla działek – kształt, pole powierzchni.

Konieczne byłoby zatem wydzielenie składowej ceny, zależnej od tych atrybutów nieprzestrzennych. Do rozwiązania tego zadania zaproponowano w pierwszej kolejności zastosowanie statystycznej metody analizy wariancji (ANOVA).

Służy ona do badania wpływu czynnika klasyfikującego (mającego kilka poziomów) na wartość nieruchomości. Przyjmując poziomy czynnika za kryterium podziału, można wyodrębnić w badanej grupie nieruchomości k populacji. Na tej podstawie weryfikowana jest hipoteza zerowa, że średnie wartości nieruchomości w grupach są jednakowe:

$$H_0: m_1 = m_2 = \dots = m_k,$$

wobec hipotezy alternatywnej:

$$H_1: \text{co najmniej dwie średnie różnią się między sobą.}$$

* Praca zrealizowana w ramach badań statutowych nr 11.11.150.006 prowadzonych w roku 2009 w Katedrze Geomatyki Akademii Górniczo-Hutniczej w Krakowie.

Pozwoli to na przetestowanie, czy wszystkie poziomy atrybutów mają ten sam wpływ na cenę. W przypadku odrzucenia hipotezy zerowej znaczące różnice wartości zostaną wyestymowane i odjęte od wartości poszczególnych nieruchomości. Zmienność wyliczonego w ten sposób przestrzennego składnika wartości nieruchomości będzie zależała wyłącznie od ich położenia (lokalizacji). Będzie go można interpolować celem wyznaczenia jego wartości dla pozostałych nieruchomości. Znajdzie tutaj zastosowanie opisana wcześniej metoda krigingu.

Następnie, uwzględniając wartości nieprzestrzennych atrybutów tych pozostałych nieruchomości oraz wpływ ich poziomów na ceny nieruchomości, będzie można wyznaczyć całkowitą wartość wszystkich nieruchomości na analizowanym obszarze. Zastosowanie tej metody może znacząco uprościć proces pozyskiwania atrybutów nieruchomości, gdyż może się on ograniczyć tylko do cech nieprzestrzennych.

Analiza wariancji

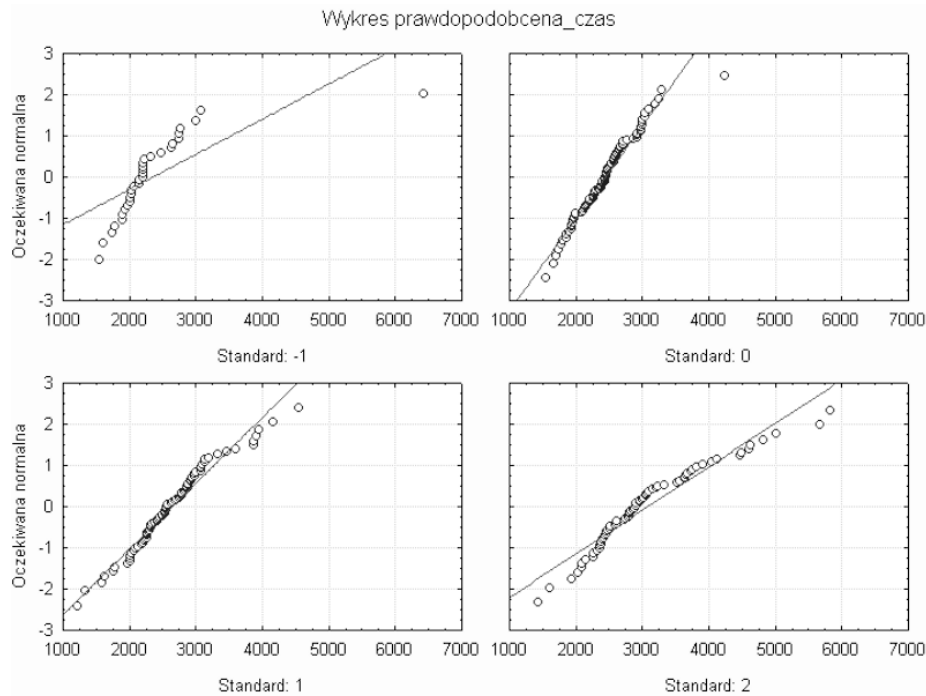
Jako pierwszy poddano analizie zbiór nieruchomości lokalowych, zawierający informacje o nieruchomościach znajdujących się w granicach byłej dzielnicy Krowodrza (obecnie dzielnice IV, VII) miasta Krakowa, obejmujący 276 transakcji zawartych między październikiem 2004 a październikiem 2005 roku. Każda ze sprzedanych nieruchomości została opisana 7 atrybutami, których nazwy oraz skale liczbowe i opisowe zaprezentowano poniżej:

1. *Data_trans* – data transakcji
2. *Pow_lok* – powierzchnia lokalu wyrażona w metrach kwadratowych
3. *Otoczenie* – wpływ otoczenia nieruchomości: bardzo korzystne (2), korzystne (1), przeciętne (0), niekorzystne (-1)
4. *Standard* – standard wyposażenia lokalu w media, urządzenia sanitarne i użytkowe: super komfort (2), komfort (1), przeciętny (0), prymitywny (-1)
5. *Komunikacja* – dostępność komunikacyjna: bardzo dobra (2), dobra (1), przeciętna (0), utrudniona (-1)
6. *Zużycie* – zużycie łączne lokalu wyrażone procentowo
7. Współrzędne X, Y.

Dodatkowo dla każdej transakcji obliczona została również cena jednostkowa sprzedaży (*Cena_jedn*) jako stosunek ceny transakcyjnej do pola powierzchni przedmiotu transakcji, a także określono czas (*Czas* – wyrażony w dniach), który upłynął od pierwszej transakcji zarejestrowanej w bazie.

