Problemy Zarządzania, vol. 10, nr 3 (38): 124–147
ISSN 1644-9584, © Wydział Zarządzania UW
DOI 10.7172.1644-9584.38.8

Sieci neuronowe Kohonena w przeprowadzaniu analiz danych. Próba wykorzystania w eksploracji danych dotyczących jednostek terytorialnych

Mirosława Lasek, Ada Myzik

W artykule przedstawiono wyniki analiz mających na celu wykrywanie zależności zawartych w nagromadzonych danych, ich graficznej reprezentacji i interpretacji za pomocą sieci neuronowych Kohonena. W celu ilustracji analiz danych za pomocą sieci Kohonena posłużono się przykładami analizy danych dotyczących jednostek terytorialnych: punktów gastronomicznych w województwach, podmiotów gospodarczych z różnych sekcji Polskiej Klasyfikacji Działalności w podregionach oraz obiektów sportowych w powiatach. Analiza została przeprowadzona od zbioru najprostszego do najbardziej złożonego, czyli zawierającego najwięcej obserwacji i zmiennych. Na potrzeby analiz danych i ich interpretacji wykorzystywano oprogramowanie firmy SAS Institute Inc.

1. Wstęp

Potrzeba analizy coraz bardziej złożonych zbiorów danych, o coraz większej liczbie obserwacji i zmiennych skłania do poszukiwania metod ich eksploracji wykraczających poza możliwości tradycyjnych metod statystycznych czy ekonometrycznych. Jedną z takich metod jest analiza danych za pomocą sztucznych sieci neuronowych Kohonena.

Działanie sztucznych sieci neuronowych, także sieci Kohonena, ma w uproszczeniu odpowiadać działaniu biologicznych struktur nerwowych złożonych z neuronów. Dają one możliwość tworzenia odwzorowań nieliniowych i wykorzystywania wielowymiarowych danych. Są metodą analizy danych stosunkowo łatwą do zrozumienia i stosowania, chociaż jeszcze niezbyt szeroko znaną i używaną. Na potrzeby analiz danych rozważono możliwość zastosowania klasycznej sieci Kohonena i dwóch metod będących jej modyfikacją: kwantowania wektorowego i uczenia wsadowego.

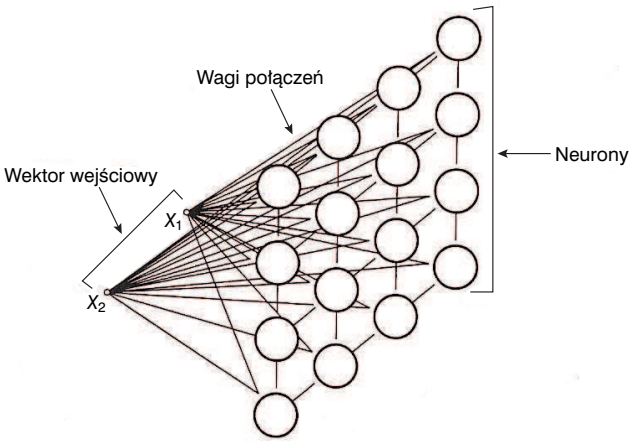
W artykule przedstawiono podstawy działania sieci Kohonena oraz próbę analizy za ich pomocą trzech zbiorów danych, dotyczących jednostek terytorialnych: różnych punktów gastronomicznych w poszczególnych województwach, podmiotów gospodarczych w podziale na sekcje Polskiej Klasyfikacji Działalności w podregionach, różnego rodzaju obiektów sportowych

w powiatach. Wszystkie zbiory dla przeprowadzenia analizy zaczerpnięto ze strony internetowej Głównego Urzędu Statystycznego. Przedstawiono możliwość geograficznej reprezentacji oraz zestawiania wyników uzyskanych na tzw. mapie topologicznej Kohonena z mapą geograficzną.

Dla każdego zbioru danych została wybrana odpowiednia struktura sieci Kohonena i jej parametry, które pozwalają uzyskać najbardziej satysfakcjonujące wyniki. Wykorzystywano oprogramowanie *SAS Enterprise Miner 6.2*, metodę *SOM/Kohonen*. Przedstawiono także mapy tworzone w programie *SAS Base 9.2*, które przedstawiają mapy Polski odpowiadające wynikom uzyskanym za pomocą sieci Kohonena i przedstawianym za pomocą mapy topologicznej Kohonena.

2. Sieci neuronowe Kohonena

W roku 1982 w artykule zatytułowanym „Self-Organized Formation of Topologically Correct Feature Maps" T. Kohonen zaproponował nowy algorytm sztucznych sieci neuronowych, który został nazwany sieciami Kohonena (Kohonen 1982: 59–69). Najkrócej można je scharakteryzować jako sieci samouczące się z wbudowaną konkurencją i mechanizmem sąsiedztwa (Lasek 2004: 17–37). Są to sieci złożone z dwóch warstw neuronów: warstwy wejściowej i warstwy wyjściowej (rysunek 1).



Rys. 1. Struktura sieci neuronowej Kohonena. Źródło: S. Osowski 2006. Sieci neuronowe do przetwarzania informacji, Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, s. 283.

Samouczenie polega na tym, że uczenie, zwane też trenowaniem sieci, odbywa się w trybie „bez nauczyciela" (*unsupervised learning*; *self-organizing*),

co oznacza, że dla podawanych danych wejściowych do treningu nie jest przedstawiana prawidłowa odpowiedź. Sieć nie jest zapoznawana z tym, jakie sygnały wyjściowe powinny odpowiadać wprowadzanym sygnałom wejściowym. Konkurencja jest mechanizmem powodującym, że neurony uczą się rozpoznawania sygnałów wejściowych i reagowania na sygnały wejściowe, konkurując ze sobą. Neuron, który najsilniej zareaguje na dany sygnał wejściowy – im bardziej wagi neuronu są podobne do sygnałów wejściowych (wartości wejściowych), tym silniejsza reakcja – „zwycięża" w konkurencji rozpoznawania określonych sygnałów wejściowych. Inne neurony zostają zwycięzcami w rozpoznawaniu innych sygnałów (wartości) wejściowych. Sąsiedztwo jest tu rozumiane jako takie nauczanie sieci, że neurony sąsiadujące z neuronem zwycięzcą w rozpoznawaniu określonych sygnałów uczą się wraz z nim, chociaż mniej intensywnie. Takie trenowanie sieci powoduje, że sąsiadujące neurony będą reagowały na podobne sygnały (wartości) wejściowe. Wynik trenowania sieci (neurony warstwy wyjściowej) jest przedstawiany na wykresie nazywanym mapą Kohonena lub mapą topologiczną. Poszczególne obserwacje nazywane są przypadkami wejściowymi lub uczącymi.

Struktura sieci Kohonena nie jest skomplikowana w porównaniu z innymi rodzajami sieci neuronowych (Myzik 2012). Sieć Kohonena składa się bowiem z warstwy wejściowej i wyjściowej, natomiast nie posiada żadnych warstw ukrytych, tak jak inne rodzaje sieci. Wymagane jest, aby dane przedstawione na wejściu sieci zostały uprzednio znormalizowane lub wystandaryzowane. S. Osowski wskazuje na konieczność nadmiaru danych wejściowych, aby metoda uczenia poprzez samoorganizację mogła wykryć istotne wzorce zawarte w danych (Osowski 2006: 282). Warstwa wyjściowa złożona jest z neuronów, które nazywane są też wektorami wagowymi lub kodowymi. Na ogół neurony przedstawiane są w sposób, który ułatwia interpretację wyników uczenia sieci (Myzik 2012: 8–12). Popularną metodą odwzorowania na płaszczyźnie warstwy wyjściowej jest nadanie jej formy dwuwymiarowej siatki składającej się z elementów (prostokątów, kółek – w zależności od oprogramowania), które odpowiadają poszczególnym neuronom.

Sieć Kohonena jest zaliczana do tzw. sieci pełnych, gdyż każdy przypadek uczący z warstwy wejściowej jest połączony z każdym neuronem z warstwy wyjściowej i nie ma powiązań między elementami tej samej warstwy. Połączeniom są przyporządkowane wagi o wartościach z przedziału [0, 1] (Larose 2006: 168–169). Wagom są nadawane wartości początkowe za pomocą różnych algorytmów. Wartości te mogą być ustalone jako losowe lub wyznaczone na podstawie algorytmu dwóch pierwszych głównych składowych.

Uczenie czy też trenowanie sieci Kohonena przebiega w sposób iteracyjny. Tzw. epoka uczenia (trenowania) sieci kończy się po zapoznaniu sieci ze wszystkimi przypadkami uczącymi i składa się z wielu pojedynczych kroków. Krok obejmuje analizę jednego przypadku uczącego, rywalizację poszczególnych neuronów o tytuł neuronu wygrywającego, a następnie modyfikację wektora wagowego zwycięskiego neuronu oraz neuronów z nim

sąsiadujących. Neuronem zwycięskim lub wygrywającym nazywany jest neuron, dla którego wartość funkcji decyzyjnej, a jest nią zazwyczaj odległość euklidesowa pomiędzy wektorem wagowym a przypadkiem wejściowym, przyjmuje najlepszą (w przypadku odległości euklidesowej – najmniejszą) wartość. Następnie wagi neuronu wygrywającego i jego sąsiadów są modyfikowane, tak aby były jeszcze bardziej podobne do przypadku uczącego.

