Sektor mikro, małych i średnich przedsiębiorstw –

opracowanie metodyki wykorzystania metod ilościowych i jakościowych *spatial data mining* do analizy danych przestrzennych MŚP w województwie kujawsko-pomorskim

Projekt „Usytuowanie na poziomie samorządów lokalnych instrumentów wsparcia dla MŚP, działających w oparciu o model wielopoziomowego zarządzania regionem”
realizowany przez konsorcjum w składzie: Województwo Kujawsko-Pomorskie, Szkoła Główna Handlowa w Warszawie i Uniwersytet Mikołaja Kopernika w Toruniu";
finansowany z Narodowego Centrum Badań i Rozwoju
w ramach Programu Strategicznego Gospostrateg –
„Społeczny i Gospodarczy Rozwój Polski
w warunkach globalizujących się rynków”
Nr umowy GOSPOSTRATEG1/385453/3/NCBR/2018

Spis treści

WPROWADZENIE....3

SPATIAL DATA MINING....6

INTELIGENCJA OBLICZENIOWA....8

WYKORZYSTYWANE METODY....13

KONCENTRACJA PRZESTRZENNA....15

METODA GŁÓWNYCH SKŁADOWYCH....17

METODA PERMUTACJI BERTINA I TYPOGRAMU UHRCZAKA....18

SAMOORGANIZUJĄCA SIEĆ KOHONENA....19

KLASYFIKACJA ROZMYTA....20

DRZEWA KLASYFIKACYJNE TYPU CART....22

LINIOWA REGRESJA WIELOWYMIAROWA (WIELORAKA)....27

REGRESJA WAŻONA GEOGRAFICZNIE (EKONOMETRIA PRZESTRZENNA)....30

REGRESJA NIELINIOWA....31

METODA MARS....32

METODA WEKTORÓW NOŚNYCH....33

DRZEWA REGRESYJNE....35

SZTUCZNE SIECI NEURONOWE....37

Wprowadzenie

Cyfrowa rewolucja technologiczna, jaka dokonała się w ostatnich latach, znacząco ułatwiła dostęp do informacji. Możliwość wymiany doświadczeń naukowych pomiędzy przedstawicielami pozornie odległych dziedzin wiedzy spowodowała gwałtowny rozwój całej nauki. Nauka o geoinformacji, podobnie jak inne nauki interdyscyplinarne, rozwija się zarówno na podstawie własnej bazy metodologicznej i teoretycznej jak również pod wpływem osiągnięć innych dyscyplin. Rozwój metodyczny nauki o geoinformacji odbywa się zwłaszcza pod wpływem rozwoju technologii informatyczno-cyfrowe, lecz także statystyki (zwłaszcza statystyki przestrzennej) oraz metod eksploracyjnej analizy danych (spatial data mining). Postępująca automatyzacja procesu pozyskiwania informacji przestrzennych pozwala na stosunkowo proste i efektywne ich gromadzenie. Zadaniem nauki o geoinformacji jest więc tworzenie lub adaptowanie z innych dziedzin wiedzy nowych metod pozwalających na optymalne wykorzystanie zgromadzonych danych.

Bezpośrednia wizualizacja nieprzetworzonych danych źródłowych może sprawić, iż uzyskany obraz będzie trudny do interpretacji. Istnieje wiele dobrze poznanych metod analizy danych źródłowych, wśród których jedną z istotniejszych jest klasyfikacja. Równie istotne znaczenie ma jednak rozwój metod szeroko rozumianej regresji wielowymiarowej – zarówno wykorzystującej metody liniowe, jak i nieliniowe metody związane z sieciami neuronowymi i systemami wnioskowania rozmytego, jak i metody ekonometrii przestrzennej (regresji ważonej geograficznie).

Przez setki lat ludzie opisywali otaczająca przestrzeń geograficzną posługując się językiem klasycznej kartografii - rysowali mapy, analizowali je i wykorzystywali do orientacji. Współcześnie językiem opisu przestrzeni geograficznej są także (a nawet głóa)nie) bazy danych przestrzennych oraz systemy informacji geograficznej (ang. Geographical Information System - GIS). Umożliwiają one nie tylko gromadzenie, ale i analizowanie danych geograficznych i związanych z nimi atrybutów opisowych.

Istotną rolę w procesie rozwoju metodyki szeroko rozumianych analiz przestrzennych odegrał w połowie XX w. rozwój tzw. kartograficznej metody badań (KMB). W 1947 roku K. A. Saliszczew w artykule „Kartografia, jej przedmiot i niektóre zadania" po raz pierwszy sformułował koncepcję wykorzystania map do badania prawidłowości w rozmieszczeniu zjawisk i ich wzajemnej współzależności. „Istota metody kartograficznej polega na włączeniu do procesu badania rzeczywistości pośredniego ogniwa – mapy geograficznej – jako modelu badanych zjawisk. Mapa występuje przy tym w podwójnej roli: jako narzędzie badania oraz jako przedmiot w postaci modelu,

zastępującego rzeczywiste zjawiska, których bezpośrednie badanie jest niemożliwe lub utrudnione" (Saliszczew, 1998). Koncepcję kartograficznej metody badań rozwinął A. M. Berlant (1986). Podana przez niego zmodyfikowana definicja KMB obejmowała wyznaczanie na podstawie map charakterystyk jakościowych i ilościowych zjawisk i procesów, określanie dynamiki ich zmian w czasie i przestrzeni, współzależności oraz tendencji rozwoju i prognozowanie jego przyszłych stanów. Współcześnie, w związku z upowszechnieniem technik cyfrowego przetwarzania informacji, klasyczna metodyka stosowania KMB jest z reguły określana jako „wzbogacenie danych" (ang. data enrichment) i/lub eksploracyjna analiza danych przestrzennych (ang. spatial data mining) (Olszewski, 2006; Fiedukowicz, Gąsiorowski, Kowalski, Olszewski, Pillich-Kolipińska, 2012).

O celowości stosowania tak rozumianej eksploracyjnej analizy danych przestrzennych i jej wiarygodności jako metody badań naukowych można mówić jedynie wówczas, gdy dane źródłowe (baza danych przestrzennych i/lub mapy analogowe) spełniają określone kryteria dokładnościowe. Na uzyskiwane rezultaty wywiera wpływ zarówno struktura bazy danych, stopień uogólnienia pojęciowego, dokładność geometryczna, jak i aktualność zgromadzonych danych.

Eksploracyjna analiza danych przestrzennych pozwala na wykorzystanie map (a współcześnie również baz danych przestrzennych) nie tylko do wizualizacji obiektów i zjawisk, lecz również do badania i ukazywania ich wzajemnych relacji. Będąca pierwotnie domeną zawodowych kartografów - analiza map stała się czynnością powszechną dla użytkowników Internetu i map cyfrowych oraz baz danych przestrzennych o charakterze ogólnogeograficznym i tematycznym. Korzystając z serwisów mapowych wyznaczamy trasę dojazdu czy oceniamy przydatność inwestycyjną terenu. Jednak prawdziwa siła analiz przestrzennych tkwi w możliwości łącznego wykorzystania informacji geograficznej, algorytmów statystyki przestrzennej, geostatystyki i metod data mining.

Dane topograficzne i tematyczne gromadzone w urzędowych bazach danych Głównego Urzędu Geodezji i Kartografii, Głównego Urzędu Statystycznego czy wielu innych organów centralnych czy też urzędów wojewódzkich warte są tyle, ile abstrahowana na ich podstawie informacja, umożliwiającą akwizycję wiedzy i tworzenie systemów wnioskowania. Dane przestrzenne (zarówno te urzędowe, jak i te gromadzone i udostępniane przez organizacje społeczne i firmy komercyjne) są więc współcześnie źródłem wiedzy o otaczającym świecie. Aby uzyskać tę wiedzę, trzeba jednak umiejętnie przetworzyć dostępne dane. Narzędziem ułatwiającym ten proces akwizycji wiedzy z danych o charakterze przestrzennym jest eksploracyjna analiza danych określana często jako spatial data mining.

Jako że większość zjawisk i procesów związanych z rozwojem przedsiębiorczości w województwie kujawsko-pomorskim ma charakter przestrzenny, a zjawiska te zazwyczaj oddziałują na

siebie tym silniej, im bliżej siebie się znajdują (tzw. „pierwsze prawo geografii" Toblera), w zaproponowanej metodyce uwzględniono przede wszystkim metody analityczne wykorzystujące odniesienie przestrzenne. Podejście to może stanowić kanwę do wszelkiego rodzaju analiz biznesowych, mających za zadanie pozyskanie nowej wiedzy zarówno w sferze przewidywania wartości określonych parametrów (ang. predictive data mining), jak również opisywania rozmaitych prawidłowości (ang. descriptive data mining).

Dla opracowania metodyki wykorzystania metod ilościowych i jakościowych data mining dla analizy geodanych, kluczowe znaczenie ma bezpośrednie lub pośrednie uwzględnienie predykatorów przestrzennych. Zadanie to może być realizowane np. poprzez wykorzystanie narzędzi GIS lub dedykowanych skryptów do wyznaczenie odległości, powierzchni, gęstości danego zjawiska itp. Wyznaczone parametry mogą być wykorzystane jako dodatkowe zmienne objaśniające w uogólnionym modelu regresji.

Kluczowe znaczenie związane z metodami analiz statystyczno-przestrzennych ma:

- modyfikacja metod analiz statystycznych tak, by w efektywny i bezpośredni sposób wykorzystywały one dane o charakterze przestrzennym (takie jak dane w postaci wektorowej),
- kodowanie zależności przestrzennych między obiektami, tak aby z jednej strony mogły być uwzględniane jako zmienne (objaśniające bądź objaśniane) w metodach analiz statystycznych, z drugiej zaś w sposób kompleksowy i wiarygodny opisywały zależności przestrzenne między obiektami,
- wizualizacja wyników analiz wraz z możliwością interaktywnej modyfikacji budowanego modelu poprzez dodanie lub usunięcie zmiennych niezależnych objaśniających rozkład przestrzenny danego zjawiska.

W opracowanej metodyce wykorzystano dane MŚP dla województwa kujawsko-pomorskiego, dane statystyczne GUS, bazy danych ZUS i REGON oraz dane ogólnogeograficzne z Bazy Danych Obiektów Topograficznych BDOT 10k. Dane MŚP (ok. 193 tys. rekordów) zostały poddane procesowi geokodowania, co pozwoliło na ich bezpośrednie wykorzystanie w prowadzonych analizach GIS. Dla wyznaczenia odległości poszczególnych MŚP od dużych miast, zjazdów z autostrady, parków technologicznych itp., a także do obliczenia gęstości sieci drogowej i gęstości zabudowy w poszczególnych gminach Wykonawca opracował autorskie skrypty osadzone w środowisku analitycznym GIS. Uzyskane wyniki obliczeń geometrycznych zostały zapisane jako atrybuty ilościowe w formacie SHP. Co pozwala na ich bezpośrednie wykorzystanie w większości narzędzi GIS>

Spatial Data Mining

Wywodząca się z kartograficznej metody badań metodyka eksploracyjnej analizy danych przestrzennych (ang. spatial data mining) pozwala zarówno na analizę informacji źródłowej – znajdowanie wzorców przestrzennych czy ekstrakcję reguł decyzyjnych, jak i na wnioskowanie prognostyczne. Przetwarzając dane demograficzne o rozmieszczeniu ludności oraz informacje charakteryzujące poziom alokacji i wydatkowania funduszy unijnych w Polsce (Gąsiorowski, Hajkowska, Olszewski, 2015) możemy np. wyznaczyć czynniki wpływające na poziom wykorzystania środków Europejskiego Funduszu Rozwoju Regionalnego, Funduszu Społecznego czy Funduszu Spójności. Analogicznie – dysponując danymi demograficznymi, informacjami podatkowymi, danymi z baz danych REGON i ZUS jesteśmy w stanie stosując metody i algorytmy spatial data mining formułować hipotezy dotyczące przyczyn wyższego/niższego rozwoju przedsiębiorczości w poszczególnych regionach oraz opracowywać modele prognostyczne zmian w tym zakresie.

Proponowane w niniejszym opracowaniu podejście łączy w sobie możliwości pakietów GIS oraz zawansowanych programów statystycznych oraz umożliwia prowadzenie bardzo złożonych analiz. Jednym ze sposobów takiej analizy jest zastosowanie tzw. „drążenia" danych (data mining) i ich „wzbogacania" (data enrichment) do wykrywania prawidłowości, reguł i struktur „ukrytych" w bazie danych. Najważniejszym celem eksploracyjnej analizy danych przestrzennych jest opracowywanie metod umożliwiających „pośrednią" analizę otaczającej przestrzeni - dokonywaną poprzez analizę modelu jakim jest poprawnie zaprojektowana baza danych przestrzennych (Gotlib, Iwaniak, Olszewski, 2007).

Dostępne w systemach informacji geograficznej funkcje przetwarzania i analizy informacji przestrzennej są dalece niewystarczające dla prowadzenia bardziej rozbudowanych badań, np. dotyczących rozwoju przedsiębiorczości w regionie, ponieważ pozwalają jedynie na dokonywanie prostych analiz przestrzennych, takich jak: wybór obiektów według określonej cechy lub wewnątrz utworzonej ekwidystanty, przecinanie warstw itp. Zadaniem eksploracyjnej analizy danych przestrzennych staje się więc poszukiwanie coraz nowszych metod pozwalających na optymalne wykorzystanie informacji zawartych w bazach danych.

Poszukiwanie relacji między zmiennymi (zarówno opisowymi, jak i przestrzennymi), jest podstawowym celem badań naukowych, w tym ekonomicznych. Kluczowe znaczenie ma „odkrywanie" relacji między zmiennymi, które stanowią istotny element otaczającego nas świata.(Litwin i in., 2005). Podchodząc do zagadnienia relacji od strony statystycznej można wyróżnić dwa rodzaje współzależności między zmiennymi: funkcyjną i stochastyczną.

Zależność funkcyjna występuję w sytuacji, gdy zmiana wartości jednej zmiennej powoduje jednoznacznie, ściśle określoną zmianę wartości drugiej zmiennej. W tym przypadku określonej wartości zmiennej objaśniającej X (reprezentowanej często w postaci wektora danych) odpowiada dokładnie jedna wartość drugiej zmiennej objaśnianej Y. Istota zależności stochastycznej polega zaś na tym, że wraz ze zmianą wartości jednej zmiennej następuje zmiana rozkładu prawdopodobieństwa drugiej zmiennej. Szczególnym przypadkiem niniejszej zależności jest zależność korelacyjna (statystyczna), w której określonym wartościom jednej zmiennej odpowiadają ściśle określone średnie wartości drugiej zmiennej.

