

Grzegorz Tarczyński

Akademia Ekonomiczna we Wrocławiu

WYKORZYSTANIE MAP CECH KOHONENA WE WSTĘPNEJ ANALIZIE DANYCH

1. Wstęp

Termin *wstępna analiza danych* (*exploratory data analysis* – EDA) po raz pierwszy użyty w 1977 r. przez J. Tukeya oznacza wykorzystanie różnych technik analizy danych, zwłaszcza graficznych, w celu umożliwienia „wglądu w dane”, odkrycia ukrytych w nich struktur, wyznaczenia istotnych zmiennych, detekcji obserwacji nietypowych.

Wstępną analizę danych stosuje się od pewnego czasu w procesie zgłębiania danych (*data mining*), będącego elementem pozyskiwania wiedzy z baz danych (*knowledge discovery in databases* – KDD). Z powodu ogromnej ilości danych zastosowanie metod statystycznej analizy danych SAD jest wówczas utrudnione. Stosuje się je zazwyczaj w celu weryfikacji wyników otrzymanych po przeprowadzeniu wstępnej analizy.

2. Różnice w podejściu do wstępnej analizy danych i statystycznej analizy danych

Istnieje wiele metod wstępnej analizy danych (WAD). Najprostsza pozwala na wizualizację pięciu wartości charakteryzujących zmienne: minimum, maksimum, mediany oraz pierwszego i trzeciego kwartyła za pomocą wykresów pudełkowych. Do metod wstępnej analizy danych zalicza się również skalowanie wielowymiarowe, rysunki symboliczne i grupowanie danych. Do tej grupy zaliczyć można również mapy cech, które łączą w sobie funkcje grupowania z prezentacją graficzną. O ich popularności decyduje również odporność na brakujące dane, możliwość analizy dowolnie dużych zbiorów danych oraz prostota działania i łatwość implementacji komputerowej.

Różnica pomiędzy wstępną analizą danych a klasyczną statystyczną analizą danych polega na kolejności wykonywania etapów oraz koncentrowaniu się na innych elementach. W statystycznej analizie danych proces przebiega według schematu:

problem → dane → model → analiza → wnioski;

we wstępnej zaś analizie danych:

problem → dane → analiza → model → wnioski.

Podczas gdy statystyczna analiza danych koncentruje się na modelu (estymacji parametrów i generowaniu na jego podstawie prognozowanych wartości), wstępna analiza danych skupia się na samych danych, ich strukturze, zależnościach, wartościach nietypowych. Pozostałe różnice pomiędzy SAD a WAD zawiera tab. 1.

Tabela 1. Różnice w podejściu do statystycznej analizy danych i wstępnej analizy danych

	Statystyczna analiza danych	Wstępna analiza danych
Koncentracja na:	modelu	danych
Techniki:	zwłaszcza ilościowe	zwłaszcza graficzne
Dokładność:	metody formalne	metody nieformalne oparte na intuicji badacza
Dane:	wykorzystane do konstrukcji modelu, „przefiltrowane”	analizowane w całości, nie traci się żadnej informacji
Założenia:	wymagają założeń, np. o normalności rozkładu	zazwyczaj brak

Źródło: opracowanie na podstawie Engineering Statistics Handbook (Internet).

3. Analiza wykresów SOM

Graficzna prezentacja map cech umożliwia przedstawienie różnego rodzaju informacji. Przegląd metod prezentacji SOM (*self-organizing map*) zawiera praca J. Vesanto [1999]. Poniżej wykorzystano niektóre z nich. Aby zilustrować możliwość wykorzystania SOM we wstępnej analizie danych wygenerowano 60 obiektów opisanych za pomocą 5 zmiennych (tab. 2). Celem analizy jest poszukiwanie w podzbiorach danych ukrytej korelacji pomiędzy pierwszą zmienną a jedną z pozostałych.

Na rysunku 1 przedstawiono wykresy zmiennych parami. Jak widać, trudno doszukać się zależności pomiędzy zmienną pierwszą a pozostałymi. O braku zależności liniowej świadczą współczynniki korelacji bliskie zeru (tab. 3). Stosunkowo silna dodatnia zależność liniowa występuje pomiędzy zmiennymi 3 i 4, 3 i 5 oraz 4 i 5.

