POLSKIE TOWARZYSTWO INFORMACJI PRZESTRZENNEJ
ROCZNIKI GEOMATYKI 2012 ❍ Tom X ❍ Zeszyt 3(53)

KARTOGRAFICZNE ASPEKTY ZASTOSOWANIA *DATA MINING* DO POZYSKIWANIA WIEDZY Z DANYCH POWSZECHNEGO SPISU ROLNEGO I NARODOWEGO SPISU POWSZECHNEGO LUDNOŚCI I MIESZKAŃ

CARTOGRAPHICAL ASPECTS OF *DATA MINING* TO GAIN KNOWLEDGE FROM THE AGRICULTURAL AND NATIONAL POPULATION AND HOUSING CENSUS DATA

Anna Fiedukowicz[1], Jędrzej Gąsiorowski[2]

[1]Zakład Kartografii Politechniki Warszawskiej, [2]Instytut Geodezji i Kartografii

Słowa kluczowe: dane statystyczne, data mining, portal geostatystyczny
Keywords: statistical data, data mining, geostatistics portal

Wprowadzenie

Postępująca dostępność danych, w tym w szczególności danych udostępnianych za pośrednictwem sieci Internet, sprawia, że coraz większym wyzwaniem staje się nie tyle sam do nich dostęp, lecz umiejętny ich wybór i takie przetworzenie, aby w wyniku uzyskać wiedzę, która jest w danych w pewien sposób „ukryta". Narzędziami służącymi tak rozumianemu wzbogacaniu danych (ang. *data enrichment*) są wszelkiego rodzaju analizy ekonometryczne oraz statystyczne, w szczególności techniki tzw. drążenia danych (ang. *data mining*). Znacząca część dostępnych danych posiada, bądź może posiadać, odniesienie przestrzenne w różnej formie. To zaś sprawia, że do ich pełniejszej analizy niezbędne wydaje się zaangażowanie narzędzi analitycznych uwzględniających przestrzenny charakter danych oraz kartograficznych metod wizualizacji.

Szczególny charakter i znaczenie wydają się mieć dane statystyczne, zwłaszcza zaś te o charakterze urzędowym. Takie właśnie dane oraz koncepcja interaktywnego atlasu statystycznego, rozwijana w Zakładzie Kartografii Politechniki Warszawskiej (Fiedukowicz i in., 2012), stały się przyczynkiem do niniejszych rozważań oraz proponowanych przykładów analiz możliwych do zaimplementowania w portalu geostatystycznym. Dane w Powszechnym Spisie Rolnym (PSR 2010) oraz Narodowym Spisie Powszechnym Ludności i Mieszkań (NSP 2011) zbierane były wraz z odniesieniem przestrzennym do punktu adresowego. Jednakże obecnie nie są jeszcze opublikowane pełne wyniki tych spisów. Dlatego w prezentowanych analizach wykorzystano zagregowane do poziomu powiatów (NTS-4) dane pochodzące z zasobów Głównego Urzędu Statystycznego, dostępne na stronie internetowej w ramach Banku Danych Lokalnych (http://www.stat.gov.pl/bdl/).

Udostępnianie danych statystycznych

Państwowe instytucje statystyczne na całym świecie udostępniają pewien zakres gromadzonych przez siebie danych obywatelom, także za pośrednictwem Internetu. Zakres tematyczny udostępnianych w ten sposób danych w poszczególnych krajach jest różny, na co mają z pewnością wpływ różnice w sposobie pozyskiwania danych, zakresie pytań spisów powszechnych (uzależnione m.in. od dominujących wydarzeń społecznych i gospodarczych w danym kraju), jak i uwarunkowania prawne, określające zakres i stopień agregacji danych objętych tajemnicą statystyczną. Sposoby udostępniania tych danych są jednak w wielu krajach zbliżone.

Zdecydowanie dominuje forma zestawień tabelarycznych, które można wygenerować wybierając odpowiednie tematy danych. Tabele można też zwykle zapisać w różnych formatach (najpopularniejszy wydaje się być format arkusza kalkulacyjnego .xls). Takie możliwości dają m.in. portale statystyczne w Wielkiej Brytanii (http://www.ons.gov.uk/ons/index.html) czy w Niemczech (https://www.destatis.de/), ale także polskie serwisy prowadzone przez GUS (np. Bank Danych Lokalnych dostępny na http://www.stat.gov.pl/). W niektórych portalach dane w formie tabelarycznej wzbogacono zbiorczymi (Wielka Brytania), a niekiedy w pewnym stopniu interaktywnymi wykresami (Niemcy), które generowane są na podstawie wybranej grupy danych. Dodatkowo w wielu portalach można znaleźć różnego typu raporty (zwykle w formacie .pdf), które zawierają analizy danych wraz ze zbiorczymi tabelami, wykresami, a także, co istotne, z opisem i interpretacją wyników tych analiz. Rozwiązanie, które wydaje się godne polecenia prezentuje portal brytyjski, gdzie obok linku do raportu w formacie .pdf, można znaleźć link do danych źródłowych, na których opierają się prezentowane w nim analizy. Pozwala to użytkownikowi prześledzić, a w razie potrzeby odtworzyć omawianą analizę.

Coraz większą popularność w serwisach statystycznych instytucji rządowych zyskują dedykowane serwisy mapowe. Jest to uzasadnione, biorąc pod uwagę fakt, że zbierane przez te instytucje dane mają odniesienie przestrzenne (najczęściej do jednostek terytorialnych NUTS różnych poziomów). Jednak w przypadku niektórych serwisów są one bardzo ubogie – jak w przypadku Wielkiej Brytanii, gdzie przeglądarka mapowa (http://www.neighbourhood.statistics.gov.uk/ dissemination/LeadBoundaryViewer.do?xW=1280&xH=1024) pozwala jedynie na podgląd granic różnych jednostek terytorialnych na tle mapy topograficznej, a mapy o charakterze statystycznym ilustrują wprawdzie wyniki niektórych analiz, ale mają one charakter typowo statyczny i często nienajlepszą jakość graficzną. Nieco bardziej interaktywne rozwiązania prezentowane są przez centralne instytucje statystyczne Węgier czy Niemiec. Jednak nawet w tych przypadkach możliwości interakcji ograniczone są do wizualizacji kartograficznej – zmiany palety barwnej, czy w najlepszym wypadku, zmiany jednostek agregacji bądź granic przedziałów klasowych zmiennych.

Na uwagę zasługuje fakt, że żaden z analizowanych przez autorów portali statystycznych nie posiada interaktywnych narzędzi, pozwalających na analizę tych danych. Oznacza to, że użytkownik może wprawdzie pobrać oryginalne, surowe dane, a niekiedy również je zwizualizować, jeśli jednak zależałoby mu na ich analizie, skazany jest na dostępne raporty zawierające gotowe wyniki, bądź zmuszony do zainstalowania i opanowania obsługi pakietu/ów statystycznych na własnym komputerze (takich jak Statistica, SPSS, czy też PSPP, będący jego otwartym odpowiednikiem) lub też wykorzystanie narzędzi *on-line* (np. proponowany przez Hansa Rollinga *Trendalyzer* dostępny na stronie fundacji http://www.gapminder.org/). Zdaniem autorów – w czasach tworzącego się obecnie społeczeństwa informacyjnego – celem popularyzacji wiedzy o analizie danych oraz wiedzy wynikającej z tej analizy, należało-

by te możliwości rozszerzyć. Z jednej więc strony zadbać o czytelną, interaktywną i poprawną kartograficznie wizualizację danych przestrzennych, z drugiej zaś zapewnić użytkownikowi narzędzia, na przykład w formie usług sieciowych, które będą dostosowane do poziomu jego aktualnej wiedzy i chęci jej poszerzenia.

Obserwowany w ostatnich latach ewolucyjny rozwój koncepcji portalu geostatystycznego GUS oraz jego wdrożenia pilotażowe, pozwalają sądzić iż docelowy serwis geoinformacyjny będzie spełniał omówione powyżej oczekiwania – zarówno „zwykłych użytkowników", jak i profesjonalistów. Już od lat GUS udostępnia bowiem dane statystyczne w formie zestawień tabelarycznych, które można generować wybierając odpowiednie tematy danych. W chwili obecnej w fazie końcowych testów znajduje się zaś dedykowany portal geostatystyczny, który będzie umożliwiał interaktywną wizualizację danych pochodzących z Powszechnego Spisu Rolnego czy Narodowego Spisu Powszechnego Ludności i Mieszkań w formie kartogramów. Kolejnym krokiem rozwoju serwisów GUS może być zaś udostępnienie w formie usług sieciowych interaktywnych narzędzi, pozwalających na przetwarzanie i analizę danych statystycznych wraz z ich późniejszą wizualizacją.

