OLSZTYN ECONOMIC JOURNAL
Abbrev.: Olszt. Econ. J., 2013, 8(4)

OBSERVATION DEPTH MEASURE IN A SAMPLE IN THE VOIVODESHIP CLASSIFICATION OF THE PROPERTY MARKET

Małgorzata Kobylińska

Department of Quantitative Methods
University of Warmia and Mazury in Olsztyn

K e y w o r d s: data classification method, outlying observation, observation depth measure in a sample.

A b s t r a c t

The problem of classification has long been an object of interest in many fields of knowledge. It allows homogeneous groups of objects to be obtained with respect to a given criterion. The selection of the appropriate distance measure, which is used in the clustering of multivariate objects, has an important effect on the obtained classification results.

This paper uses observation depth measure in a sample of voivodeship classification, with respect to selected features concerning the property market in 2011. Voivodeships characterized by typical values for all analysed features were distinguished and those which could be considered outliers were so designated because of high or low values for the studied variables.

MIARA ZANURZANIA OBSERWACJI W PRÓBIE W KLASYFIKACJI WOJEWÓDZTW NA RYNKU NIERUCHOMOŚCI

Małgorzata Kobylińska

Katedra Metod Ilościowych
Uniwersytet Warmińsko-Mazurski w Olsztynie

S ł o w a k l u c z o w e: metody klasyfikacji danych, obserwacje nietypowe, miara zanurzania obserwacji w próbie.

A b s t r a k t

Problem klasyfikacji jest od dawna przedmiotem zainteresowań w wielu dziedzinach wiedzy. Pozwala ona na uzyskanie jednorodnych grup obiektów ze względu na dane kryterium. Wybór odpowiedniej miary odległości, która jest wykorzystywana w grupowaniu obiektów wielowymiarowych, ma istotny wpływ na uzyskane wyniki klasyfikacji.

W pracy zastosowano miarę zanurzania obserwacji w próbie do klasyfikacji województw, ze względu na wybrane cechy dotyczące rynku nieruchomości w 2011 roku. Wyodrębniono województwa charakteryzujące się typowymi wartościami wszystkich analizowanych cech oraz te, które można uznać za odstające ze względu na osiąganie w nich wysokich lub niskich wartości badanych zmiennych.

Introduction

An important problem in the analysis of socioeconomic data is their proper classification. This can be defined as the division of a given set of objects into disjointed and exhaustive subsets (classes, groups) with regard to a specific criterion, based on the features of the classified objects. A class is understood as a set of objects characterized by certain common properties (GRABIŃSKI et al., 1989). The aim of classification is to examine the similarity or difference between objects with respect to a given criterion.

Many criteria for the division of classification methods can be encountered in the literature. Area methods can be distinguished, including similarity-based methods, in which a defined measure of similarity is assigned to an individual pairs of objects, clustering the most similar objects. Other classification method groups include, among others, hierarchical methods, non-hierarchical methods and graphical presentation methods, including multivariate scaling and correspondence analysis (GATNAR, WALESIAK 2004). Algorithms applied in the clustering of multivariate objects use defined distance measures. In practice, different geometric measures of distance in the feature space are most often used. The selection of such a measure has a decisive influence on classification results and is usually connected with a particular clustering method. It should be remembered that the selection of the proper measure is usually determined by the nature of available data and the object and method of classification.

The aim of the paper is to analyse the utility of statistical methods based on observation depth measures in a classification sample of multivariate data. Numerical data for voivodeships with respect to selected features of the property market in 2011 were used for this purpose. The description was primarily limited to the presentation of observation classification using an observation depth measure in a sample. The result of the final analysis is the grouping of voivodeships characterized by similar levels of the studied features.

Description of the classification method

Tukey's paper (TUKEY 1975) introduced into statistical theory the concept of observation depth in a sample, which refers to the central cluster of multivariate observations. The depth measurement value lies within the closed interval [0,1], with the observations to which its higher values correspond located more centrally in the data set.