W pierwszej kolejności za pomocą macierzy korelacyjnej sprawdzono, czy istnieje związek pomiędzy wartościami poszczególnych atrybutów a ceną jednostkową. Z cech o charakterze nieprzestrzennym tylko atrybut *Standard* wykazał korelację na poziomie średnim, pozostałe zaś na niskim lub wręcz nieistotnym. Zwrócono natomiast uwagę na dodatnią korelację pomiędzy ceną jednostkową a czasem, który upłynął od pierwszej transakcji. Świadczy to o generalnym wzroście cen nieruchomości w tym okresie. Aby uwolnić się od wpływu tego czynnika na wyniki analiz, wyznaczono parametry prostej regresji liniowej, a następnie dokonano korekty cen jednostkowych lokali. Po wykonaniu tych działań można było przejść do przeprowadzenia analizy wariancji (ANOVA). Jako jedyną istotną cechą nieprzestrzenną wybrano *Standard*.

ANOVA jest zestawem technik pozwalających na porównywanie kilku średnich równocześnie. Jeśli badana populacja podzielona jest na kilka grup, to umożliwia stwierdzenie róż-



Rys. 2. Skategoryzowane wykresy prawdopodobieństwa cechy *cena_czas* w grupach określonych przez *Standard*

Kolejny warunek do sprawdzenia to równość wariancji, a więc i odchyłeń standardowych rozkładów cechy w każdej grupie. Tu także za wystarczającą uznano niemożliwość odrzucenia hipotezy zerowej mówiącej o zachodzeniu takiej równości. Sprawdzenie tego warunku umożliwił test Levene'a. Testuje on hipotezę zerową o braku różnic między wariancjami przeciwko hipotezie alternatywnej o istnieniu takiej różnicy. Dla poziomu istotności przyjęto typową wartość 0,05. Pożądana jest w tym przypadku odpowiedź negatywna, to znaczy brak możliwości odrzucenia hipotezy zerowej. Jej odrzucenie bowiem oznacza istotne różnice wariancji i uniemożliwia stosowanie metod analizy wariancji.

Niestety okazało się, że otrzymane p jest mniejsze od 0,05, co skutkuje odrzuceniem hipotezy zerowej, oznaczając istotne różnice wariancji i uniemożliwia stosowanie metod analizy wariancji, gdyż ewentualne uzyskane wyniki byłyby pozbawione wartości. Oznacza to, że poszczególne grupy nieruchomości nie są jednakowo zróżnicowane. Jedną z możliwych przyczyn takiego wyniku analizy może być stwierdzenie, że rynek nieruchomości w Polsce jest wciąż jeszcze słabo rozwinięty, niestabilny i mało przejrzysty (Rekomendacja, 2000).

Dlatego postanowiono przetestować proponowaną procedurę na danych z innego, rozwiniętego rynku. Za taki można uznać rynek Stanów Zjednoczonych Ameryki Północnej. Wykorzystano dostępną w sieci Internet (pod adresem <http://sal.agecon.uiuc.edu/datasets/baltimore.zip>) listę 211 transakcji sprzedaży budynków mieszkalnych z miasta Baltimore w stanie Maryland, które miały miejsce w roku 1978, opisanych następującymi 17 atrybutami (Dubin, 1992):

1. STATION – identyfikator transakcji
2. PRICE – cena sprzedaży budynku wyrażona w tysiącach dolarów

3. NROOM – liczba pokoi
4. DWELL – 1 jeżeli budynek wolnostojący, 0 w przeciwnym wypadku
5. NBATH – liczba łazienek
6. PATIO – 1 jeżeli budynek posiada patio, 0 w przeciwnym wypadku
7. FIREPL – 1 jeżeli budynek jest wyposażony w kominek, 0 w przeciwnym wypadku
8. AC – 1 jeżeli budynek jest wyposażony w klimatyzację, 0 w przeciwnym wypadku
9. BMENT – liczba pomieszczeń piwnicznych
10. NSTOR – liczba kondygnacji
11. GAR – liczba miejsc garażowych (0 jeżeli brak garażu)
12. AGE – wiek wyrażony w latach
13. CITCOU – 1 jeżeli budynek jest położony w hrabstwie Baltimore, 0 w przeciwnym wypadku
14. LOTSZ – powierzchnia działki wyrażona w setkach stóp kwadratowych
15. SQFT – powierzchnia użytkowa budynku wyrażona w setkach stóp kwadratowych
16. X – współrzędna X
17. Y – współrzędna Y

Dodatkowo dla każdej transakcji obliczona została również cena jednostkowa sprzedaży (PRICE_SQFT), jako stosunek ceny transakcyjnej do pola powierzchni przedmiotu transakcji. Ponadto, aby możliwe było wykonanie analizy wariancji, konieczny był podział wartości atrybutów SQFT, AGE i LOTSZ na grupy. Posłużono się klasyfikacją metodą naturalnych przerw. W przypadku SQFT kolejnym trzem przedziałom przyporządkowano malejące wartości atrybutu SQFT_GRP od 2 do 0. W przypadku AGE kolejnym czterem przedziałom przyporządkowano malejące wartości atrybutu AGE_GRP od 2 do -1. W przypadku LOTSZ kolejnym pięciu przedziałom przyporządkowano malejące wartości atrybutu SQFT_GRP od 3 do -1.

W pierwszej kolejności za pomocą macierzy korelacyjnej sprawdzono, czy istnieje związek pomiędzy wartościami poszczególnych atrybutów a ceną jednostkową. Najmocniej skorelowany z ceną jest atrybut NSTOR (liczba kondygnacji) i od niego rozpoczęto analizę wariancji. Sprawdzono normalność rozkładu i usunięto jedną wartość odstającą. Uzyskane z testu Levene'a p jest znacznie większe od 0,05, co pozwoliło wykorzystać metodę ANOVA. Otrzymano p mniejsze od 0,05, zatem można było odrzucić hipotezę zerową i przyjąć hipotezę alternatywną o istnieniu istotnej statystycznie różnicy między średnią ceną jednostkową nieruchomości w każdej kategorii cechy NSTOR.

Dla zniwelowania wpływu zmienności liczby kondygnacji na cenę jednostkową nieruchomości obliczono różnice średnich cen jednostkowych w grupie drugiej i w pozostałych grupach, a następnie dodano te różnice do cen jednostkowych w poszczególnych grupach, zapisując wynik tej operacji jako atrybut P_NSTOR. Tak zmodyfikowana cena jednostkowa najsilniej koreluje z cechą CITCOU. Jednak jest to cecha o charakterze przestrzennym, dlatego wybrano kolejną z następnym niższym poziomem korelacji AGE_GRP. Jednak przeprowadzone przygotowania do wykonania w oparciu o nią jednoczynnikowej analizy wariancji pokazały, że wynik testu Levene'a ($p=0,028968$) nie pozwala na jej wykonanie. Wybrano zatem kolejną zmienną AC. W tym przypadku wszystkie prekwizyty do wykonania analizy wariancji były spełnione ($p=0,560091$) i jej wynik ($p=0,000000$) upoważnia do stwierdzenia występowania istotnych różnic ceny P_NSTOR pomiędzy grupami wyznaczonymi przez wartości cechy AC.