Dla opisanego przykładu wygenerowano mapę cech z pięcioma wierszami i pięcioma kolumnami. Heksagonalna mapa cech z polami wyróżnionymi za pomocą wartości z macierzy U i ze zmienną grubością linii separujących przedstawiona jest na rys. 2a. Numery znajdujące się wewnątrz pól mapy odpowiadają numerom obiektów przypisanych do danego pola. Histogram przedstawiony na rys. 2b informuje nas, ile obiektów ze zbioru uczącego przypisanych zostało do określonego pola mapy. Większe koła oznaczają większą liczbę obiektów. Jak widać, obiekty

Tabela 2. Dane wygenerowane w arkuszu kalkulacyjnym Excel

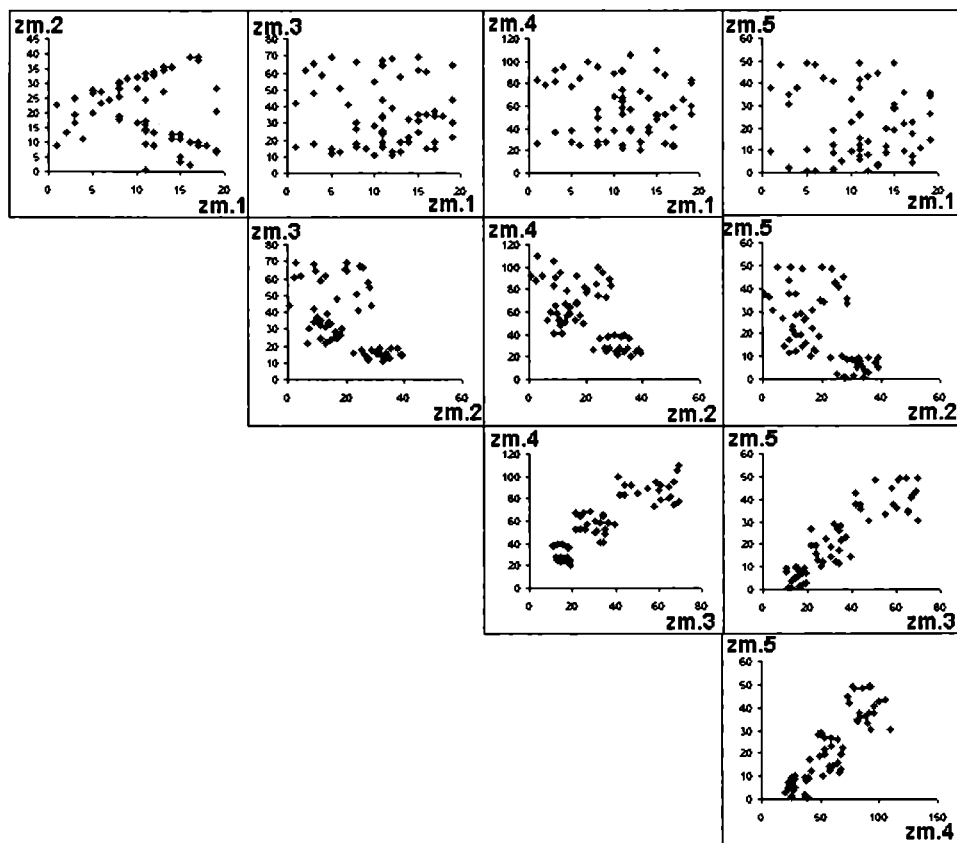
Zmienna 1	Zmienna 2	Zmienna 3	Zmienna 4	Zmienna 5	Zmienna 1	Zmienna 2	Zmienna 3	Zmienna 4	Zmienna 5
11	33,56	15,66	25,06	6,02	16	10,1	34,71	52,75	21,59
13	35,51	18,9	20,26	3,15	12	13,2	39,33	57,68	14,25
8	30,8	15,95	24,83	1,26	14	12,87	21,38	67,35	19,61
11	31,44	18,15	22,6	6,97	15	12,65	31,56	50,65	29,01
5	26,48	14,68	28,41	9,94	11	16,97	24,79	67,61	12,75
13	34,68	12,25	27,15	3,55	15	10,99	23,92	53,36	19,19
16	38,77	15,02	26,63	9,31	11	14,51	23,37	64,57	15,59
14	35,44	18,22	36,46	9,14	19	6,52	21,36	52,89	26,63
1	22,63	15,74	25,86	9,22	11	14,37	33,05	58,47	26,65
9	31,94	14,82	27,95	5,09	17	8,92	34,07	40,8	17,32
17	39,08	14,42	22,86	4,7	2	13,13	61,01	78,84	48,47
10	32,47	10,5	38,72	9,63	11	24,5	67,12	74,26	41,78
8	30,25	15,51	39,07	8,47	19	28,6	44,01	83,02	35,66
3	25,04	17,42	36,77	2,37	13	27,26	57,73	72,75	44,81
12	32,59	10,34	37,79	7,97	10	28,24	54,78	89,37	32,92
8	28,45	17,76	27,56	8,75	19	20,33	64,84	80,65	34,2
12	33,64	12,73	39,11	0,93	15	4,76	61,28	92,41	49,26
5	27,57	11,28	37,44	0,69	3	19,36	65,22	81,39	34,95
6	27,31	12,86	25,29	0,48	4	10,98	58,76	95,66	37,9
17	37,58	18,93	25,08	7,02	11	0,36	43,99	91,7	37,74
8	18,76	30,33	49,57	18,9	3	16,7	47,42	92,51	30,55
18	9,03	34,07	65,84	11,26	8	25,82	66,82	94,63	40,61
14	11,28	32,52	41,4	11,94	16	2,03	60,24	88,51	36,18
17	9,97	36,75	58,4	22,92	12	8,88	68,62	105,68	43,47
11	14,06	34,26	64,48	25,96	7	24,22	41,45	99,28	42,37
19	7,24	30,56	59,98	14,69	6	23,4	50,26	85,11	48,3
15	11,13	35,06	48,29	28,22	1	8,82	41,56	82,82	37,73
10	16,48	28,4	69,36	22,72	15	3,07	69,63	109,61	30,72
11	16,36	25,7	52,51	10,11	5	20,07	69,4	77,37	48,91
8	17,85	26,71	57,67	12,59	11	9,37	64,76	91,27	49,01