Let there be given a set of n objects $O = \{O_1, O_2,..., O_n\}$, which are the object of classification and let $C = \{X_1, X_2,..., X_p\}$, be the set of p features (variables) which characterize the classification space. A row vector, $x_i = [x_{i1}, x_{i2},..., x_{ip}]$, for $i = 1, 2,..., n$ defines the values of the features $X_1, X_2, ..., X_p$ for the object O_i. The set of n vectors determines a p-variate sample with a size of n, denoted as P_n^p. The classification method uses the Mahalanobis observation depth measure in a sample (LIU et al. 1999, ROUSSEEUW, RUTS 1996), which is based on the following definition:

Definition 1. We call the Mahalanobis depth measure (Mzan_p) of point θ in the sample P_n^p the function

$$\text{Mzan}_p\ (\theta;\ P_n^p) = [1 + Q(\theta,\ P_n^p)]^{-1} \tag{1}$$

where $Q(\theta, P_n^p) = (\theta - \bar{x})^T\ S^{-1}\ (\theta - \bar{x})$ is the Mahalanobis distance of the vector θ from the mean vector $\bar{x}$, with $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ ... \\ \theta_p \end{bmatrix}$, $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ ... \\ \bar{x}_p \end{bmatrix}$, $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, S the matrix of covariance between the considered p vectors and S^{-1} its inverse matrix.

The classification method based on the observation depth measure in the sample proceeds according to the steps presented below.

Step 1. The set of p features is divided into two-element subsets $\{X_j, X_k\}$ $j, k = 1, 2,..., p, j \neq k$. Each of these subsets is treated as a bivariate sample, which is denoted by $\text{PD}^{(m)}$. The total number of all bivariate samples is $m = \frac{1}{2}\,p(p - 1)$.

The observation matrix for the pair (j, k), which forms the bivariate sample $\text{PD}^{(m)}$, is defined as

$$\text{PD}^{(m)} = \begin{bmatrix} x_{1j} & x_{1k} \\ x_{2j} & x_{2k} \\ ... & ... \\ x_{nj} & x_{nk} \end{bmatrix}, j,\ k = 1,\ 2,..., p,\ j \neq k,$$

where $x_i^{(m)} = [x_{ij}, x_{ik}]$ is the vector of the values of the jth and kth features corresponding to the object O_i.

Step 2. The values of the Mahalanobis depth measure $\mathrm{Mzan}_2^{(m)}$ ($x_i^{(m)}$, $\mathrm{PD}^{(m)}$ are computed for each vector $x_i^{(m)}$ of the sample $\mathrm{PD}^{(m)}$, $i = 1, 2,..., n$. Depth measure values ordered non-decreasingly enable ranking observations with respect to their distance from the central cluster, which is defined by the observations with the highest depth measure values. The centroid of the sample P_n^p is determined by the observation to which the highest depth measure value corresponds. It determines the median vector $\mathrm{WM} = [\mathrm{Me}_1, \mathrm{Me}_2,..., \mathrm{Me}_p]$ of the sample P_n^p. Lower depth measure values indicate a higher distance from a given observation to the sample centre, or the median vector. The observations to which the lowest depth measure values correspond are the most distant from the sample's central cluster. They can be treated as outlying observations with respect to the values of the studied variables.

Step 3. Classes containing the observations with the lowest depth measure values are created for each bivariate sample. They are denoted as $\mathrm{NMZan}_2^{(m)}$.

Step 4. The observations belonging to the classes specified in step 3 are reviewed to establish for which diagnostic features observations are more distant from the central cluster of P_n^p.

Step 5. To determine which observations in the sample P_n^p assume typical numerical values for all studied variables, subsets containing the 50% of observations with the greatest depth are determined for each of the m $PD^{(m)}$ samples. A group of objects $\mathrm{NWZan}_p = \{x_i^{(m)} : \mathrm{NWZan}_2^{(1)} \cap \mathrm{NWZan}_2^{(2)} \cap ... \cap \mathrm{NWZan}_2^{(m-1)} \cap \mathrm{NWZan}_2^{(m)}\}$ is created. They define the class of observations of the sample P_n^p for which each of the studied features assumes typical values.

The presented classification method was described in detail by KOBYLIŃSKA and WARNER 2010. Discussion of observation depth in a sample can be found, among others, in the papers by DONOHO and GASKO (1992), HEand WANG (1997) and LIU et al. (1999).

Voivodeship classification with respect to selected features concerning the property market

The property market is one of the most dynamically-developing segments of the Polish economy. One of the main values used to characterize the property market is the housing stock in individual regions, which can be applied to determine the dwelling rate per 1000 residents. This allows the housing stock of the studied regions to be objectively compared. The demographic and economic situation of a given region influences the housing development rate. The housing stock in Poland consists mainly of flats in

multi-family houses located in larger towns and cities and single-family houses, which predominate in the countryside and in smaller towns. The dwelling saturation rates are more favourable in larger towns and cities. The main reason is the significant activity of developers, a lower unemployment rate, a higher pay level and a substantial number of settlers coming to academic centres. A measure for the dwelling purchase potential is the average monthly pay. Banks offer home loans, which constituted 36% of bank loan portfolios in 2011. It should be remembered that the property market is a local market and its activity depends on a given region.