Dla zniwelowania wpływu wyposażenia budynku w klimatyzację na cenę jednostkową nieruchomości obliczono różnice średnich cen jednostkowych w grupie drugiej i w grupie

pierwszej, a następnie dodano te różnice do cen jednostkowych w grupie drugiej, pozostawiając ceny w grupie pierwszej bez zmian i zapisano wynik tej operacji jako atrybut P_AC. Teraz dla tak zmodyfikowanej ceny jednostkowej wyznaczono korelacje. Cechą wykazującą największe skorelowanie z ceną P_AC jest (oprócz pomijanego CITYCOU) cecha BMENT. W tym przypadku wszystkie prerekwizyty do wykonania analizy wariancji były spełnione ($p=0,264169$) i jej wynik ($p=0,001474$) upoważnia do stwierdzenia występowania istotnych różnic ceny P_AC pomiędzy grupami wyznaczonymi przez wartości cechy BMENT.

Dla zniwelowania wpływu zmienności liczby pomieszczeń piwnicznych na cenę jednostkową nieruchomości obliczono różnice średnich cen jednostkowych w grupie drugiej i w pozostałych grupach, a następnie dodano te różnice do cen jednostkowych w poszczególnych grupach, zapisując wynik tej operacji jako atrybut P_BMENT. Teraz dla tak zmodyfikowanej ceny jednostkowej wyznaczono korelacje. Pozostałe cechy nieprzestrzenne istotnie skorelowane z ceną P_BMENT spełniały warunki testu Levene'a, ale wyniki analizy wariancji nie upoważniały do stwierdzenia występowania istotnych różnic ceny P_BMENT.

Zakończono więc na tym analizy wariancji uznając, że zmodyfikowana w powyżej opisanym sposobie cena jednostkowa (zapisana ostatecznie jako atrybut P_NS_AC_BM) została pozbawiona istotnych zależności od cech nieruchomości o charakterze nieprzestrzennym. Uznano, że uzyskana w ten sposób jednostkowa cena nieruchomości zależy wyłącznie od jej lokalizacji i będzie ją można interpolować metodą krigingu celem wyznaczenia jej wartości dla pozostałych nieruchomości.

Interpolacja przestrzennego składnika wartości nieruchomości

Do realizacji tego zadania wykorzystano funkcję *kriging zwykły* dostępną w pakiecie oprogramowania ArcGIS (ESRI, 2004). Wynikiem jej działania jest powierzchnia (zapisana w rastrowym modelu danych), której wysokości odpowiadają wartościom wyinterpolowanym we wszystkich punktach regularnej siatki.

Pierwszą taką powierzchnię wygenerowano z domyślnymi parametrami automatycznie dobranymi przez program (rys. 4). Okazało się, że dopasowanie tego modelu nie jest najlepsze, na co wskazują parametry dokładnościowe zaprezentowane na rysunku 8.

Podjęto zatem próbę znalezienia bardziej optymalnych parametrów modelu. Narzędziem *analiza trendu* zidentyfikowano niewielki trend U-kształtny na kierunku północ południe, który można było zamodelować krzywą stopnia drugiego (rys. 3). Jednak próby dobrania lepszych wartości tych parametrów dały w wyniku model (rys. 5) o niewiele tylko lepszych parametrach dokładnościowych (rys. 9).

Postanowiono więc poszukać przyczyn takiego stanu rzeczy. Sporządzona metodą sygnatur o zmiennej kolorystyce mapa rozkładu cen jednostkowych P_NS_AC_BM (rys. 12) pokazała, że w zbiorze mogą istnieć wartości odstające. Ich formalnej identyfikacji dokonano za pomocą diagramu *Voronoi* z parametrem *klaster* (rys. 13).

Wyróżnił on z wszystkich wieloboków podzielonych na pięć klas te, które należą do innej klasy niż jego sąsiedzi, tym samym pokazując punkty (nieruchomości) o odstających cenach. Punkty te zostały usunięte ze zbioru danych, w oparciu o który powtórnie wyinterpo-

lowano powierzchnię metodą *krigingu zwykłego* (rys. 6). Parametry dokładnościowe tego modelu w dalszym ciągu nie są satysfakcjonujące (rys. 10). Uzyskane metodą krosvalidacji odchyłki w punktach o znanych wartościach wynoszą od -1,459164 do 1,598858, co stanowi nawet do 260% cen jednostkowych poszczególnych budynków.

Dodatkowo wygenerowano jeszcze jeden model w oparciu o pierwotne, niemodyfikowane ceny jednostkowe budynków (rys. 7). Porównanie parametrów dokładnościowych z najlepszym z wcześniejszych modeli zaprezentowano na rysunku 11. Pokazuje ono, że zaproponowana metoda daje lepsze wyniki niż uzyskane w oparciu o dane surowe. Niestety nie jest jednak wystarczająco dokładna, żeby można było ją wykorzystać do taksacji nieruchomości.

Wnioski

Próby zwiększenia dokładności predykcji wartości nieruchomości z zastosowaniem metod analizy wariancji i geostatystyki zakończyły się częściowym powodzeniem. W przypadku danych pochodzących z krakowskiego rynku nieruchomości istotne różnice wariancji cen jednostkowych nieruchomości pomiędzy grupami wyznaczonymi przez różne wartości ich atrybutów uniemożliwiły stosowanie metod analizy wariancji, gdyż ewentualne uzyskane wyniki byłyby pozbawione wartości. Oznacza to, że poszczególne grupy nieruchomości nie są jednakowo zróżnicowane. Można podejrzewać, że rynek nieruchomości w Polsce jest wciąż jeszcze słabo rozwinięty, niestabilny i mało przejrzysty. Cenne byłoby powtórzenie takiej analizy w momencie, gdy rynek będzie już wystarczająco rozwinięty celem zweryfikowanie jej użyteczności.