Źródło: opracowanie własne.

Tabela 3. Macierz korelacji

	Zmienna 1	Zmienna 2	Zmienna 3	Zmienna 4	Zmienna 5
Zmienna 1	1,00	-0,13	-0,04	-0,07	-0,08
Zmienna 2	-0,13	1,00	-0,51	-0,61	-0,52
Zmienna 3	-0,04	-0,51	1,00	0,87	0,90
Zmienna 4	-0,07	-0,61	0,87	1,00	0,87
Zmienna 5	-0,08	-0,52	0,90	0,87	1,00

Źródło: opracowanie własne.

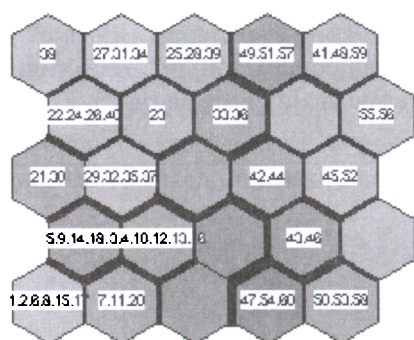
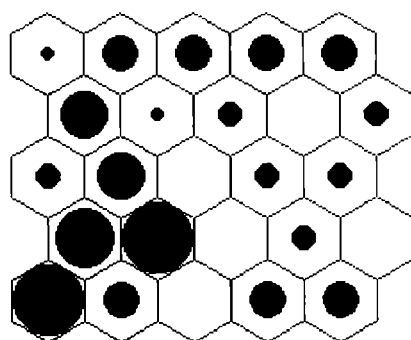


Rys. 1. Wykresy par zmiennych

Źródło: opracowanie własne.

zajął niemal całą powierzchnię mapy cech (tylko pięć pól nie reprezentuje w tej chwili żadnych obiektów ze zbioru uczącego). Lewy dolny róg mapy zawiera skupienie większej liczby obiektów. Puste pola na środku wyraźnie separują obiekty znajdujące się po lewej stronie od obiektów będących po prawej stronie mapy cech. Na podstawie przedstawionych rysunków trudno jednak wyciągnąć daleko idące wnioski. Dogłębniejsza analiza będzie możliwa na podstawie kolejnych prezentacji.