W procedurze modyfikacji stosowany jest tzw. współczynnik uczenia, który maleje wraz ze wzrostem liczby kroków, aby zmiany na początku były duże (gwałtowne), a w kolejnych krokach uczenia neuronów coraz mniejsze (Larose 2006: 169–170). Drugim współczynnikiem, którego wartość jest zmniejszana wraz z kolejnymi krokami algorytmu, jest promień sąsiedztwa, tak aby początkowo wartości wag neuronów sąsiadujących z neuronem zwycięzcą szybko (gwałtownie) się dostosowywały do zwycięzcy, a następnie stopniowo zmiany stawały się coraz mniejsze. Jednakże uwzględnienie podczas modyfikacji wag neuronów zachodzenia sąsiedztwa powoduje, że neurony położone obok siebie na mapie topologicznej są często do siebie podobne.

Cały algorytm trenowania sieci Kohonena można uporządkować w proces złożony z następujących kroków (Larose 2006: 170; Myzik 2012: 10–11):

1) inicjalizacja początkowych wektorów wagowych, np. w sposób losowy lub opierając się na dwóch pierwszych składowych głównych;
2) wybranie przypadku uczącego (obserwacji);
3) obliczenie wartości funkcji decyzyjnej dla wszystkich neuronów i wybranie neuronu wygrywającego;
4) określenie neuronów sąsiadujących z neuronem zwycięskim na podstawie wartości funkcji sąsiedztwa;
5) dostosowanie wag neuronów sąsiadujących przy wykorzystaniu współczynnika uczenia (tzw. adaptacja);
6) modyfikacja współczynnika uczenia i rozmiaru sąsiedztwa;
7) powrót do realizacji punktu 2, jeżeli nie zostały spełnione warunki zakończenia uczenia sieci.

Największą zaletą sieci neuronowych, także Kohonena, jest umożliwienie przedstawiania zależności nieliniowych oraz rozwiązywanie problemów, dla których nie umiemy dokładnie zdefiniować zależności przyczynowo-skutkowych.

R. Tadeusiewicz (2001: 48–49) zwraca uwagę, że sieć Kohonena potrafi wykrywać powiązania, które zostałyby pominięte, gdyby zastosowano tradycyjną sieć i sposób uczenia (tj. uczenie z „nauczycielem", gdy w danych wejściowych do uczenia sieci zawarte są poprawne odpowiedzi).

R.S. Collica (2007: 212) w odniesieniu do praktycznej strony zastosowania sieci Kohonena wśród zalet wymienia łatwość zrozumienia algorytmu, jego logiczną strukturę, natomiast za główne zagrożenia stosowania algorytmu uważa jego problemy w radzeniu sobie z brakującymi obserwacjami, duże zapotrzebowanie na moc obliczeniową oraz zmienne rezultaty w zależności od wielkości próby, która zostanie wykorzystana w badaniu.

W artykule M. Lasek jako podstawowe wady sieci Kohonena wymienia się (Lasek 2004: 26–27):

– brak reguł budowy architektury sieci, takich jak określenie liczby neuronów czy wymiaru sieci, oraz trudności wyznaczania parametrów trenowania, takich jak wartości początkowe wag, współczynnik uczenia, promień sąsiedztwa, porządek prezentowania danych w procesie uczenia, które są znajdywane metodą prób i błędów; nieodpowiedni dobór parametrów może być przyczyną braku sukcesu w trenowaniu sieci;
– wynik trenowania sieci zależy od porządku, w jaki pobierane są przykłady do trenowania, ponieważ wagi neuronów są modyfikowane w sposób wskazany przez algorytm po zaprezentowaniu każdego kolejnego przykładu;
– zbieżność procesu uczenia nie opiera się na kryterium optymalizacji, lecz jest wymuszana zmniejszającym się współczynnikiem uczenia i zasięgiem uczenia; dlatego też nie ma gwarancji, że sieć może nauczyć się prawidłowo rozpoznawać sygnały wejściowe oraz że będzie przedstawiać w pełni złożoność problemu.

Aby przezwyciężyć wymienione powyżej wady, podjęto próby polepszenia metody trenowania sieci Kohonena, m.in. poprzez włączenie algorytmu rozmytych k-średnich, gdzie „rozmytość" jest rozumiana zgodnie z teorią zbiorów rozmytych L.A. Zadeha (Lasek 22–24, 27). Dzięki włączeniu algorytmu rozmytych k-średnich uzyskano szereg korzyści, jak np. większą odporność wyników na początkowe wartości wprowadzanych parametrów, uniezależnienie od kolejności wprowadzania przykładów do trenowania sieci, zmniejszenie koniecznej liczby cykli trenowania (Lasek 2007: 113–134). Jednakże opisywana wcześniej dla przypadku nierozmytej sieci Kohonena adaptacja sąsiedztwa na mapie Kohonena nie ma już miejsca. Neurony reprezentujące podobne obserwacje (dane wejściowe) nie muszą sąsiadować ze sobą na mapie topologicznej Kohonena. Zalety wizualne mapy zwykłych sieci Kohonena zostają utracone przy wykorzystywaniu rozmytego algorytmu trenowania.

Uproszczoną wersją algorytmu Kohonena jest tzw. kwantowanie wektorowe (*Vector Quantization – VQ*). W przypadku tego algorytmu zbiór danych zostaje podzielony na zdefiniowaną liczbę skupień, tak że każde z nich jest reprezentowane przez jeden wektor wagowy. Kwantowanie wektorowe nie wykorzystuje koncepcji sąsiedztwa. Modyfikowane są jedynie wagi neuronów zwycięzców. Przy zastosowaniu takiego algorytmu końcowe rezultaty są w znacznej mierze uzależnione od początkowych wartości wektorów wagowych, a algorytm cechuje się skłonnością do wybierania minimów lokalnych zamiast globalnych (Kohonen 2008: 310–311).

Inną modyfikacją algorytmu Kohonena jest tzw. uczenie wsadowe sieci. W przeciwieństwie do oryginalnego algorytmu Kohonena, wektory wagowe w przypadku uczenia wsadowego są dostosowywane po wczytaniu nie kolejno rozpatrywanych obserwacji, ale całego zbioru danych. Po każdym przypadku uczącym zapamiętywane są informacje o wektorze zwycięskim i parametrze

sąsiedztwa, a na koniec całej epoki uczenia na podstawie tych danych są obliczane nowe wektory wagowe (Kohonen 1998: 3–4). Następnie przy zastosowaniu nowych wektorów wagowych ponownie analizowane są wszystkie przypadki wejściowe, aż do spełnienia warunków zaprzestania uczenia sieci. Obliczanie wektorów wagowych może się odbywać na podstawie wartości nieparametrycznej funkcji regresji (SAS 2010). Metoda ta jest szybsza w działaniu, bardziej stabilna i nie wymaga stosowania współczynnika uczenia.

Utrudnieniem stosowania algorytmu Kohonena może być pojawianie się, i to już w trakcie uczenia, tzw. martwych neuronów, tj. takich, które nie reprezentują żadnych danych wejściowych. Przyczyną ich pojawiania się może być losowość wag początkowych lub zbyt mały rozmiar sąsiedztwa. Pojawianie się martwych neuronów na mapie topologicznej Kohonena zwiększa tzw. błąd kwantyzacji, mierzony jako średnia wartość odległości wyliczanych dla każdego przypadku wejściowego do najbardziej podobnego do niego reprezentanta na mapie (można zastąpić średnią inną miarą, np. medianą) i niekorzystnie wpływa na możliwości interpretowania wyników, gdyż dane będą odzwierciedlone przez mniejszą liczbę neuronów (Osowski 2006: 286).

Nauczona sieć Kohonena może zostać zastosowana nie tylko do analizy zależności występujących w zbiorze danych wejściowych, które są wykorzystywane do jej trenowania, czego próbę przedstawiamy w tym artykule, lecz także przeznaczona do klasyfikowania nowych obiektów, tzn. takich, z którymi sieć nie była jeszcze zapoznawana. Jeżeli przedstawiona wyuczonej sieci nowa obserwacja nie zostaje przydzielona do żadnej z grup, wskazuje to, że jest odmienna od obserwacji, na których przeprowadzono jej trenowanie.

3. Budowa i wykorzystywane sieci Kohonena przy zastosowaniu programu SAS Enterprise Miner (procedura SOM/Kohonen)

SAS Enterprise Miner 6.2 jest komponentem pakietu *SAS*, opracowanym specjalnie na potrzeby przeprowadzania eksploracji danych (*Data Mining*). Jest wyposażony w intuicyjny, graficzny interfejs, ułatwiający i znacznie przyspieszający budowę modeli eksploracji danych. Prostokąty odpowiednio oznaczone (tekst i grafika), reprezentujące działania (tzw. węzły) dla przeprowadzania kolejnych kroków analizy danych, użytkownicy *SAS Enterprise Miner* wprowadzają na diagram i łączą strzałkami wskazującymi kolejność przetwarzania. Przetwarzanie odbywa się zgodnie z założeniami metodyki *SEMMA* w toku realizacji pięciu kolejnych etapów: próbkowanie, eksploracja, modyfikacja, modelowanie i ocena (*Sampling*, *Exploration*, *Modification*, *Modeling*, *Assessment*) (Lasek i Pęczkowski 2010: 117–133). Dla zbudowania i wykorzystywania sieci neuronowych Kohonena trzeba wykorzystać węzeł (narzędzie) *SOM/Kohonen*.