Istnieje wiele analiz, które mogą być realizowane przez takie usługi. Wśród nich wyróżnić można grupę metod realizujących zadania regresyjne oraz rozmaite metody klasyfikacyjne. W artykule zaprezentowano wyniki dwóch przykładowych analiz z tych grup: regresję wieloraką uwzględniającą zależności przestrzenne oraz grupowanie metodą *k*-średnich, w tym z ustalaniem optymalnej liczby klas metodą *v*-krotnej oceny krzyżowej.

Propozycje funkcjonalności analitycznych

Krajowy portal geostatystyczny powinien z jednej strony czerpać z najlepszych doświadczeń portali już istniejących – światowych o podobnym charakterze, ale także rozwiązań regionalnych, takich jak serwis Monitorowanie Rozwoju Mazowsza, z drugiej jednak strony powinien być miejscem rozwijania i testowania nowych funkcjonalności analitycznych, wizualizacyjnych czy społecznościowych.

Rozwiązaniami już wykorzystywanymi, a wartymi implementacji także w Polsce, są możliwość interaktywnego generowania wizualizacji w formie kartogramów, w których użytkownik ma możliwość określenia liczby klas, sposobu podziału na klasy, czy wreszcie palety barwnej. Sama bowiem wizualizacja zgeneralizowanych, podzielonych na klasy danych jest elementem ułatwiającym interpretację przestrzennego rozkładu zjawiska i dającym jego całościowy obraz. Przyczynia się zatem do wytworzenia kartograficznej wartości dodanej. Wyróżnianie pozycji legendy odpowiadającej wybranej na mapie jednostce terytorialnej dodatkowo ułatwia odczytywanie informacji i interpretację mapy. Połączenie mapy z danymi źródłowymi o charakterze tabelarycznym umożliwia dalsze analizy, a odnośniki do komentarzy ekspertów oraz zapewnienie wysokiej jakości metadanych przyczynić się mogą do pełniejszego zrozumienia danych. Portal geostatystyczny rozwijany obecnie przez GUS implementuje znaczącą część wymienionych powyżej rozwiązań. Poza udostępnieniem zaawansowanych narzędzi do prezentacji szerokiego spektrum danych statystycznych, zapewnia on dodatkowo możliwość pracy na dwóch poziomach: podstawowym, który dostępny jest wszystkim użytkownikom oraz eksperckim, udostępnionym na zasadzie uwierzytelniania bardziej zaawansowanym i świadomym użytkownikom.

Tworzenie się społeczeństwa informacyjnego pozwala jednak na zdefiniowanie nowej roli interaktywnego atlasu statystycznego – roli edukacyjnej i zarazem kształtującej postawy

społeczne. Aby wyjść naprzeciw tego typu oczekiwaniom, zasadnym wydaje się być zrozumiałe opisywanie dostępnych funkcji analitycznych powstającego systemu. Ponadto przydatna mogłaby się okazać możliwość zapisania wyników gotowych analiz w formie swego rodzaju pliku konfiguracyjnego (skryptu), który umożliwiłby nie tylko odtworzenie, ale przede wszystkim prześledzenie działań użytkownika (tego typu narzędzie mogłoby służyć np. mediom, jeśli te chciałyby udowodnić swoją rzetelność prezentując analizy danych statystycznych i ich wizualizacje). Cele społeczno-edukacyjne mogłyby być też realizowane przez włączenie internetowych narzędzi pozwalających na dzielenie się w portalach społecznościowych wynikami analiz, a jednocześnie przyczyniające się do popularyzacji i szerszego wykorzystania serwisu.

Inną istotną kwestią jest dobór podkładu referencyjnego, który wzbogaca możliwości interpretacji wyników analiz, czy łączenie wyników analiz z danymi tematycznymi. Integracja wynikowego kartogramu np. z siecią drogową może dostarczyć dodatkowych walorów interpretacyjnych, ale także stać się punktem wyjścia do dalszych analiz sprawdzających w sposób formalny (statystyczny) prawidłowości dostrzeżone na wizualizacji. Szeroki obecnie wybór dostępnych treści podkładowych jest elementem sprzyjającym tego typu analizom.

Istotę interaktywnego atlasu statystycznego powinien stanowić moduł analityczny (zintegrowany z modułem wizualizacji danych). Moduł ten może oferować rozmaite rodzaje analiz, zapewniające zróżnicowany poziom „wydobywania wiedzy" z danych. Od najprostszych – umożliwiających obliczenie pewnych wskaźników, poprzez operacje matematyczne na atrybutach odpowiadających sobie jednostek terytorialnych (jak podzielenie przez siebie wartości dwóch zmiennych), poprzez analizy klasycznej statystyki (jak obliczanie korelacji czy regresji między zmiennymi), aż po bardziej zaawansowane, uwzględniające przestrzennych charakter danych już na etapie samej analizy, nie zaś dopiero w momencie wizualizacji danych.

Do realizacji tych zadań niezbędna jest z jednej strony wiedza dotycząca metod statystyki (np. data mining i sztucznej inteligencji) czy ekonometrii, z drugiej zaś określenie narzędzi technologicznych, które mogłyby posłużyć zaproponowanym analizom. Istotny jest sposób implementacji tych narzędzi, który pozwoli docelowym użytkownikom na efektywną i łatwą w zrozumieniu i obsłudze realizację analiz, jak i udostępnienie narzędzi zapewniających kartograficznie poprawną wizualizację ich wyników.

Regresja wieloraka z uwzględnieniem sąsiedztwa

Analizowane w poniższych przykładach dane pochodzą w przeważającej większości z Banku Danych Lokalnych GUS, są więc powszechnie dostępne. Dodatkowo wykorzystano w sposób pośredni informacje o charakterze przestrzennym (odległości), uwzględniając je dodatkowo jako atrybuty – zmienne objaśniające modelu regresji (x11, x12). Analizy prowadzono na poziomie powiatów. Badano wpływ wybranych zmiennych objaśniających (tab. 1) na wartość bezrobocia w Polsce ogółem (rys. 2A), bezrobocia kobiet i bezrobocia mężczyzn dla poszczególnych powiatów, konstruując różne warianty modeli regresji.

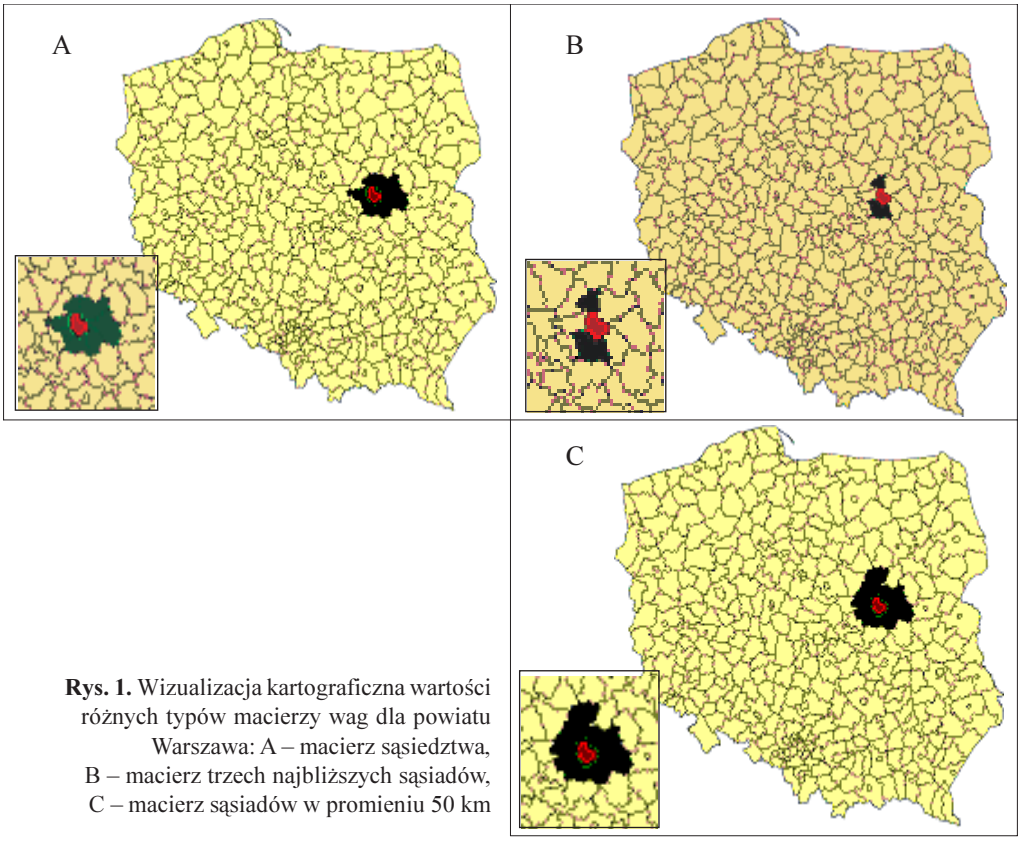
W pierwszej kolejności dokonano analizy przestrzennego rozmieszczenia zmiennych objaśnianych obliczając wartości statystyki I Morana dla bezrobocia ogółem, kobiet i mężczyzn. Obliczenia te wskazują, że przestrzenny rozkład tej zmiennej jest nielosowy tj. występują skupiska małych i dużych wartości zmiennej częściej niż wynikałoby to z przypadku. Świadczy o tym statystyka I Morana większa od zera (wartość statystyki równa 1 świadczy o idealnym skupieniu, zaś -1 o idealnie równomiernym rozkładzie zmiennej).