The classification was carried out for 16 voivodeships in 2011 with respect to three features concerning the property market:

X_1 – mean dwelling unit purchase/sale transaction price (PLN for m^2),

X_2 – gross average monthly pay (PLN),

X_3 – number of new dwellings put into use per 1000 residents.

Table 1 contains the numerical values of the studied features in the voivodeships in the year. A statistical description of the features is provided in Table 2.

Table 1

Numerical values of the features X_1, X_2, X_3 in voivodeships in 2011

Voivodeship	X_1	X_2	X_3	Voivodeship	X_1	X_2	X_3
Lower Silesia	3,641	3,587.25	3.62	Subcarpathia	3,760	3,023.21	2.68
Kuyavia-Pomerania	3,087	3,062.32	3.06	Podlasie	3,764	3,178.15	3.69
Lublin	3,477	3,257.14	2.92	Pomerania	4,079	3,567.49	5.22
Lubusz	2,429	3,073.95	3.24	Silesia	2,572	3,794.62	2.00
Łódź	3,028	3,245.97	2.37	Świętokrzyskie	3,039	3,137.91	2.00
Lesser Poland	4,295	3,332.98	3.78	Warmia-Mazury	2,886	3,019.37	3.15
Masovia	6,701	4,504.66	4.76	Greater Poland	3,710	3,284.41	4.00
Opole	2,745	3,249.58	1.69	West Pomerania	3,451	3,289.56	3.46

Source: stat.gov.pl.

The computed high range values indicate a wide spread in the values of the studied features in voivodeships. The coefficients of variation are 28.19%, 11.23% and 30.03%, respectively. It can be assumed that the variation in gross average monthly pay is low and variation in the other features is moderate (WASILEWSKA 2009). 87.50% of the observations for the features X_1 and X_2 and 68.75% of observations for the feature X_3 fall within the range of one standard deviation for these features, i.e. (2543.12; 4539.88), (2974.43; 3726.64) and (2.26; 4.20). The first value differs considerably from 68%, which is assumed for symmetrical distributions. This is also indicated by high skewness values.

Mean dwelling unit purchase/sale transaction prices and gross average monthly pay levels below the mean value were recorded in most voivodeships in 2011.

Table 2

Numerical characteristics of the studied features

Statistics	X_1	X_2	X_3	Statistics	X_1	X_2	X_3
Lowest observation	2,429.00	3,019.37	1.69	Variance	996,762.53	141,458.10	0.95
Highest observation (max)	6,701.00	4,504.66	5.22	Standard deviation	998.38	376.11	0.97
Range	4,272.00	1,485.29	3.53	Median	3,464.00	3,253.36	3.20
Arithmetic mean	3,541.50	3,350.54	3.23	Skewness	2.20	2.17	0.34

Source: own work.

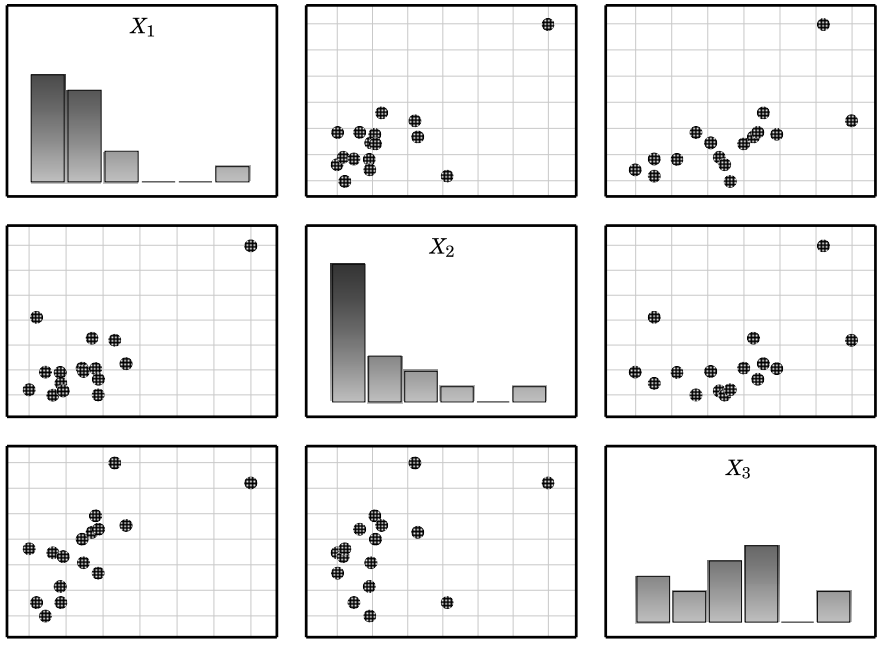

Fig. 1. Matrix diagram for the studied features

Source: own work.