Możliwe jest zastosowanie różnych algorytmów tworzenia modelu sieci Kohonena: kwantowania wektorowego (*Kohonen VQ*), samoorganizującej

się mapy Kohonena (*Kohonen SOM*) oraz uczenia wsadowego *SOM* (*Batch SOM*) z możliwością wyboru wygładzania *Nadaraya-Watsona* (*Nadaraya-Watson smoothing*) i wygładzania lokalnego liniowego (*local-linear smoothing*). Ponadto węzeł *SOM/Kohonen* umożliwia określenie wielu szczegółowych parametrów (opcji programu), które umożliwiają dostosowanie modelu do potrzeb analizowanych danych. Parametry te w programie *SAS Enterprise Miner 6.2* zostały podzielone na cztery podstawowe sekcje: *Ogólne*, *Uczenie*, *Ocena punktowa*, *Status*. Szczegółowo są opisane w dokumentacji programu, a także w pracy A. Myzik (2012: 18–24).

Podstawowe opcje określające sposób budowy modelu i jego wykorzystywania zawiera sekcja *Uczenie*. W tej sekcji można wyselekcjonować zmienne do badania, określić metodę budowy sieci oraz wskazać sposób standaryzacji lub zadecydować, że nie będzie przeprowadzana. Sekcja *Uczenie* zawiera aż dziesięć podsekcji: *Segment*, *Opcje ziarna*, *Uczenie wsadowe SOM*, *Opcje lokalne liniowe*, *Opcje Nadaraya-Watsona*, *Kohonen VQ*, *Kohonen*, *Opcje sąsiedztwa*, *Kodowanie zmiennych klasyfikujących* i *Braki danych*. Szkolenie sieci można przeprowadzić, przyjmując domyślne, proponowane w programie wartości parametrów, jednakże w takim przypadku możemy otrzymać rezultaty w znacznym stopniu odbiegające od oczekiwanych. Trzeba też pamiętać, że wiele opcji jest dostępnych tylko dla konkretnych metod (*Kohonen VQ*, *Kohonen SOM* lub *Batch SOM*), dlatego podstawową rolę odgrywa wybór metody.

W programie *SAS Enterprise Miner* stosowane są pojęcia krok oraz iteracja. Termin „krok" oznacza działanie polegające na zapoznaniu sieci z jednym przypadkiem wejściowym (obserwacją) i zaktualizowaniu na tej podstawie wektorów wagowych. „Iteracja", podobnie jak epoka, jest odnoszona do zapoznania sieci z całym dostępnym zbiorem danych.

Metodą domyślną wskazywaną w programie jest wsadowe *SOM* (*Batch SOM*). W przypadku uczenia wsadowego dla właściwego zbudowania modelu istotne jest wskazanie właściwej liczby kolumn i wierszy (podsekcja *Segment*), które definiują rozmiar mapy topologicznej Kohonena oraz określenie rozmiaru sąsiedztwa (podsekcja *Opcje sąsiedztwa*). Rozmiar mapy topologicznej i rozmiar sąsiedztwa powinny zostać utrzymane w odpowiedniej proporcji, tak że np. dwukrotne zwiększenie liczby kolumn i wierszy powinno narzucać dwukrotne zwiększenie rozmiaru sąsiedztwa. Zachowanie tej proporcji jest istotne, ponieważ od tych wartości zależy poziom wygładzania. Dla przyjęcia odpowiednich wartości najczęściej niezbędne jest przeprowadzenie wielu prób.

R. Matignon wskazuje, że jeżeli przyjmiemy zbyt małą mapę topologiczną, to nie uzyskamy dobrego odwzorowania nieliniowości zawartych w danych, natomiast zbyt duży rozmiar spowoduje wydłużenie czasu uczenia sieci, wymaga większej mocy obliczeniowej komputera oraz może powodować występowanie pustych segmentów na mapie. Stwierdza, że lepiej sprawdzają się większe mapy, o ile każdemu skupieniu odpowiada wystarczająco

duża liczba przypadków wejściowych (Matignon 2007: 230). W dokumentacji programu *SAS Enterprise Miner 6.2* wskazuje się, że powinno ich być pięć lub dziesięć. Tylko dla przypadku uczenia wsadowego jest możliwe wskazanie opcji wygładzania i wyznaczanie ich parametrów. Program *SAS Enterprise Miner 6.2* umożliwia zastosowanie opcji wygładzania lokalnego liniowego i wygładzania *Nadaraya-Watsona*, które regulują formę funkcji sąsiedztwa. Ich wykorzystanie daje możliwość odpowiedniego dostosowania parametru sąsiedztwa i obszaru wokół neuronu wygrywającego (Matignon 2007: 235). Wygładzanie *Nadaraya-Watsona* zmniejsza prawdopodobieństwo, że algorytm zatrzyma się w minimum lokalnym. Wygładzanie lokalne liniowe natomiast umożliwia wyeliminowanie tzw. efektu granicznego (*border effect*) (Matignon 2007: 230), powodującego, że wektory wagowe z obrzeży mapy topologicznej są ściągane do jej centrum.

Wybór metody *Kohonen SOM* wymaga spełnienia tych samych zasad dotyczących liczby wierszy oraz kolumn mapy i rozmiaru sąsiedztwa, jak w przypadku metody wsadowej *Batch SOM*. W przypadku *Kohonen SOM* trzeba dodatkowo ustalić właściwy poziom współczynnika uczenia (poziom domyślny wskazywany w programie wynosi 0,9). R. Matignon proponuje wybranie wysokiej początkowej wartości, zwłaszcza w przypadku gdy początkowe wektory wagowe zostały wyznaczone w sposób losowy. Natomiast, jeżeli wektory wagowe zostały ustalone w oparciu o wcześniejsze analizy przed przystąpieniem do uczenia sieci, stwierdza, że zasadne może być przyjęcie niższej wstępnej wartości współczynnika uczenia. Niezbędne jest też przyjęcie odpowiednich kryteriów zakończenia uczenia lub pozostawienie wartości domyślnych. Program daje możliwość wskazania końcowego poziomu współczynnika uczenia, maksymalnej liczby kroków, maksymalnej liczby iteracji i kryterium zbieżności.

Jeżeli zostanie wybrana metoda *VQ*, wymagane jest określenie maksymalnej liczby skupień oraz współczynnika uczenia. Zasada wyboru parametru uczenia jest taka sama jak w przypadku *SOM Kohonen*. Jednak dla kwantowania wektorowego już wartość 0,5 może być uznana za wystarczająco wysoki poziom początkowy. Liczba skupień może przyjmować dowolne wartości dodatnie, a jedyną wskazówkę doboru właściwej ich liczby może dać metoda prób i błędów.

Należy podkreślić, że bardzo istotny wpływ na działanie algorytmu ma właściwy dobór parametrów sąsiedztwa. Nie dotyczy to metody Kohonen VQ, ponieważ nie stosuje się tu koncepcji sąsiedztwa.

Podstawowym atrybutem sąsiedztwa jest jego docelowy rozmiar, który w programie jest jako domyślny, przyjęty według wzoru: *max*(*5*, *max(liczba wierszy, liczba kolumn)*/*2*. Parametr ten może przyjmować dowolną wartość ze zbioru liczb całkowitych dodatnich. Jednakowoż istotnym ograniczeniem możliwej do przyjęcia wartości jest zależność pomiędzy docelowym rozmiarem sąsiedztwa a liczbą wierszy i kolumn mapy topologicznej. W dokumentacji programu *SAS* podaje się, że wybór końcowej wielkości

sąsiedztwa odgrywa decydującą rolę w trenowaniu mapy samoorganizującej. Drugim istotnym parametrem jest parametr, który określa postać sąsiedztwa, w programie nazywany kształtem jądra. Może ono przyjmować następujące formy: *jednostajne*, *Epanechnikov*, *dwuwagowe* (*biweight*), *trójkątne* (*triweight*). Kolejny istotny parametr nosi nazwę metryki jądra i przedstawia miarę odległości neuronów. Na podstawie metryki jądra dostosowywane są poszczególne wektory wagowe do neuronu zwycięskiego oraz określana jest odległość poszczególnych ziaren. Stosowanymi metrykami są m.in. odległość maksymalna, miejska, euklidesowa. Dodatkowe parametry sąsiedztwa stanowią wartości, które pozwalają regulować przebieg budowy mapy, np. regulować tempo zmniejszania parametru sąsiedztwa do końcowej wartości.