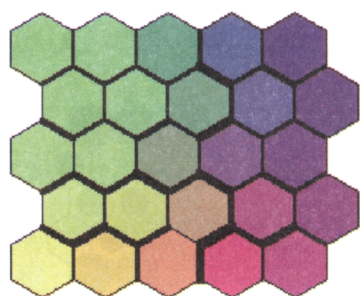
Mapy cech mogą być również przedstawione za pomocą kodowania kolorowego, gdzie podobne obszary mapy są zobrazowane za pomocą zbliżonych kolorów. Taki sposób prezentacji połączony z wykresami par zmiennych ułatwia zaobserwowanie zależności, które mogą występować w pewnych podzbiorach. Rysunek 3 przedstawia mapę cech przedstawioną za pomocą kodowania kolorowego i wykres pary zmiennych 1 i 2.

a) macierz U 

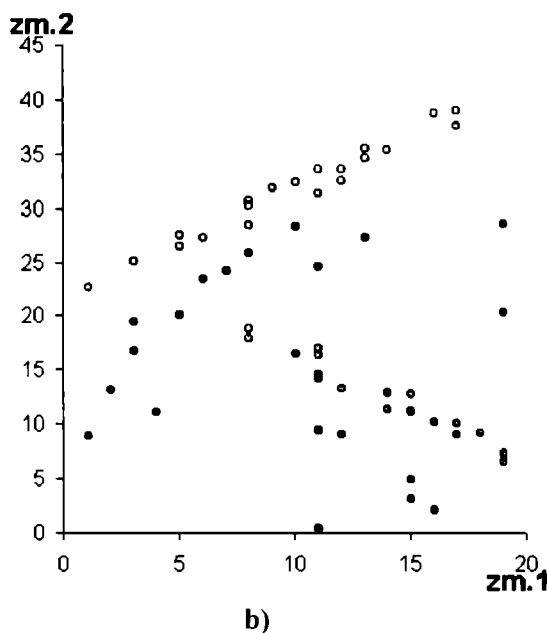
b) histogram

Rys. 2. Mapa cech

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.



a)



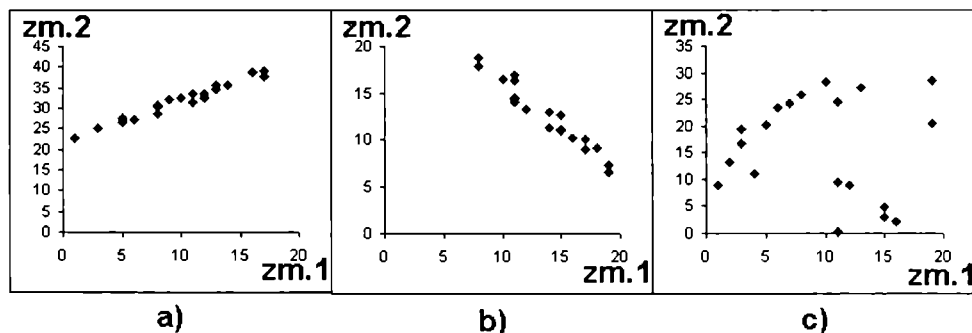
b)