Obliczenie statystyki I Morana wymaga uwzględnienia modelu sąsiedztwa, opisywanego przez macierz wag. Macierze wag można definiować w rozmaity sposób, najczęściej wykorzy-

stując w tym celu elementy geograficzno-geometryczne. W niniejszej pracy wykorzystano kilka, najbardziej popularnych: macierz sąsiedztwa – uwzględniająca wspólne granice (zero-jedynkowa lub standaryzowana wierszami), macierz k-najbliższych sąsiadów (przyjęto k=3), macierz sąsiadów w odległości d (w tym badaniu d = 50 km) oraz macierz odwrotnej odległości (rys. 1). Statystyka Morana dla bezrobocia wynosi więc dla trzech pierwszych macierzy ok. 0,54, dla macierzy sąsiadów w odległości 50 km ok. 0,45, a dla macierzy odwrotnych odległości ok. 0,10.

Tabela 1. Zmienne objaśniające wyjściowego modelu regresji wielorakiej

Nr	Objaśnienie zmiennej
x1	% osób z wyższym wykształceniem w populacji
x2	% osób z wykształceniem podstawowym lub niższym w populacji
x3	współczynnik skolaryzacji netto dla szkół podstawowych
x4	% dzieci poniżej 24 lat pozostających na utrzymaniu w populacji
x5	liczba rozwodów na 1000 osób
x6	liczba rodzin z trójką i więcej dzieci na 100 osób
x7	odsetek dzieci w wieku 3-5 lat w przedszkolach
x8	liczba żłobków na 1000 dzieci w wieku 0-4 lat
x9	odsetek osób zagrożonych w pracy
x10	liczba zarejestrowanych regon na 10tys. ludności
x11	odległość od zachodniej granicy
x12	odległość w km od miasta wojewódzkiego



Rys. 1. Wizualizacja kartograficzna wartości różnych typów macierzy wag dla powiatu Warszawa: A – macierz sąsiedztwa, B – macierz trzech najbliższych sąsiadów, C – macierz sąsiadów w promieniu 50 km

Widać więc, że wybór sposobu modelowania sąsiedztwa może mieć ogromne znaczenie dla wyników analiz, z drugiej jednak strony większość (4 z 5) sprawdzanych macierzy wag wskazują na nielosowy, przestrzenny rozkład zmiennej, co pozwala przypuszczać, że klasyczny model regresji dla tej zmiennej może okazać się niewystarczający. Wygenerowany wykres punktowy Morana pokazuje jednostki sąsiadujące z tymi o podobnych (ćwiartki H-H i L-L) oraz różnych (H-L i L-H) wartościach zmiennej (rys. 2B). Wizualizacja przynależności do tych ćwiartek na mapie wykazała wyraźne wykazanie skupisk wartości niskich, wysokich oraz stref przejściowych (rys. 2C).

Kolejnym krokiem analizy było zbudowanie modeli regresji. Należy zaznaczyć, że pełna interpretacja osiąganych wyników wymagałaby współpracy socjologa lub/i ekonomisty. Niniejszy przykład ma zaś jedynie pokazać rozwiązania analityczne możliwe do zaimplementowania w docelowym portalu, a próby interpretacji wyników są niezwykle uproszczone.

Jako pierwsze zbudowano klasyczne modele regresji wielorakiej oparte na założeniu o 12 zmiennych objaśniających, a następnie zawężono je do zmiennych wykazujących najwyższą istotność. Czynnikami istotnie wpływającymi na bezrobocie w tym modelu okazały się być: procentowy udział osób z wyższym wykształceniem w populacji, udział w populacji dzieci do lat 24 pozostających na utrzymaniu (rys. 4A), liczba rozwodów przypadająca na 1000 osób (rys. 4D) (wzrost tych czynników zwiększał bezrobocie) a także odsetek dzieci w wieku 3-5 lat objętych edukacją przedszkolną (rys. 4B) oraz odsetek osób zagrożonych w pracy (rys. 4E) (ujemna korelacja). Dodatkowo, w przypadku kobiet znaczenie miała także odległość od zachodniej granicy (rys. 4C) (im dalej tym mniejsze bezrobocie), a w przypadku mężczyzn udział w populacji osób z wykształceniem podstawowym i niższym (rys. 4F) (im więcej takich osób, tym bezrobocie bardziej rośnie).

Wizualna analiza przestrzennego rozmieszczenia reszt z modeli regresji (różnic pomiędzy wartością pomierzoną a wartością estymowaną z modelu; rys. 3A) pozwala zauważyć występowanie pewnych skupisk reszt, zarówno silnie dodatnich, jak i silnie ujemnych. Potwierdzają to także wartości statystyki I Morana obliczonej dla residuów modeli, które wahają się w okolicach 0,30 dla 3 pierwszych modeli, 0,25 dla modelu uwzględniającego sąsiadów w odległości d=50km. Jedynie macierz odwrotnej odległości wydaje się wskazywać niemal równomierny rozkład reszt (ok. 0,04). Skupienia reszt są też widoczne na wykresie punktowym Morana (rys. 3B), a ich rozkład przestrzenny ukazuje rysunek 3C. Na rysunku 3 przedstawiono jedynie wyniki dla modelu regresji bezrobocia ogółem, tendencje modeli bezrobocia kobiet i bezrobocia mężczyzn, jeśli chodzi o nierównomierność rozkładu przestrzennego, są jednak zbliżone.

W związku z brakiem losowości przestrzennej reszt z modeli klasycznych, uzasadnione wydaje się być modelowanie regresji z uwzględnieniem przestrzennego sąsiedztwa jednostek terytorialnych. Ekonometria przestrzenna wyróżnia kilka typów modelowania przestrzennego oraz ich modyfikacje. Najbardziej popularnym jest model Cliffa i Orda (Witkowski, 2010), którego dwa szczególne typy wykorzystano w niniejszym badaniu:

- model typu *spatial lag* (opóźnienia przestrzennego)
- model typu *spatial error* (błędu przestrzennego).

Model *spatial lag* zakłada, że na wartość zmiennej objaśnianej w rozpatrywanej jednostce mają wpływ nie tylko zmienne objaśniające dla tej jednostki, ale także wartość jaką przyjmuje zmienna objaśniana dla jednostek sąsiednich (przy czym stopień sąsiedztwa – bliskości zależy od sposobu określenia macierzy wag). W badanym przypadku oznacza to, że na bezrobocie danego powiatu wpływ mają nie tylko wytypowane w badaniu zmienne objaśniające, ale także bezrobocie w powiatach sąsiednich. Model *spatial error* zakłada zaś, poza wpływem

zmiennych objaśniających, wpływ wartości składnika losowego modelu dla sąsiednich jednostek na wartość zmiennej zależnej w danej jednostce.