4. Zastosowanie sieci Kohonena na potrzeby analizy danych o punktach gastronomicznych w województwach

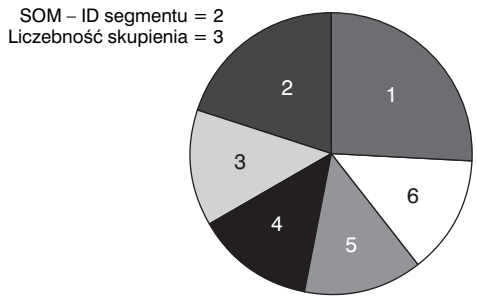
Dane o punktach gastronomicznych w województwach pobrano ze strony internetowej Głównego Urzędu Statystycznego (dane z 2009 r.). Zbiór danych składa się z szesnastu obserwacji i czterech zmiennych, które zostaną wykorzystane do podziału obserwacji na skupienia. Każda obserwacja (przypadek wejściowy) przedstawia informacje dotyczące jednego województwa na temat liczebności punktów gastronomicznych czterech typów: restauracji, barów, stołówek, pozostałych. Ten zbiór danych wybrano ze względu na jego prostotę, co powoduje, że stanowi dobrą ilustrację sposobu działania metody *SOM/Kohonen* oraz dobrą podstawę dla przeprowadzenia analizy wpływu zmiany parametrów modelu na wyniki. Ponadto analiza przeprowadzona dla mało złożonego zbioru może być pomocna przy budowie modelu, gdy będą wykorzystywane bardziej złożone dane.

Już wstępna analiza danych wskazuje wyróżniającą się obserwację dla województwa mazowieckiego. Niezależnie od tego, jaką wybierzemy metodę badania (jaki model sieci Kohonena), poszczególne parametry, architekturę sieci, algorytm (*Kohonen VQ*, czy też samoorganizująca się mapę Kohonena), obserwacja dla województwa mazowieckiego tworzy odrębne skupienie. Ponadto ten jednoelementowy segment jest zawsze znacząco oddalony od pozostałych skupień. W Mazowieckiem jest ponad dwukrotnie więcej punktów gastronomicznych niż w jakimkolwiek innym województwie. Odmienność tego przypadku wejściowego od pozostałych powoduje w budowanej mapie topologicznej powstawanie pustych skupień i wybór wektorów wzorcowych o ujemnych wartościach. Na podobne zjawisko wskazuje G. Tarczyński w pracy, w której analizował jednostki terytorialne (powiaty) pod względem rozwoju gospodarczego (Tarczyński 2011: 90–94). W dalszej analizie tymczasowo usunięto, zgodnie z sugestią G. Tarczyńskiego, obserwację dla województwa mazowieckiego, aby sprawdzić, czy sieć stanie się wówczas bardziej wrażliwa na subtelniejsze zależności.

W przypadku analizowanego zbioru danych potwierdziła się zależność wskazywana przez S. Osowskiego, orzekająca, że sieć Kohonena wymaga nadmiaru danych do wykrywania wzorców w nich zawartych (Osowski 2010: 282).

Dla analizowanego przez nas niewielkiego zbioru danych, niezależnie od dobieranej konfiguracji, mapa samoorganizująca się nie sprawdza się najlepiej. Wiele skupień zawiera po jednej obserwacji. Ograniczanie liczby grup przekłada się na gorsze odwzorowanie zróżnicowania segmentów mapy, a ponadto pojawia się problem martwych neuronów. Lepsze efekty pozwala uzyskać kwantowanie wektorowe – zgodnie z terminologią przyjętą w oprogramowaniu *SAS Enterprise Miner*: metoda *Kohonen VQ*. Założono utworzenie sześciu skupień. Mniejsza liczba sprawia, że skupienia nie różnią się co do rodzaju punktów gastronomicznych, a jedynie co do ich liczby. Przyjęcie większej liczby powoduje z kolei nadmierne rozdrobnienie zbioru danych i powstawanie licznych pustych segmentów. Jako metodę standaryzacji danych przyjęto przeprowadzenie normalizacji. Początkowe wartości neuronów wybrano, opierając się na metodzie składowych głównych. Podstawowa analiza wykazała, że dwie pierwsze składowe główne wyjaśniają niemal 97% zmienności, wykres osypiska także sugeruje wybór dwóch składowych. Pozostałe parametry zostały przyjęte jako domyślne, zgodnie z sugestią wykorzystywanego zaprogramowanego algorytmu.

Metoda *Kohonen VQ* dzieli zbiór danych na siedem skupień, jeżeli uwzględnimy wyodrębnione wcześniej województwo mazowieckie. Rysunek 2 przedstawia zbiór piętnastu województw w rozbiciu na sześć zbiorów. Tabela 1 zawiera statystyki dla utworzonych skupień. Rysunek 3 przedstawia wizualizację na mapie Polski uzyskanego podziału na grupy za pomocą kwantowania wektorowego.

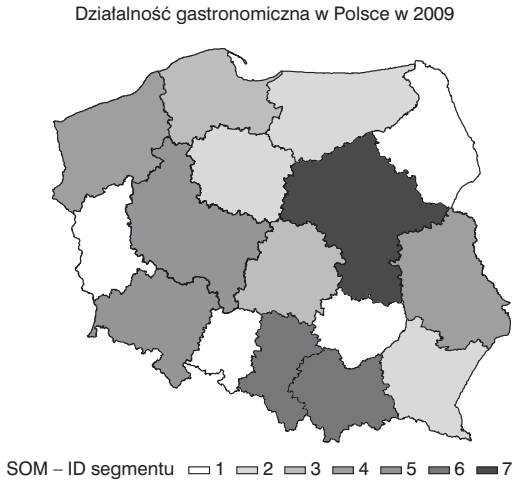


Rys. 2. Podział województw na skupienia według liczebności punktów gastronomicznych. Źródło: opracowanie własne.

Dane zamieszczone w tabeli 1 wskazują, że najbardziej jednorodne jest skupienie pierwsze, w przypadku którego odchylenie standardowe wynosi 0,04, a największa odległość od ziarna skupienia jest równa 0,08. Najmniej homogeniczne jest skupienie czwarte. Istotne jest, że w żadnym przypadku odległość od najbliższego skupienia nie jest mniejsza niż największa odległość od ziarna skupienia, co potwierdza rozdzielność utworzonych grup.

Lp.	Segment	Liczebność	Odchylenie standardowe	Największa odległość od ziarna skupienia	Najbliższe skupienie	Odległość od najbliższego skupienia
1.	1	4	0,04	0,08	2	0,26
2.	2	3	0,09	0,17	1	0,26
3.	3	2	0,10	0,14	2	0,29
4.	4	2	0,15	0,21	2	0,66
5.	5	2	0,14	0,20	3	0,50
6.	6	2	0,09	0,14	5	0,74

Tab. 1. Statystyki dla skupień utworzonych na podstawie danych dotyczących punktów gastronomicznych w województwach. Źródło: opracowanie własne.



Rys. 3. Mapa Polski podzielona na województwa z naniesionymi skupieniami utworzonymi na podstawie danych dotyczących punktów gastronomicznych. Źródło: opracowanie własne.

Lp.	Segment	Restauracje	Bary	Stołówki	Pozostałe
1.	1	92,3	94,5	47,3	45,8
2.	2	156,7	137,7	75,0	71,7
3.	3	230,0	201,5	83,5	117,5
4.	4	150,5	118,0	186,0	113,5
5.	5	381,5	277,0	136,5	151,5
6.	6	531,5	404,0	158,0	312,5

Tab. 2. Wektory wagowe dla skupień utworzonych na podstawie danych dotyczących punktów gastronomicznych w województwach. Źródło: opracowanie własne.

Dane z tabeli 2 wskazują, że poszczególne skupienia różnią się między sobą wyraźnie liczbą punktów gastronomicznych rozmaitych typów. Rozpatrując segmenty 1, 2, 3, 5, 6, można zauważyć stopniowy wzrost liczby restauracji, barów, stołówek i pozostałych. Na podstawie tej zależności w jedno skupienie zostały połączone województwo śląskie z małopolskim, wielkopolskie z dolnośląskim, łódzkie z pomorskim, a świętokrzyskie, podlaskie, lubelskie i opolskie w jedną, najliczniejszą grupę. Województwo lubelskie i zachodniopomorskie zostały zaliczone do odrębnego skupienia (czwartego) ze względu na wykryty przez algorytm rozkład punktów gastronomicznych poszczególnych typów różniący się od charakteryzującego pozostałe województwa: są to dwa województwa o najniższym udziale barów i najwyższym udziale stołówek. Wyjaśnienie przyczyny tego faktu wymagałoby przeprowadzenia dalszej, szczegółowej analizy i zapewne odwołania się do wiedzy osób zajmujących się specyfiką tych regionów.

5. Zastosowanie sieci Kohonena na potrzeby analizy danych o podmiotach gospodarczych z różnych sekcji Polskiej Klasyfikacji Działalności w podregionach

Analizowany zbiór danych pochodzi z Krajowego Rejestru Urzędowego Podmiotów Gospodarki Narodowej (REGON) i zawiera informacje o liczbie podmiotów gospodarki narodowej w 2010 r. według poszczególnych sekcji PKD 2007 w podziale na podregiony. Podregiony są tu rozumiane jako jednostki terytorialne utworzone do celów statystycznych poprzez połączenie powiatów w większe grupy (Centrum Informatyki Statystycznej 2011).