Wykorzystanie danych z rozwiniętego rynku Stanów Zjednoczonych Ameryki Północnej pozwoliło zaobserwować wzrost dokładności prognozy wynikający z redukcji wpływu nieprzestrzennych cech nieruchomości na ich ceny jednostkowe. Niestety jednak zaproponowana metoda nie jest wystarczająco dokładna, żeby można ją było bezpośrednio wykorzystać do taksacji nieruchomości.

Literatura

- Cichociński P., 2007: Zastosowanie metod kartograficznych i geostatystycznych do wstępnej analizy rynku nieruchomości. *Studia i materiały Towarzystwa Naukowego Nieruchomości*, vol. 15, nr 3-4, 155-166.
- Dubin R.A., 1992: Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics* 22(3), 433-452.
- ESRI, 2004: ArcDoc Version 9.0. Environmental Systems Research Institute, Redlands.
- Rekomendacja J. z dnia 20 lipca 2000 r. dotycząca tworzenia przez banki baz danych odnoszących się do rynku nieruchomości, http://www.knf.gov.pl/Images/rekomendacja_j_tcm20-8560.pdf

Abstract

The starting point for the proposed valuation method were observations made during works on the review of application of cartographic methods to analyze the real estate market. Examples presented in the author's previous publications have shown the possibility of using the simplified methodology, while revealing its disadvantages. In examining the spatial distribution of real estate values on the basis of unit prices, an error is made due to the fact that the variability of these prices depends not only on the difference in the position (location) of the real estates, but also on other non-spatial characteristics, such as area, the standard of equipment, the floor, number of rooms – for dwellings and the shape, surface area – for parcels.

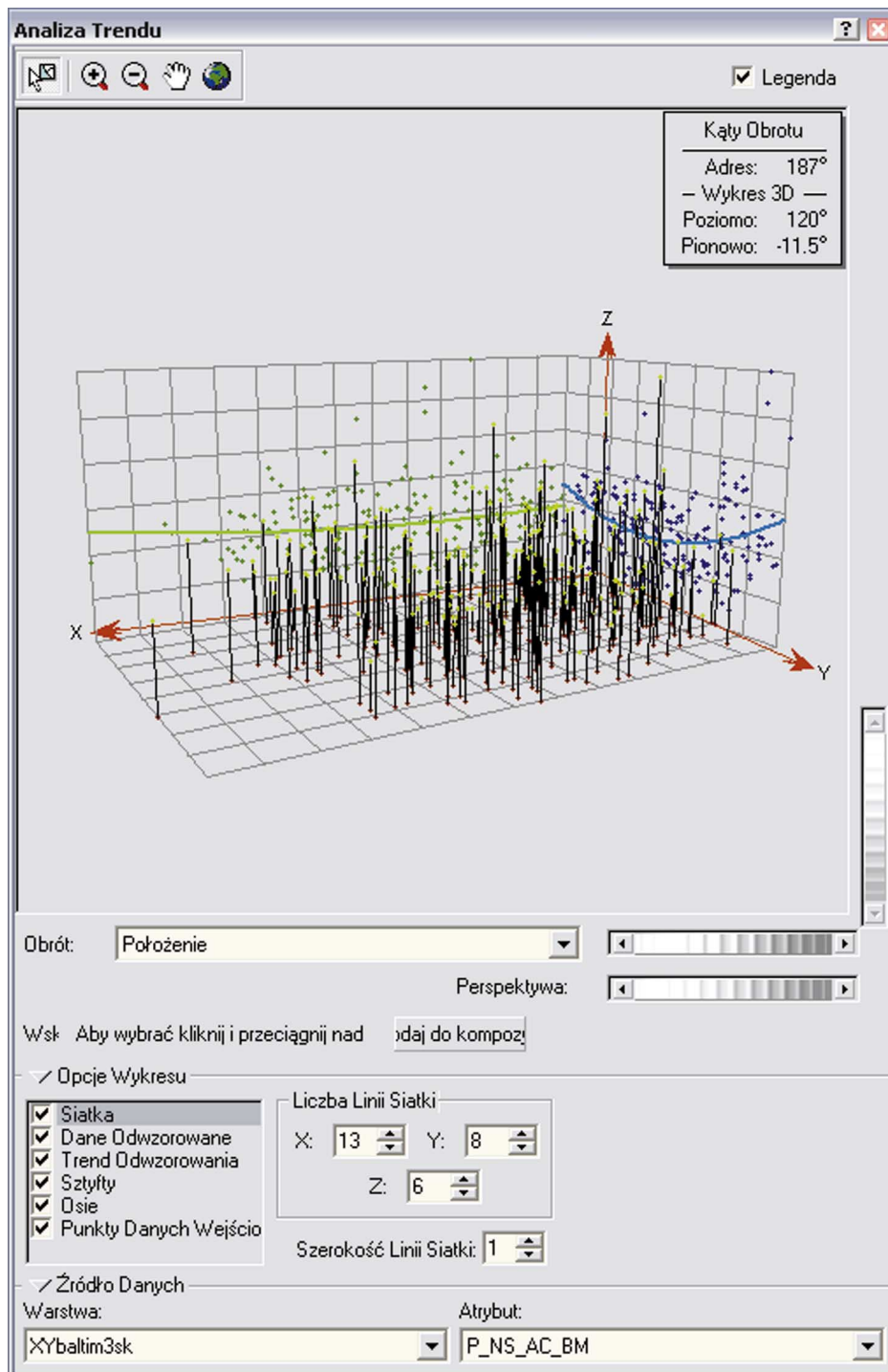
So, it would be necessary to separate the price component, depending on these non-spatial attributes. To solve this task, the application of analysis of variance (ANOVA) statistical method was proposed first. It allows testing the influence of the classifying factor (having several levels) on the value of the real estate. This allows to test whether all levels of the attributes have the same impact on the price. In case of their appearance, significant differences will be estimated and values will be subtracted from the real estate value. Variability of the spatial component of the real estate value calculated this way depends only on its location. It is possible to interpolate it to calculate its value to other real estates. The kriging method is applied here. Then, taking into account the values of non-spatial attributes of these other real estates and the impact of their levels on the price of real estate, it is possible to calculate the total value of all real estates on a given area. Application of this method can significantly simplify the process of acquiring the real estate attributes, as it may be limited to the non-spatial characteristics. The set of dwellings from the Krakow area was analyzed first. Each sold dwelling was described by 7 attributes. First, it was examined by means of correlation matrix whether a relationship exists between the values of individual attributes and the unit price. From the set of non-spatial characteristics only standard of equipment showed a correlation at the medium level, and the others were at a low or even negligible level. Unfortunately, it appeared that there were significant differences of variance between groups of dwellings distinguished by the value of this attribute, and thus it was impossible to apply analysis of variance methods, as any results obtained would be worthless. This means that different groups of dwellings are not equally diversified. One of the possible reasons for this result of analysis may be the statement that the property market in Poland is still underdeveloped, unstable and not transparent.