W definiowaniu podejścia wykorzystującego metody spatial data mining istotne znaczenie ma swego rodzaju synergia pomiędzy „czystą" statystyką, a geoinformacją. Statystyka dzięki swojemu wyrafinowanemu aparatowi obliczeniowemu pozwala wyliczyć niezwykle interesujące wielkości lub ich zestawy, a nauka o geoinformacji umożliwia analizę kontekstu przestrzennego oraz rezentację wyniki obliczeń w sposób dający lepsze pojęcie o ich lokalizacji w przestrzeni i o ich znaczeniu (Berlant, 1973; Boczarow, 1976). Takie wykorzystanie modeli statystycznych i algorytmów uczenia maszynowego wykorzystywanych do analizy danych przestrzennych pozwala na uzyskanie odpowiedzi na pytanie np. jak silne, a jednocześnie skomplikowane mogą być związki przestrzenne pomiędzy zjawiskami (Gotlib, Iwaniak, Olszewski, 2007). Czy kształt gminy ma związek z równomiernością rozkładu zarobków? Czy odległość od siedziby władz wojewódzkich ma związek z wielkością przyznanych dotacji? Czy da się objaśnić rozwój przedsiębiorczości, zwłaszcza zaś MSP odległością od dużych miast? zjazdów z autostrady? Liczbą parków technologicznych w danym regionie? Czy poziom bezrobocia wpływa na wartość dotacji uzyskiwanych z funduszy unijnych i jeśli tak to w jaki sposób? Pytania można mnożyć, a odpowiedzi na nie mogą dać zarówno odpowiednio dostosowane — rozbudowane do właściwego „rozumienia" związków przestrzennych znane ze statystyki algorytmy (korelacji, regresji, analizy skupień, reguły asocjacyjne, etc.), jak również zupełnie nowe algorytmy wykorzystujące metody uczenia maszynowego i spatial data mining – sztuczne sieci neuronowe, metody wnioskowania rozmytego, drzewa klasyfikacyjne i regresyjne itp.

Celem niniejszego opracowania jest zarówno dyskusja na poziomie teoretycznym na temat możliwości i korzyści płynących z połączenia analiz statystycznych i przestrzennych do oceny stopnia rozwoju MSP w województwie kujawsko-pomorskim, jak również przykłady wykorzystania metod statystycznych i algorytmów spatial data mining tak, by mogły one w sposób efektywny uwzględniać zależności przestrzenne.

Inteligencja obliczeniowa

Do grupy algorytmów inteligencji obliczeniowej zalicza się między innymi systemy wnioskowania rozmytego (fuzzy inference systems – FIS), sztuczne sieci neuronowe (artificial neural networks), zbiory przybliżone (rough sets) i drzewa decyzyjne (decision trees).

Do podstawowych problemów rozwiązywanych przez CI należą (Duch, 2005):

- klasyfikacja struktur,
- odkrywanie wiedzy w bazach danych: zrozumienie struktury danych,
- selekcja cech, redukcja wymiarowości problemu,
- inteligentne szukanie z uwzględnieniem semantyki pytania,
- kontrola jakości rozwiązań,
- optymalizacja wielokryterialna, inteligentne wspomaganie decyzji,
- detekcja regularności, analiza interesujących skupień,
- askrypcja danych: łączenie informacji z kilku źródeł,
- wizualizacja informacji ukrytej w bazach danych.

Poniżej omówiono podstawowe założenia i kanwę metodyczną sieci neuronowych oraz systemów wnioskowania rozmytego, gdyż te metody inteligencji obliczeniowej (oraz pokrewne im algorytmy data mining) zostały wykorzystane w dalszej części opracowania do opisu metodyki analiz ilościowych i jakościowych.

Zarówno systemy wnioskowania rozmytego, jak i sztuczne sieci neuronowe mogą być stosowane w szeroko rozumianej analizie i eksploracji danych przestrzennych. Każde z tych podejść ma określone zalety. Sztuczne sieci neuronowe charakteryzuje naturalna zdolność do generalizacji wiedzy uzyskanej poprzez analizę podanych przykładów, jednakże proces uzyskiwania wyników generowanych przez odpowiednio wytrenowaną SSN jest trudny do interpretacji. Dla ułatwienia interpretacji stosowany jest zatem często proces ekstrakcji reguł opisujących w sposób jawny działanie modelowanego systemu. Jest to podejście zbliżone do definiowania relacji IF-THEN w metodzie sterowania rozmytego.

Systemy wnioskowania rozmytego

Początki tzw. logiki rozmytej i teorii zbiorów rozmytych związane są z nazwiskiem Lofti Zadeha (1965): „ogólnie rzecz biorąc, złożoność i precyzja są ze sobą w relacji odwrotnej w tym sensie, że jeżeli złożoność rozpatrywanego problemu wzrasta, to zmniejsza się możliwość jego precyzyjnej analizy”. Logika rozmyta umożliwiająca rozpatrywanie w sposób formalny

sytuacji „niedokładnych”, spełnia tym samym postulat adekwatności w opisie rzeczywistości. Teoria zbiorów i systemów rozmytych znalazła zastosowanie także w analizie i klasyfikacji danych o charakterze przestrzennym.

W klasycznej, arystotelesowskiej logice dwuwartościowej zdanie może przybierać wartości ze zbioru {0,1} często opisywane jako prawda lub fałsz. W logice wielowartościowej zdanie jest prawdziwe bądź fałszywe w określonym stopniu – przyjmuje dowolną wartość z przedziału [0,1]. Natomiast w logice rozmytej zdanie przybiera wartości logiczne lingwistyczne np. całkiem prawdziwe, w większości prawdziwe, częściowo prawdziwe, raczej nieprawdziwe, absolutnie fałszywe itp. Granice tych wartości nie są ściśle określone. Często przedziały zachodzą na siebie, przenikają się lub pokrywają (Yager, Filev, 1994).

Istotnym pojęciem stosowanym w logice rozmytej jest tzw. funkcja przynależności. Funkcja przynależności jest to krzywa, która obrazuje stopień przynależności każdego punktu przestrzeni rozważań do danego zbioru rozmytego.

Zaproponowana przez Zadeha (1965, 1973) teoria zbiorów rozmytych zakłada, iż w opisie działania systemu wykorzystywane są:

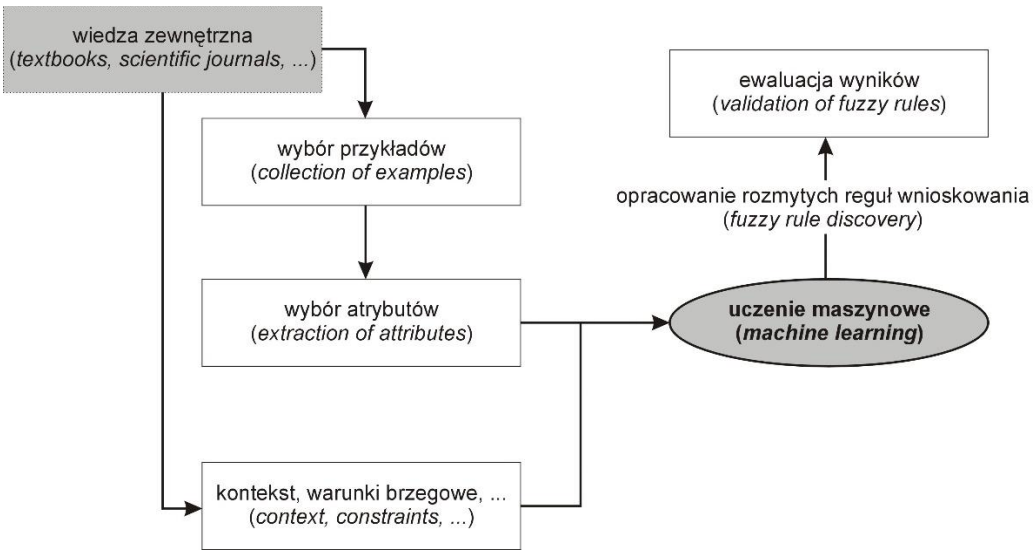
- tzw. zmienne lingwistyczne (np. duży, mały, około połowy, dość wysoki, raczej istotny, smaczny itp.),
- tzw. rozmyte zdania warunkowe wyrażające zależności pomiędzy zmiennymi lingwistycznymi w postaci reguł IF-THEN (jeżeli – to), np. jeżeli potencjał demograficzny jest duży to rozwój przedsiębiorczości następuje dość szybko,
- tzw. złożeniowe reguły wnioskowania pozwalające na indukowanie wartości wynikowej na podstawie znajomości zmiennej pierwotnej oraz zależności między.

Wymienione elementy są wystarczające do opracowania systemu wnioskowania rozmytego, ponieważ umożliwiają reprezentację wiedzy eksperta (wyrażonej w języku naturalnym) i jej praktyczne wykorzystanie (Kacprzyk, 2001). Opisane podejście jest przykładem metody deskryptywnej – nie opartej na modelu, jako że w sposób jawny opisywany jest sposób sterowania procesem wnioskowania rozmytego.

Istotnym elementem systemu wnioskowania rozmytego jest baza reguł. Baza ta zawiera zdania warunkowe określające zależności przyczynowe pomiędzy zmiennymi wejściowymi i wynikowymi. Celem procesu wnioskowania realizowanego w systemie FIS jest znalezienie wartości zmiennej wynikowej indukowanej przez aktualne wartości zmiennych pierwotnych oraz bazę reguł. Jeżeli wynikiem wnioskowania jest wartość rozmyta, należy przeprowadzić proces tzw. defuzyfikacji (wyostrzenia). Polega on na przekształceniu wartości rozmytych w ścisłe wartości liczbowe.

Użycie reguł rozmytych umożliwia uwzględnienie nieprecyzyjności, co pozwala na podniesienie efektywności rozwiązania. Według Zadeha kluczowe znaczenie dla budowy systemów wykorzystujących logikę rozmytą ma stosowanie języka naturalnego

pozwalającego na kompleksowy opis modelowanego systemu (Rys. 1) dokonany przez eksperta w danej dziedzinie (ludzka percepcje rzeczywistości).



Rys. 1. Architektura systemu wnioskowania rozmytego

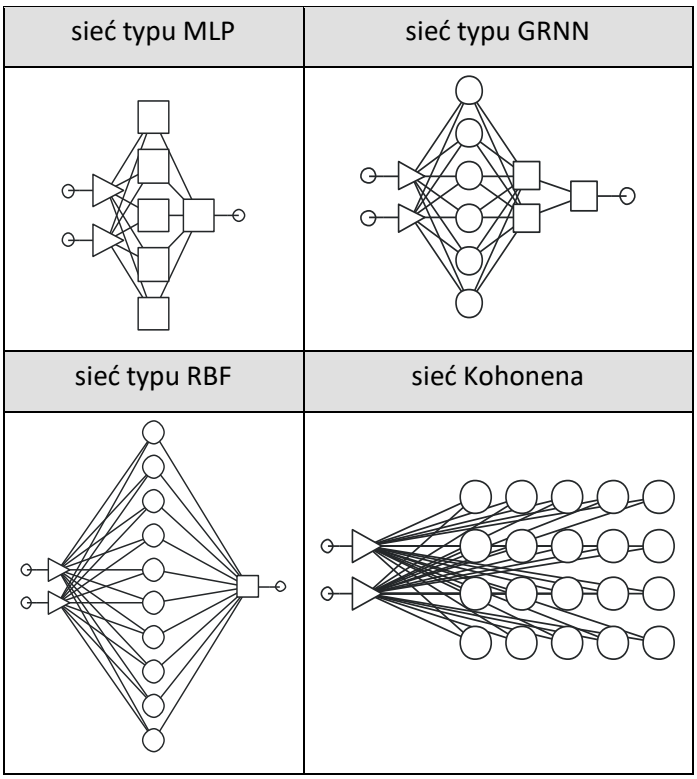
Sztuczne sieci neuronowe

Sztuczne sieci neuronowe (SSN) rozwinęły się w wyniku badań prowadzonych w dziedzinie inteligencji obliczeniowej, a zwłaszcza prac dotyczących funkcjonowania podstawowych struktur mózgu (Tadeusiewicz, 1998). Systemy oparte na wykorzystaniu sieci neuronowych znalazły szerokie zastosowanie w tych dziedzinach wiedzy, których przedmiotem badania są zagadnienia predykcji i klasyfikacji. SSN pozwalają na stosunkowo proste tworzenie złożonych modeli nieliniowych, „ucząc się” na przedstawianych przykładach (Fausett, 1994). Efektywność stosowania sieci neuronowych zależy od trafnego skojarzenia ich właściwości z charakterem problemu, który za ich pomocą chcemy rozwiązywać. SSN mogą być użyteczne w sytuacjach, w których struktura problemu jest słabo rozpoznana, informacje wejściowe niekompletne, a przybliżone wyniki końcowe są zadowalające (Domański, 2000).

Do sieci neuronowych wprowadza się pewne zmienne wejściowe (dane), zaś jej wyjścia definiują zmienne wyjściowe (rozwiązania). Dlatego też mogą one być stosowane wszędzie tam, gdzie na podstawie pewnych znanych informacji (np. liczba ludności w powiatach, odległość od stolicy województwa, gęstość sieci drogowej) wnioskuje się o pewnych nieznanych informacjach (np. rozwój przedsiębiorczości charakteryzowany przez liczbę MŚP na 10 tys. mieszkańców) (Fausett, 1994).

Sieci neuronowe pozwalają na modelowanie dowolnie złożonych funkcji nieliniowych. Proces „uczenia” sieci typu regresyjnego i większości sieci klasyfikacyjnych odbywa się przy

użyciu tzw. algorytmu „z nauczycielem” – tj. poprzez wielokrotne „pokazywanie” sieci danych źródłowych wraz z poprawną odpowiedzią (Patterson, 1996).



Rys. 2. Architektura sieci MLP, GRNN, RBF i sieci Kohonena

Istnieje wiele typów sieci neuronowych, różniących się strukturą i zasadami działania (Rys. 2). Do najczęściej stosowanych sieci regresyjnych należą tzw. perceptrony wielowarstwowe – MLP (multilayer perceptron) oraz sieci o radialnych funkcjach bazowych – RBF (radial basis functions) oraz sieci realizujące regresję uogólnioną – GRNN (generalized regression neural network).

Z punktu widzenia zagadnienia klasyfikacji danych źródłowych interesujące jest zastosowanie SSN uczonych w trybie „bez nauczyciela”. Przykładem takiej sieci klasyfikacyjnej jest tzw. sieć Kohonena. W procesie uczenia sieć ta nie otrzymuje informacji zwrotnej o poprawności udzielanej odpowiedzi. Sieć Kohonena nie znając pożądanej informacji wyjściowej samoczynnie uczy się rozpoznawania struktury danych (Kohonen, 1982). W procesie tym następuje porządkowanie danych wejściowych według kryterium ich wzajemnego podobieństwa. Sieci Kohonena znajdują zastosowanie w eksploracyjnej analizie danych, przy rozpoznawaniu skupień zmiennych wejściowych, a także w klasyfikacji danych (Tadeusiewicz, 1998). Klasyfikacja ta oparta jest na wewnętrznym podobieństwie struktury zmiennych, bez konieczności zakładania a priori kryteriów podobieństwa i liczby wydzielanych

klas. Najistotniejszą cechą charakteryzująca sieć Kohonena jest możliwość odzwierciedlenia N-wymiarowej relacji podobieństwa obiektów w przestrzeni dwu- a nawet jednowymiarowej. Sieć Kohonena złożona jest tylko z dwóch warstw: wejściowej i wyjściowej (Osowski, 1996). Warstwa wyjściowa sieci tworzy tzw. mapę topologiczną (tzw. mapę cech Kohonena), odwzorowującą w przestrzeni jedno- lub dwuwymiarowej relacje podobieństwa cech analizowanych obiektów. Uczenie sieci Kohonena ma charakter iteracyjny. Początkowe, dobrane losowo, wartości wagowe poszczególnych neuronów są w poszczególnych etapach modyfikowane tak, aby sąsiadujące ze sobą neurony „specjalizowały" się w rozpoznawaniu podobnych obiektów.