Przed przystąpieniem do modelowania dokonano jednak oceny *a priori* modeli typu *error* i *lag* z różnymi macierzami wag. Spośród testowanych modeli regresji wybrano te najbardziej wiarygodne statystycznie (eliminując modele oparte na macierzy odwrotnej odległości). Ograniczono też liczbę testowanych modeli, testując z każdą z pozostałych macierzy wag tylko ten model (*error* bądź *lag*), który okazał się bardziej wiarygodny. Dla każdego z testowanych modeli obliczono też statystykę I Morana dla reszt, modelując sąsiedztwo w ten sam sposób jak w modelu, którego rozkład reszt sprawdzano. W większości przypadków odnotowano wyraźny spadek tej statystyki, co oznacza, że wyeliminowano lub znacząco zmniejszono nierównomierność rozkładu przestrzennego reszt.

Znaczące zmniejszenie wartości statystyki I Morana było zdecydowanie największe dla modelu opóźnienia przestrzennego *(lag)* uwzględniającego macierz sąsiedztwa standaryzowaną wierszami, dlatego też wyniki tego modelowania przedstawiono na rysunku 5. Wzrost równomierności rozkładu przestrzennego reszt widać też na wykresach punktowych Morana (rys. 5D-F). Warto jednak zwrócić również uwagę na fakt spadku wartości bezwzględnych reszt z regresji w porównaniu z klasycznym modelem, nieuwzględniającym sąsiedztwa. Oznacza to, że modele uwzględniające sąsiedztwo lepiej tłumaczą badane zjawiska (bezrobocie). Zmniejszenie rozrzutu reszt widoczne jest w przypadku wszystkich modeli przestrzennych. Dotyczy to także modelu prezentowanego na rysunku 5. Na rysunkach 3A oraz 5A,B,C zastosowano tę samą skalę kolorystyczną przyjmując wartości przedziałów do 1, do 2 i powyżej dwóch odchyleń standardowych pierwszego modelu (odcienie czerwieni to reszty dodatnie, odcienie niebieskiego – reszty ujemne).

Także zmienne istotne w modelu zmieniają się w zależności od wariantu: płci, macierzy wag i rodzaju modelu. Jedynie dwie zmienne objaśniające pozostają zawsze istotne (a kierunek ich oddziaływania nie zmienia się): procent jaki w populacji stanowią dzieci do lat 24 pozostające na utrzymaniu oraz odsetek dzieci w wieku 3 do 5 lat objętych edukacją przedszkolną. Niemal zawsze istotne znaczenie mają też liczba rozwodów na 1000 osób oraz odsetek osób zagrożonych w pracy (każda z tych zmiennych jest eliminowana jedynie z jednego z modeli dla bezrobocia mężczyzn). Dodatkowo, zawsze przy modelowaniu bezrobocia mężczyzn, na znaczeniu zyskuje udział osób z wykształceniem podstawowym lub niższym, przyczyniając się do wzrostu bezrobocia w tej grupie (czynnik ten pojawia się także dla niektórych modeli bezrobocia ogółem). W dwóch z czterech modeli przestrzennych dotyczących bezrobocia kobiet istotna okazuje się zaś odległość od zachodniej granicy, która rosnąc przyczynia się do spadku bezrobocia w tej grupie.

Rozkłady przestrzenne zmiennych, które wykazują istotny wpływ na bezrobocie w Polsce zilustrowano na rysunku 4. Tak jak wspomniano na wstępie bardziej precyzyjna analiza znaczenia tych czynników wymagałaby współpracy socjologa lub/i ekonomisty. Wydaje się jednak, że kierunek ich działania na zmienną objaśnianą (bezrobocie) jest zgodny z oczekiwaniami i intuicją. Procentowy udział w populacji dzieci do lat 24 na utrzymaniu zwiększa bezrobocie, bo z jednej strony może powodować konieczność opieki, która uniemożliwia podjęcie pracy zawodowej, z drugiej zaś, niepracująca młodzież po zakończeniu nauki sama staje się bezrobotna, zwiększając stopę bezrobocia w regionie. Zwiększanie się bezrobocia wraz ze względną liczbą rozwodów można tłumaczyć np. zwiększonymi obowiązkami związanymi z gospodarstwem domowym oraz skutkami emocjonalnymi rozwodu, które utrudniają znalezienie, bądź utrzymanie, pracy. Większy odsetek dzieci objętych edukacją przedszkolną pozwala z kolei na większą

aktywność zawodową i skutkuje zmniejszeniem stopy bezrobocia. Stopę bezrobocia zmniejsza także odsetek osób pracujących w warunkach zagrożenia związanego ze środowiskiem pracy, co można tłumaczyć tym, że takie warunki wynikają zwykle ze specyfiki działających na danym terenie przedsiębiorstw, które jednak mogą być znaczącym pracodawcą w regionie. Dodatni wpływ rosnącego odsetka osób z najniższym wykształceniem na poziom bezrobocia mężczyzn, może oznaczać, że rynek pracy dla takich osób jest w dużej mierze nasycony, a co za tym idzie zwiększenie ich udziału w społeczeństwie zwiększa poziom bezrobocia. Z kolei spadek bezrobocia kobiet, wraz z oddalaniem się od zachodniej granicy, tłumaczyć można kwestiami związanymi z emigracją zarobkową.

Grupowanie metodą *k*-średnich

Innym zadaniem, na którym mógłby zależeć użytkownikowi portalu geostatystycznego jest klasyfikacja (grupowanie) jednostek administracyjnych w grupy homogeniczne pod względem wybranych przez niego cech. Aby zrealizować to zadanie, musi mieć do dyspozycji odpowiedni algorytm klasyfikacyjny bądź grupujący. Poniżej przedstawione zostanie grupowanie metodą *k*-średnich (ang. *k-means clustering algorithm*), które jest jednym z algorytmów analizy skupień (ang. *cluster analysis*). Istotą analizy skupień jest pogrupowanie przypadków (w omawianym przykładzie będą do powiaty) w taki sposób, aby przypadki należące do tej samej grupy charakteryzowały się jak największym stopniem podobieństwa, przy równoczesnym jak najmniejszym stopniu podobieństwa z przypadkami sklasyfikowanymi w innych grupach. Istotą analizy skupień, zwaną również klasyfikacją bez nadzoru jest fakt, iż charakter wynikowych klas (grup) nie jest w żaden sposób definiowany *a priori* przed wykonaniem analizy. W analizie wykorzystywany jest tylko zbiór wektorów wejściowych (zmiennych objaśniających), przy braku wektorów wyjściowych (zmiennych objaśnianych). Jest to więc taki rodzaj analizy, który odkrywa pewną wiedzę ukrytą w danych, a więc jest techniką deskrypcyjnego drążenia danych (ang. *descriptive data mining*) (Kantardzic, 2003).

W przypadku algorytmu *k*-średnich, użytkownik definiuje wynikową liczbę klas (grup), a następnie algorytm identyfikuje tyle skupień przypadków, ile założył użytkownik. Istotą tego algorytmu jest fakt wykorzystania jako miary podobieństwa przypadków odległości (zazwyczaj euklidesowej) w wielowymiarowej przestrzeni, w której wymiarami są wybrane przez użytkownika cechy o charakterze ilościowym (Hartigan, Wong, 1979). Idea algorytmu *k*-średnich jest stosunkowo prosta i ma charakter iteracyjny. Polega na przypisaniu na podstawie kryterium najmniejszej odległości wszystkich wektorów wejściowych do centroidów każdej grupy (przy czym początkowe centroidy wyznaczane są w sposób mniej lub bardziej losowy), a następnie ponownym obliczeniu centroidów na podstawie przydzielonych do nich wektorów wejściowych. Te dwa kroki wykonywane są w określonej przez użytkownika liczbie iteracji. Zaletą algorytmu *k*-średnich jest jego prostota i szybkość, co ma niebagatelne znaczenie w kontekście jego ewentualnej implementacji w portalu statystycznym.

