PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Tomasz Klimanek, Marcin Szymkowiak

Uniwersytet Ekonomiczny w Poznaniu

TAKSONOMICZNE ASPEKTY ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ KORELACJĘ PRZESTRZENNĄ

Streszczenie: Głównym celem artykułu jest prezentacja metod i technik estymacji pośredniej uwzględniających przestrzeń. Wykorzystując dane z Narodowego Spisu Rolnego 2002 oraz mierniki autokorelacji przestrzennej, autorzy podejmują próbę oceny obciążenia estymatora EBLUP oraz estymatora EBLUP uwzględniającego korelację przestrzenną SEBLUP. Wyniki przeprowadzonych symulacji wskazują, że wykorzystanie informacji *a priori* o występowaniu bądź braku autokorelacji przestrzennej badanego zjawiska może w znaczący sposób poprawić jakość uzyskanych oszacowań (obciążenie estymatora).

Słowa kluczowe: spis rolny 2002, statystyka małych obszarów, autokorelacja przestrzenna, estymator EBLUP wykorzystujący korelacje przestrzenne

1. Wstęp

Od połowy lat 90. XX wieku metodologia estymacji pośredniej zyskuje coraz większe uznanie w polskim środowisku naukowym z zakresu metod ilościowych. Jedną z większych zalet statystyki małych obszarów jest fakt, że oferuje ona metody szacowania parametrów rozkładu cech w domenach/małych obszarach, dla których stosowanie klasycznych estymatorów, znanych z metody reprezentacyjnej, skutkowałoby błędami szacunku o nieakceptowanym poziomie, a wynikającymi ze zbyt małych prób w tych domenach/małych obszarach. Metody te opierają się na analogicznej, jak w przypadku taksonomii, idei podobieństwa obiektów (w przypadku estymacji pośredniej podobieństwa domen/małych obszarów).

W statystyce małych obszarów ważną rolę odgrywają estymatory oparte na modelach (*model based estimators*), których popularność w ostatnich latach znacząco wzrosła [Rao 2003]. Należy jednakże zauważyć, że znacząca liczba dotychczasowych aplikacji w estymacji pośredniej polegała na stosowaniu modeli przekrojowych. Dopiero na początku XXI wieku pojawiły się pierwsze prace, głównie o charakterze teoretycznym i metodologicznym, w zakresie wykorzystania modeli

przestrzennych, modeli czasowych oraz mieszanych modeli przestrzenno-czasowych w estymacji pośredniej [Saei, Chambers 2003, 2004]. Zaczęły się również pojawiać pierwsze prace aplikacyjne, w których proponowano zastosowanie modeli wykorzystujących autokorelację przestrzenną w estymacji pośredniej, w badaniach związanych z rolnictwem [Petrucci, Salvati 2006, s. 169-182; Chandra, Salvati, Chambers 2007, s. 887-906; Pratesi, Salvati 2008, s. 113-141].

Jednym z głównych problemów oraz wyzwań stojących przed metodologią estymacji pośredniej jest brak wyraźnych wskazań, który z estymatorów w danym badaniu należałoby zastosować celem spełnienia oczekiwań odbiorcy co do jakości oszacowań. Ważne jest, aby potencjalny badacz dysponował jednolitą teorią przed podjęciem badań, a nie określał własności estymacji na podstawie szeregu symulacji. Niestety, statystyka małych obszarów w dalszym ciągu dopracowuje się tej teorii. Niniejszy artykuł stanowi próbę wypełnienia istniejącej luki. Przedstawiona została w nim autorska propozycja wykorzystania pewnych informacji *a priori* o występowaniu bądź braku autokorelacji przestrzennej badanego zjawiska, które mogą sugerować wybór danego estymatora oraz w znaczący sposób poprawić jakość uzyskanych oszacowań w sensie obciążenia estymatora.