Zastosowanie sieci neuronowych upowszechniło się dzięki ich właściwościom użytecznym w modelowaniu i obliczeniach. Zalety SNN można ująć następująco (Tadeusiewicz, 1998):

- SSN pozwalają na stosunkowo proste tworzenie dowolnie złożonych modeli nieliniowych, „ucząc się" na przedstawianych przykładach. Metoda ta nie wymaga przyjmowania a priori żadnych założeń co do kształtu i stopnia nieliniowości tworzonej funkcji regresji,
- sieci neuronowe charakteryzuje wysoki stopień odporności na błędy. Informacje na podstawie których uczymy sieć mogą być niepełne lub obarczone błędem. Odpowiednio nauczona sieć neuronowa potrafi skutecznie odfiltrować szum i prowadzić obliczenia w oparciu o wypadkowe tendencje i trendy,
- istotną zaletą stosowania sieci neuronowych jest łatwość użycia. SSN działają na zasadzie „czarnej skrzynki": pytanie – odpowiedź. Sieci neuronowe w praktyce same konstruują potrzebne użytkownikowi modele, ponieważ automatycznie uczą się na podanych przykładach. Proces uczenia sieci zastępuje programowanie. Dzięki temu można rozwiązywać problemy obliczeniowe nie znając algorytmu, dysponując jedynie testowym zbiorem zawierającym „pytania" i „odpowiedzi".
- właściwie przygotowaną (nauczoną) SSN cechuje zdolność do generalizacji, tj. umiejętność uogólniania zdobytej wiedzy.

Wykorzystywane metody

Wobec złożoności informacji zawartych w udostępnionej bazie REGON, ZUS, danych statystycznych oraz informacji topograficznej zawartej w bazie danych BDOT 10 K, Wykonawca podjął starania opracowania metodyki analiz, która uwzględniając dane lokalizowane przestrzennie, umożliwi zarazem analizę wszystkich zawartych w tabeli źródłowej informacji opisowych dotyczących MŚP, np. formy własności, stanu aktywności, liczby pracowników itp. Kluczowe znaczenie ma jednak opracowanie metodyki eksploracyjnej analizy danych przestrzennych, która uwzględnia nie tylko „surowe" dane, lecz także kontekst przestrzenny – odległość od ośrodka decyzyjnego, firm partnerskich, autostrad, a także szeregu parametrów demograficznych.

Jako kanwa analityczna wykorzystywana niniejszym opracowaniu zostały zaproponowane metody jakościowe i ilościowe – zarówno „klasyczne", jak i wywodzące się z tzw. metod inteligencji obliczeniowej i data mining (Rys. 3).

Rys. 3. Schemat zależności pomiędzy metodami data mining, statystyką, kartograficzną metodą badań, sztuczną inteligencją oraz uczeniem maszynowym.

W niniejszym opracowaniu przedstawiono możliwości wykorzystania w analizie rozwoju MŚP w województwie kujawsko-pomorskim metod jakościowych związanych z klasyfikacją danych, jak i metod ilościowych umożliwiających realizację szeroko rozumianych zagadnień o charakterze regresyjnym. Poniżej wymieniono omówione w dalszej części opracowania metody.

Metody klasyfikacji:

- koncentracja przestrzenna
- główne składowe
- permutacje Bertina
- typogram Uhorczaka
- klasyfikacja rozmyta
- sieci samoorganizujące Kohonena
- drzewa klasyfikacyjne typu CART

Metody regresji uogólnionej:

- liniowa regresja wielowymiarowa
- regresja ważona geograficznie (ekonometria przestrzenna)
- regresja nieliniowa
- MARS
- wektory nośne
- drzewa regresyjne
- sieci neuronowe

Możliwości wykorzystania poszczególnych metod analitycznych opisanych poniżej zilustrowano wykorzystując zbiór danych charakteryzujących gminy województwa kujawsko-pomorskiego. Zestaw danych zawiera 144 rekordy odpowiadające 144 gminom tego województwa. Poszczególne obiekty charakteryzowane są przez zestaw kilkudziesięciu parametrów (cech) wyrażonych zarówno w ilościowych, rangowych, jak i jakościowych skalach pomiarowych. Przykładowo zestaw danych zawiera informacje na temat typu gminy (miejska, miejsko-wiejska, wiejska), jej wielkości, gęstości zaludnienia, procencie zabudowy, liczbie ludności, poziomie bezrobocia, wpływach z podatków, liczbie MŚP w poszczególnych gminach itp.

Koncentracja przestrzenna

W ujęciu strukturalnym pod pojęciem koncentracji rozumie się pewien stan jaki osiągnął badany układ przestrzenny. Zjawisko koncentracji można zdefiniować jako nierównomierny podział ogólnej sumy wartości cechy pomiędzy poszczególne jednostki badanej zbiorowości, a do oceny jej stopnia można zastosować różne metody:

- graficzną, w której wykreśla się i analizuje tzw. krzywą koncentracji Lorenza,
- numeryczną, w której wyznacza się odpowiednie wskaźniki liczbowe, np.indeks Giniego.

Aby opracować krzywą koncentracji należy pogrupować dane i zestawić je w dwa szeregi kumulacyjne. Pierwsza z kolumn zawiera skumulowaną częstość dla liczebności zjawiska (np. liczby MŚP w poszczególnych powiatach), drugi zaś określa skumulowaną częstość dla wartości cechy. Obie częstości wyrażane są w procentach, co pozwala na normalizację wartości liczbowych i opracowanie krzywej koncentracji w tzw. kwadracie jednostkowym.

Dwukrotność pola pomiędzy krzywą Lorenza a przekątną kwadratu jednostkowego nazywana jest wskaźnikiem Giniego. Wskaźnik ten stosowany jest w statystyce jako miara koncentracji (nierównomierności) rozkładu zmiennej losowej. Przyjmuje on wartości od 0 do 1, przy czym 0 określa pełną równomierność rozkładu, a wartość 1 oznacza, że tylko jedna obserwacja z nieskończonego zbioru uzyskała dodatnią wartość zmiennej. Wskaźnik ten często jest wyrażany także w %.

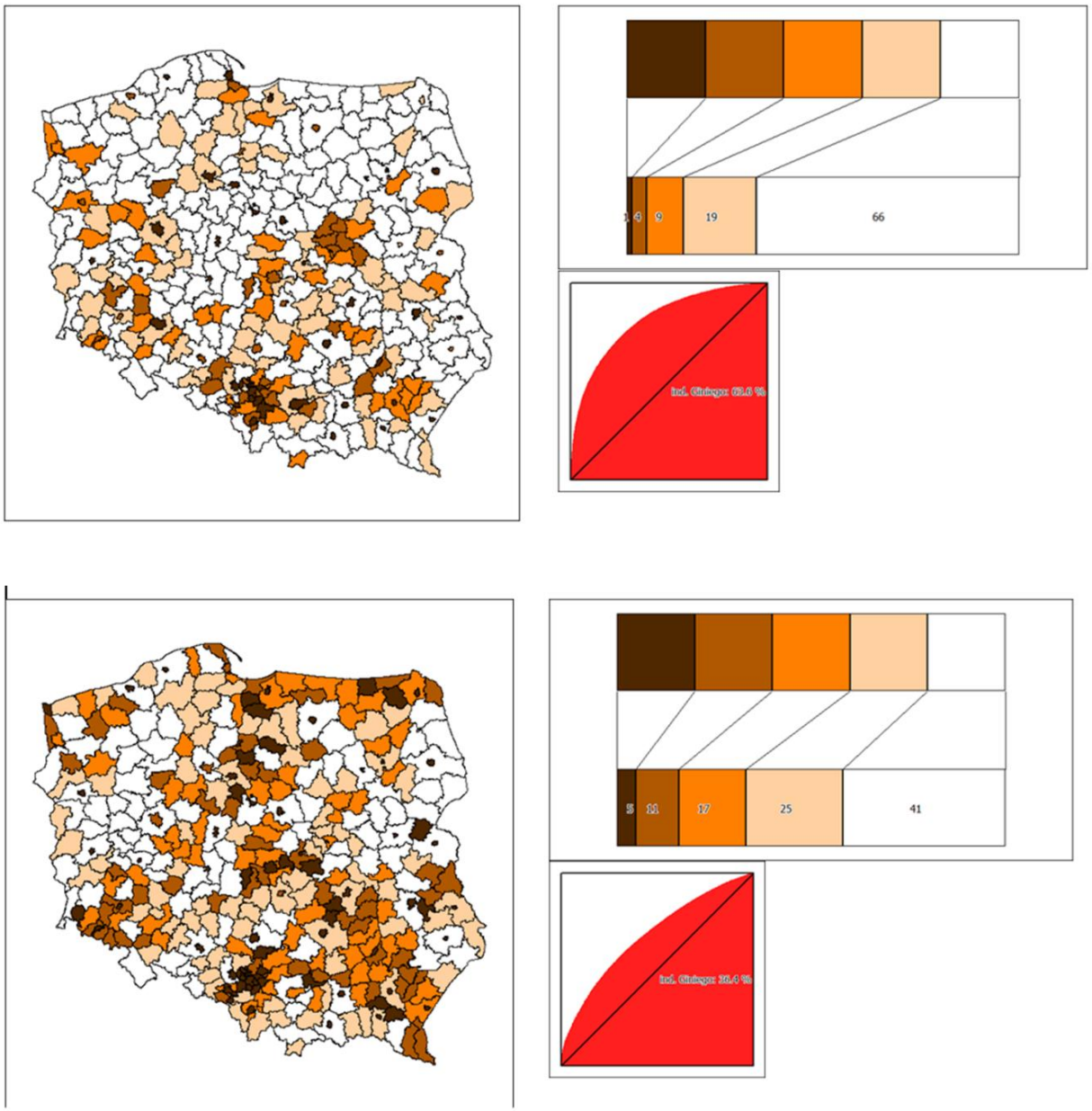
W przypadku analizy danych przestrzennych analiza krzywej Lorenza pozwala na opracowanie mapy koncentracji nazwanej przez F. Uhorczaka mapą koncentracji mozaikowej. Metoda wykorzystująca pojęcie koncentracji powierzchniowej umożliwia badanie nierównomierności przestrzennego rozmieszczenia zjawisk bądź obiektów w ramach przyjętego układu powierzchniowych jednostek odniesienia, np. podziału administracyjnego kraju czy województwa.

Mapa koncentracji powierzchniowej powstaje w wyniku wyróżnienia jednostek odniesienia, do których przyporządkowane są określone wartości liczbowe cechy badanego obiektu lub zjawiska. Materiałami źródłowymi niezbędnymi do określenia koncentracji powierzchniowej są:

- mapa z zaznaczonymi granicami jednostek układu terytorialnego,
- dane statystyczne wyznaczone dla tych jednostek.

Do prezentacji wyników na mapie stosuje się metodę kartogramu, jednakże dobór przedziałów klasowych w tej metodzie wynika wprost z analizy krzywej koncentracji Lorenza.

Na poniższych rysunkach zaprezentowano koncentrację przestrzenną funduszy unijnych uzyskanych w ramach Programu Operacyjnego Innowacyjna Gospodarka oraz Programu Operacyjnego Kapitał Ludzki. Obie mapy obrazują rozkład przestrzenny funduszy unijnych w 2014 r., wskazują zarazem na znacznie większą koncentrację przestrzenną środków przeznaczonych na rozwój gospodarczy niż społeczny (Rys. 4).



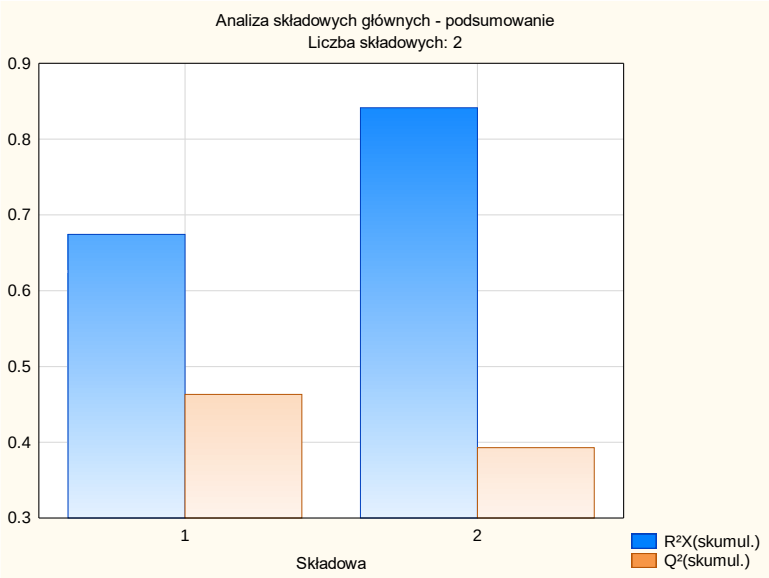
Rys. 4. Koncentracja przestrzenna środków unijnych uzyskanych przez poszczególne powiaty w ramach Programu Operacyjnego Innowacyjna Gospodarka (góra) oraz Programu Operacyjnego Kapitał Ludzki (dół) w 2014 r. (analiza własna)

Metoda głównych składowych

Analiza czynnikowa jako metoda badania układów wielozmiennych jest często stosowana w geografii ekonomicznej. Pojęcie „analiza czynnikowa" jest wieloznaczne i odnosi się do wielu metod. Najbardziej znane spośród nich to analiza czynnikowa w węższym znaczeniu i tzw. analiza głównych składowych (Z. Chojnicki, T. Czyż, 1977).

Za twórcę analizy składowych głównych uważa się H. Hotellinga, który zastosował opracowaną przez siebie metodę w dziedzinie psychologii. Popularność stosowania algorytmu głównych składowych jako metody klasyfikacji wynika z faktu, iż umożliwia on sprowadzenie uzyskanych wyników z przestrzeni N wymiarowej do przestrzeni trój-, dwu-, a nawet jednowymiarowej (Z. Kaczmarek, J. Parysek, 1977). Istotą metody jest transformacja wyjściowych zmiennych Yj (j = 1, 2, ..., n) – atrybutów charakteryzujących badane jednostki przestrzenne w nowe zmienne Vj. Zmienne te zwane są składowymi głównymi, przy czym V1 zawiera największy ładunek informacji, V2 mniejszy, zaś Vn najmniejszy. Do ostatecznej analizy klasyfikacyjnej przyjmuje się zwykle jednostki przestrzenne uporządkowane liniowo względem jednej, dwóch lub trzech pierwszych składowych, o czym decyduje ładunek informacyjny wnoszony przez poszczególne składowe.