Analiza skupień metodą *k*-średnich wykorzystana zostanie do pokazania, w jaki sposób użytkownik portalu geostatystycznego może wykorzystać określone dane do sklasyfikowania powiatów dla obszaru całego kraju w grupy pod względem sytuacji społecznej, ze szczególnym uwzględnieniem rynku pracy, profilu rodzin oraz dostępności palcówek edukacyjnych. Zmiennymi objaśniającymi, a więc wymiarami będzie część danych wykorzystywanych w poprzedniej analizie: procent zarejestrowanych bezrobotnych, odsetek osób zagrożonych w

pracy, procent osób z wykształceniem podstawowym lub niższym, liczba rozwodów na 1000 osób, współczynnik skolaryzacji netto dla szkół podstawowych, odsetek dzieci w wieku 3-5 lat w przedszkolach, liczba żłobków na 1000 dzieci w wieku 0-4 lat, procent dzieci poniżej 24 lat pozostających na utrzymaniu rodziców oraz liczba rodzin z trójką dzieci lub więcej na 100 osób. Prócz stopy bezrobocia, która stanowiła zmienną objaśnianą w poprzedniej analizie, są to zmienne od x2 do x9 (tab. 1). Mamy więc do czynienia łącznie z dziewięcioma zmiennymi objaśniającymi. Na rysunku 6 przedstawiono wyniki analiz dla dwóch różnych zdefiniowanych liczb skupień. Rysunek 6A ilustruje podział powiatów na trzy grupy, natomiast rysunek 6B na sześć. W pierwszym przypadku zaobserwować można wyraźny podział na powiaty o charakterze miejskim bądź wchodzące w skład aglomeracji (kolor czerwony), powiaty zlokalizowane w zachodniej części kraju (kolor niebieski) oraz powiaty zlokalizowane we wschodniej części kraju (kolor żółty). Jednakże rodzi się pytanie, czy trzy skupienia są wystarczającym podziałem, czy może na podstawie tych danych nie dałoby się wydobyć więcej wiedzy o wzajemnym podobieństwie powiatów i ich przestrzennym rozmieszczeniu. Z drugiej strony, analizując drugi przypadek, również można zaobserwować przestrzenne uwarunkowanie podziału powiatów na sześć grup, jednak nie ma pewności, czy pewne grupy nie zostały utworzone sztucznie (np. jedno z naturalnych skupień zostało podzielone na dwa) tylko dlatego, że użytkownik ustalił taką a nie inną liczbę grup wynikowych.

Dlatego też – mimo zalet jakimi charakteryzuje się grupowanie metodą *k*-średnich, w szczególności szybkiego i nieskomplikowanego działania – niesie ona ze sobą wadę, jaką jest wymóg określenia z góry liczby skupień (grup). W praktyce użytkownik nie ma wiedzy na ile naturalnych skupień dzielą się przypadki w zależności od wybranych zmiennych objaśniających. Stoi więc przed problemem zdefiniowania optymalnej liczby klas (Koronacki, Ćwik, 2008). Z pomocą może przyjść algorytm, który na podstawie danych samodzielnie proponowałby liczbę skupień. Przykładem takiego algorytmu jest v-krotna ocena krzyżowa (ang. *v-fold cross-validation*), a ściślej jej modyfikacja przystosowana do analizy skupień (Tibshirani, Walther, 2005). Jej istotą jest podzielenie, najczęściej w sposób losowy, wszystkich obserwacji na podzbiory uczące oraz testowe. Następnie określona analiza, a więc w omawianym przypadku analiza skupień metodą *k*-średnich, wykonywana jest osobno na przypadkach z podzbioru uczącego i testowego (wyznaczane są centroidy skupień). W kolejnym kroku przypadki z podzbioru testowego porównywane są z centroidami wyliczonymi na podstawie przypadków z podzbioru uczącego. Procedura ta powtarzana jest dla różnej liczby skupień (których zakres określa użytkownik), a optymalna jest wyznaczana na podstawie najmniejszej średniej odległości przypadków próby testowej od centroidów wyznaczonych przez próbę uczącą. O ile metoda ta wymaga stosunkowo dużej liczby obliczeń (analiza skupień wykonywana jest wielokrotnie, ponadto wykonywane muszą być niezbędne porównania), obliczenia te nie charakteryzują się wysokim stopniem złożoności i wydaje się, że mogą być z powodzeniem zastosowane w portalu geostatystycznym.

Wyznaczenie optymalnej liczby skupień metodą *v*-krotnej oceny krzyżowej zostało wykonane dla omawianego wyżej przykładu. W wyniku przeprowadzonej analizy okazało się, że powiaty – ze względu na wymienione wyżej kryteria (zmienna objaśniające) – w sposób najbardziej naturalny dzielą się na cztery grupy. Ich rozkład przestrzenny zilustrowano na rysunku 7. Prócz grup zidentyfikowanych przy trzech grupach (miasta, zachodnia i wschodnia część kraju) zaobserwować można jeszcze grupę powiatów otaczających duże aglomeracje miejskie (kolor żółty).

Warto pamiętać, że prócz samej klasyfikacji, a więc przypisania każdego powiatu do określonej grupy, w wyniku przeprowadzenia analizy skupień metodą *k*-średnich użytkownik otrzymuje znacznie więcej informacji, jak np. standaryzowane odległości pomiędzy centroidami skupień, średnie arytmetyczne wartości wszystkich zmiennych objaśniających dla poszczególnych skupień, czy odległości poszczególnych przypadków od centroidów skupień, do których zostały zaklasyfikowane. Wszystkie te informacje mogą być przedstawione użytkownikowi portalu geostatystycznego w postaci tabelarycznej, jakkolwiek można się pokusić o próbę kartograficznej prezentacji niektórych zjawisk. Poniżej zaproponowano wizualizację odległości przypadków od centroidów skupień. Do tego celu wykorzystano dwie zmienne wizualne: kolor (jak w poprzednich przykładach – do różnicowania powiatów ze względu na przyporządkowanie do odpowiednich grup) oraz jasność (do różnicowania powiatów ze względu na odległość powiatów od centroidów grup). Przykład takiej wizualizacji pokazano na rysunku 8.

Odległości od centroidów podzielono na trzy klasy, w ramach których powiaty zlokalizowane najbliżej centroidów przedstawiono najciemniej, natomiast te, które zlokalizowane są najdalej centroidów – najjaśniej. W ten sposób użytkownik portalu geostatystycznego, mając do dyspozycji surowe dane oraz odpowiedni algorytm grupujący zaimplementowany w portalu, ma możliwość uzyskania wiedzy na temat zarówno podziału powiatów w zależności od wybranych danych, jak również stopnia przynależności powiatów do poszczególnych grup.

Wykorzystane narzędzia analityczne

Do realizacji zadań analiz statystycznych może służyć wiele aplikacji pozwalających na pracę z danymi statystycznymi. Większość z nich używa jednak klasycznych metod statystycznych, które nie uwzględniają przestrzennego charakteru danych. Narzędziem, które uwzględnia ten aspekt jest pakiet R, a konkretnie jego biblioteka predefiniowana do celów ekonometrii przestrzennej – *spdep*. Język R jest językiem programowania oraz środowiskiem obliczeń statystycznych i wizualizacji ich wyników, działającym na licencji GNU (zapewniająca jego darmowość i możliwość wprowadzania własnych modyfikacji). Jego niewątpliwą zaletą jest otwarta forma, możliwość darmowego używania, ale także możliwość tworzenia spersonalizowanych pakietów i bibliotek. Ponadto możliwe jest wykorzystywanie funkcji R z poziomu innych języków, co może okazać się przydatne w kontekście próby implementacji tych rozwiązań w ramach atlasu statystycznego, np. w postaci usług sieciowych i ich integracji z pozostałymi elementami atlasu.

Szerokie możliwości pakietu R zostały w niniejszej pracy wykorzystane w przykładzie analizy regresji wielorakiej. Przeprowadzona ona została przy wykorzystaniu bibliotek:

- *spdep (spatial dependence)* – umożliwiającej modelowanie zależności przestrzennych,
- *maptools* – umożliwiającej pracę z danymi przestrzennymi w formatach ESRI .shp,
- *sp* – zawierającej klasy i metody dla danych przestrzennych, w tym umożliwiające ich wizualizację w formie map,
- *RColorBrewer* – zawierającej palety do rysowania map oraz *classInt* umożliwiającej podział zmiennych na przedziały klasowe (Kopczewska i in., 2009).

Należy zauważyć, że pakiety do wizualizacji mają w środowisku R dość ograniczone możliwości jeżeli chodzi o redakcję kartograficzną. Z tego względu docelowo wizualizacja powinna być raczej realizowana w środowisku bazy danych przestrzennych przez narzędzia GIS, przy wykorzystaniu wyników uzyskanych dzięki funkcjom języka R.