2. Opis procedury badawczej i wykorzystanych źródeł danych

Zamieszczone wyniki stanowią pewną część prac podgrupy roboczej do spraw metod statystyczno-matematycznych na rzecz spisów[1], powołanej przez prezesa GUS. Prace te, w pierwszym etapie, koncentrowały się na testowaniu estymatorów pośrednich w badaniach symulacyjnych, polegających na losowaniu prób z bazy spisu rolnego 2002 w oparciu o różne schematy losowania i szacowaniu wybranych charakterystyk gospodarstw rolnych. Uzyskane wyniki pozwoliły sformułować nowe kierunki badań (zob. [Klimanek 2009]), które należałoby podjąć w dalszych pracach badawczych nad rozwojem metodologii statystyki małych obszarów w odniesieniu do danych z zakresu rolnictwa. Jednym z ważniejszych kierunków prac okazało się wykorzystanie estymatorów pośrednich, wykorzystujących informację przestrzenną (współrzędne **x**, **y** charakterystycznych punktów małych obszarów).

Procedura badawcza oparta została na podejściu symulacyjnym metodą Monte Carlo. Z bazy danych jednostkowych pochodzących ze spisu rolnego w 2002 roku dla województwa wielkopolskiego wylosowano 200 prób o liczebności względnej 1%[2]. Zastosowano schemat losowania warstwowego proporcjonalnie do pierwiastka *k*-tego stopnia z liczby gospodarstw rolnych w warstwie. Warstwy (domeny/małe

[1] W dalszej części artykułu, zamiast pełnej nazwy – podgrupa do spraw metod statystyczno-matematycznych na rzecz spisów, będziemy stosować skróconą nazwę: podgrupa.

[2] W stosunku do wcześniejszych prac zmieniono względną wielkość próby z 5% na 1%, dopasowując tę wielkość do bardziej realistycznych sytuacji (zbyt małych liczebności prób w domenach), czego oczekuje się od metodologii statystyki małych obszarów.

obszary) stanowiło 35 powiatów województwa wielkopolskiego, natomiast stopień pierwiastka ustalono jako równy 2. Funkcję zmiennej objaśnianej pełniła powierzchnia użytków rolnych w gospodarstwach rolnych wyrażona w arach, natomiast zmiennymi objaśniającymi były: powierzchnia gruntów ornych, produkcja zbóż oraz liczba ciągników własnych ogółem w gospodarstwie rolnym[3].

Ogólny wzór na wyznaczenie wielkości próby w warstwie *d* ma postać następującą:

$$n_d = \frac{\sqrt[k]{N_d}}{\sum_{d=1}^{D} \sqrt[k]{N_d}} \cdot N \cdot r \,, \tag{1}$$

gdzie: N – liczebność populacji (liczba gospodarstw rolnych w województwie wielkopolskim),
N_d – liczebność populacji w warstwie d (liczba gospodarstw rolnych w powiecie d),
d – warstwa (powiat),
D – liczba warstw (35 powiatów województwa wielkopolskiego),
n_d – liczebność próby z warstwy d (liczba gospodarstw rolnych losowanych z powiatu d),
k – stopień pierwiastka (w niniejszej publikacji przyjęto 2),
r – względna wielkość próby (w niniejszej publikacji przyjęto 1%).

Szczegółowy opis procedury badawczej omówiony jest w pracy Klimanka [2009], zatem opis ograniczony zostanie do tych nowych elementów, które wynikają z uwzględnienia danych przestrzennych w procesie estymacji.

Zastosowanie estymatora EBLUP, uwzględniającego korelację przestrzenną, wynikało z naturalnej konsekwencji uchylenia krępującego założenia o niezależności efektów związanych z każdą parą obszarów (zob. [Saei, Chambers 2004]). Okazuje się, że w tego rodzaju podejściu model może zostać wzmocniony informacjami na temat odległości między obszarami w przestrzeni geograficznej. Jak wskazuje się w literaturze, naturalnym wyborem jest uwzględnienie w procedurach estymacyjnych współrzędnych **x** i **y** dla każdego małego obszaru. Najczęściej jako punkt charakterystyczny małego obszaru wybiera się jego środek, tzw. centroidę[4].

Spośród szeregu miar oceny jakości estymacji analiza wyników zastosowanego w niniejszym artykule podejścia została ograniczona do obciążenia estymatorów.

[3] Wyboru zmiennych dokonano spośród 11 *kandydatek* za pomocą procedury *proc reg* w pakiecie SAS oraz z wykorzystaniem metody *stepwise* (parametr *method = stepwise*).