Przykładem zastosowania tej metody może być analiza zestawu czterech zmiennych charakteryzujących dochody poszczególnych gmin województwa kujawsko-pomorskiego (zmienne: dochody1, dochody2, dochody3, dochody4). Analiza głównych składowych pozwala na uzyskanie ponad 2/3 wartości informacyjnej wektora danych wejściowych na podstawie pierwszej składowej głównej (Rys. 5). Dodanie drugiej składowej zwiększa wartość informacyjną modelu wynikowego do poziomu blisko 85%.



Skła...	R²X	R²X(...	Wart...	Q²	Gran...	Q²(s...	Istot...
1	0.675	0.675	2.698	0.463	0.255	0.463	Tak
2	0.167	0.842	0.668	-0.100	0.338	0.393	Nie

Rys. 5. Analiza głównych składowych dla 4 typów dochodów gmin województwa kujawsko-pomorskiego.

Metoda permutacji Bertina i typogramu Uhrczaka

Zarówno metoda permutacji, jak i wywodząca się z kartografii tematycznej metoda typogramu ma na celu dokonanie klasyfikacji jednostek przestrzennych na podstawie charakteryzujących je cech. Klasyfikację można określić jako systematyczne grupowanie danych na podstawie jednej lub kilku cech (M-J. Kraak, F. Ormeling, 1998). Prowadzi to do generalizacji tematycznej danych źródłowych, poprawia jednak znacznie czytelność przekazu informacji. Istotnym celem klasyfikacji jest zatem porządkowanie obiektów w przestrzeni fizycznej, a także spełnianie funkcji poznawczej – zrozumienie struktury danych.

Jednym ze sposobów prowadzenia klasyfikacji jest typologia – wyróżnianie typów, czyli wyznaczenie pewnego zbioru, którego elementy wykazują w określonym stopniu podobieństwo cech do wzorca. Istotnym kryterium wydzielania typów jest ich jednorodność. Algorytmy klasyfikacji obiektów wielocechowych oparte są na metodzie „odkrywania" podobieństwa wewnętrznej struktury tych obiektów (W.S. Tikunow, 1997).

Wyniki klasyfikacji obiektów przedstawiane są zwykle w postaci zestawień tabelarycznych. W przypadku klasyfikacji jednostek przestrzennych jest to także mapa. Mapy tematyczne pozwalają bowiem prezentować nie tylko fakty, lecz także hipotezy i rezultaty analiz.

Kartograficznym przykładem prezentacji tematycznej obiektów wielocechowych jest opracowany przez F. Uhorczaka kartotypogram (F. Uhorczak, J. Ostrowski, 1972). Podstawą konstrukcji typogramu jest układ współrzędnych. Każda z osi reprezentuje intensywność występowania określonego atrybutu charakteryzującego własności jednostki przestrzennej. Najczęściej stosowany jest układ ortogonalny, umożliwiający prezentację zjawisk czterocechowych, możliwe jest także użycie układów wieloosiowych. Głównym celem typogramu jest określanie typu obiektu poprzez porównanie kształtów figur reprezentujących poszczególne jednostki przestrzenne. Współcześnie metoda ta określana jest tzw. wykres radarowy. Jego odniesienie do poszczególnych jednostek przestrzennych (np. poszczególnych powiatów w danym województwie) pozwala na opracowanie wielocechowej

mapy tematycznej obrazującej np. liczbę MŚP, liczbę ludności, gęstość sieci drogowej itp. na jednej mapie tematycznej charakteryzującej całe województwo.

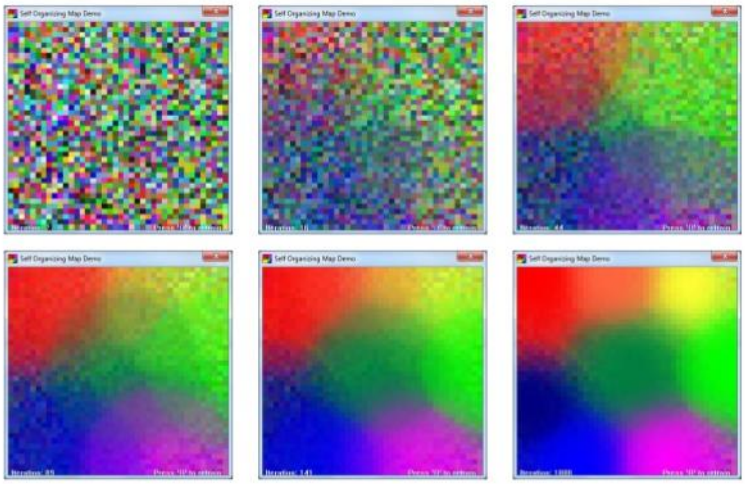
Innym przykładem niestandardowej klasyfikacji danych przestrzennych jest opracowana przez J. Bertina (1981) metoda permutacji. Jest to metoda graficzna oparta na zdolności percepcyjnej oka ludzkiego do postrzegania relacji w układzie przestrzennym. Dla analizowanych obiektów i charakteryzujących je atrybutów budowana jest macierz informacji. W kolumnach tej macierzy zawarte są informacje o poszczególnych obiektach, zaś w wierszach o ich cechach. Porządkowaniu podlegają zarówno obiekty jak i cechy. Współcześnie metoda ta została w pełni zautomatyzowana i pozwala na wyznaczanie „klastrów przestrzennych" zarówno w przestrzeni dwuwymiarowej, jak i N-wymiarowej przestrzeni cech. Rozwinięcie tej metody stanowi samoorganizująca się sieć neuronowa Kohonena. Zagadnienie to został omówione poniżej.

Samoorganizująca sieć Kohonena

Z punktu widzenia zagadnienia klasyfikacji danych źródłowych interesujące jest zastosowanie SSN uczonych w trybie „bez nauczyciela". Przykładem takiej struktury jest tzw. sieć Kohonena. W procesie uczenia sieć ta nie otrzymuje informacji zwrotnej o poprawności udzielanej odpowiedzi. Sieć Kohonena nie znając pożądanej informacji wyjściowej samoczynnie uczy się rozpoznawania struktury danych (T. Kohonen, 1982). W procesie tym następuje porządkowanie danych wejściowych według kryterium ich wzajemnego podobieństwa. Sieci Kohonena znajdują zastosowanie w eksploracyjnej analizie danych, przy rozpoznawaniu skupień zmiennych wejściowych, a także w klasyfikacji danych (R. Tadeusiewicz, 1998). Klasyfikacja ta oparta jest na wewnętrznym podobieństwie struktury zmiennych, bez konieczności zakładania a priori kryteriów podobieństwa i liczby wydzielanych klas. Najistotniejszą cechą charakteryzująca sieć Kohonena jest możliwość odzwierciedlenia N-wymiarowej relacji podobieństwa obiektów w przestrzeni dwu- a nawet jednowymiarowej.

Sieć Kohonena złożona jest tylko z dwóch warstw: wejściowej i wyjściowej. Warstwa wyjściowa sieci tworzy tzw. mapę topologiczną (mapę cech Kohonena), odwzorowującą w przestrzeni jedno- lub dwuwymiarowej relacje podobieństwa cech analizowanych obiektów. Uczenie sieci Kohonena ma charakter iteracyjny. Początkowe, dobrane losowo, wartości wagowe poszczególnych neuronów są w poszczególnych etapach modyfikowane tak, aby sąsiadujące ze sobą neurony „specjalizowały" się w rozpoznawaniu podobnych obiektów (Rys. 6).

Zastosowanie sztucznych sieci neuronowych można rozumieć jako rozwinięcie tradycyjnych metod klasyfikacji danych przestrzennych. Zastosowanie sieci Kohonena pozwala na porządkowanie struktury wielowymiarowych danych źródłowych i jej wizualizację w przestrzeni jedno- lub dwuwymiarowej. Zastosowanie jednowymiarowej warstwy wyjściowej SSN pozwala na bezpośrednie prezentowanie wyników klasyfikacji na mapie tematycznej, co jest podejściem optymalnym z kartograficznego punktu widzenia.



Rys. 6. Etapy działania sieci Kohonena

Dla danych testowych zawierających 144 obiekty reprezentujące gminy województwa kujawsko-pomorskiego opracowana została samoorganizująca sieć Kohonena zawierająca 8 neuronów w warstwie wejściowej oraz trzy neurony w warstwie wyjściowej. Danymi wejściowymi były np. liczba ludności, powierzchnia gminy, gęstość zabudowy, gęstość sieci drogowej itp. Warstwa wyjściowa umożliwiła klasyfikację obiektów na trzy kategorie, które pozwoliły na dokonanie podziału danych źródłowych na gminy „duże", „średnie" i „małe" na podstawie jednoczesnej analizy 8 cech (zmiennych niezależnych). Dane te zapisane jako cecha WielGmin o charakterze rangowym zostały wykorzystane w dalszych analizach spataial data mining, np. klasyfikacji danych wykorzystującej drzewa decyzyjne typu CART.

Klasyfikacja rozmyta

Zagadnienie klasyfikacji jednostek przestrzennych stanowi jeden z istotniejszych problemów nauki o geoinformacji, może być także stosowane w celu tworzenia i obrazowania modeli ekonomicznych. Metody taksonomiczne zapożyczone z nauk niegeograficznych traktują poszczególne obiekty jako niezależne całości w oderwaniu od ich położenia przestrzennego. Metody taksonomiczne grupują jednostki przestrzenne w klasy (lub typy) bez

zachowania warunku ciągłości przestrzennej charakterystycznego dla procesu regionalizacji. Procedura typologii w ujęciu taksonomicznym uwzględnia cztery podstawowe etapy (Z. Chojnicki, T. Czyż, 1973):

1. opracowanie macierzy informacji,
2. redukcję wymiarów przestrzeni wielocechowej,
3. estymację wielocechowego podobieństwa,
4. grupowanie lub podział przestrzenny.

Procedura ta może być realizowana na wiele sposobów. Jedną z nich jest wykorzystanie ogólnej metodyki wnioskowania rozmytego do klasyfikacji wielocechowych obiektów mających odniesienie przestrzenne.

Teoria zbiorów i systemów rozmytych znalazła zastosowanie w analizie i klasyfikacji danych o charakterze przestrzennym. Koncepcja rozmytości stosowana w typologii wykorzystuje pojęcie rozmytej relacji podobieństwa. Kryterium podziału jednostek przestrzennych na określone klasy jest ich wzajemne podobieństwo rozpatrywane ze względu na charakteryzujące je atrybuty. Podobieństwo to ujęte jest w pewien przedział wartości liczbowych.

Utożsamienie idei rozmytości i prawdopodobieństwa jest przedmiotem dyskusji i sporów prowadzonych przez badaczy od początków powstania teorii zbiorów rozmytych (K. Jajuga, 1984). „Rozmyte" funkcje podobieństwa obiektów można interpretować bowiem jako podobieństwo obiektu O do klasy rozmytej K. W klasyfikacji rozmytej dany obiekt należy do różnych klas z różnymi stopniami przynależności.

W literaturze przedmiotu wyróżnia się kilka metod tworzenia klasyfikacji rozmytej. Jedną z częściej stosowanych jest iteracyjny algorytm wykorzystujący pojęcie rozmytego środka ciężkości. W kolejnych krokach obliczeniowych dokonuje się zmiany stopnia przynależności danego obiektu do poszczególnych klas. Procedurę kontynuuje się aż do momentu, gdy wyznaczane wartości stopnia przynależności przestaną zmieniać się w stopniu istotnym.

Na użyteczność stosowania w ekonomii i nauce o geoinformacji metody klasyfikacji opartej na zbiorach rozmytych zwrócił uwagę W.S. Tikunow (1997). Opracowany przez niego niestandardowy algorytm typologii pozwolił na dokonanie „rozmytej" klasyfikacji państw świata według kryterium ich potencjału gospodarczego. Autor ten szeroko omawia także użyteczność różnych metod wydzielania grup taksonomicznych na przykładzie klasyfikacji państw Europy według kryterium stosunków handlowych z krajami byłego ZSRR.

Drzewa klasyfikacyjne typu CART

Metodą wywodząca się z grupy algorytmów data mining, która pozwala na uwzględnienie przestrzennego charakteru danych jest zastosowanie drzew klasyfikacyjnych z predykatorami przestrzennymi, co pozwala na realizowanie zadań klasyfikacyjnych w oparciu nie tylko o zmienne predykcyjne o charakterze jakościowym lub ilościowym, ale również o zmienne określające przestrzenne zależności między obiektami. Przestrzenne zależności mogą być definiowane za pomocą trzech typów relacji: relacji topologicznych (określonych przez operatory przestrzenne typu: obiekt A przecina obiekt B, obiekt A zawiera obiekt B, etc.), relacji odległości (określającej odległości między obiektami), relacją kierunku (określającą wzajemne położenie obiektów względem stron świata). Zdefiniowane w ten sposób relacje przestrzenne pozwolą na uzyskanie kompleksowej informacji o przestrzennych zależnościach między obiektami. Klasyfikacja obiektów za pomocą metody drzew klasyfikacyjnych z wykorzystaniem predykatorów przestrzennych daje realną szansę uzyskania efektywniejszych wyników, aniżeli klasyczne drzewa klasyfikacyjne, gdyż w sposób bezpośredni (w postaci niezagregowanej w żaden sposób) uwzględnia przestrzenny układ obiektów.

Chcąc przeprowadzić podział dowolnego zbioru obiektów przeprowadzamy kolejne grupowania zakresu zmienności cech opisujących obiekty, wówczas najwygodniej jest przedstawić ów proces w postaci drzewa. Hierarchiczna budowa takiego drzewa klarownie oddaje proces podejmowania decyzji. Drzewo stanowi graf złożony z węzłów i krawędzi łączących wybrane węzły. Początek drzewa reprezentuje węzeł zwany korzeniem drzewa, który stanowi cały zbiór obserwacji. Z każdego węzła może wychodzić dwie lub więcej krawędzie. W tym pierwszym przypadku mówi się o tzw. drzewach binarnych, w drugim zaś o drzewach dowolnych.

Węzeł, z którego nie wychodzą żadne krawędzie nazywany jest liściem drzewa. Każdy węzeł prezentuje decyzję o podziale zbioru (lub podzbioru) obiektów na podzbiory z uwagi na jedną z cech objaśniających. Dzielony zbiór (lub podzbiór) nosi nazwę węzła macierzystego, natomiast wydzielone podzbiory to węzły potomki. Wielkość drzewa określona jest przez liczbę liści, natomiast głębokość drzewa to liczba krawędzi między korzeniem drzewa, a najbardziej odległym liściem.