Therefore, the author decided to test the proposed procedure on data from another developed market. The market of the United States of America can be considered as such. Available on the Internet list of sale transactions of houses, described by 17 attributes, was used.

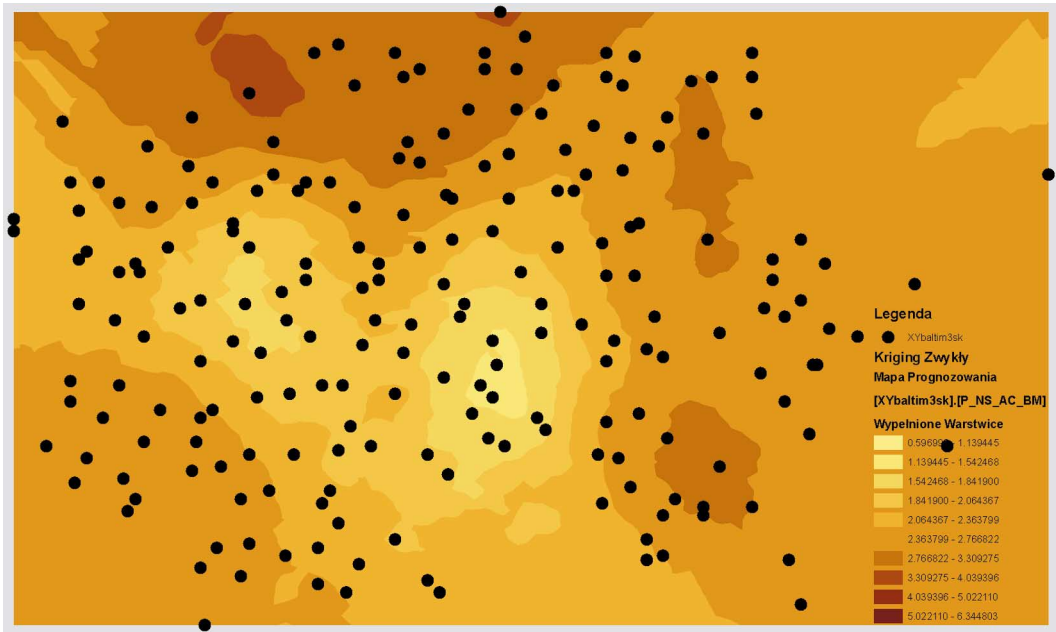
Using correlation matrix it was examined whether there were relationships between the individual attribute values and the unit price. For consecutive attributes strongly correlated to the price, the existence of statistically significant differences between the average unit house prices in categories set by the values of this characteristic was examined. If the difference occurred, the differences of average unit prices for individual groups were calculated and were used to offset the impact of the variability of any given characteristic on the value of the house. It was assumed that the unit price of the building obtained this way, did not essentially depend on non-spatial real estate characteristics, relying solely on its location and it could be interpolated using kriging method to determine values of other buildings. Ordinary Kriging function was used to accomplish this task. The result of this function is the surface (in the raster data model), having „height” values interpolated at all points of a regular grid.

The first surface was generated with default parameters automatically selected by the program. It turned out that the fit of this model was not good. Attempts undertaken to select better values of these parameters gave as a result the model with only slightly better accuracy parameters. Deviations at the points with known values, obtained by cross-validation method were up to 260% of the unit prices of individual buildings. Comparison of accuracy parameters with the model generated basing on the original unit prices showed that the proposed method gave better results than those obtained on the basis of the raw data. Unfortunately, however, it is not sufficiently accurate to be used for real estate valuation.

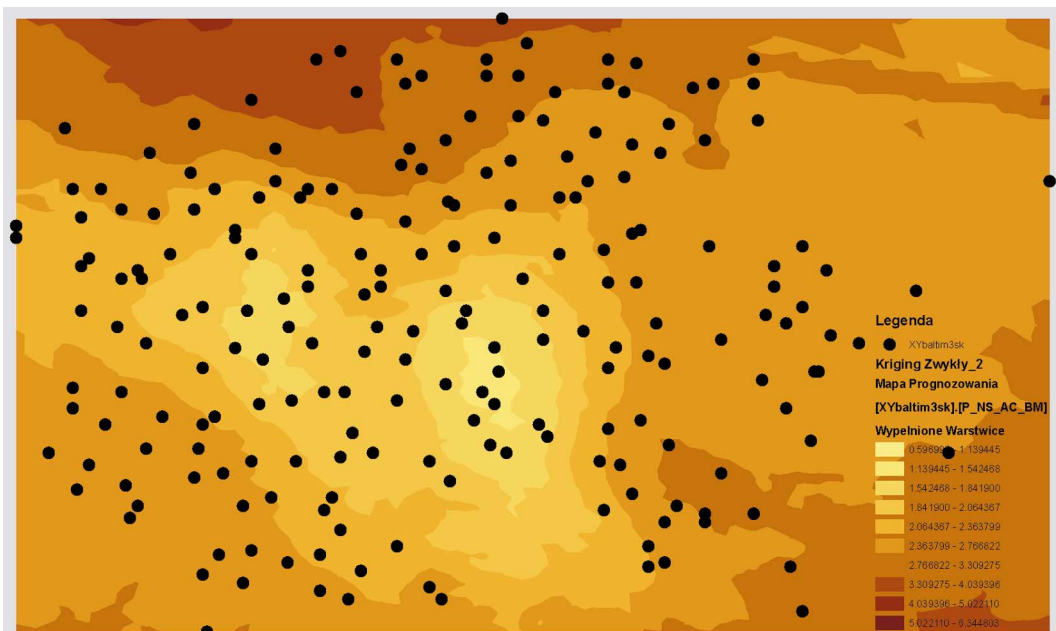
dr inż. Piotr Cichociński
Piotr.Cichocinski@agh.edu.pl
tel. +48 12 617 34 31