Rys. 3. Zależność pomiędzy zmiennymi 1 i 2 – kodowanie kolorowe

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

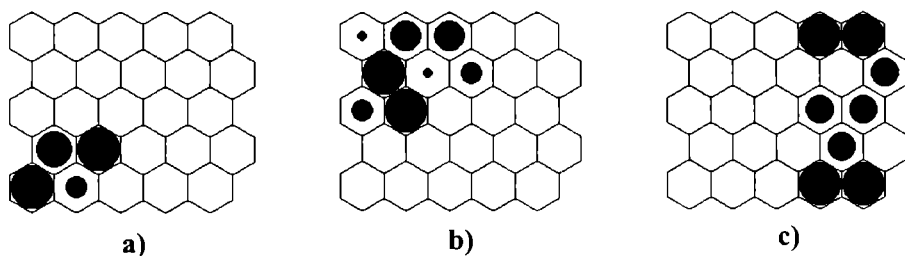
Kolory punktów przedstawionych na rys. 3b odpowiadają kolorom pól mapy cech z rys. 3a. Zbliżone kolory dwuwymiarowych punktów odpowiadają podobnym obiektom w przestrzeni 5-wymiarowej. Na rysunku 3b bez trudu można dostrzec, że pomiędzy podzbiorem zmiennych 1 i 2 przedstawionych za pomocą żółtych kropek istnieje dodatnia zależność liniowa, natomiast pomiędzy obiektami

zaprezentowanymi w odcieniach zieleni występuje korelacja ujemna. Rysunek 4 ilustruje zależności pomiędzy zmienną 1 i 2 w podzbiorach: „żółtym”, „zielonym” i pozostałym. Histogramy dotyczące tych podzbiorów znajdują się na rys. 5.



Rys. 4. Wykresy zmiennych 1 i 2: a) „klasa żółta”, b) „klasa zielona”, c) pozostałe obiekty

Źródło: opracowanie własne z wykorzystaniem programu Microsoft Excel.



Rys. 5. Histogram: a) „klasa żółta”, b) „klasa zielona”, c) pozostałe obiekty

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

Istnieje bardzo silna korelacja pomiędzy pierwszą i drugą zmienną. Występuje ona jednak w pewnych podzbiorach zbioru uczącego. Informacja o tym, do której klasy należy przypisać obiekt, a więc czy występuje pomiędzy dwoma pierwszymi zmiennymi liniowa zależność dodatnia, ujemna, czy też tej zależności nie da się ustalić, ukryta jest w wartościach pozostałych zmiennych.

Zaobserwowanie takiej zależności na danych rzeczywistych jest zadaniem bardzo trudnym. W zaprezentowanym przykładzie, na specjalnie przygotowanym zbiorze danych, relacje pomiędzy zmiennymi w wyróżnionych klasach były bardzo silne. Analiza prawdziwych danych wymaga dużo większego nakładu czasu i doświadczenia osoby przeprowadzającej badanie.

W kolejnym przykładzie mapy cech wykorzystane zostały do poszukiwania zależności ukrytych w danych obrazujących jakość życia w 373 powiatach opisanych za pomocą 27 zmiennych.

Wygenerowana została mapa cech 16·16, przedstawiona na rys. 6. Zielony kolor prawej strony mapy świadczy o dużym podobieństwie znajdujących się tam obiektów. Obiekty znajdujące się w lewym górnym rogu mapy są bardziej zróżnicowane, co przejawia się w czerwonym kolorze komórek, jak również w grubszych krawędziach linii separujących.