Drzewo jest graficznym modelem, którego konstrukcja opiera się na zasadzie rekurencyjnego podziału zbioru obserwacji A na n rozłącznych podzbiorów A1, A2, A3, …An. Celem budowy modelu jest otrzymanie podzbiorów maksymalnie jednorodnych, w odniesieniu do wartości zmiennej objaśnianej. Jest to proces wieloetapowy, w którym w każdym następnym kroku może być wykorzystywana inna zmienna niezależna. Na każdym etapie

analizuje się bowiem wszystkie predykatory, a następnie wybiera ten, dzięki któremu uzyskamy najlepszy podział węzła, a więc najbardziej homogeniczne podzbiory.

Drzewa klasyfikacyjne stosowane są do określania przynależności przypadków lub obiektów do klas jakościowej zmiennej objaśnianej w oparciu o pomiary jednej lub więcej zmiennych objaśniających. Priorytetem analizy opartej na drzewach klasyfikacyjnych jest prognozowanie lub objaśnianie odpowiedzi (reakcji) ukrytych w jakościowej zmiennej zależnej.

Do najbardziej znanych drzew decyzyjnych należą drzewa klasyfikacyjne i regresyjne typu CART (Classification and Regression Trees). Modele drzew klasyfikacyjnych i regresyjnych CART umożliwiają budowę zarówno modeli służących do rozwiązywania zagadnień klasyfikacyjnych (zmienna predykcyjna jest cechą jakościową) jak i regresyjnych (ilościowa zmienna zależna). W problemach klasyfikacyjnych głównym celem jest znalezienie drzewa, którego liście byłyby możliwie jak najbardziej jednorodne, tj. zwierały obserwacje, które należałyby do tej samej klasy lub kategorii.

Do kluczowych zalet metody CART należą:

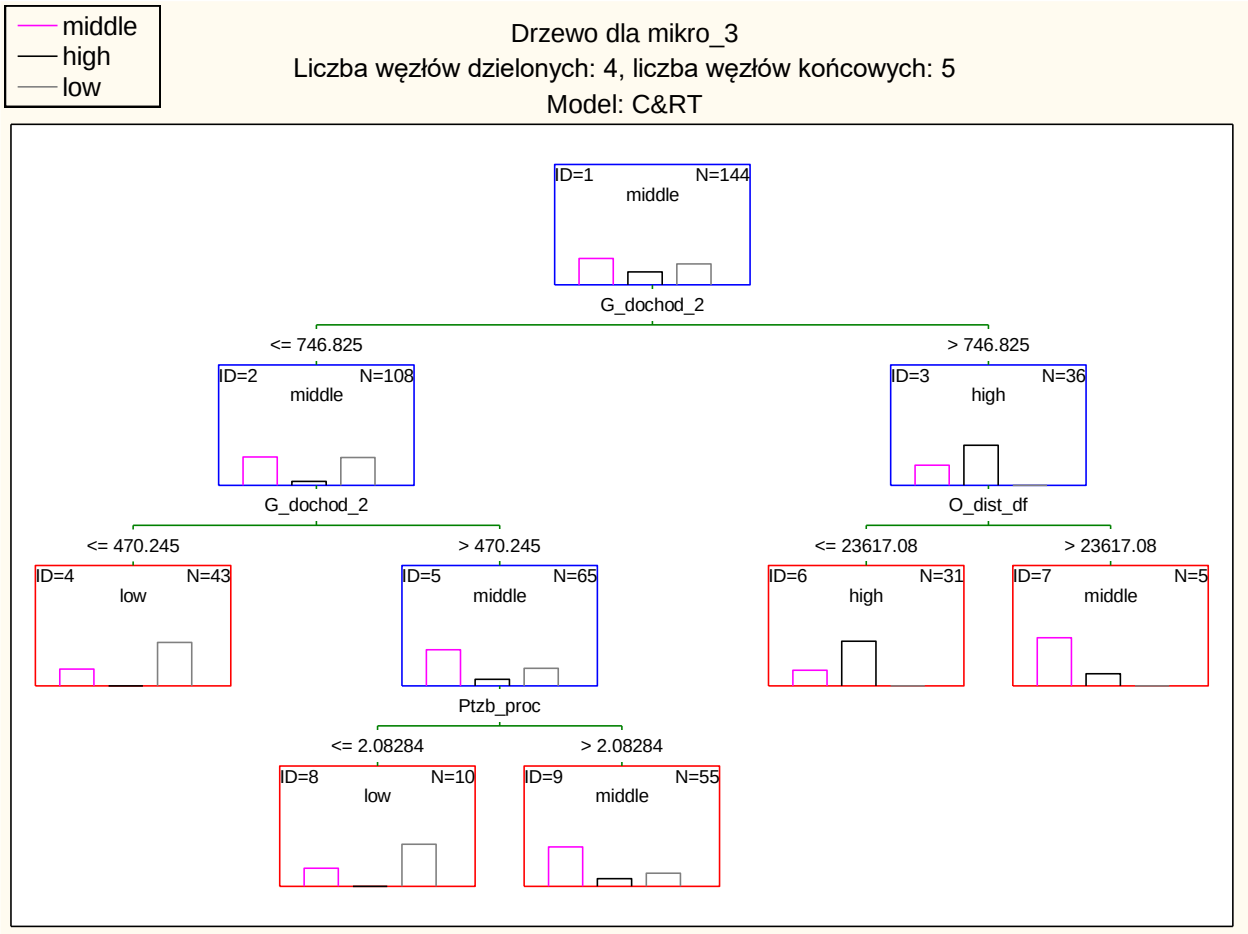
- wykorzystywanie każdej kombinacji zmiennych ciągłych i kategorialnych
- umożliwienie analizy zbiorów danych o złożonej strukturze
- brak czułości na występowanie obserwacji nietypowych.

Istotną zaletą modeli typu CART jest także brak założeń wstępnych co do liniowości modeli. CART jest przykładem metody nieparametrycznej, w której uzyskiwane wyniki mogą być zaprezentowane w postaci kilku warunków logicznych typu IF-THEN. Algorytmy wykorzystujące drzewa klasyfikacyjne są często stosowane w zagadnieniach data mining, gdzie wiedza a priori jest relatywnie niska, i trudno jest postawić klasyczne hipotezy statystyczne. Metody drzew klasyfikacyjnych (w tym algorytmy CART) potrafią „ujawnić" związki pomiędzy kilkoma zmiennymi, które nie zostałyby wykryte przez inne techniki analityczne.

Przykładem zastosowania drzew CART do klasyfikacji wielocechowych zjawisk odniesionych przestrzennie jest analiza poziomu przedsiębiorczości w gminach województwa kujawsko-pomorskiego. Jako zmienne objaśniane w modelu regresji wielorakiej stosowane były liczebności firm (mikro, małych i średnich) przypadające na 10 000 mieszkańców. Dla zachowania spójności oraz uzyskiwania porównywalności wyników uzyskiwanych dzięki zastosowaniu drzew klasyfikacyjnych typu CART oraz liniowej regresji wielorakiej przy podziale danych na klasy zastosowano jednakowy podział trójdzielny w obu tych metodach.

Wykorzystując zgromadzone dane źródłowe dokonano klasyfikacji eksperckiej mikro przedsiębiorczości na trzy kategorie (Rys. 7): wysoka, średnia i niska przedsiębiorczość w poszczególnych gminach wyrażona jako stosunek liczby firm na 10 000 mieszkańców.

Przedstawione poniżej drzewo klasyfikacyjne CART wskazuje, iż dla objaśnienia zmiennej zależnej kluczowe znaczenie mają dochody własne gmin uzyskiwane z podatków, odległość od dużych firm, z którymi kooperują mikroprzedsiębiorcy, a także procent zabudowy i gęstość sieci drogowej.



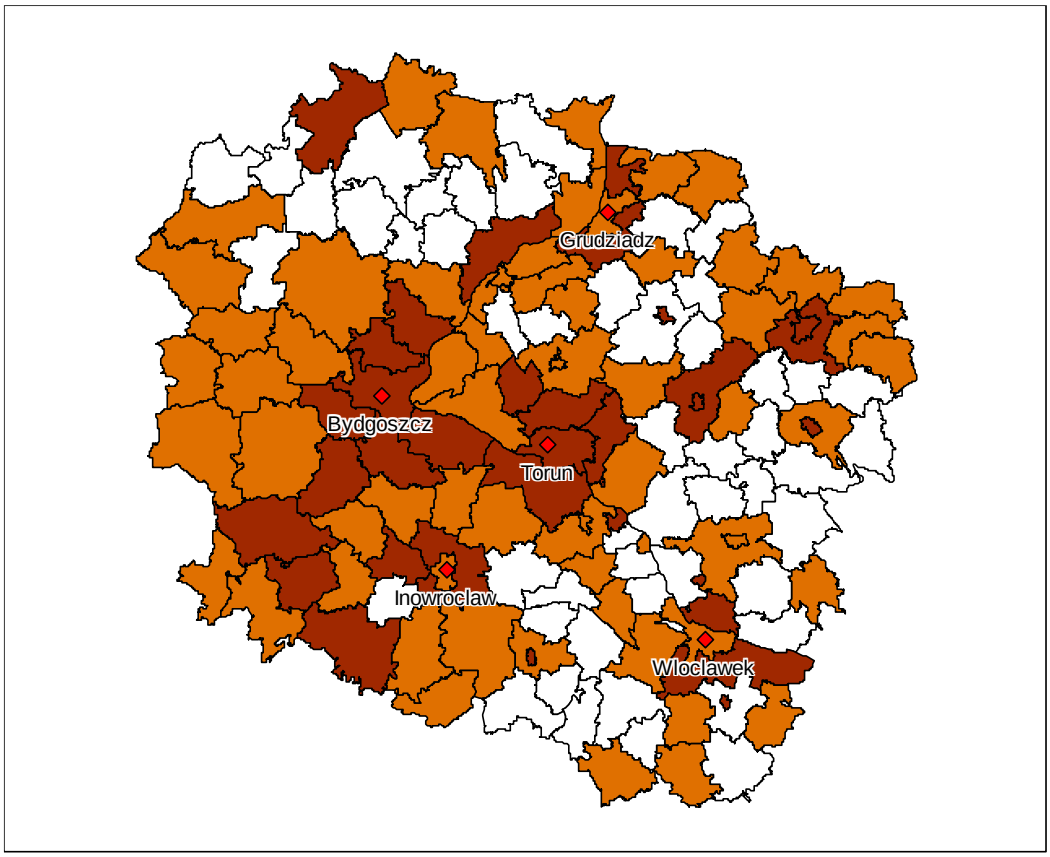
Rys. 7. Drzewo klasyfikacyjne CART opracowane dla firm typu mikro w województwie kujawsko-pomorskim (analiza własna).

Interpretacja sposobu klasyfikacji realizowanej przez drzewo CART wyrażona w języku naturalnym ma postać:

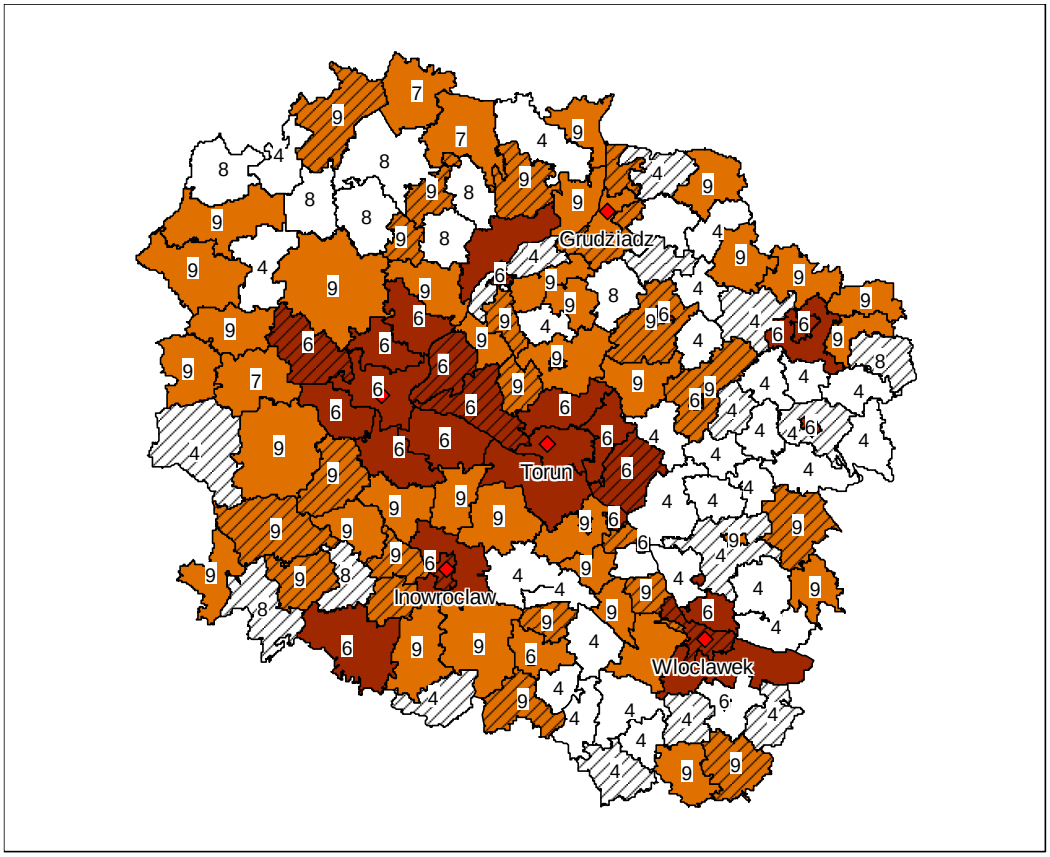
- przedsiębiorczość (mikro) w danej gminie jest wysoka jeśli dochody przypadające na osobę w tej jednostce są większe niż 747 zł, zaś odległość od dużych firm jest mniejsza niż 24 km (liść 6).
- przedsiębiorczość (mikro) w danej gminie jest średnia jeśli dochody przypadające na osobę w tej jednostce są większe niż 747 zł, zaś odległość od dużych firm jest większa niż 24 km (liść 7) lub dochody są niższe niż 747 zł (aczkolwiek wyższe niż 470 zł), zaś procent zabudowy w tej gminie jest wyższy niż 2% (liść 9)

- przedsiębiorczość (mikro) w danej gminie jest niska jeśli dochody przypadające na osobę w tej jednostce są mniejsze niż 470 zł (liść 4) lub dochody są wyższe niż 470 zł (aczkolwiek niższe niż 747 zł), zaś procent zabudowy w tej gminie jest niższy niż 2% (liść 8).