Analizę skupień metodą *k*-średnich wykonano w środowisku Statistica Data Miner. Jego zaletą – prócz prostej i intuicyjnej obsługi poprzez okna dialogowe – jest możliwość definiowania skryptów i makr w języku Visual Basic. Makra takie mogą na przykład odtwarzać zapisy całych sesji analitycznych, na które składają się powiązane analizy statystyczne korzystające wzajemnie ze swych wyników. Makra mogą być uruchamiane również w innych narzędziach wykorzystujących środowisko programistyczne Visual Basic, w szczególności w oprogramowaniu GIS (np. ArcGIS, czy MapInfo). Daje to możliwość zdefiniowania zależności przestrzennych (np. analiz sąsiedztwa) w analizach statystycznych oraz bezpośredniego wykorzystania dostępnych w tych narzędziach zaawansowanych metod wizualizacji kartograficznej do prezentacji wyników analiz.

Znaczącą przeszkodą w wykorzystaniu środowiska Statistica Data Miner jest jego komercyjny charakter, a w konsekwencji potencjalnych trudności natury prawnej i organizacyjnej przy implementacji funkcji analitycznych tego oprogramowania w portalu geostatystycznym.

Podsumowanie i perspektywy

Zaproponowane przykłady analiz z pewnością nie wyczerpują bogatych możliwości, jakie Główny Urząd Statystyczny mógłby zaproponować odbiorcom swoich danych. Co więcej, nawet opisane analizy mogą być rozwijane i ulepszane, np. macierze wag w modelu regresji wielorakiej, w zależności od modelowanego zjawiska, mogą przybierać różne formy, uzależnione nie tylko od geometrii, ale również od wartości ekonomicznych łączących jednostki terytorialne, czy też od obecności infrastruktury, takiej jak sieć drogowa. Zarówno jednak opisane w niniejszym artykule, jak i zaledwie zasugerowane możliwości analityczne wymagają rozwiązania problemów natury techniczno-organizacyjnej związanej z ich wdrożeniem. Kluczowy będzie tu więc wybór oprogramowania realizującego określone funkcje oraz sposób jego implementacji w podstawowym interfejsie użytkownika, dostępnym przez stronę internetową portalu udostępniającego dane. Wydaje się, że najbardziej obiecującym środowiskiem jest pakiet R, gdyż charakteryzuje się stosunkowo dużym potencjałem implementacyjnym w ramach portalu geostatystycznego oraz brakiem ograniczeń natury prawno-organizacyjnej z uwagi na jego otwarty charakter. Niezbędnym jednak krokiem byłoby utworzenie i zaimplementowanie w portalu graficznego interfejsu użytkownika (GUI), który w intuicyjny i interaktywny sposób pozwalałby użytkownikowi na manipulowanie parametrami i – poprzez automatyczne uruchamianie odpowiednich kodów języka R – wykonywanie udostępnionych analiz. Na korzyść środowiska R działa również znaczna dostępność bibliotek i pakietów. Pozwalają one na wykonywanie zaawansowanych analiz statystycznych, w szczególności z zakresu *data mining*, np. implementację drzew decyzyjnych i regresyjnych (pakiety *tree, rpart, randomForest*), reguł asocjacyjnych (pakiet *arules*), czy sztucznych sieci neuronowych (np. pakiety *nnet, neural, kohonen*), które mogą realizować zarówno zadania regresyjne, jak również klasyfikacyjne w postaci analizy skupień (samoorganizująca sieć Kohonena).

W opinii autorów największym wyzwaniem, w obliczu szerokiej dostępności rozmaitych narzędzi analitycznych, jest z jednej strony odpowiedni ich wybór i implementacja w sposób zapewniający mniej lub bardziej zaawansowanym użytkownikom portalu geostatystycznego możliwość pozyskania na podstawie danych użytecznej wiedzy niedostępnej *explicite*, z drugiej zaś odpowiednie wykorzystanie metod wizualizacji, aby przyczyniały się one do powstania „kartograficznej wartości dodanej". Na uwagę zasługuje fakt, że realizacja powyższych

zadań wpisywałaby się w ideę infrastruktury wiedzy przestrzennej (ang. *spatial knowledge infrastructure*) (Iwaniak, 2011).

Literatura

Fiedukowicz A., Gąsiorowski J., Kowalski P. J., Olszewski R., Pillich-Kolipińska A., 2012: The statistical geoportal and the cartographic "added value"– creation of the spatial knowledge infrastructure. *Geodesy and Cartography*, Vol. 61, No. 1, zaakceptowany w redakcji.

Hartigan J. A., Wong M. A., 1979: A K-Means Clustering Algorithm. *Applied Statistics* Vol. 28, No. 1, 100-108.

Iwaniak A., 2011: Inteligentny geoportal, III Konferencja z cyklu „Wolne oprogramowanie w geoinformatyce", Wrocław.

Kantardzic M., 2003: Data mining: Concepts, Models, Methods and Algoritms. John Wiley & Sons, New York.

Kopczewska K., Kopczewski T., Wójcik P., 2009: Metody ilościowe w R. Aplikacje ekonomiczne i finansowe, CeDeWu.pl, Warszawa.

Koronacki J., Ćwik J., 2008: Statystyczne systemy uczące się. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Tibshirani R., Walther G., 2005: Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics*, Vol. 14, Issue 3, 511-528.

Witkowski B., 2010: Zastosowanie metod ekonometrii przestrzennej. Prace Instytutu Ekonomii, Szkoła Główna Handlowa, Kolegium Analiz Ekonomicznych.

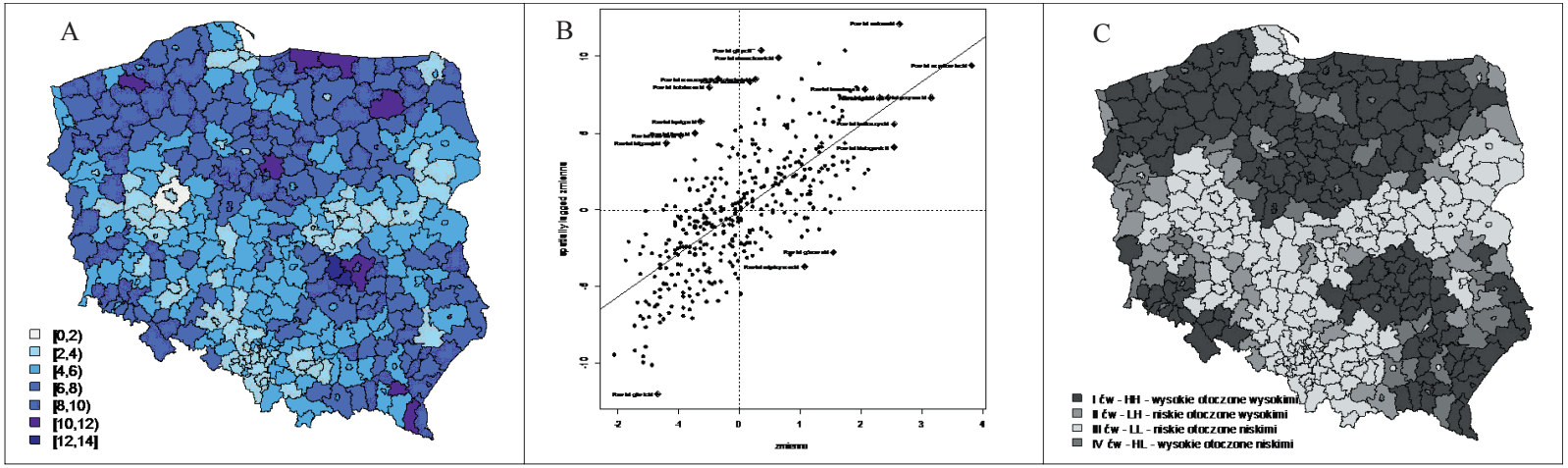
Abstract

In the face of ubiquitous data availability, it becomes a challenge to process data in such a way that allows to gain useful knowledge based on the analysis of source information. The aim of the authors was to discuss the use of advanced spatial data mining techniques to data collected by the Central Statistical Office interviewers in two censuses: Agricultural Census and National Census of Population and Housing and of data enrichment. Using this approach, which is a modern equivalent of the cartographic research method, allows not only to discover spatial patterns and regularities, but above all to reveal some knowledge contained in the database. Taking into account the scope and level of detail (the lowest available level of aggregation by the Central Statistical Office are communes) in the data obtained in the two censuses a number of relationships between data may be expected – both intuitive, requiring only statistical confirmation and cartographic visualization, as well as more complex and „hidden" in the data. Identification, analysis and visualization of these dependencies will allow to gain additional knowledge that can be used to develop national spatial planning policy.