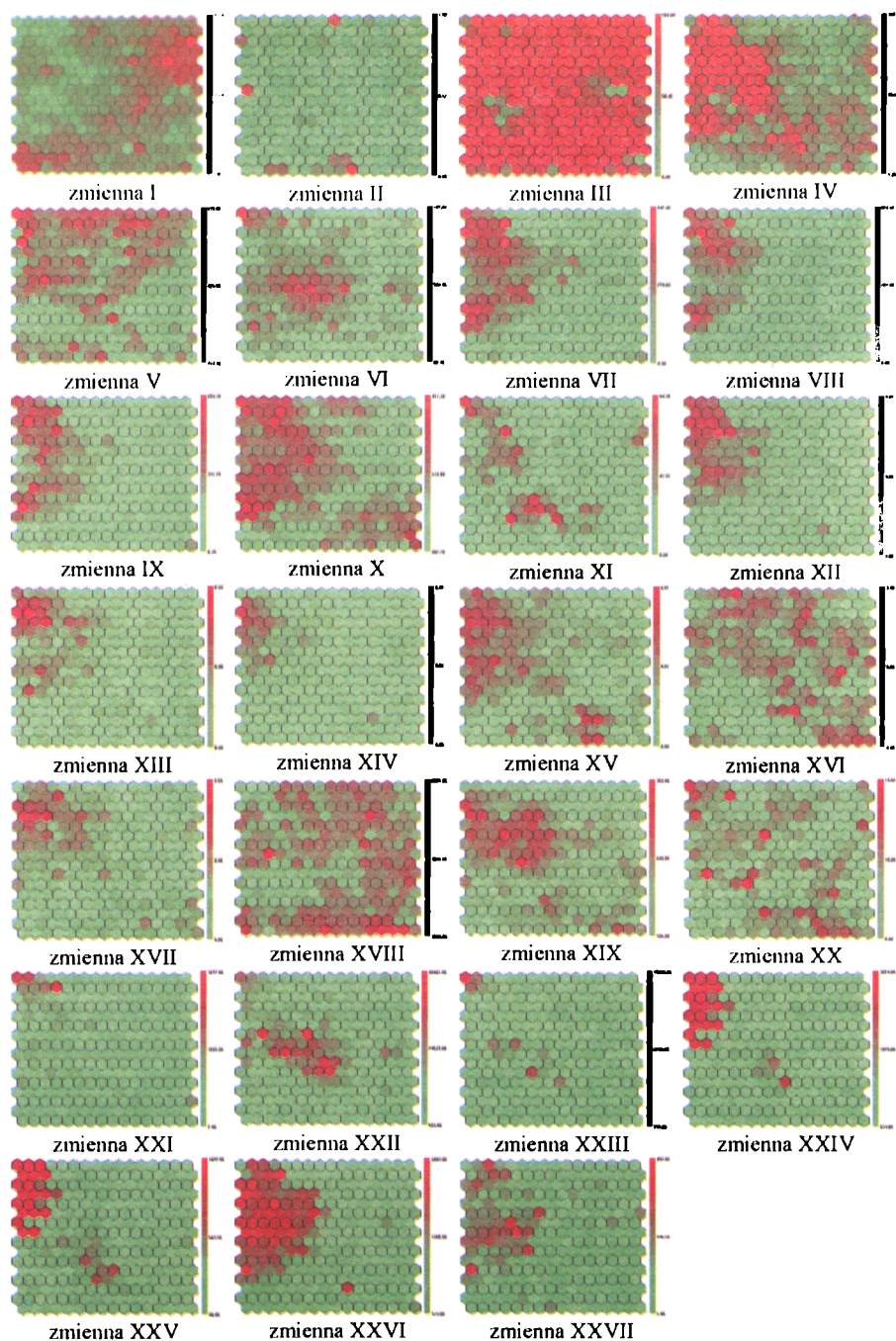


Rys. 6. Mapa cech (16·16) 373 powiatów

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

Poszukując zależności zachodzących pomiędzy cechami opisującymi obiekty, często przedstawia się poszczególne zmienne na osobnych rysunkach. Rysunek 7 zawiera wszystkie (27) zmienne wykorzystane w badaniu. Po prawej stronie każdego wykresu znajduje się skala. W prezentowanych przykładach przyjęto ogólnie stosowaną konwencję zaznaczania małych wartości zmiennych za pomocą koloru zielonego, dużych zaś – kolorem czerwonym.