Umieszczone poniżej wynikowe mapy tematyczne pokazują (odpowiednio) źródłową klasyfikację tj. podział gmin w województwie kujawsko-pomorskim na wysoką, średnią i niską przedsiębiorczość w zakresie mikro firm (Rys. 8) oraz wyniki klasyfikacji realizowanej przez drzewo decyzyjne CART (rys. 9). Etykiety gmin wskazują na numer liścia w drzewie CART, według którego klasyfikowane były poszczególne gminy. Czerwona szrafura wskazuje na gminy, które zostały błędnie zaklasyfikowane. Liczba gmin błędnie zaklasyfikowanych wynosi 43, co stanowi 29.9%. Błąd ten wynika z uwzględnienia w drzewie decyzyjnym zaledwie trzech zmiennych objaśniających.



Rys. 8. Klasyfikacja ekspercka gmin w województwie kujawsko-pomorskim na wysoką, średnią i niską przedsiębiorczość w zakresie mikro firm



Rys. 9. Wyniki klasyfikacji realizowanej przez drzewo decyzyjne CART

Uzyskane wyniki wskazują, iż użycie zaledwie trzech zmiennych objaśniających pozwala na utworzenie klasyfikatora CART, który pozwala na przypisanie poszczególnych gmin do grupy jednostek administracyjnych o relatywnie wysokiej/średniej/niskiej mikro przedsiębiorczości w odniesieniu do 10 000 mieszkańców. Rozkład przestrzenny wyników klasyfikacji CART jest zbliżony do źródłowej klasyfikacji eksperckiej, jednak należy podkreślić, iż wartością dodaną zastosowania drzewa decyzyjnego jest przypisanie poszczególnym gminom identyfikatora liścia CART, co pozwala na objaśnienie wyników klasyfikacji w postaci składni SQL, a także języka naturalnego. Umożliwia to poszukiwanie odpowiedzi nie tylko na pytanie KTÓRE gminy charakteryzuje np. wysoka przedsiębiorczość, lecz także DLACZEGO w tych gminach rozwój przedsiębiorczości jest wysoki.

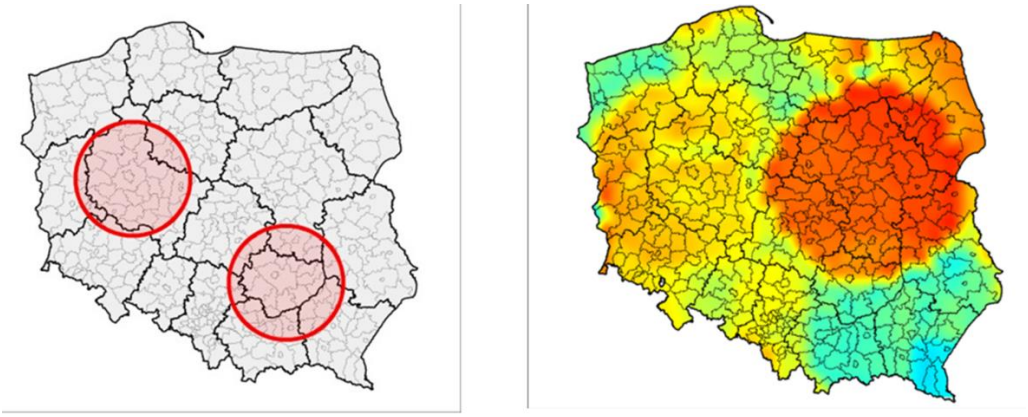
Ze względu na prostotę działania oraz dużą komunikatywność uzyskiwanych wyników, a także bezpośrednią ekstrakcję reguł decyzyjnych objaśniających model klasyfikacji (lub regresji), Wykonawca szczególnie poleca prowadzenie analiz spatial data minng z wykorzystaniem drzew decyzyjnych typu CART.

Liniowa regresja wielowymiarowa (wieloraka)

Tworzenie modelu regresji wielowymiarowej wymaga postawienia hipotezy badawczej o istnieniu zależności przyczynowo skutkowej pomiędzy zmiennymi niezależnymi oraz zmienną zależną. Co więcej zależność ta, której miarą jest wartość współczynnika korelacji Pearsona nie może mieć charakteru pozornego, dane powinny być wyrażone w ilościowych skalach pomiarowych o rozkładzie zbliżonym do normalnego. Są to bardzo silne założenia, które sprawiają, iż wielowymiarowe modele regresji liniowej (tzw. regresji wielorakiej) są stosowane głównie jako tzw. pierwsze przybliżenie złożonego modelu budowanego w oparciu o funkcje nieliniowe (np. sztuczne sieci neuronowe). Regresja liniowa ma jednak istotne zalety – zbudowany model w bardzo prosty sposób objaśnia złożone zjawiska. Przykładowo jeżeli model regresji wyrażony jest w postaci Y = 2*X + 10, gdzie Y oznacza liczbę MŚP w danym powiecie, X zaś liczbę ludności w powiatach w tysiącach osób, to interpretacja tego modelu jest niezwykle prosta: każdy tysiąc mieszkańców „generuje" powstanie jednej firmy typu MŚP, zaś niezależnie od liczby mieszkańców w każdym powiecie w bazie REGON odnaleźć można 10 MŚP. Miarą, która określa wiarygodność modelu regresji jest tzw. współczynnik determinacji R^2, który stanowi procentowe ujęcie kwadratu wartości współczynnika korelacji Pearsona. Wartość współczynnika determinacji zbliżona do 100% wskazuje na to, iż znacząca część zmienności zjawiska została objaśniona przez model. Niska wartość tego współczynnika mówi z kolei o braku istotnej zależności pomiędzy zestawem zmiennych objaśniających, a zmienną objaśnianą.

Tworząc model regresji liniowej (jedno- lub wielowymiarowej) należy pamiętać, iż dane źródłowe powinny być nie tylko wyrażone w ilościowych skalach pomiarowych, lecz także wolne od tzw. wartości odstających. Analiza i ewentualne eliminacja tych wartości stanowi istotny etap tworzenia modelu regresji.

Tworzenie modelu regresji dla danych źródłowych mających odniesienie przestrzenne powinno być dodatkowo związane z analizą **lokalnej** wartości współczynnika korelacji pomiędzy wartością objaśniającą i objaśnianą. Przykładowo, globalna wartość współczynnika korelacji Pearsona pomiędzy liczbą ludności, a wartością funduszy unijnych pozyskanych w poszczególnych powiatach w Polsce wynosi ok. 0.9. Jednak niezwykle interesująca analiza lokalna tej zależności np. realizowana w tzw. polu podstawowym, np. kole o określonym promieniu). Takie podejście pozwoli na uzyskanie nie jednej wartości charakteryzującej średnią korelację (np. dla całej Polski), a lokalną korelację w danej gminie, powiecie, województwie bądź dowolnej innej jednostce odniesienia.



Rys. 10. Analiza lokalnej korelacji pomiędzy liczbą ludności a wartością dopłat unijnych do projektów realizowanych w poszczególnych powiatach w latach 2007-14. (analiza własna)

Metodyka ta powinna uwzględniać lokalizowane w przestrzeni zjawiska określone w różnych skalach pomiarowych: ilościowej (np. korelacja lokalna pomiędzy dochodami gminy, a poziomem rozwoju sieci drogowej), rangowej (np. korelacja lokalna pomiędzy gęstością zabudowy, a poziomem uciążliwości hałasu), jakościowej (np. korelacja lokalna pomiędzy występowaniem gleby określonego gatunku, a rodzajem roślinności).

Ponadto niezwykle interesujące wyniki może dać budowa modelu regresji odniesionego do danych przestrzennych, który wyjaśniłby charakter poszczególnych zależności. Wizualizacja reszt z regresji dla każdego pola podstawowego pozwala na identyfikację rozkładu przestrzennego zależności pomiędzy badanymi zjawiskami. Obserwując rozkład przestrzenny wielkości reszt z regresji, a następnie dobierając odpowiednio zmienne niezależne tak, aby rozkład przestrzenny reszt był możliwie równomierny, a reszty zbliżone do zera, analityk uzyskuje możliwość budowania modeli regresji, które w większym stopniu tłumaczą zależność przestrzenną danego zjawiska.

Przykładem zastosowania regresji wielowymiarowej jest utworzenie modelu regresji wielorakiej z wykorzystaniem zestawu danych dla województwa kujawsko-pomorskiego opisanych w pierwszej części niniejszego opracowania.

Do budowy modelu regresji objaśniającej liczbę mikro firm z grupy MŚP na osobę wykorzystano zestaw kilku zmiennych objaśniających (analogiczny jak w przypadku analizy klasyfikacyjnej wykorzystującej drzewa CART):

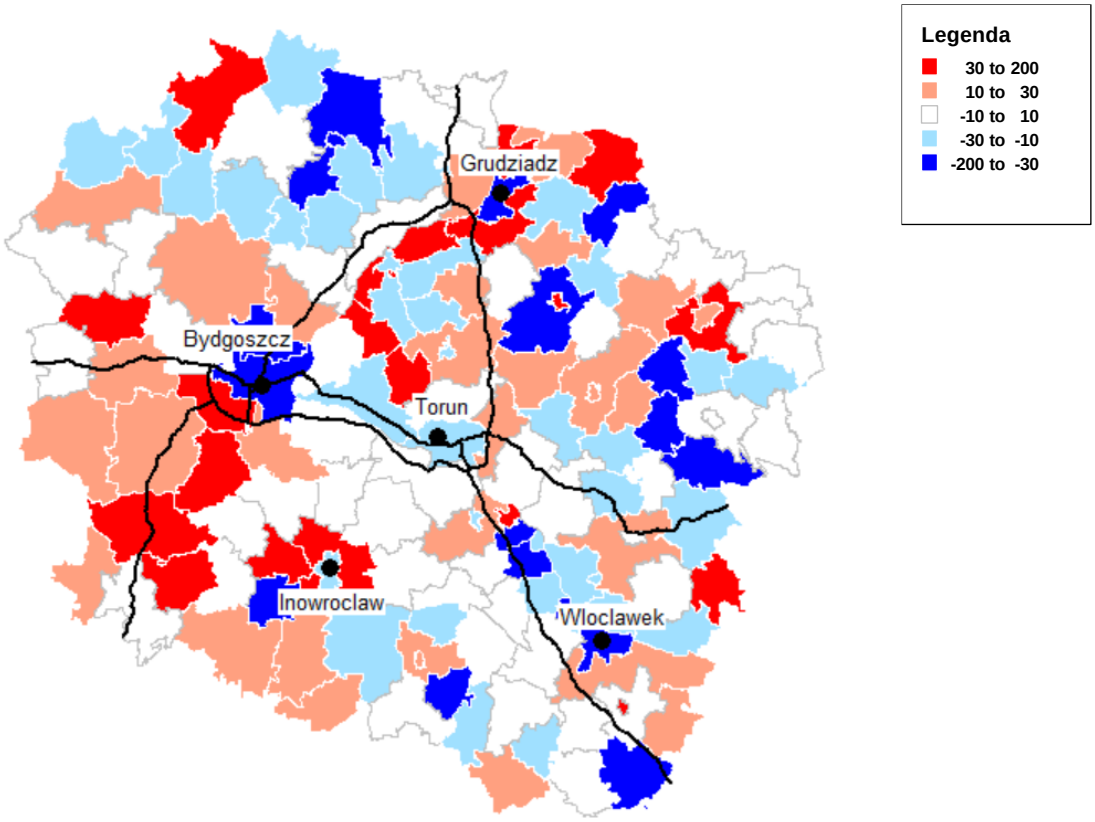
- dochody gmin z podatków
- procent zabudowy w gminach
- odległość od dużych firm, z którymi kooperują MŚP

```
Wyniki regresji wielorakiej

Zmn. zależ Mikro_os          Wielor. R =    .78074565      F = 72.85776
                                     R^2=   .60956377     df =    3,140
Liczba przyp. 144            Popraw. R^2=   .60119728      p = 0.000000
              Błąd standardowy estymacji 26.995868190
Wyr. wolny 77.112131883  Błąd std    8.041573  t(  140) = 9.5892  p =  .0000
```

```
    Ptzb_proc b*=.155      G_dochod_2 b*=.691       O_dist_df b*=-.04
```

N=144	b*	Bł. std z b*	b	Bł. std z b	t(140)	p
W. wolny			77.11213	8.041573	9.58919	0.000000
Ptzb_proc	0.155187	0.058211	0.98899	0.370974	2.66593	0.008581
G_dochod_2	0.691416	0.060340	0.10227	0.008925	11.45857	0.000000
O_dist_df	-0.035264	0.057998	-0.00013	0.000222	-0.60801	0.544163



Rys. 11. Rozkład przestrzenny reszt z regresji dla modelu liniowej regresji wielorakiej objaśniającej liczbę mikro firm na osobę (analiza własna)

Uzyskane wyniki wskazują, iż wynikowy model w 60% objaśnia złożoność zjawiska wykorzystując zaledwie dwie zmienne niezależne – dochody podatkowe oraz procent zabudowy. Podczas budowy modelu regresji wartość charakteryzującą średnią odległość od dużych firm została uznana za nieistotną statystycznie. Wartość błędu standardowego dla tego modelu wynosi 26.6, zaś rozkład przestrzenny reszt z regresji wskazuje na potrzebę

zastosowania modelu regresji ważonej geograficznie. Zagadnienie to zostało opisane w odrębnym opracowaniu Wykonawcy.

Regresja ważona geograficznie (ekonometria przestrzenna)

Zagadnienie regresji ważonej geograficznie zostało szerzej omówione w osobnym opracowaniu Wykonawcy, poniżej podano jedynie podstawowe założenia tej metody.

Klasyczna regresja wielowymiarowa stosowana jest w przypadku gdy nie jest uwzględniany kontekst przestrzenny analizowanych danych, np. model zależności pomiędzy wzrostem a wagą osób dorosłych może być zbudowany poprawnie (aczkolwiek będzie on miał raczej charakter nieliniowy) bez uwzględnienia miejsca zamieszkania poszczególnych osób.

W przypadku, w którym na wartość zmiennej zależnej (np. liczba MŚP w danej gminie) wpływa nie tylko wektor zmiennych niezależnych (liczba mieszkańców tej gminy, współczynnik charakteryzujący poziom bezrobocia w tej gminie itp.), lecz także charakterystyka sąsiadujących jednostek przestrzennych, wymaga zastosowania podejścia zwanego regresją ważoną geograficznie lub ekonometrią przestrzenną.

W literaturze przedmiotu podejście to nazywane jest także „modelem Cliffa i Orda" (1973). Praktyczne wykorzystanie ekonometria przestrzennej stało się możliwe wraz z rozwojem technologii ICT i wzrostem mocy obliczeniowej komputerów.