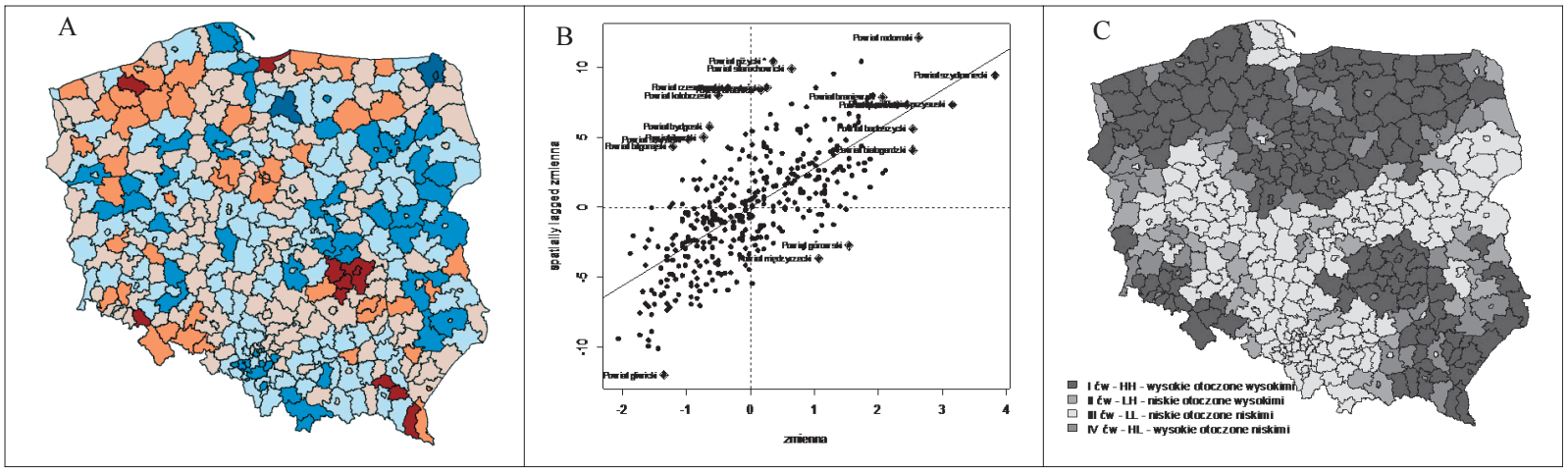
The authors presented proposals of either statistical analyses or cartographic presentation of the results of analyses, which may be useful in achieving objectives set by the statistical geoportal. The article describes two examples of such analyses. The first one is based on multiple regression analysis taking into account the neighborhood relationships. The model describing the relationships between variables gathered for the administrative units was constructed in the result of the analysis. The second example described in the article is a cluster analysis performed by the k-means algorithm. This method was used for statistical classification of administrative units allowing to extract homogeneous groups with regard to multi-factor similarity determined in a non-metric feature space.

mgr inż. Anna Fiedukowicz
a.fiedukowicz@gik.pw.edu.pl

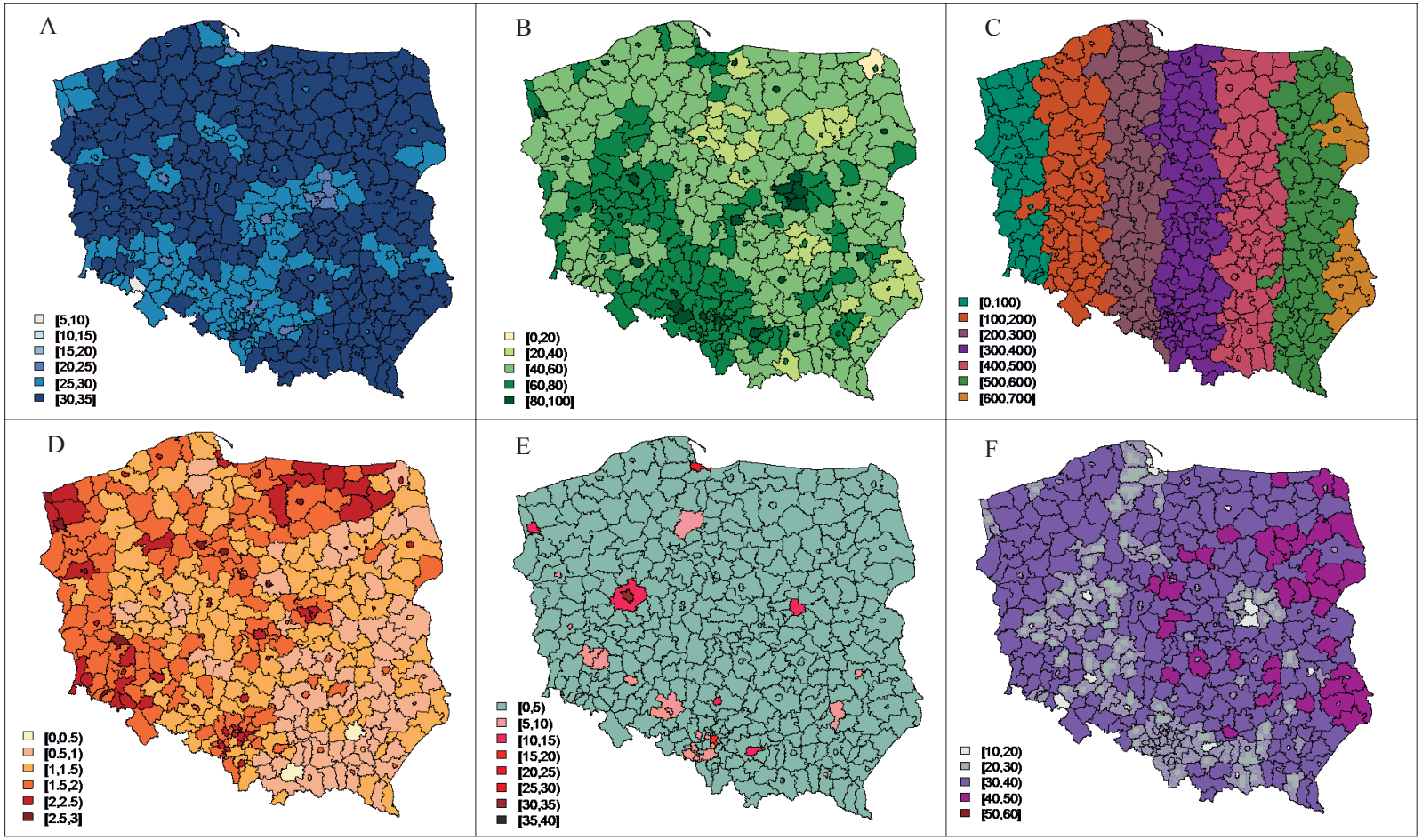
mgr inż. Jędrzej Gąsiorowski
jedrzej.gasiorowski@igik.edu.pl