W przypadku teraz rozpatrywanego zbioru danych najmniej satysfakcjonujące wyniki, odznaczające się bardzo małą jednorodnością skupień, uzyskiwano, posługując się metodą kwantowania wektorowego (metoda *Kohonen VQ*). Powodem jest zapewne nieuwzględnianie w tej metodzie

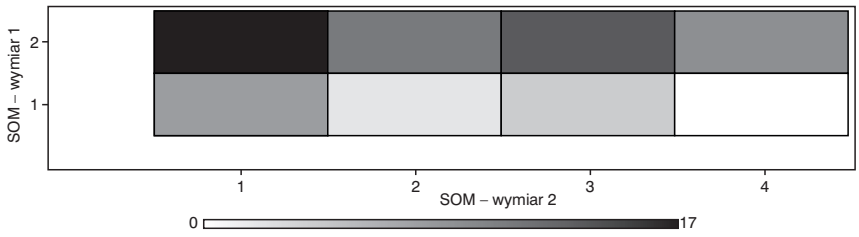
sąsiedztwa, które w przypadku większych zbiorów danych odgrywa znacząca rolę. Lepsze wyniki dawała samoorganizująca się mapa Kohonena (metoda *SOM Kohonen*), jak też uczenie wsadowe. Aby znaleźć najodpowiedniejszą metodę i parametry, wygenerowano szereg map. Ostatecznie wybrano metodę *SOM Kohonen*, ponieważ pozwalała ona na znaczące zmniejszenie błędu kwantyzacji, przypomnijmy, liczonego jako średnia odległość przypadków wejściowych od ich reprezentantów na mapie oraz umożliwiała uzyskiwanie skupień odznaczających się jednorodnością.

Pewnym problemem okazał się wybór rozmiaru mapy topologicznej, wynikły z ograniczonych możliwości, stwarzanych przez wykorzystywany program komputerowy: konieczność wybrania liczby wierszy i kolumn spośród wartości dostępnych na rozwijanej liście (wielokrotności dwójki poniżej dziesięciu wierszy/kolumn, a powyżej – wielokrotności dziesięciu). Tak więc, gdy osiem skupień (mapa 2 na 4) okazywało się zbyt małą liczbą, następna możliwa do utworzenia była mapa topologiczna o rozmiarach 4 na 4 i szesnastu skupieniach. Po próbach wygenerowania i przeanalizowania map o różnych rozmiarach, zdecydowano, że najwygodniejsza dla prowadzenie analizy danych będzie mapa z ośmioma skupieniami. Na takiej mapie nie ma pustych skupień, a wielkość mapy okazuje się mieć wystarczające rozmiary, by nie ograniczać odwzorowania różnorodności. Przykładowo w sytuacji szesnastu skupień na mapie, aż trzy czwarte z nich liczy mniej niż pięć obiektów, co zgodnie z zaleceniami programu *SAS Enterprise Miner*, przedstawionymi w dokumentacji, nie powinno być akceptowane, utrudnia bowiem poprawność wnioskowania.

Porównanie map topologicznych otrzymywanych za pomocą różnych metod wyboru początkowych wektorów wagowych prowadzi do wniosku, że w rozpatrywanej tu analizie danych metoda głównych składowych nie zapewnia dobrych wyników ze względu na wyższą wartość błędu kwantyzacji niż w przypadku innych metod wyboru startowych wag. Dwie pierwsze składowe główne wyjaśniają co prawda aż 85% zmienności, jednak dogodniejsze do analiz okazują się mapy, w których początkowe wektory wagowe są wybierane za pomocą opcji programu *Wartość odstająca*. Znaczną poprawę wyników trenowania sieci, ujawniającą się zmniejszeniem błędu kwantyzacji i eliminacją pustych skupień, można otrzymać poprzez modyfikację promienia, określającego minimalną odległość pomiędzy wyjściowymi wektorami wagowymi. Po przeprowadzeniu wielu prób przyjęto wartość 0,1. Podregion miasta Warszawa, również jak w przypadku poprzednio analizowanego zbioru danych, tworzy odrębne skupienie. Jednakże w obecnym badaniu przeprowadzanie uczenia sieci na zbiorze po usunięciu tego podregionu nie daje poprawy wyników, więc nie jest uzasadnione i nie zostało przeprowadzone.

Ostatecznie wybrano mapę topologiczną złożoną z ośmiu neuronów, do których zostały przypisane wszystkie przypadki wejściowe – podregiony. Na rysunku 4 przedstawiono liczebności obserwacji przypisane poszczególnym

segmentom mapy. Najliczniejszy segment obejmuje 17 obserwacji, następny według liczebności – 14, a kolejne odpowiednio 10 i 8. Dwa najmniej liczne segmenty mają po mniej niż 5 przypadków wejściowych.



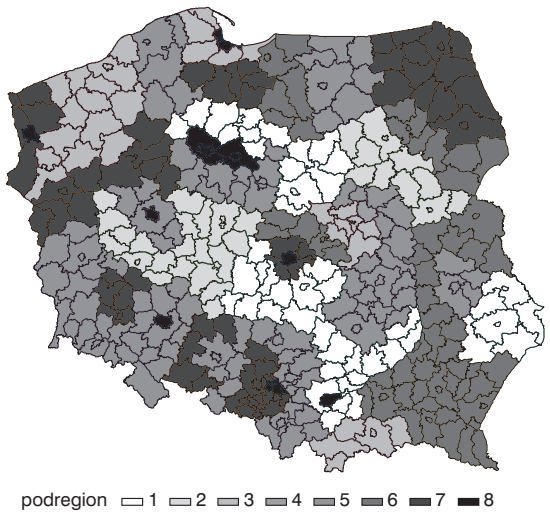
Rys. 4. Liczba podregionów przyporządkowanych do segmentów mapy topologicznej utworzonej na podstawie danych dotyczących liczebności podmiotów gospodarczych z różnych sekcji PKD. Źródło: opracowanie własne.

W tabeli 3 przedstawiono statystyki uzyskane dla mapy topologicznej przedstawionej na rysunku 4, tj. dla każdego jej segmentu: liczebność, odchylenie standardowe obiektów od ziarna skupienia, największą odległość obiektów od ziarna skupienia, najbliższe skupienie i odległość od najbliższego skupienia. Średnie odchylenia standardowe są zbliżone do siebie we wszystkich segmentach, co wskazuje że segmenty są jednorodne. Jedynie segment 1:3 jest mniej jednolity. Jeżeli pominiemy segment 1:4 złożony z jednej obserwacji (jest nią różniąca się znacząco od innych zgrupowań Warszawa), najniższe wartości odchylenia standardowego mają segmenty 2:2 (0,04), 2:3 (0,05) i 1:1 (0,05). Pewien niepokój może budzić fakt, że w kilku segmentach znajdują się podregiony znacznie oddalone od wektorów wagowych, do których zostały przypisane. Ponadto odległość od najbliższego skupienia dla wielu zgrupowań nie jest zbyt wysoka, a w przypadku niektórych segmentów mniejsza od najbardziej oddalonego od ziarna przypadku wejściowego. Rozwiązaniem tego mankamentu mogłoby być zwiększenie rozmiarów mapy topologicznej, lecz wówczas zwiększa się liczba segmentów reprezentujących małą liczbę podregionów, np. 2 lub 3. Odchylenia standardowe, a także odległości każdego ze skupień od najbliższego do niego można, tak jak ich liczebności, również przedstawić na mapie topologicznej (Myzik 2012: 35).

Na rysunku 5 ukazano mapę Polski przedstawiającą powiaty odpowiadające poszczególnym segmentom mapy topologicznej. Program *SAS Base 9.2* nie umożliwia w sposób bezpośredni przedstawienia mapy z podregionami, toteż przedstawiono mapę z mniejszymi jednostkami terytorialnymi. W tej sytuacji pewną niedogodnością jest brak oznaczeń granic podregionów.

Lp.	Wiersz: Kolumna	Liczebność	Odchylenie standardowe	Największa odległość od ziarna skupienia	Najbliższe skupienie	Odległość od najbliższego skupienia
1.	1:1	7	0,05	0,28	6	0,31
2.	1:2	4	0,07	0,39	1	0,52
3.	1:3	5	0,12	0,57	5	0,48
4.	1:4	1	–	0,00	8	3,44
5.	2:1	17	0,07	0,50	1	0,42
6.	2:2	10	0,04	0,28	7	0,29
7.	2:3	14	0,05	0,34	6	0,29
8.	2:4	8	0,07	0,48	5	0,59

Tab. 3. Statystyki dla mapy topologicznej utworzonej na podstawie danych dotyczących liczebności podmiotów gospodarczych z różnych sekcji PKD. Źródło: opracowanie własne.



Rys. 5. Mapa Polski podzielona na powiaty z naniesionymi segmentami mapy topologicznej utworzonej na podstawie danych dotyczących liczebności podmiotów gospodarczych z różnych sekcji PKD. Źródło: opracowanie własne.

Analiza liczebności podmiotów gospodarczych z poszczególnych sekcji PKD w poszczególnych segmentach umożliwia dostrzeżenie wielu interesujących zależności, czasem zgodnych z naszą wiedzą, co potwierdza poprawność działania sieci Kohonena. Na przykład zdecydowanie od pozostałych

odbiega rozkład w segmentach podmiotów gospodarczych z sekcji PKD: administracja publiczna i rolnictwo. Przeprowadzając szczegółową analizę, można zauważyć, że liczebność podmiotów gospodarczych z sekcji administracji publicznej przyjmuje w zdecydowanej większości skupień zbliżone wartości. Natomiast rozkład liczby podmiotów rolniczych między segmentami, jak można się było spodziewać, jest odmienny od rozkładu podmiotów przemysłowych czy usługowych. Dokładna interpretacja otrzymanych wyników wymaga pogłębionej analizy, do której należałoby zaangażować specjalistów zajmujących się analizami regionalnymi oraz zaznajomienia się z fachową literaturą z zakresu analiz regionalnych, co jednak zmusiłoby do przekroczenia zamierzonych badań i dopuszczalnej zawartości tego artykułu.