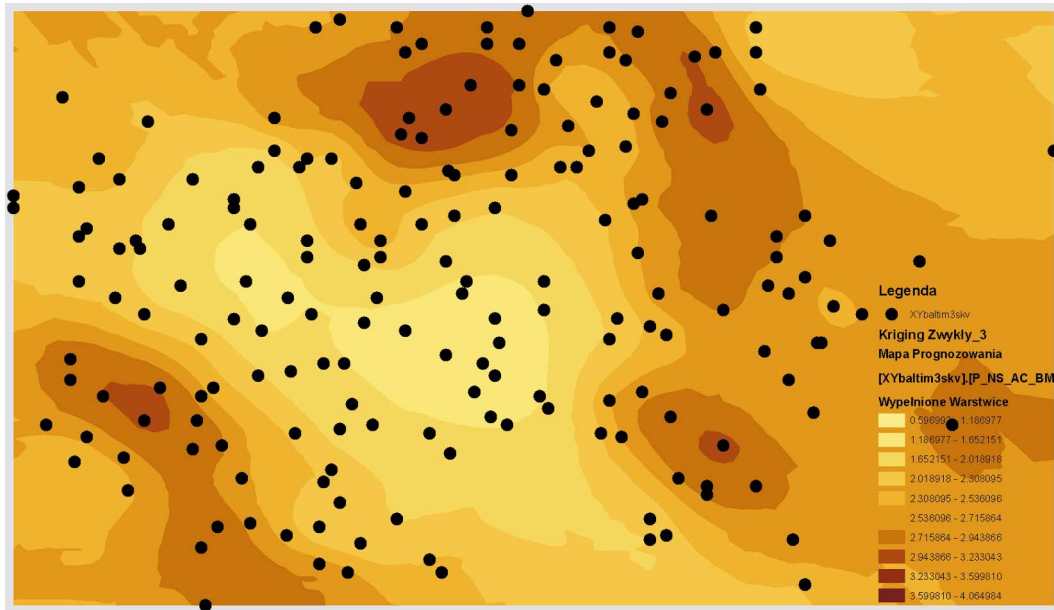
Rys. 3. Analiza trendu



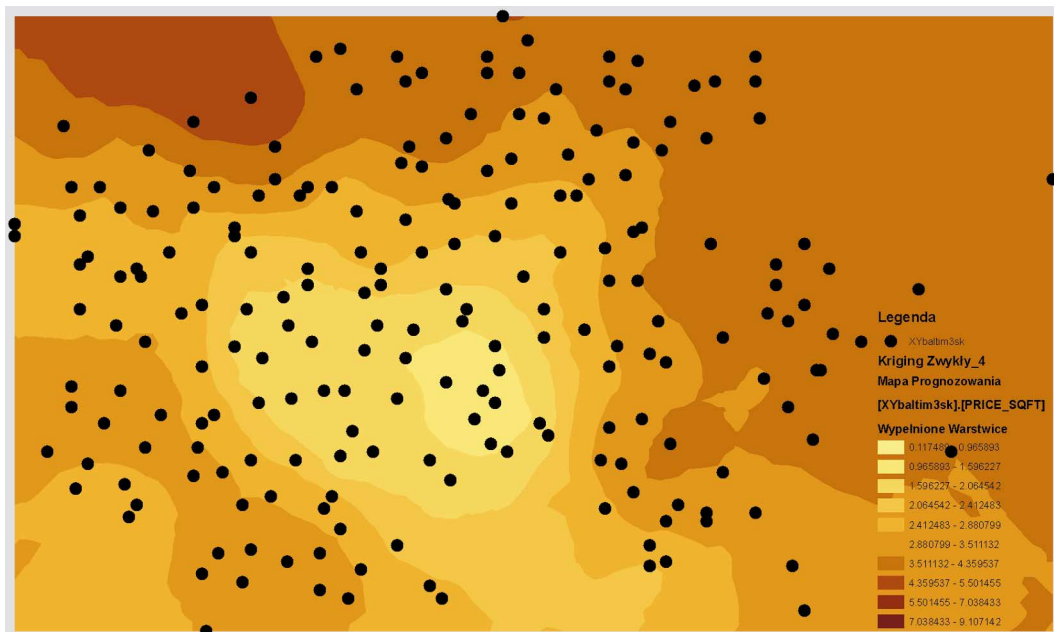
Rys. 4. Kriging zwykły z domyślnymi parametrami



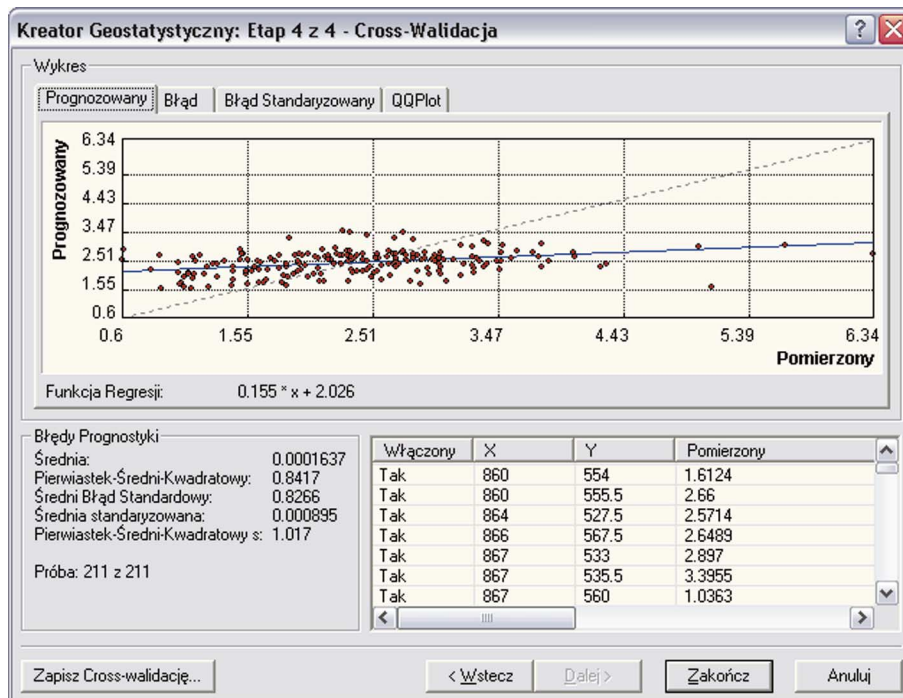
Rys. 5. Kriging zwykły z optymalnymi parametrami