[4] Centroidy powiatów województwa wielkopolskiego zostały uzyskane w oparciu o oprogramowanie MapInfo. Obrazem centroid są punkty leżące wewnątrz obszarów, jakimi są powiaty województwa wielkopolskiego, i powinny one wyznaczać geometryczny środek figury, której obraz, w rzucie na płaszczyznę, stanowi dany powiat.

Punktem wyjścia dla autorów była analiza występowania autokorelacji przestrzennej badanej zmiennej z wykorzystaniem statystyki globalnej *I* Morana oraz statystyk lokalnych *I* Morana[5]. Pierwsza z nich służy do testowania obecności globalnej autokorelacji przestrzennej w oparciu o macierz wag **W** (wagi pierwszego rzędu standaryzowane rzędami do jedności), natomiast druga jest bardzo użyteczna przy identyfikowaniu skupień dużych lub małych wartości badanej zmiennej oraz obserwacji nietypowych (zob. [Suchecki 2010, s. 112-125]).

Globalna statystyka Morana:

$$I_w = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{*}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (2)$$

Lokalna statystyka Morana:

$$I_w = \frac{(x_i - \bar{x})\sum_{j=1}^{n} w_{ij}^{*}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (3)$$

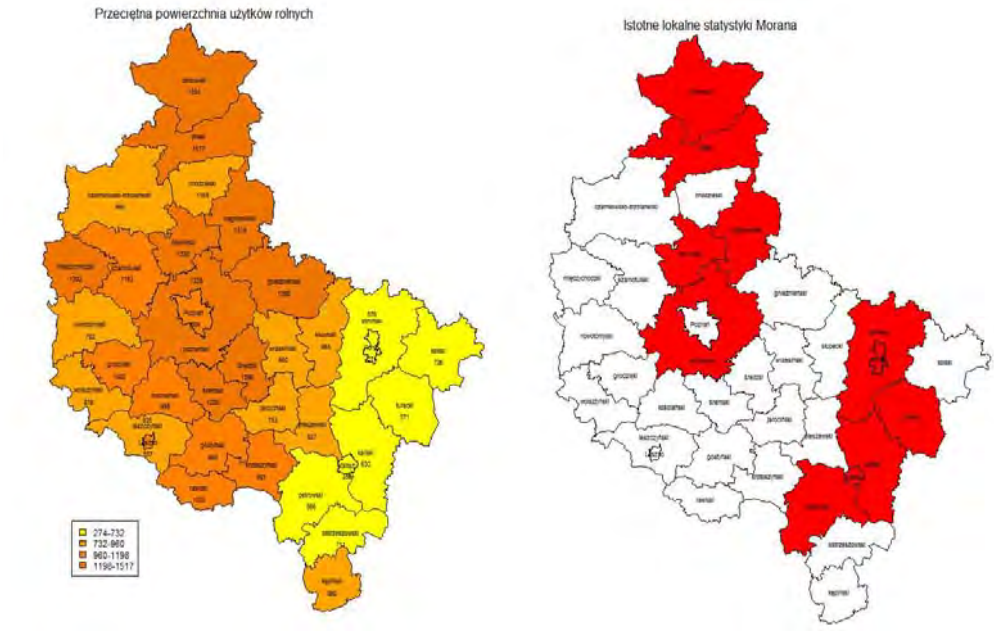
Globalna statystyka Morana osiągnęła wartość równą 0,5 przy wartości prawdopodobieństwa krytycznego *p-value* = $2{,}58 \times 10^{-7}$. Oznacza to występowanie istotnej współzależności przestrzennej badanej zmiennej. Można się zatem spodziewać, że zastosowanie estymatora uwzględniającego korelację przestrzenną powinno przynieść poprawę jakości oszacowań.