Mapa Polski sporządzona na podstawie mapy topologicznej przedstawiającej liczbę podmiotów gospodarczych różnych sekcji PKD podzielonych na możliwie jednorodne, ale różniące się między sobą segmenty, chociaż nie opiera się na tradycyjnych wskaźnikach rozwoju gospodarczego, pozwala na wizualizację stopnia rozwoju poszczególnych regionów. Analiza obu map, topologicznej i geograficznej stwarza możliwość rozważania różnic w rodzajach działalności gospodarczej podregionów i wydobycia o nich wielu szczegółowych informacji.

6. Zastosowanie sieci Kohonena na potrzeby analizy danych o obiektach sportowych w powiatach

Analiza dotyczy obiektów sportowych w powiatach w roku 2010. Liczba obserwacji – powiatów – wynosi 379. Ze zbioru dostępnego na stronie internetowej Głównego Urzędu Statystycznego (Centrum Informacji Statystycznej 2011) wybrano obiekty sportowe do analizy. Uwzględniono do przeanalizowania rozmieszczenia w powiatach następujące obiekty sportowe: stadiony, boiska do gier wielkich, boiska do gier małych, boiska uniwersalne – wielozadaniowe, duże wielofunkcyjne hale sportowe o wymiarach 44 m × 22 m i większych, średnie hale sportowe o wymiarach od 36 m × 19 m do 44 m × 22 m, sale gimnastyczne o wymiarach od 24 m × 12 m do 36 m × 19 m, małe sale pomocnicze o wymiarach poniżej 24 m × 12 m, korty tenisowe otwarte, korty tenisowe kryte, pływalnie kryte, pływalnie otwarte, tory, strzelnice, lodowiska sztuczne mrożone. Zbiór danych, podobnie jak poprzednie, zawiera także tylko dane numeryczne. Ma bardziej złożoną strukturę przede wszystkim ze względu na większą liczbę obserwacji: 379, w tym 65 miast na prawach powiatu i 314 pozostałych powiatów.

W rozpatrywanym przypadku zrezygnowano z zastosowania metody *Kohonen VQ*, ponieważ jest ona zalecana do niezbyt dużych i nie nadto złożonych zbiorów danych. Przeprowadzono próby zastosowania do utworzenia mapy topologicznej pozostałych dwóch metod: *SOM Kohonena* i *Uczenia wsadowego SOM*, w obu przypadkach przyjmując różne parametry algorytmów. Z uwagi na uzyskiwanie lepszych efektów w przypadku stosowania

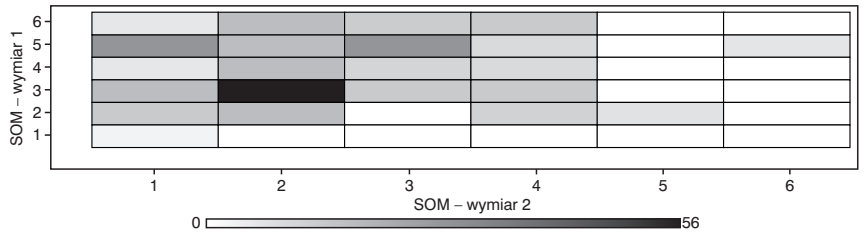
metody *SOM Kohonen*, tę metodę postanowiono wykorzystywać w dalszych badaniach.

Z uwagi na fakt analizy zbioru złożonego ze znacznej liczby obserwacji zakłada się tym razem utworzenie dużej mapy topologicznej o rozmiarach 6 wierszy na 6 kolumn, tak aby wypadło średnio 10–11 przypadków wejściowych na jeden wektor wzorcowy. Pomimo mapy złożonej aż z 36 segmentów, nie pojawiają się puste segmenty, a jedynie 9 segmentów liczy poniżej 5 przypadków wejściowych.

Przeprowadzając próby budowy mapy, przy założeniach różnej wielkości parametrów, nie dostrzeżono, aby ich modyfikacje wywoływały znaczące zmiany wyników. Wyniki skupiania są podobne, niezależnie od metody wyboru początkowych ziaren: *Wartości odstającej* lub *Głównych składowych*. Jedynie dopuszczenie zbyt dużego promienia rozmieszczenia neuronów pogarsza rezultaty i powoduje powstawanie pustych segmentów.

Pewnym zaskoczeniem jest fakt, że nawet dość znaczne zmiany opcji sąsiedztwa nie mają wpływu w analizowanym przypadku na otrzymywane wyniki. Modyfikacje jego początkowej wartości, a także liczby iteracji nie zmieniają zasadniczo wyglądu mapy topologicznej. Istotne okazują się natomiast zmiany opcji dotyczących współczynnika uczenia. Te opcje przyjęto po wielu próbnych modyfikacjach i analizie generowanych kolejno map.

Na rysunku 6 przedstawiono mapę topologiczną złożoną z 36 segmentów i wybraną do dalszej analizy rozmieszczenia obiektów sportowych w powiatach.



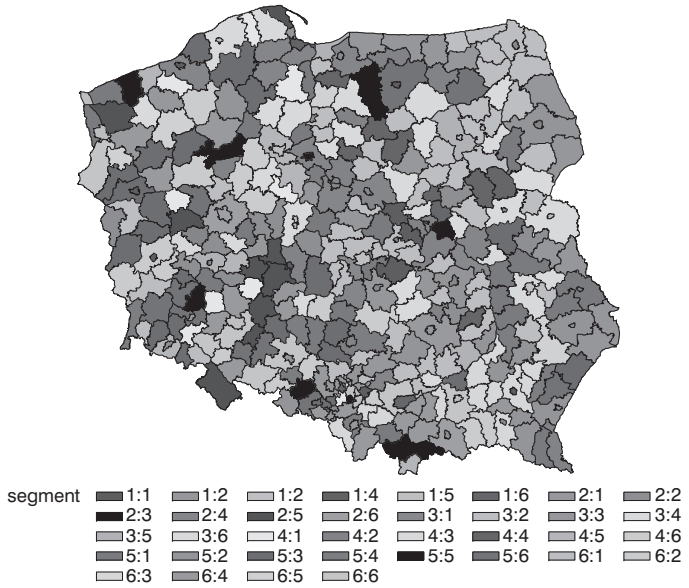
Rys. 6. Liczba powiatów przyporządkowanych do segmentów mapy topologicznej utworzonej na podstawie danych o obiektach sportowych w powiatach. Źródło: opracowanie własne.

Mapa topologiczna przedstawiona na rysunku 6 ilustruje, że spośród wszystkich segmentów zdecydowanie wyróżnia się co do liczebności obserwacji, którymi są tym razem powiaty, segment 3:2 (obejmuje 56 powiatów). Następne w kolejności co do liczby obserwacji są segmenty 5:3 (23 powiaty) oraz 5:1 (21 powiatów). Najmniej liczny jest segment 1:4 z jednym przypadkiem wejściowym. Na przedstawianej mapie wyraźnie uwidocznione jest sąsiedztwo neuronów, reprezentowanych w postaci segmentów (prostoką-

tów). Segmenty liczniejsze usytuowane są po lewej stronie i nieco przesunięte w górę. W prawym dolnym rogu mapy usytuowane są segmenty z niewielką liczbą powiatów.

Średnie odchylenie standardowe dla wszystkich segmentów wynosi 0,056. Najbardziej jednorodny, charakteryzujący się najmniejszym średnim odchyleniem standardowym okazuje się segment, który jest jednocześnie najliczniejszy – segment 3:2. Najmniej spośród wszystkich jednorodne skupisko to 1:6, które jest jednocześnie najbardziej oddalone od pozostałych (znajduje się w prawym dolnym rogu mapy topologicznej).

Na rysunku 7 przedstawiono mapę Polski z naniesionymi wynikami podziału na segmenty mapy topologicznej z rysunku 6.



Rys. 7. Mapa Polski podzielona na powiaty z naniesionymi segmentami mapy topologicznej utworzonej na podstawie danych o obiektach sportowych. Źródło: opracowanie własne.

Na rysunku przedstawiającym mapę Polski można zauważyć, że powiaty z segmentu 3:2 są umiejscowione po wschodniej stronie Polski. Szczególnie licznie jest on reprezentowany przy północno-wschodniej granicy. Podobnie sytuacja wygląda w przypadku segmentu 3:3, do którego należy wiele powiatów przygranicznych. Wiele miast zawartych jest w segmentach 1:2, 1:3, 1:4, 1:5, 1:6, czyli generalnie w pierwszym wierszu mapy topologicznej. Wiersz ten wyróżnia się od pozostałych większym odchyleniem standardowym i mało licznymi zgrupowaniami. Rozmiar miast stopniowo wzrasta w miarę prze-

suwania się bliżej prawej strony i rośnie również zróżnicowanie od innych segmentów. Powiaty miejskie są także skupione w segmentach 2:1, 2:2 i 3:1. Należą do nich głównie mniejsze miasta ze wschodniej części Polski.