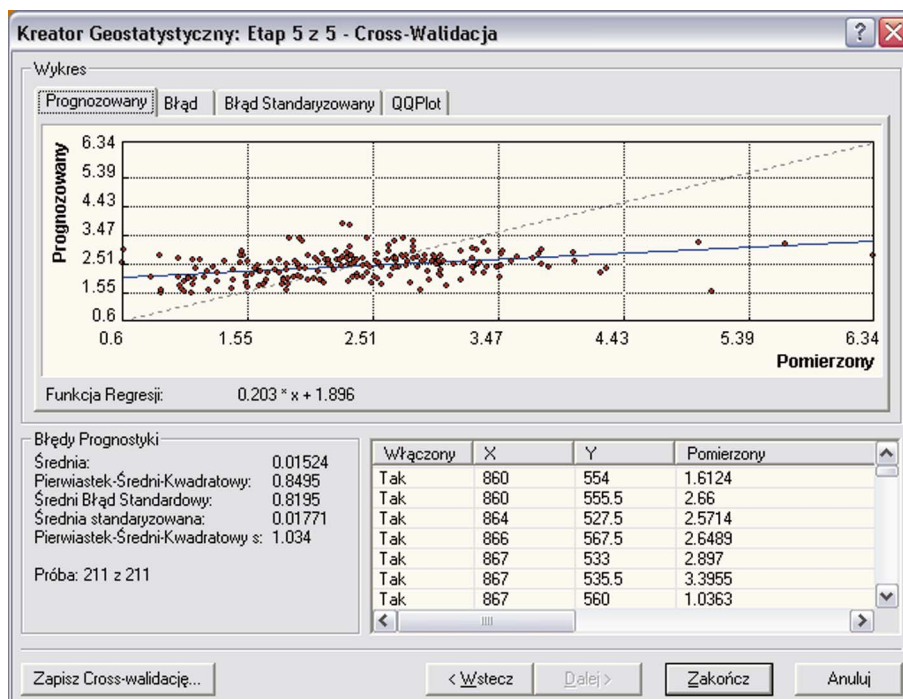
Rys. 6. Kriging zwykły z optymalnymi parametrami – usunięte obserwacje odstające



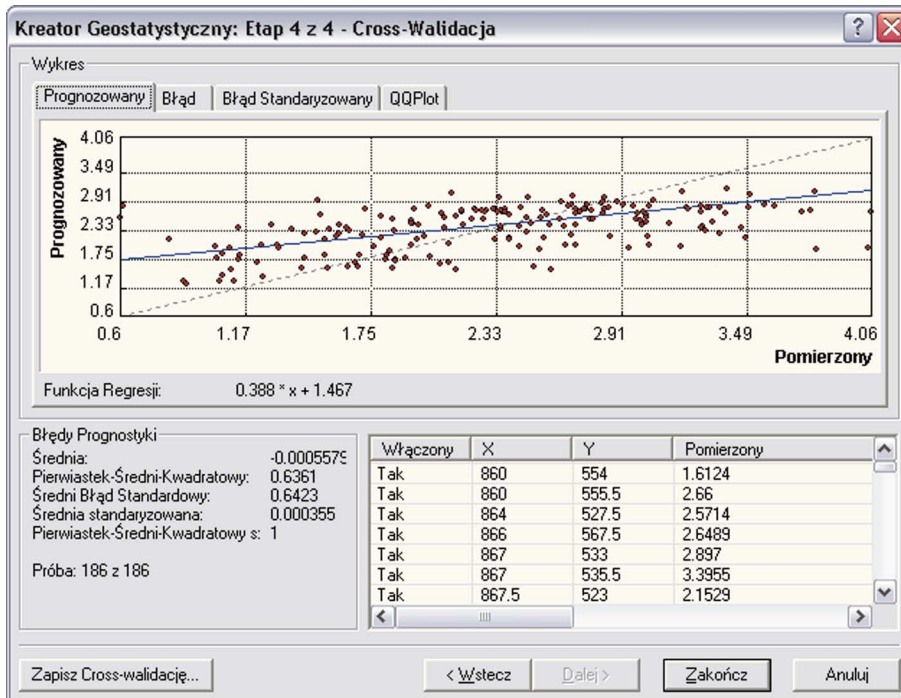
Rys. 7. Kriging zwykły z optymalnymi parametrami – ceny niemodyfikowane



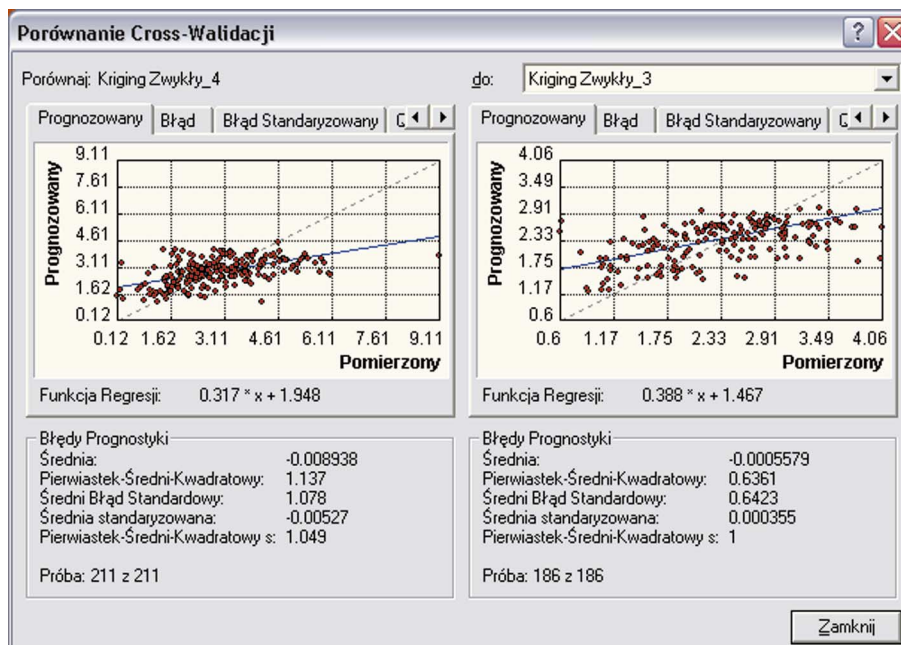
Rys. 8. Kriging zwykły z domyślnymi parametrami – dopasowanie modelu