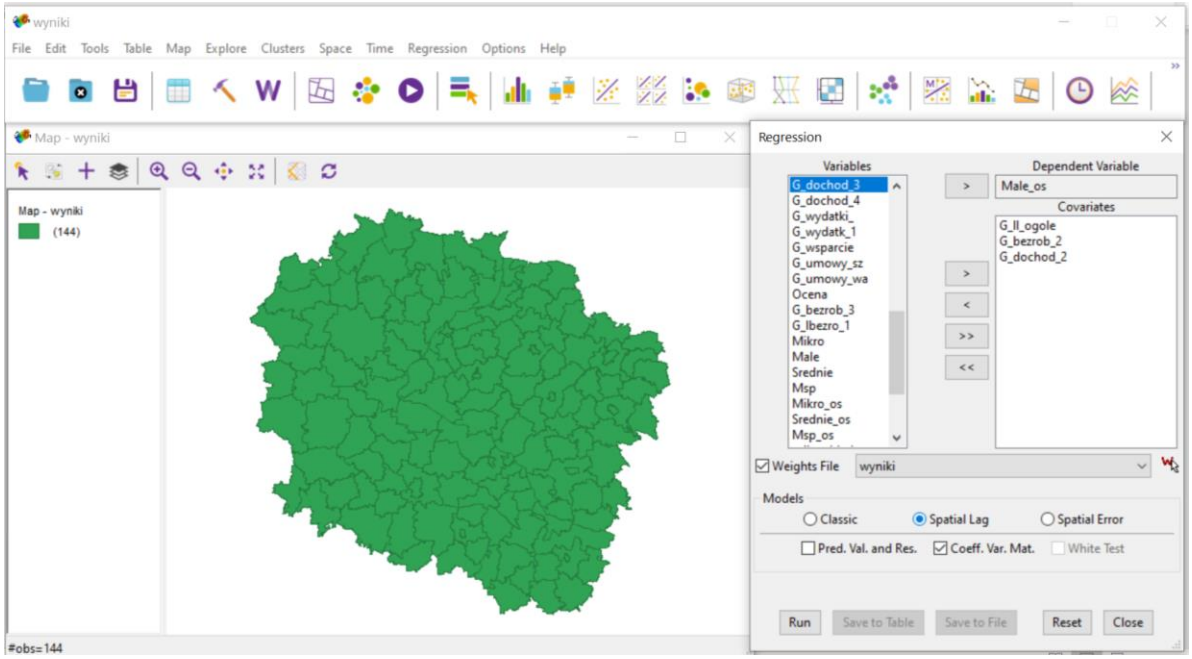
W regresji ważonej geograficznie analizowana jest niezależność przestrzenna reszt z regresji „klasycznych" modeli regresji wielowymiarowej. W przypadku stwierdzenia istotności rozkładu przestrzennego tych reszt wskazane jest opracowanie macierzy sąsiedztwa i użycie jej w budowie modelu ekonometrii przestrzennej. Wynika to z faktu, iż sąsiednie jednostki przestrzenne (np. gminy czy powiaty) mogą mieć wpływ na wartość zmiennej zależnej. Przykładowy rozwój przedsiębiorczości w danej gminie może być warunkowany liczbą ludności, infrastrukturą drogową i liczbą przedsiębiorstw w gminach sąsiednich. Za „sąsiadów" można uznać jednostki przylegające, odległe nie więcej niż X km, mające co najmniej Y km wspólnej granicy itp. Dobór macierzy sąsiedztwa ma kluczowe znaczenie dla budowy modeli regresji ważonej geograficznie. Wiarygodność modeli ekonometrii przestrzennej testowana jest na podstawie zmodyfikowanej wartości współczynnika R^2 obliczanej dla poszczególnych wariantów.

Niezwykle wygodnym narzędziem łączącym zalety GIS i pakietów statystycznych umożliwiającym tworzenie modeli ekonometrii przestrzennej jest GeoDa. Narzędzie to typu open-source można pobrać ze strony: http://geodacenter.github.io/download.html

Stosując narzędzie GeoDa można nie tylko wyznaczyć model ekonometrii przestrzennej typu „spatial lag" oraz „spatial error", lecz przede wszystkim wyznaczyć

istotność poszczególnych zmiennych niezależnych oraz statystyki Morana – zarówno globalną, jak i lokalną.

Zagadnienie parametryzacji modelu regresji ważonej geograficznie zostało szerzej opisane w osobnym opracowaniu Wykonawcy.



Rys. 12. Okno programu GeoDa oraz paraetryzacja modelu regresji ważonej geograficznie

Regresja nieliniowa

Przyjęcie założenia o liniowości zależności pomiędzy zestawem zmiennych niezależnych, a zmienną zależną (objaśnianą) bardzo silnie wpływa na kształt modelu funkcji regresji. Jeżeli analiza korelacji cząstkowych poszczególnych zmiennych objaśniających ze zmienną zależną wskazuje na nieliniowy kształt tej zależności, zasadne jest zastosowanie modelu regresji uogólnionej. Model ten może mieć dowolnie nieliniowy charakter. Estymacja nieliniowa stanowi ogólną procedurę dopasowania, która służy do oszacowania dowolnego rodzaju zależności między zmienną zależną (lub objaśnianą) a zmiennymi niezależnymi. Przykładem tego typu podejścia może być także zastosowanie sztucznych sieci neuronowych, które zostało omówione poniżej.

Estymacja nieliniowa umożliwia zdefiniowanie zasadniczo dowolnego typu ciągłego lub nieciągłego modelu regresji. Najczęściej stosowane są modele nieliniowe takie jak probit, logit oraz wzrostu wykładniczego.

W uogólnionym modelu regresji wyróżnia się dwa podejścia, w których zmienna zależna jest typu binarnego lub ciągłego. Przykładem pierwszego podejścia jest budowa

modelu objaśniającego zależność typu dychotomicznego (np. duży czy też znaczący poziom rozwoju przedsiębiorczości w danej gminie lub niski poziom przedsiębiorczości). Do realizacji zadań tego typu stosowana jest np. regresja logistyczna (logit).

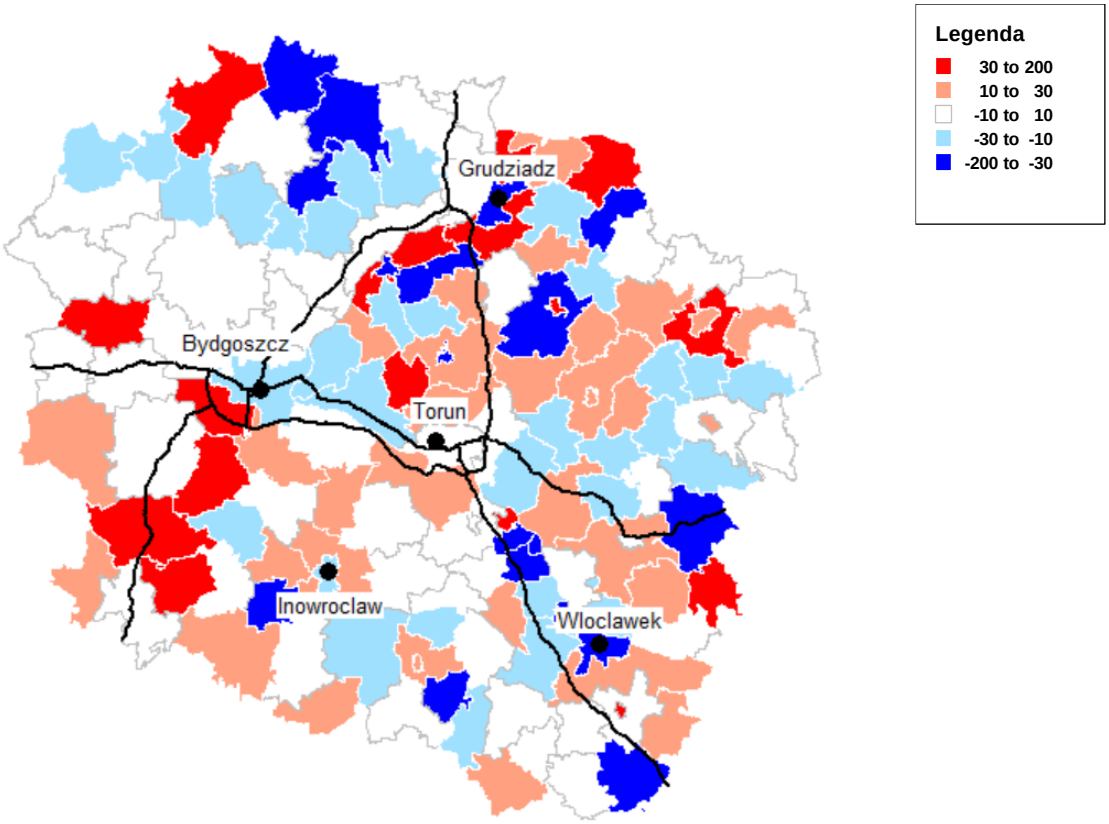
Innym typem modelu regresji uogólnionej jest tzw. regresja fragmentami liniowa (regresja segmentowa). Swego rodzaju uogólnieniem tego podejścia jest model MARS.

Metoda MARS

Metoda MARS (Multivariate Adaptive Regression Splines) wykorzystywana jest do prognozowania lub klasyfikacji zjawisk, które można opisać przy pomocy wielu zmiennych, lub które zależą od wielu zmiennych objaśniających. Metoda ta rozszerza tradycyjne ujęcie zmiennych objaśniających w modelu regresyjnym. Istotą tego podejścia jest uwzględnienie potencjalnych interakcji pomiędzy zmiennymi. W metodzie MARS analizowane są wszystkie obserwacje danej zmiennej objaśniającej i obszar zmienności cechy dzielony jest na przedziały, w których ma ona różny wpływ na badane zjawisko. Pozwala to na opracowanie tzw. krzywych składanych (ang. splines) do estymacji regresji uogólnionej.

W metodzie MARS interakcję zmiennych opisywane są za pomocą iloczynu funkcji bazowych, co pozwala na opracowanie wynikowego modelu regresji. Algorytm MARS wykorzystuje metodę rekurencyjnego podziału przestrzeni cech do budowy modelu regresyjnego w postaci krzywych składanych (Friedman, 1991).

Błąd standardowy dla modelu regresji uogólnionej wykorzystującego algorytmy MARS wynosi 25.3 i jest nieznacznie niższy niż błąd dla modelu liniowej regresji wielorakiej. Zupełnie odmienny jest jednak rozkład przestrzenny reszt z regresji dla tego modelu (Rys. 13), co każe zastanowić się nad koniecznością uwzględnienia w modelach nieparametrycznych regresji uogólnionej także innych zmiennych niezależnych, np. zmiennych o charakterze rangowym i jakościowym, takich jak opracowana w metodą sieci Kohonena zmienna WielGmin.



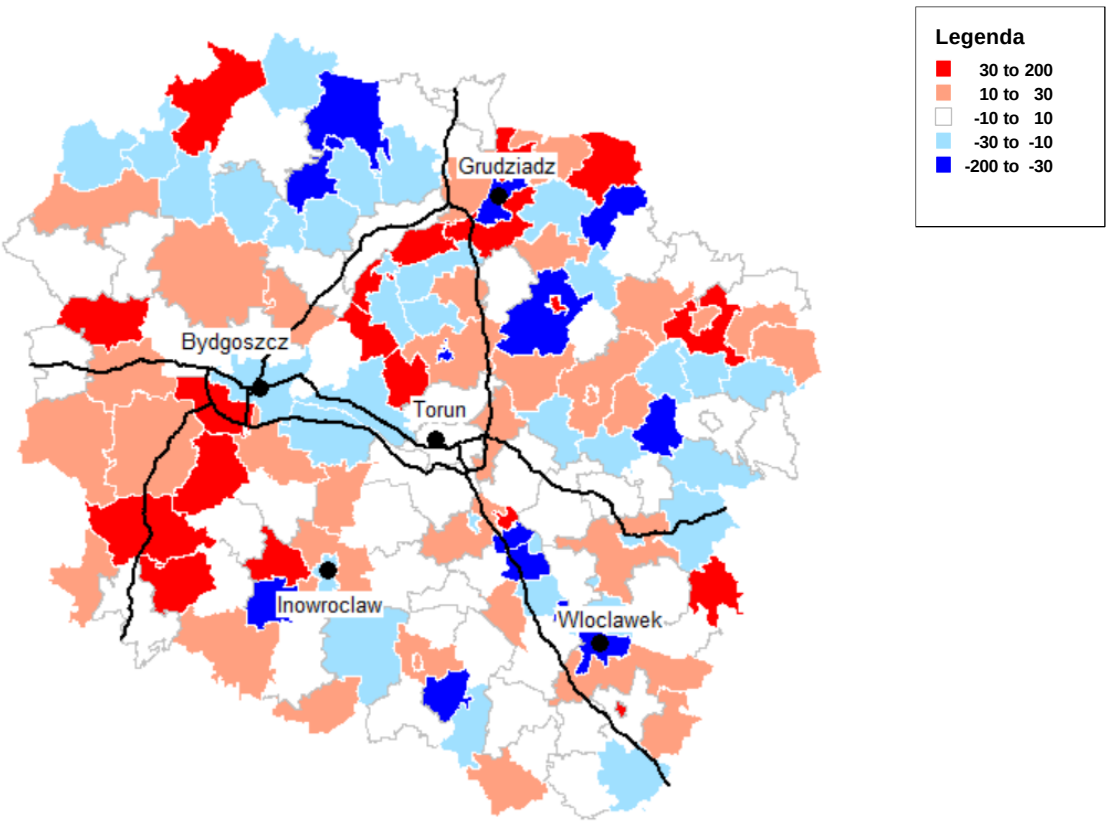
Rys. 13. Rozkład przestrzenny reszt z regresji dla modelu MARS objaśniającego liczbę mikro firm na osobę (analiza własna)

Metoda wektorów nośnych

Metoda wektorów nośnych (Support Vector Machines - SVM) stosowana jest głównie w zagadnieniach o charakterze klasyfikacyjnym, umożliwia bowiem uzyskanie separowalności klas. Podejście to może być jednak stosowane także w zagadnieniach o charakterze regresyjnym. Podobnie jak w innych modelach regresji uogólnionej, metoda SVM umożliwia konstruowanie modelu, w którym zestaw zmiennych niezależnych objaśnia zmienną zależną. W modelu SVM stosowane są w tym celu tzw. funkcje jądrowe różnej postaci: liniowe, wielomianowe, radialne funkcje bazowe (RBF) i funkcje sigmoidalne.