Powiaty ziemskie ze wschodniej Polski są w większości umieszczone na mapie topologicznej w kolumnach z lewej strony, głównie w środkowych wierszach. Powiaty ziemskie z zachodniej części Polski znajdują się w górnej części mapy topologicznej, umiejscowione ku prawej stronie.

Jeżeli przeanalizujemy jakiego typu obiekty sportowe występują w jakiej liczbie w poszczególnych segmentach, to najwięcej różnych typów obiektów sportowych znajdziemy w segmencie 1:6, który odpowiada tylko dwóm obserwacjom, którymi są Warszawa i Kraków. Segment 1:6 nie ma jedynie (w porównaniu z pozostałymi segmentami) największej liczby boisk do gier wielkich, których liczy najwięcej segment 6:6, oraz boisk uniwersalnych, których więcej mają segmenty 2:6 i 3:6. Segment (neuron zgodnie z nomenklaturą sieci neuronowych) czy też skupienie 6:6 odpowiada przede wszystkim regionom położonym w zachodniej części Polski (wraz ze Szczecinem), blisko granicy z Niemcami.

Jeżeli przeanalizujemy segmenty mapy topologicznej pod względem liczebności stadionów, to możemy stwierdzić, że występuje mniej więcej równomierne ich rozłożenie w różnych segmentach, z jedynie kilkoma wyróżniającymi się skupieniami. Analiza występowania boisk do gier wielkich wskazuje, że segmenty o większej ich liczbie są skoncentrowane w górnej części mapy topologicznej. Interesujące jest rozmieszczenie na mapie topologicznej boisk uniwersalnych, ukazujący wyraźną koncentrację segmentów z ich dużą liczbą po prawej stronie mapy i stopniowe „promieniste" rozmieszczenie segmentów z coraz mniejszą ich liczbą w miarę oddalania się od tego punktu. Zauważalny jest fakt, że w decydującej mierze segmenty z największą liczbą różnego typu obiektów sportowych znajdują się po prawej stronie mapy topologicznej. Generalnie najczęściej usytuowane są one w ostatniej kolumnie mapy, do której zaklasyfikowane są duże miasta: Warszawa, Kraków, Poznań, Szczecin.

Jeżeli rozważymy liczbę pływalni odkrytych, to wyróżniający się pod tym względem okazuje się segment 1:3, który odpowiada trzem powiatom: Katowice, Lublin i Tarnów. W tych miastach znajduje się ponadprzeciętna liczba pływalni odkrytych w porównaniu z innymi miastami. Segment 1:3 charakteryzuje się również wysoką liczbą lodowisk sztucznych mrożonych. Próba wyjaśnienia tej zależności, która zapewne nie jest przypadkowa, wymagałaby szczegółowej, pogłębionej analizy podobieństw pomiędzy tymi miastami. Najwięcej strzelnic mają powiaty z segmentu 1:5. Skupienie to wyróżnia się także pod względem znacznej liczby innych obiektów sportowych, na przykład sal gimnastycznych, a w przypadku liczby wielu obiektów zajmuje pośrednie miejsca wśród innych wyodrębnionych segmentów. Do skupienia 1:5 należą duże miasta: Łódź, Wrocław, Bydgoszcz.

Zauważalny, interesujący i warty podkreślenia wydaje się fakt, że w przypadku analizy rozmieszczenia obiektów sportowych wiele powiatów ziem-

skich, które znajdują się obok siebie na mapie Polski, sąsiaduje na mapie topologicznej. W przypadku miast można zauważyć inną zależność. Wydaje się, że decydującymi czynnikami determinującymi liczbę obiektów sportowych w przypadku miast są: wielkość miasta oraz stopień jego rozwoju, a prawdopodobnie także inne czynniki, jak na przykład świadomość społeczna na danym terenie. Miasta o podobnej wielkości często znajdują się we wspólnych grupach – segmentach mapy topologicznej, jednakże segmentów z powiatami miejskimi jest utworzonych kilka, co pozwala zaryzykować wniosek o ich zróżnicowaniu pod względem obiektów sportowych w różnych segmentach, ale podobieństwie posiadanych obiektów sportowych – takich, które należą do tych samych segmentów.

Sieć Kohonena i mapa topologiczna, dając możliwość wizualizacji różnych cech dotyczących obiektów sportowych wspólnych dla powiatów, jak i je odróżniających, pozwalają na dostrzeżenie wielu różnych zależności, z których niektóre są trudne do zauważenia na podstawie samego zbioru danych.

7. Zakończenie

Przedstawiona w artykule analiza za pomocą sieci Kohonena dotyczyła jednostek terytorialnych opisywanych za pomocą różnych charakterystyk. Algorytm Kohonena może być stosowany w celu lepszego zrozumienia struktury danych, wydobycia z nich nieoczywistych informacji. Badacz nie musi posiadać rozległej wiedzy na temat analizowanego zjawiska przed przystąpieniem do analizy, aby stworzyć kryteria podziału zbioru na skupienia lub wybrać odpowiednią formę modelu. Klasyfikacja danych na grupy, przeprowadzona przez sieć Kohonena jedynie na podstawie ich wewnętrznej struktury, przedstawiona na mapie topologicznej, może pomóc lepiej zrozumieć zależności zawarte w danych wejściowych, zwrócić uwagę na nietypowe lub niespodziewane zależności. Dopiero wówczas wskazane jest odwołanie się do wiedzy merytorycznej (specjalistów), aby zadecydować, czy powstałym skupieniom można przypisać sensowną interpretację i nadać stosowne etykietki, tj. nazwy czy oznaczenia (Tadeusiewicz 2001: 49–51).

W literaturze można znaleźć analizy danych dotyczących jednostek terytorialnych. Dogodna dla uzyskania wartościowych wniosków z badania jednostek terytorialnych za pomocą sieci Kohonena, co potwierdziły także nasze analizy, jest możliwość tworzenia graficznych odwzorowań wyników zarówno na mapie topologicznej, jak i geograficznej, a następnie ich zestawienie ze sobą, co znacznie zwiększa możliwości interpretacji. W artykule „Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World", jego autorzy – S. Kaski i T. Kohonen (1996: 498–507) – wykorzystują zestaw wskaźników opisujących różne aspekty standardu życia, np. oczekiwana dalsza długość trwania życia w momencie narodzin, analfabetyzm wśród dorosłych wyrażony w procentach, udział opieki medycznej w wydatkach gospodarstw domowych, aby uzyskać wyniki odwzorowujące

rozkład dobrobytu dla ludności na świecie. Mapa zbudowana na podstawie wskaźników ukazuje zależności pomiędzy różnymi państwami, np. kraje należące do Organizacji Współpracy Gospodarczej i Rozwoju (OECD) znalazły się w jednej grupie, a kraje Europy Wschodniej w drugiej.

Zastosowanie sieci Kohonena do analizy danych dotyczących jednostek terytorialnych przedstawiają także autorzy artykułu: „Living Standards of Vietnamese Provinces: a Kohonen Map" (Nguyen, Haughton i Hudson 2009: 109–113). W artykule zbudowano mapę topologiczną Kohonena w celu ukazania rankingu prowincji wietnamskich na podstawie wskaźników, takich jak PKB, dochód gospodarstwa domowego na jednego członka rodziny, miara ubóstwa, miernik rozwoju społecznego.

G. Tarczyński, stosując sieć Kohonena, przeprowadza analizę polskich powiatów, zmierzającą do oceny i porównania standardu życia ludności różnych powiatów na podstawie różnorodnych wskaźników. Wyniki wskazują na wyraźną odrębność powiatu warszawskiego od pozostałych. Uzyskane mapy nie ukazują wyraźnego podziału powiatów na zgrupowania. Zdaniem autora, aby uzyskać bardziej jednoznaczne wyniki, należy wyodrębnić zupełnie odmienną obserwację, jaką jest powiat warszawski, oraz przeprowadzić analizę oddzielnie dla powiatów ziemskich i miast na prawach powiatu (Tarczyński 2011: 90–94).

Powróćmy do analizy danych krótko przedstawianych w tym artykule, a szczegółowo opisanych w pracy A. Myzik (2012) na podstawie badań z wykorzystaniem sieci Kohonena i dotyczących jednostek terytorialnych. Zastosowanie sieci Kohonena i utworzonej dzięki niej mapie topologicznej w znacznym stopniu ułatwia zrozumienie zależności zawartych w zbiorze z dużą liczbą zmiennych dotyczących jednostek terytorialnych, takich jak liczba różnych obiektów sportowych w powiatach czy liczba podmiotów gospodarczych z różnych sekcji PKD. Analiza bez zapewnionej dzięki algorytmom Kohonena redukcji wymiaru i klarownej formy wizualizacji nie może być tak wyrazista i czytelna. Niewątpliwą zaletą jest to, że stosowanie sieci Kohonena nie wymaga żadnej specjalistycznej wiedzy na temat dziedziny, z jakiej analizowane są dane, ani związków między zmiennymi. Możliwe jest, opierając się na rezultatach zastosowania sieci, postawienie rozmaitych, często budzących zaskoczenie hipotez, które mogą być dopiero dalej weryfikowane na podstawie wiedzy fachowej.