Rysunek 1 przedstawia istotne wartości statystyk lokalnych Morana oraz przestrzennego rozkładu przeciętnej powierzchni użytków rolnych w przeliczeniu na gospodarstwo rolne w powiatach województwa wielkopolskiego. Statystycznie istotna autokorelacja przestrzenna badanej zmiennej występowała w 2002 roku w następujących powiatach: złotowskim, pilskim, wągrowieckim, obornickim, konińskim, tureckim, kaliskim, ostrowskim, oraz dwóch miastach na prawach powiatu: Koninie i Kaliszu. Wzrokowa analiza kartogramu przedstawiającego kształtowanie się przestrzennego rozkładu przeciętnej powierzchni użytków rolnych w przeliczeniu na gospodarstwo rolne w powiatach Wielkopolski wykazuje, że są to faktycznie powiaty, które sąsiadują z powiatami o podobnych wartościach badanej cechy.

Wykrycie powiatów charakteryzujących się istotną statystycznie autokorelacją przestrzenną można wykorzystać do wyodrębnienia grup „podobnych" domen/małych obszarów:

- powiaty o braku istotnej statystycznie autokorelacji przestrzennej,
- powiaty o istotnej statystycznie autokorelacji przestrzennej dodatniej,
- powiaty o istotnej statystycznie autokorelacji przestrzennej ujemnej.

[5] Obliczenia wykonano w pakiecie R.



Rys. 1. Przestrzenny rozkład przeciętnej powierzchni użytków rolnych w powiatach województwa wielkopolskiego oraz istotne lokalne statystyki Morana (na szaro zaznaczono powiaty, w których stwierdzono występowanie istotnych lokalnych statystyk Morana)

Źródło: opracowanie własne w pakiecie R.

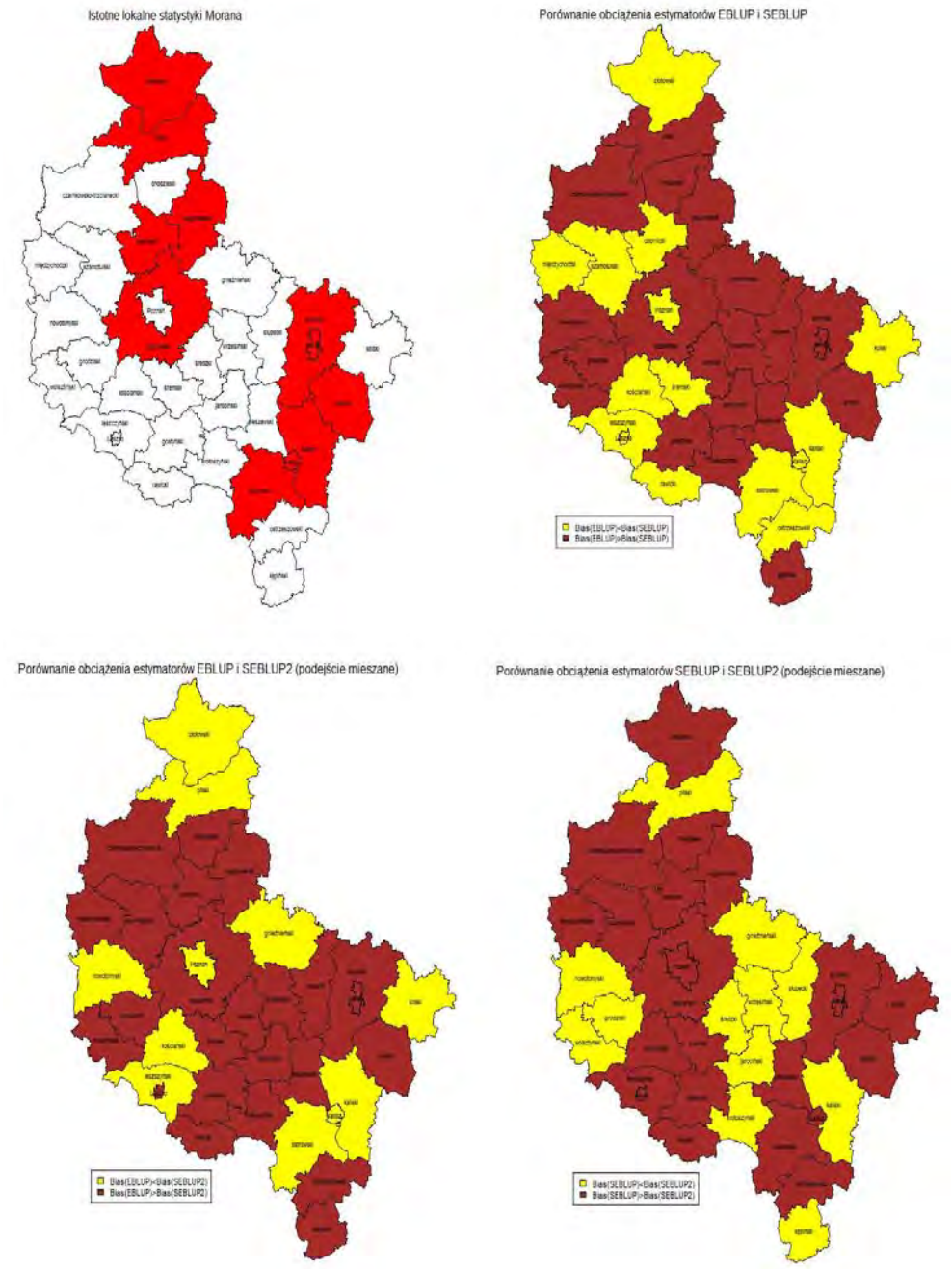
W analizowanym przez nas przypadku mieliśmy do czynienia jedynie z dwiema grupami podobnych domen (charakteryzującymi się brakiem autokorelacji oraz autokorelacją dodatnią). Zatem w celu zweryfikowania przydatności informacji *a priori* o występowaniu zależności przestrzennych do poprawienia jakości estymacji pośredniej zaproponowane zostało następujące postępowanie badawcze:

– zastosowanie do wszystkich domen estymatora EBLUP, który nie uwzględnia w swojej konstrukcji istnienia korelacji przestrzennej (EBLUP);
– zastosowanie do wszystkich domen estymatora EBLUP, który uwzględnia w swojej konstrukcji istnienia korelacji przestrzennej (SEBLUP);
– zastosowanie do domen charakteryzujących się brakiem istotnej statystycznie autokorelacji przestrzennej estymatora EBLUP, natomiast w przypadku domen charakteryzujących się istotną autokorelacją przestrzenną – estymatora SEBLUP[6].

3. Uzyskane wyniki

Jak pokazują wyniki przeprowadzonych analiz, uwzględnienie w modelu informacji o zależnościach małych obszarów w przestrzeni w znaczący sposób wpływa na

[6] Szczegółowy opis rozważanych w artykule estymatorów wraz z wzorami i ich własnościami można znaleźć w dokumentacji projektu EURAREA, znajdującej się na stronie http://www.statistics.gov.uk/eurarea.



Rys. 2. Porównanie obciążenia analizowanych estymatorów oraz istotne lokalne statystyki Morana

Źródło: opracowanie własne w pakiecie R.

redukcję obciążenia estymatorów. Wzrokowa analiza kartogramów zamieszczonych na rys. 2 wskazuje, że EBLUP, uwzględniający korelacje przestrzenne (SEBLUP), w zdecydowanej większości powiatów charakteryzował się mniejszym obciążeniem w porównaniu z estymatorem EBLUP wyznaczonym przy założeniu braku zależności przestrzennych. Przewaga estymatora SEBLUP nad estymatorem EBLUP uwidacznia się zwłaszcza w podejściu mieszanym, w którym stworzone zostały skupiska powiatów podobnych ze względu na występowanie bądź nie autokorelacji przestrzennej. Takie grupowanie w większości przypadków prowadziło do zmniejszenia obciążenia estymatora SEBLUP w porównaniu z EBLUP. Metody geostatystyki można zatem traktować jako pewną odmianę metod taksonomicznych stosowanych w estymacji pośredniej do pożyczania mocy w przestrzeni.

Należy ponadto oczekiwać, że uwzględnienie w tworzeniu skupisk − oprócz faktu występowania (bądź nie) korelacji przestrzennych − również informacji na temat sąsiedztwa i oddalenia w przestrzeni małych obszarów (powiatów) dodatkowo może przyczynić się do redukcji obciążenia estymatora SEBLUP. W analizowanym na potrzeby artykułu przykładzie empirycznym mogłoby to oznaczać stworzenie czterech skupisk, które tworzyłyby odpowiednio: skupisko 1 (powiaty: złotowski i pilski), skupisko 2 (powiaty: wągrowiecki, obornicki i poznański), skupisko 3 (powiaty: koniński, kaliski, turecki, ostrowski oraz miasta Konin i Kalisz), skupisko 4 (pozostałe powiaty charakteryzujące się brakiem występowania autokorelacji przestrzennych). Weryfikacja tej hipotezy wykracza jednak poza ramy niniejszego artykułu i będzie przedmiotem dalszych dociekań autorów.