Specyfikacja modelu	**Wartość**
Liczba predyktorów	3
Typ SVM	Regresja typu 1
Typ jądra	Radialne funkcje bazowe (RBF)
Liczba WN	80 (76 związane)

Podsumowanie regresji	Mikro_os
Średnia obserwowanych	144.5180
Średnia przewidywanych	143.4738
Obserwowane odch.std.	42.7482
Odch. std. przewidywanych	32.6523
Błąd średniokwadratowy	663.5448
Średnia błędu	1.0443
Odch. std. błędu	25.8280
Średni błąd bezwzględny	19.6684
Stosunek odch.std.	0.6042
Korelacja	0.7976



Rys. 14. Rozkład przestrzenny reszt z regresji dla modelu SVM objaśniającego liczbę mikro firm na osobę (analiza własna)

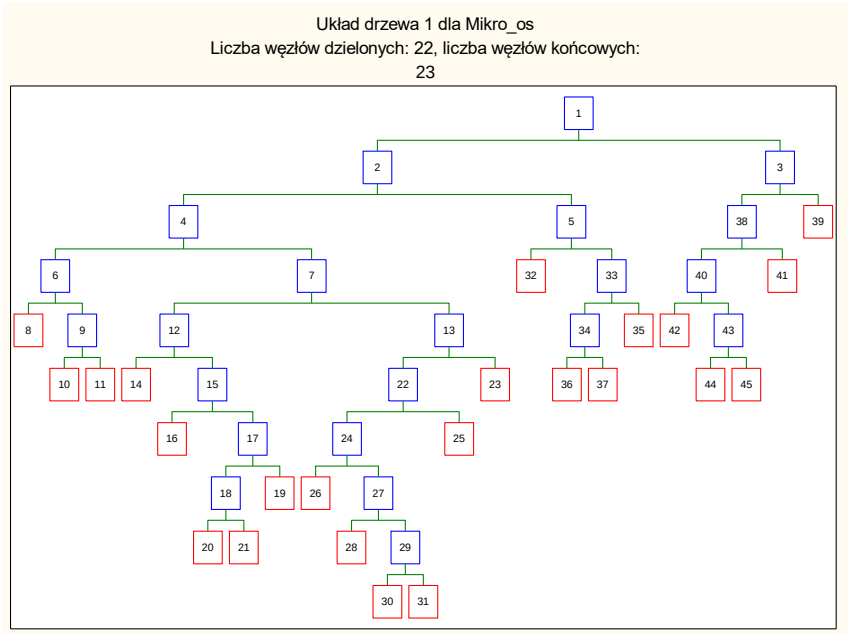
Błąd standardowy dla modelu regresji uogólnionej wykorzystującego algorytmy SVM wynosi 25.7 i jest tylko nieznacznie niższy niż błąd dla modelu liniowej regresji wielorakiej. Zupełnie odmienny jest jednak rozkład przestrzenny reszt z regresji dla tego modelu (Rys. 14). Rozkład ten różni się także istotnie od analogicznego opracowania dla modelu regresji wykorzystującego metodykę MARS.

Drzewa regresyjne

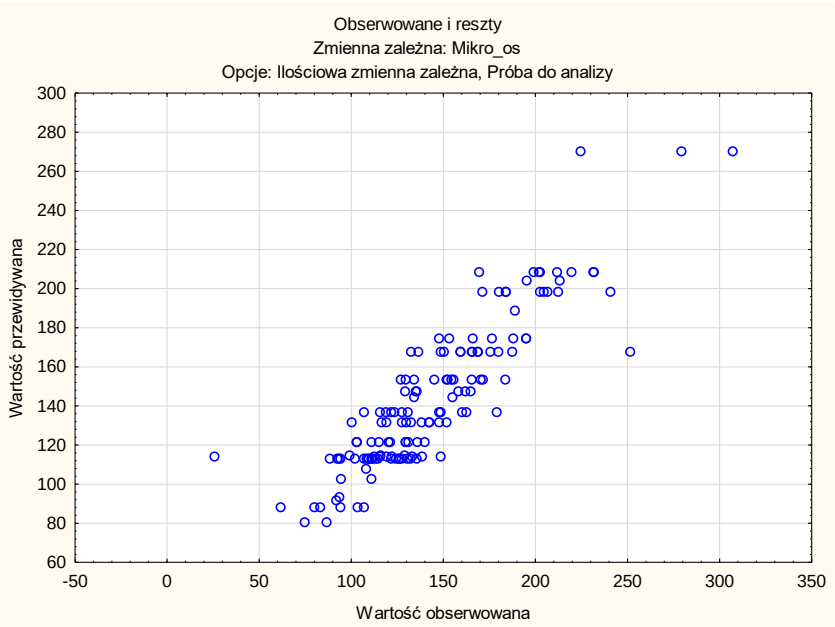
Ogólne zasady działania drzew decyzyjnych typu CART zostały omówione w części opracowania poświęconej zagadnieniom klasyfikacji danych przestrzennych. Metody CART można jednak stosować także do budowy uogólnionych, nieliniowych modeli regresyjnych. Idea drzew opiera się na rekursywnym podziale danych na mniejsze i bardziej homogeniczne grupy w celu jak najlepszego dopasowania. Początkowo zestaw danycy (węzeł macierzysty, korzeń) dzielona jest na dwa podzbiory. Węzeł optymalny wyszukuje się na podstawie wszystkich punktów węzłowych dla każdej zmiennej. Następnie proces jest powtarzany dla każdego węzła potomnego, a te podczas dzielenia traktowane są jak węzły macierzyste.

Drzewa regresyjne - służą do przewidywania wartości zmiennej zależnej, która wyrażona jest w ciągłej skali ilościowej. W podejściu CART, podobnie jak w opisanych poniżej sztucznych sieciach neuronowych, do budowy modelu regresyjnego mogą być wykorzystywane zmienne niezależne dowolnego typu – ilościowe, rangowe i jakościowe.

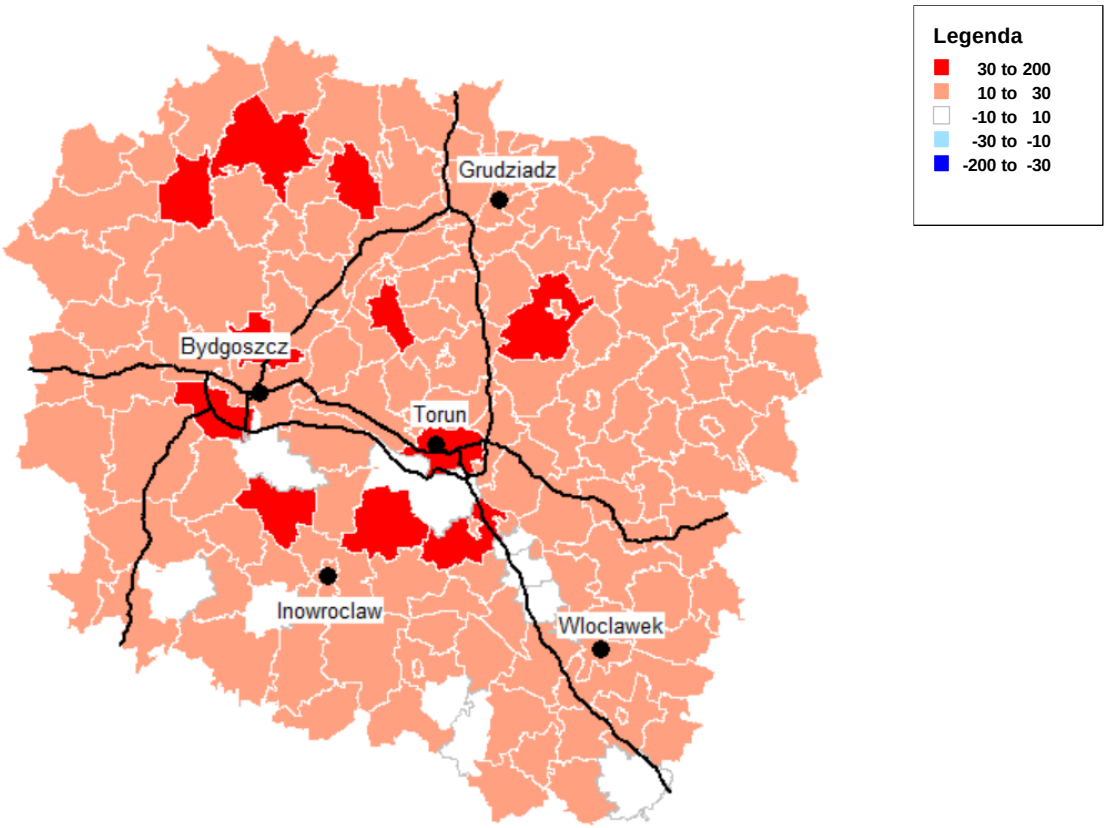
Należy podkreślić, iż metody drzew typu CART są nieparametryczne i nieliniowe. Wyniki wykorzystujące metody drzew regresyjnych dają się ująć w postaci kilku warunków logicznych IF-THEN. Podejście to nie zakłada zatem na wstępie liniowości związku pomiędzy predyktorami, a zmienną zależną. Założenie takie jest natomiast podstawą konstruowania modeli liniowej regresji wielorakiej.



Rys. 15. Schemat drzewa regresyjnego CART objaśniającego liczbę mikro firm na osobę (analiza własna)



Rys. 16. Zestawienie wartości obserwowanych oraz estymowanych dla drzewa regresyjnego CART objaśniającego liczbę mikro firm na osobę (analiza własna)



Rys. 17. Rozkład przestrzenny reszt z regresji dla modelu regresyjnego CART objaśniającego liczbę mikro firm na osobę (analiza własna)

Sztuczne sieci neuronowe

Opisane we wstępnej części tego opracowania sztuczne sieci neuronowe (SSN) stanowią niezwykle efektywne narzędzie tworzenia uogólnionych modeli regresji nieliniowej. Sztuczne sieci neuronowe z definicji posiadają nieliniowy charakter i pozwalają na tworzenie modeli o dowolnie nieliniowych zależnościach pomiędzy zmiennymi niezależnymi, a zmienną zależną. SSN uczą się na przedstawionych przykładach, pozwalają na uogólnienie uzyskanej wiedzy, są relatywnie nieczułe na występowanie wartości odstających, umożliwiają także wykorzystywanie danych wyrażonych nie tylko w ilościowych, lecz także rangowych i jakościowych skalach pomiarowych.

Podejście to posiada jednak także pewne wady:

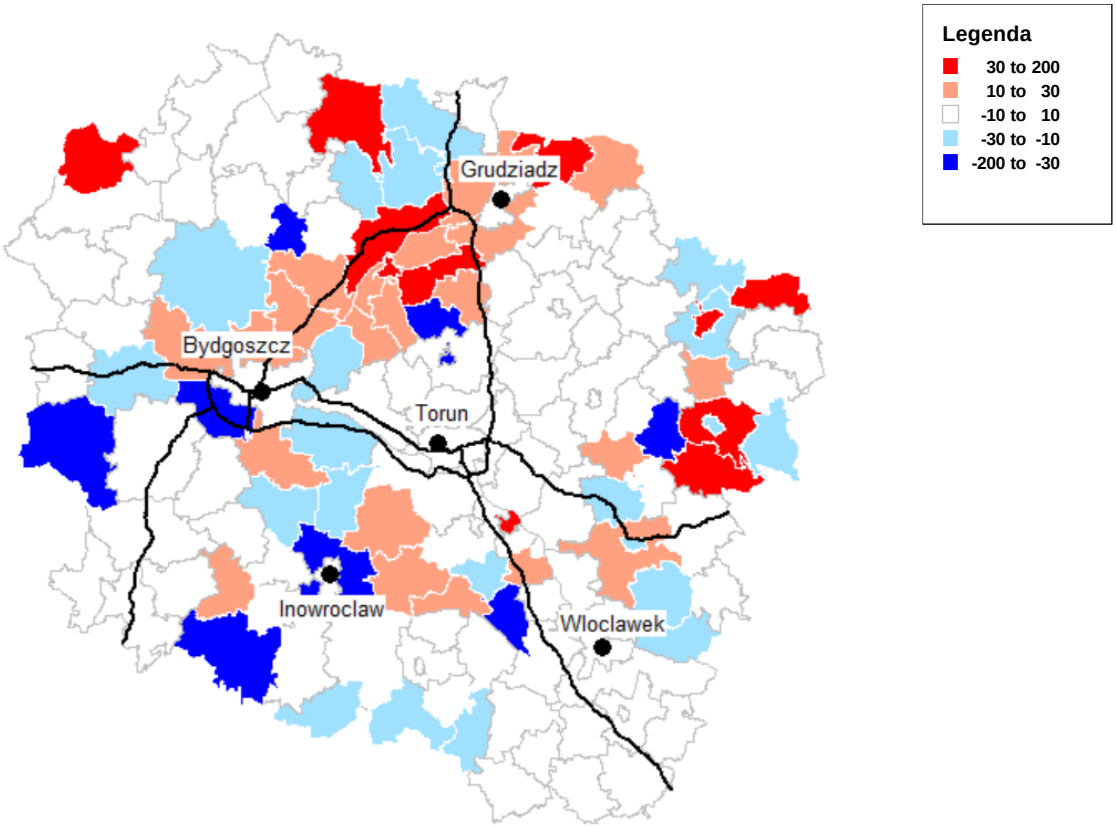
- wymaga dużego zestawu danych uczących (z reguły znacznie liczniejszego niż w przypadku budowy modeli regresji wielorakiej typu liniowego)

- opracowany model stanowi rodzaj tzw. czarnej skrzynki (ang. black box), który nie jest prosty w interpretacji.

Zwłaszcza druga z wymienionych cech SSN stanowi istotny problem. Model regresji uogólnionej zbudowany z wykorzystaniem SSN pozwala z reguły na uzyskanie wyższej wartości współczynnika objaśnienia zależności przez model (odpowiednik zmodyfikowanej wartości R^2 w modelu regresji wielorakiej) w stosunku do regresji liniowej, jednak realizowane jest to kosztem utraty prostoty modelu.

W modelu neuronowym nie sposób stwierdzić, iż np. każdy tysiąc mieszkańców „generuje" jedną firmę typu MŚP, zaś trzykrotny wzrost odległości od miasta wojewódzkiego zmniejsza poziom przedsiębiorczości o połowę. Model tego typu pozwala uzyskać wiarygodną wartość zmiennej zależnej na podstawie przetworzenia zmiennych niezależnych przez setki czy tysiące współczynników wagowych, jednak nie „tłumaczy" wzajemnych zależności zmiennych.

Błąd standardowy dla modelu regresji uogólnionej wykorzystującego sztuczną sieć neuronową typu MLP wynosi 19.5 (Rys. 18) i jest znacząco niższy niż błąd dla modelu liniowej regresji wielorakiej. Dla znacznej części gmin wartość błędu tylko nieznacznie różni się od zera.



Rys. 18. Rozkład przestrzenny reszt z regresji dla modelu wykorzystującego sieć neuronową typu MLP objaśniającego liczbę mikro firm na osobę (analiza własna)

Zestawiając wartości błędów standardowych modeli opracowanych z wykorzystaniem pięciu różnych modeli regresji (tabela poniżej) można stwierdzić, iż najniższa wartość błędu predykcji (najwyższa dokładność modelu) charakteryzuje drzewo regresyjne typu CART oraz siec neuronową typu MLP. Należy jednak pamiętać, iż charakteryzujący się najwyższą wartością błędu model regresji wielorakiej jako jedyny pozwala na prostą interpretację uzyskanych wyników.

Model regresji	**Wartość błędu predykcji**
Liniowa regresja wieloraka	26.6
Metoda wektorów nośnych (SVM)	25.7
Drzewo regresyjne CART	19.3
Metoda MARS	25.3
Sieć neuronowa typu MLP	19.5