W pierwszym z przedstawionych w artykule badaniu, dotyczącym punktów gastronomicznych w województwach, wykorzystano kwantowanie wektorowe, a w dwóch następnych, dotyczących podmiotów gospodarczych z różnych sekcji Polskiej Klasyfikacji Działalności oraz obiektów sportowych w powiatach, podstawową wersję metody Kohonena. Niespodziewanym ograniczeniem wykorzystywania metod Kohonena, choć niewielkim i w niedużym stopniu uciążliwym, okazuje się fakt, że program *SAS Enterprise Miner* pozwala na wybór liczby wierszy i liczby kolumn budowanej mapy topologicznej z proponowanej narzuconej listy wartości. Nie stwarza to większego problemu

w przypadku dużych zbiorów danych, ale gdy dysponujemy nieco mniejszym zbiorem, utrudnia to przyjęcie najbardziej dogodnych rozmiarów mapy.

We wszystkich przedstawianych w tym artykule analizach danych obserwacja dla województwa mazowieckiego, powiatu lub podregionu miasta Warszawa bardzo wyraźnie odróżnia się od pozostałych największą liczbą punktów gastronomicznych, podmiotów gospodarczych z większości sekcji Polskiej Klasyfikacji Działalności i różnego typu obiektów sportowych. Jedynie w przypadku analizy liczby obiektów sportowych w powiatach Warszawa nie tworzy całkiem odrębnego, oddalonego od innych segmentu na mapie topologicznej, ale oddzielny segment, do którego zaliczony jest także Kraków.

Analiza dotycząca punktów gastronomicznych w województwach, ze względu na niewielką liczbę obserwacji, nie wymaga redukcji wymiaru, jednak w takim przypadku zastosowanie kwantowania wektorowego pozwala na usystematyzowanie informacji. Liczba punktów gastronomicznych jest główną cechą różnicującą skupienia. Z pewnością bardziej interesującą wykrytą zależnością jest mała liczba barów, a duża stołówek w województwach lubelskim i zachodniopomorskim, zapewne możliwa do wyjaśnienia po rozważeniu specyfiki tych województw i być może, jeżeli nie jest łatwa do wyjaśnienia, stwarzająca inspirację dla bardziej szczegółowej analizy.

Analiza podmiotów gospodarczych z poszczególnych sekcji PKD (2007 r.) w podziale na podregiony bardzo wyraźnie uwidacznia powiązania między sekcjami PKD, jak i podobieństwa (lub nie) podregionów. W omawianym przypadku wybrano jako najodpowiedniejszą do analizy mapę topologiczną o dwóch wierszach i czterech kolumnach. Utworzona mapa topologiczna budzi skojarzenia z geograficzną mapą rozwoju gospodarczego. Na mapie topologicznej wyodrębniają się segmenty reprezentujące duże miasta, segmenty reprezentujące słabiej rozwinięte tereny po wschodniej stronie Polski, segmenty regionów turystycznych. Podmioty gospodarcze z większości sekcji mają podobny, odpowiadający sekcjom rozkład na segmentach mapy topologicznej. Bliższa analiza prowadzi do ciekawych wniosków co do powiązań podmiotów gospodarczych z sekcji wprawdzie różnych, ale które znalazły się w tych samych lub sąsiadujących segmentach mapy topologicznej.

Do przeprowadzenia analizy obiektów sportowych w powiatach utworzono mapę topologiczną złożoną z sześciu wierszy i sześciu kolumn (36 segmentów mapy). Segmenty o największej liczbie obiektów sportowych skupiają się w pierwszym i szóstym wierszu oraz ostatniej kolumnie mapy. Natomiast skupienia odpowiadające większej liczbie przypadków wejściowych (powiatów) gromadzą się po lewej stronie mapy, podczas gdy dolny prawy róg zajmują segmenty reprezentujące większe miasta, grupując je po kilka. Znaczna część powiatów ze wschodniej części kraju zgrupowana jest w jeden najbardziej jednorodny segment o stosunkowo dużej liczebności. Mapa topologiczna i utworzona na jej podstawie mapa geograficzna Polski rozmieszczenia obiektów sportowych wyraźnie ilustrują różnice w rozmiesz-

czeniu obiektów sportowych między wschodnią a zachodnią częścią Polski, pomiędzy powiatami miejskimi i ziemskimi, między mniejszymi i większymi miastami. Potwierdzają znany fakt, że kultura sportowa jest lepiej rozwinięta w większych miastach oraz w zachodniej części Polski.

Jak zwraca się uwagę w pracy A. Myzik (2012), w celu dokładnego przeanalizowania wszystkich zależności i wyprowadzenia bardziej wartościowych wniosków dotyczących rozmieszczenia obiektów sportowych w powiatach (tworzących duży i złożony zbiór danych) i uchwyconych przez sieć Kohonena należałoby podjąć obszerne oddzielne badania, wymagające zapewne zaangażowania specjalistów z różnych dziedzin. Pomocne mogłoby się także okazać podzielenie zbioru na odrębne części, np. miasta na prawach powiatu i powiaty ziemskie, jak proponował w swoich badaniach G. Tarczyński (2011) lub region wschodni i zachodni Polski.

Informacje o autorkach

Prof. dr hab. Mirosława Lasek – Katedra Informatyki Gospodarczej i Analiz Ekonomicznych, Wydział Nauk Ekonomicznych Uniwersytetu Warszawskiego. E-mail: mlasek@wne.uw.edu.pl.

Mgr Ada Myzik – Wydział Nauk Ekonomicznych Uniwersytetu Warszawskiego. E-mail: ada.myzik@gmail.com.

Bibliografia

Centrum Informacji Statystycznej 2011. *Bank Danych Lokalnych*, http://www.stat.gov.pl/bdl/app/strona.html?p_name=indeks, dostęp: 28.11.2011.

Centrum Informatyki Statystycznej 2011. *Przewodnik po Banku Danych Lokalnych*, http://www.stat.gov.pl/bdl/docs/opisy_bdl.pdf, dostęp: 28.11.2011.

Collica, R.S. 2007. *CRM Segmentation and Clustering Using SAS Enterprise Miner*, Cary: SAS Publishing, http://books.google.com/books?id=6IHA2amBGxwC&printsec=frontcover&hl=pl&source=gbs_atb#v=onepage&q&f=false, dostęp: 27.10.2011.

Kaski, S. i T. Kohonen 1996. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World, w: A.-P. N. Refenes, Y. Abu-Mostafa, J. Moody i A. Weigend (red.) *Neural Networks in Financial Engineering, Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, Singapure, http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.53.3954&rep=rep1&type=pdf, dostęp: 19.09.2011.

Kohonen, T. 1982. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, nr 43.

Kohonen, T. 1998. The Self-organizing Map. *Neurocomputing*, nr 21, https://han.buw.uw.edu.pl/han/atoz/v1s1.icm.edu.pl/pdflinks/11092521434903141.pdf, dostęp: 22.08.2011.

Kohonen, T. 2008. Data Management by Self-Organizing Maps, w: J.M. Zurada, G.G. Yen i J. Wang (red.) *Computational Intelligence: Research Frontiers, Lecture Notes in Computer Science 5050*, Hongkong: Springer, https://han.buw.uw.edu.pl/han/atoz/www.springerlink.com/content/8147v650748n2740/fulltext.pdf, dostęp: 9.11.2011.

Larose, D.T. 2006. *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Warszawa: Wydawnictwo Naukowe PWN.

Lasek, M. 2004. Od danych do wiedzy. Metody i techniki „Data Mining". *Optimum. Studia ekonomiczne*, nr 2 (22).

Lasek, M. 2007. *Metody Data Mining w analizowaniu i prognozowaniu kondycji ekonomicznej przedsiębiorstw. Zastosowania SAS Enterprise Miner*, Warszawa: Difin, rozdział 4.2.

Lasek, M. i M. Pęczkowski 2010. Metodyka procesu eksploracji danych SEMMA. *Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego*, nr 3.

Matignon, R. 2007. *Data Mining Using SAS Enterprise Miner*, New Jersey: John Wiley & Sons.

Myzik, A. 2012. *Analiza przydatności sieci neuronowych Kohonena do eksploracji zbiorów danych dotyczących jednostek terytorialnych*, praca magisterska pod kierunkiem M. Lasek, Warszawa: Wydział Nauk Ekonomicznych, Uniwersytet Warszawski.

Nguyen, P., Haughton, D. i I. Hudson 2009. Living Standards of Vietnamese Provinces: a Kohonen Map. *Case Studies in Business, Industry and Government Statistics* (*CS-BIGS*), nr 2 (2), http://legacy.bentley.edu/csbigs/documents/nguyen.pdf, dostęp: 26.10.2011.

Osowski, S. 2006. *Sieci neuronowe do przetwarzania informacji*, Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.

SAS 2010. *Dokumentacja programu SAS Enterprise Miner 6.2*, Cary: SOM/Kohonen Node, SAS Institute Inc.

Tadeusiewicz, R. 2001. *Wprowadzenie do sieci neuronowych*, Kraków: StatSoft Polska.

Tarczyński, G. 2011. *Algorytm Kohonena w analizie danych ekonomicznych*, Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.