4. Wnioski

Przeprowadzone badania symulacyjne oraz uzyskane wyniki wskazują, że zaproponowane podejście może stanowić podstawę wyboru postaci estymatora w przypadku posiadania pewnej wstępnej wiedzy na temat kształtowania się badanego zjawiska. Taką wiedzą możemy dysponować w oparciu o informacje uzyskane od ekspertów w danej dziedzinie lub na podstawie wyników wcześniejszych badań. Zastosowanie statystyk globalnych i lokalnych Morana pozwala z kolei zweryfikować hipotezę o zależności przestrzennej i na tej podstawie wyodrębnić grupy obiektów, dla których zostaną stworzone odrębne modele estymacji pośredniej (model dla domen charakteryzujących się brakiem istotnej autokorelacji przestrzennej, model dla domen charakteryzujących się istotną dodatnią autokorelacją przestrzenną oraz model dla domen charakteryzujących się istotną ujemną autokorelacją przestrzenną).

Wiele problemów wymaga jednak dalszych prac badawczych. Jedną z najważniejszych kwestii jest pytanie o uniwersalność zastosowanego podejścia w odniesieniu do innych kryteriów oceny jakości estymacji. W przedstawionej aplikacji wybranym kryterium oceny jakości zastosowanych estymatorów było ich obciążenie. Jest to jedno z ważniejszych, ale nie jedyne kryterium. Należy podjąć dalsze prace badawcze, które

wskazałyby, w jakim stopniu zastosowane podejście ma charakter uniwersalny, tzn. odnosi się także do błędu średniokwadratowego, względnych miar oceny estymacji i na przykład zewnętrznych kryteriów oceny jakości estymacji.

Literatura

Chandra H., Salvati N., Chambers R., *Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods*, Statistics in Transition 2007, vol. 8, no. 2, s. 887-906.

Klimanek T., *Estymacja pośrednia charakterystyk gospodarstw rolnych na podstawie Spisu Rolnego z 2002 roku*, Zeszyty Naukowe UEP nr 128, Poznań 2009, s. 88-101.

Petrucci A., Salvati N., *Small Area Estimation considering spatial correlation in watershed erosion assessment survey*, „Journal of Agricultural, Biological, and Environmental Statistics" 2006, vol. 11, no. 14, s. 169-182.

Pratesi M., Salvati N., *Small Area Estimation: The EBLUP estimator based on spatially correlated random area effects*, „Statistical Methods & Applications" 2008, vol. 17, no. 1, s. 113-141.

Rao J.N.K., *Small Area Estimation*, Wiley, Hoboken, New Jersey 2003.

Saei A., Chambers R., *Small Area Estimation*: *A Review of Methods Based on the Application of Mixed Models*, EURAREA, 2003.

Saei A., Chambers R., *Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects*, University of Southampton, Southampton 2004,

Suchecki B. (red.), *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*, Wydawnictwo C.H. Beck, Warszawa 2010.

Źródło internetowe

http://www.statistics.gov.uk/eurarea − EURAREA_Project_Reference_Volume.

TAXONOMIC ISSUES OF INDIRECT ESTIMATION WITH SPATIAL AUTOCORRELATION

Summary: The main purpose of the paper is to present methods and techniques of indirect estimation which take the space into account. Using the data from National Agricultural 2002 Census and some measures of spatial autocorrelation, the authors attempt to assess the bias of empirical best linear unbiased predictor – EBLUP and empirical best linear unbiased predictor based on spatially correlated area effects – SEBLUP. The results obtained via simulation approach indicate that *a priori* information about the presence or the lack of spatial autocorrelation in a phenomenon under study can significantly improve the quality of estimates (the bias of the estimator).