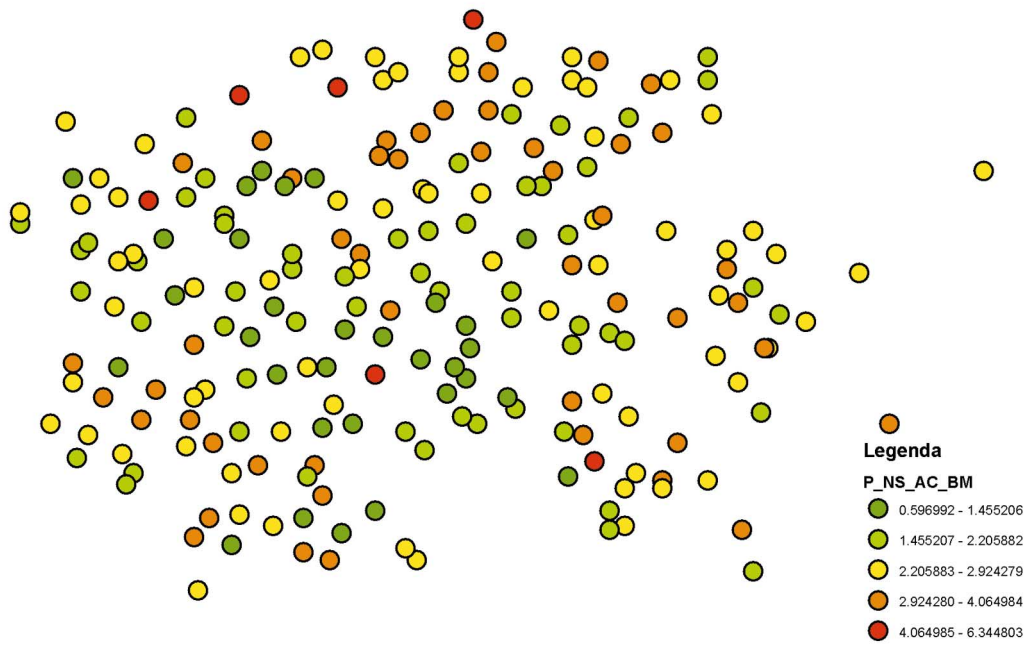
Rys. 9. Kriging zwykły z optymalnymi parametrami – dopasowanie modelu



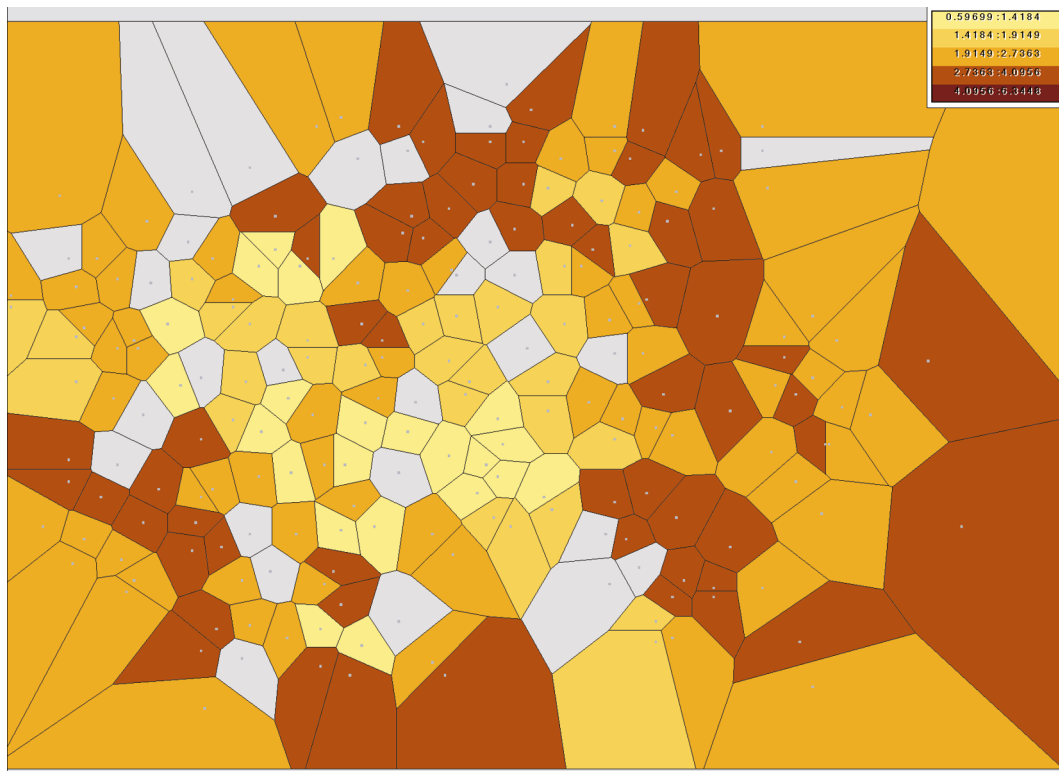
Rys. 10. Kriging zwykły z optymalnymi parametrami – usunięte obserwacje odstające – dopasowanie modelu



Rys. 11. Porównanie parametrów dokładnościowych modelu zbudowanego w oparciu o ceny niemodyfikowane i modelu najlepszego



Rys. 12. Mapa rozkładu cen jednostkowych P_NS_AC_BM



Rys. 13. Diagram Voronoi z parametrem *klaster*