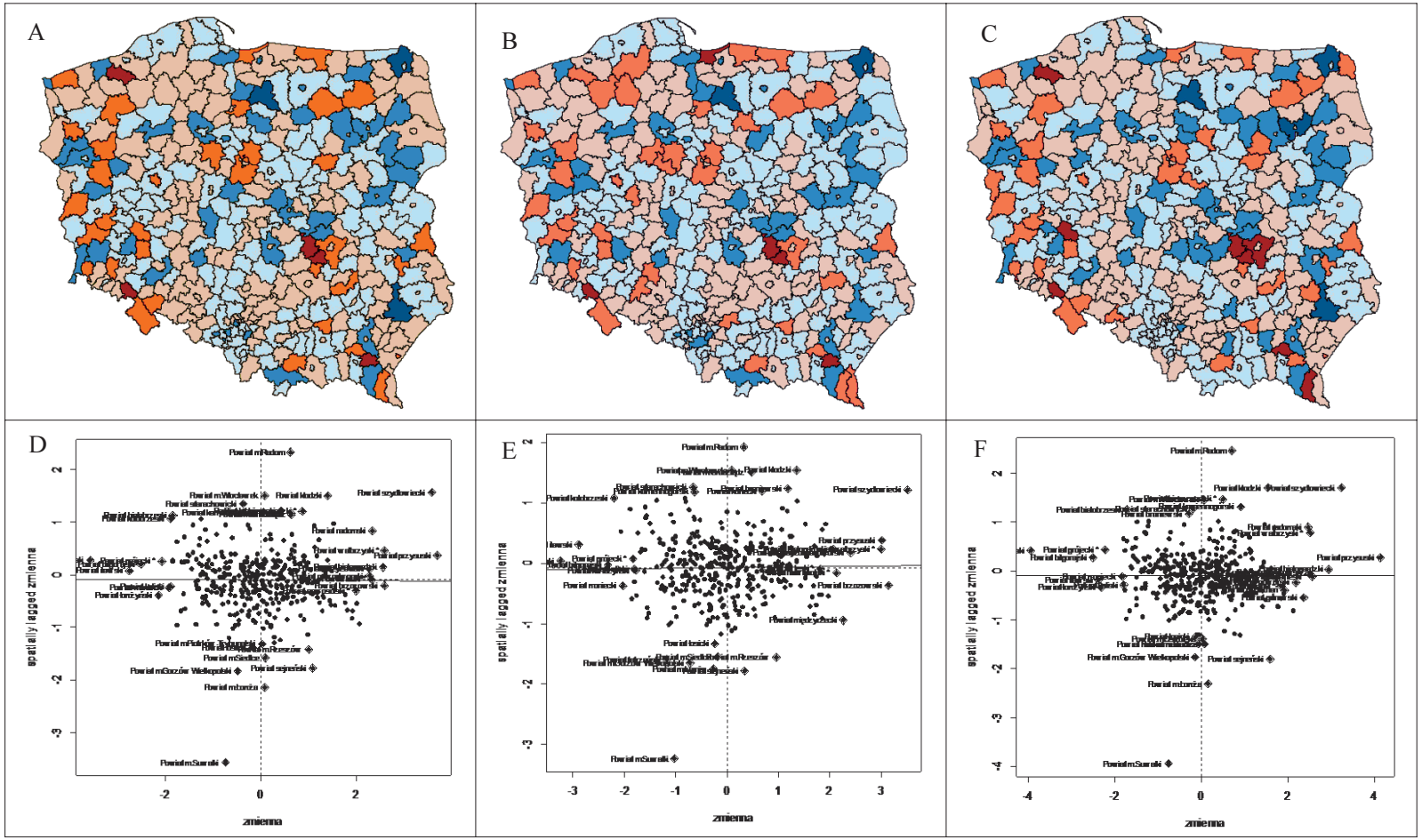
Rys. 2. Analiza zmiennej objaśnianej: A – wizualizacja kartograficzna rozkładu przestrzennego bezrobocia ogółem, B – wykres punktowy Morana dla bezrobocia ogółem, C – wizualizacja przedstawiająca obszary skupienia się wysokich i niskich wartości bezrobocia ogółem



Rys. 3. Analiza klasycznego modelu regresji dla bezrobocia ogółem opartego na zmiennych x1, x4, x5, x7, x9: A – wizualizacja kartograficzna reszt z regresji, B – wykres punktowy Morana dla reszt, C – skupienia wysokich i niskich reszt z regresji

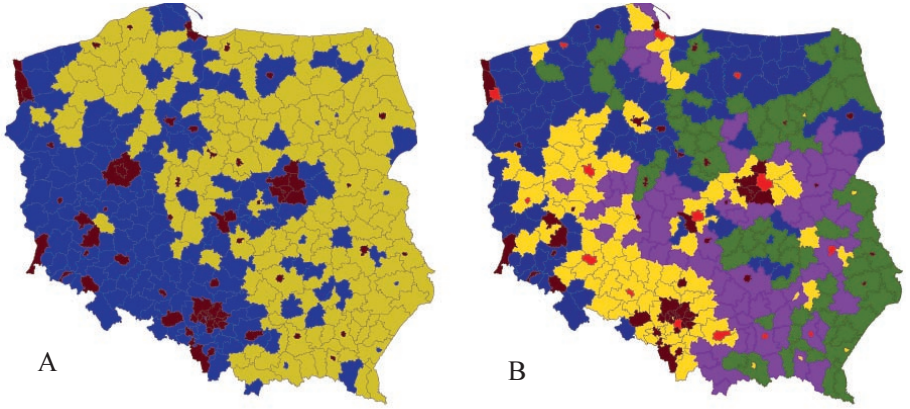


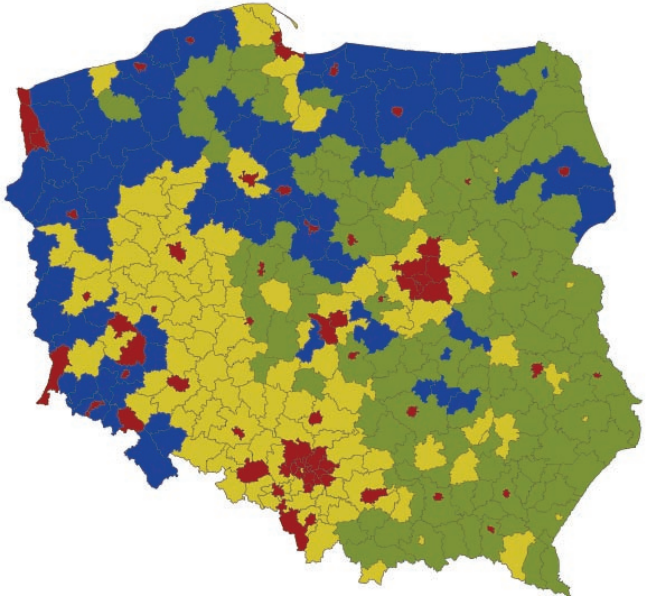
Rys. 4. Zmienne objaśniające istotne dla modelowania bezrobocia (w nawiasach kierunek zależności): A – udział w populacji dzieci do lat 24 pozostających na utrzymaniu (+), B – odsetek dzieci w wieku 3-5 lat objętych edukacją przedszkolną (-), C – odległość od zachodniej granicy (-), D – liczba rozwodów przypadająca na 1000 osób (+), E – odsetek osób zagrożonych w pracy (-), F – udział w populacji osób z wykształceniem podstawowym i niższym (+). Czynniki A, B, D, E wpływają na bezrobocie ogółem, czynnik C na bezrobocie kobiet, D na bezrobocie mężczyzn



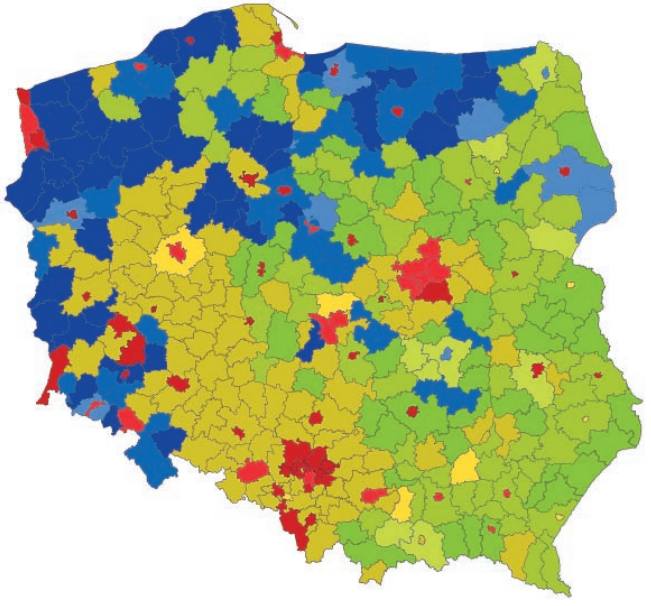
Rys. 5. Reszty z regresji (A, B, C) oraz wykresy punktowe Morana (D, E, F) dla modeli przestrzennych typu lag wykorzystujących macierz sąsiedztwa standaryzowaną wierszami; A, D – bezrobocie ogółem, B, E – bezrobocie kobiet, C, F – bezrobocie mężczyzn

Rys. 6. Powiaty sklasyfikowane na podstawie wybranych zmiennych objaśniających metodą *k*-średnich w: A – trzy grupy, B – sześć grup





Rys. 7. Powiaty sklasyfikowane na podstawie wybranych zmiennych objaśniających metodą *k*-średnich w optymalnej liczbie czterech grup, wyznaczonej metodą v-krotnej oceny krzyżowej



Rys. 8. Powiaty sklasyfikowane na podstawie wybranych zmiennych objaśniających metodą *k*-średnich w cztery grupy, z informacją o stopniu przynależności do poszczególnych grup (odległości od centroidów)