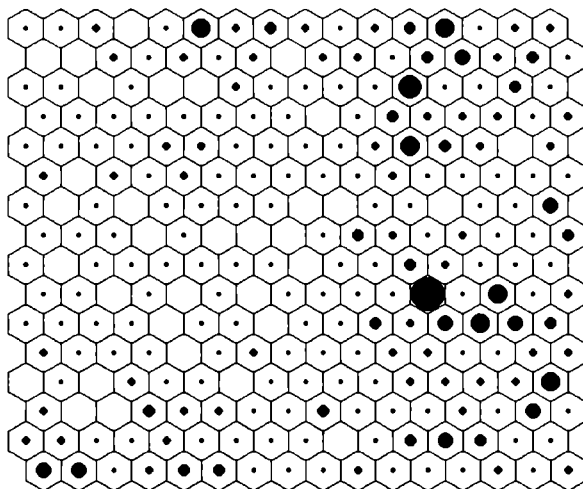
Przed przystąpieniem do analizy warto jeszcze przyjrzeć się histogramom (rys. 8-10). Rysunek 8 przedstawia rozmieszczenie wszystkich powiatów na mapie cech. Kolejne rysunki przedstawiają podzbiory danych. Powiaty grodzkie (rys. 9a) w przeważającej większości skupione są w lewym górnym rogu mapy (wyjątek stanowią: Mysłowice i Świnoujście). Powiaty ziemskie (rys. 9b) zajmują pozostałą część mapy (wyjątek to powiat ziemski warszawski usytuowany w lewym górnym rogu). Pozostałe histogramy (rys. 10) przedstawiają powiaty należące do poszczególnych województw. Bardzo często powiaty należące do jednego województwa zajmują na mapie cech sąsiadujące obszary. Oznacza to, że wykazują one podobny



Rys. 7. Prezentacja zmiennych na mapie cech

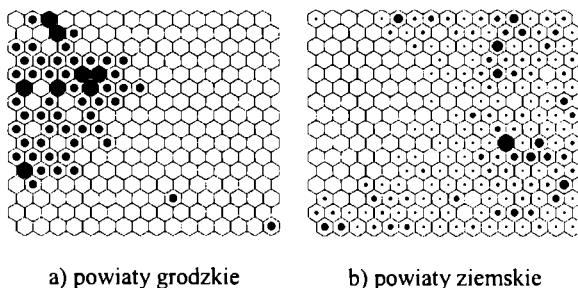
Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

poziom badanego zjawiska. W ramach województw występuje też pewne zróżnicowanie poziomu życia ludzi, co wynika z wyższego standardu w powiatach grodzkich niż w ziemskich.



Rys. 8. Rozłożenie powiatów na mapie cech – histogram

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.



a) powiaty grodzkie

b) powiaty ziemskie

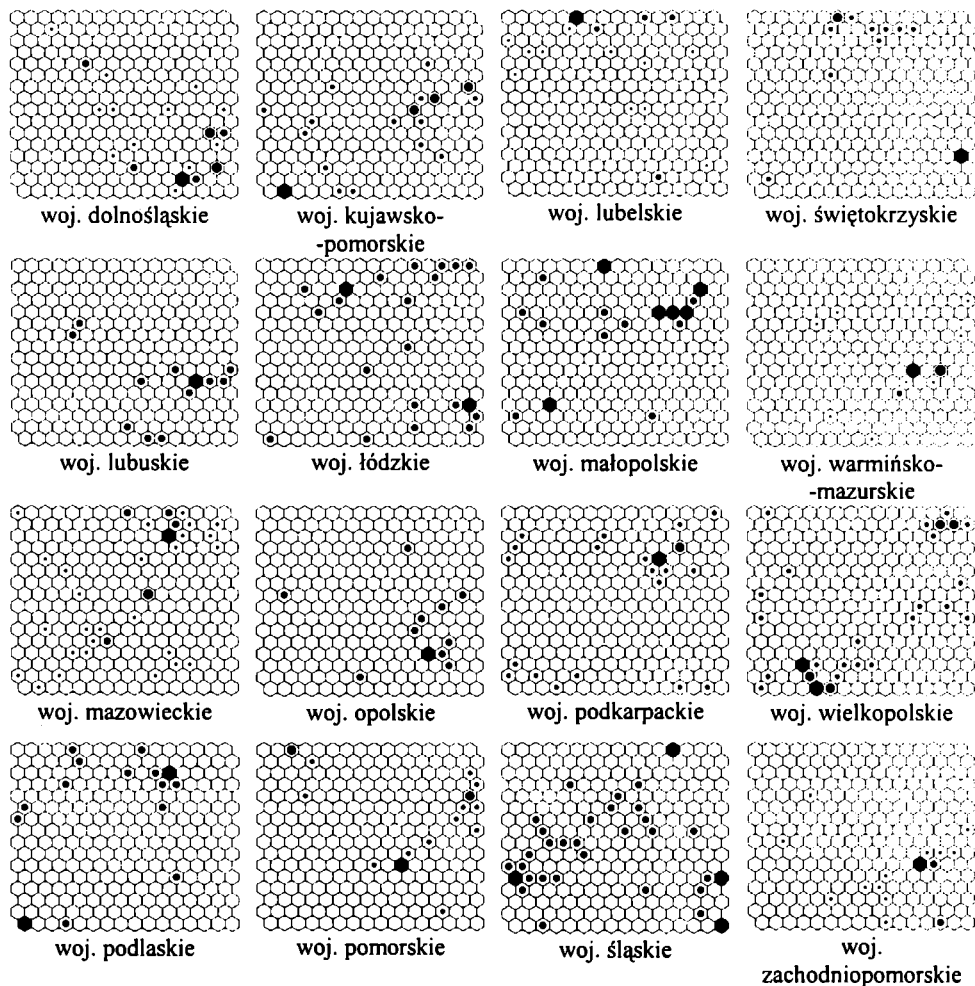
Rys. 9. Histogramy przedstawiające usytuowanie na mapie cech powiatów grodzkich i ziemskich

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

Badając wykresy zmiennych (rys. 7), szuka się zależności pomiędzy cechami. Wizualna analiza przedstawionych rysunków pozwala na zaobserwowanie pewnych podobieństw pomiędzy zmiennymi: np. zmienne VII, VIII i IX są prawdopodobnie silnie ze sobą skorelowane, zwłaszcza w podzbiorze powiatów grodzkich (poza nietypowymi). Współczynniki korelacji wynoszą odpowiednio: $r(\text{VII}, \text{VIII})=0,91$, $r(\text{VII}, \text{IX})=0,90$, $r(\text{VIII}, \text{IX})=0,90$.

Na prezentowanych rysunkach widać również, jakie wartości przyjmują zmienne w określonych miejscach mapy cech: np. zmienne XXIV i XXV przyjmują duże

wartości tylko w pewnym podzbiorze powiatów grodzkich. Natomiast zmienna XXVI (dochody budżetów powiatów i miast na prawach powiatu) w zasadzie dzieli powiaty na dwie klasy: powiaty grodzkie o wysokich dochodach i powiaty ziemskie o dużo niższych dochodach.



Rys. 10. Histogramy przedstawiające usytuowanie na mapie cech powiatów w poszczególnych województwach

Źródło: opracowanie własne z wykorzystaniem programu „Mapy cech”.

4. Wnioski

Mapy cech Kohonena są narzędziem, które z powodzeniem może być wykorzystywane we wstępnej analizie danych. Ich użycie jest jednak uzasadnione w odniesieniu do dużych zbiorów danych, kiedy zastosowanie innych metod jest

utrudnione. Interpretacja wygenerowanych wykresów nie jest łatwa, wymaga dużego doświadczenia osoby przeprowadzającej badanie. Warto również podkreślić, że uzyskane wyniki powinny zostać zweryfikowane za pomocą tradycyjnych metod statystycznej analizy danych.

Literatura

- Kaski S., *Data Exploration Using Self-organizing Maps*, PhD thesis, Helsinki University of Technology, Helsinki 1997.
- Kohonen T., *Self-organization and Associative Memory*, Springer Verlag, Berlin 1989.
- Tarczyński G., *Wykorzystanie map cech (SOM) i odwzorowania Sammona do prezentacji wielowymiarowych danych ekonomicznych*, [w:] *Zastosowania statystyki i matematyki w ekonomii*, Wyd. AE, Wrocław 2003, s. 245-253.
- Vesanto J., *SOM-based Data Visualization Methods*, „Intelligent Data Analysis” 1999.

THE USE OF KOHONEN'S SELF ORGANIZING MAPS IN EXPLORATORY DATA ANALYSIS

Summary

This paper describes some differences between Exploratory Data Analysis and Statistical Data Analysis. The advanced projection technics of Kohonen's Self Organizing Maps is used for measuring the life quality in Polish districts.