The relationship between the analysed variables was illustrated in correlation diagrams arranged in the form of a matrix diagram. Considerable clustering of observations in the low and average value ranges can be observed along with the occurrence of observations much different from the others because of the high values of the analysed variables. The respective computed correlation

coefficients are $r_{X_1X_2} = 0{,}73$, $r_{X_1X_3} = 0{,}69$ and $r_{X_2X_3} = 0{,}43$. The mean dwelling unit purchase/sale transaction price is strongly correlated with the other variables. Higher gross monthly pay was recorded and more new dwellings were put into use in the voivodeships with higher mean transaction prices.

The set of features $\{X_1, X_2, X_3\}$,was divided into three two-element subsets $X^{(1)} = \{X_1, X_2\}$, $x^{(2)} = \{X_1, X_3\}$, $X^{(3)} = \{X_2, X_3\}$. Each of these subsets defines a bivariate sample with a size of 16. The Mahalanobis depth measures presented in Table 3 were computed for each of these samples. The number of observations according to Table 1 was placed in columns next to the non-decreasingly ordered depth measure values.

The observation number 16 (West Pomeranian voivodeship) reached the highest depth measure value in each bivariate sample. It is located centrally in the studied data set and defines the median vector of the sample P_{16}^3, whose coordinates are WM = (3451,00; 3289,56; 3,46).

Table 3

Mahalanobis observation depth measures for bivariate samples

Obs. no.	$Mzan_2^{(1)}$ (X_2X_2)	Obs. no.	$Mzan_2^{(2)}$ (X_1X_3)	Obs. no.	$Mzan_3^{(3)}$ (X_2X_3)
7	0.077	7	0.0803	7	0.090
12	0.097	11	0.1355	12	0.149
9	0.286	8	0.2395	11	0.173
6	0.410	4	0.2825	8	0.239
4	0.427	12	0.3650	13	0.485
10	0.514	9	0.4465	15	0.492
8	0.524	13	0.4748	5	0.513
14	0.547	15	0.5054	14	0.518
1	0.580	5	0.5086	9	0.534
2	0.605	14	0.5829	10	0.553
11	0.722	6	0.6211	4	0.587
13	0.736	10	0.7985	2	0.606
5	0.766	2	0.8008	6	0.691
15	0.808	1	0.8161	1	0.694
3	0.911	3	0.8350	3	0.875
16	0.971	16	0.8495	16	0.876

Source: own work.

Using the data contained in Table 3, it can be established which voivodeships are in the most central range of the sample PR_{16}^3, with respect to all numerical values of the variables. Subsets containing the 50% of observations with the greatest depth were determined for this purpose for each bivariate sample. This is presented graphically in Figure 2. The triangle

contains the observations with the greatest depth in all bivariate feature subsets. Observations belonging to each of these subsets were placed inside the triangle. The Lower Silesia, Kuyavia-Pomerania, Lublin and West Pomerania voivodeships can be considered typical for all diagnostic features.

The values of the Mahalanobis observation depth measures in the sample presented in Table 3 determine which voivodeships are outliers due to their very low or very high values of the studied variables. The lowest depth measure values correspond to the Masovia voivodeship (observation 7), in which the highest values of the X_1 and X_2 variables were recorded in 2011. The Silesia voivodeship (observation 12) can be considered outlying because of the low values of both the dwelling purchase/sale transaction price and the number of new dwellings put into use. High values of gross average monthly pay and the number of new dwellings put into use per 1000 residents were recorded in the Pomerania voivodeship (observation 11) in 2011.

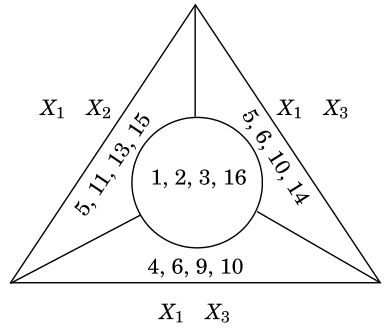

Fig. 2. Triangle containing the observations with the greatest depth in all bivariate feature subsets
Source: own work.

Table 4

Mahalanobis observation depth measures in the sample PR^3_{16}

Obs. no.	Mzan	Obs. no.	Mzan	Obs. no.	Mzan
7	0.095	14	0.530	6	0.700
12	0.159	9	0.556	1	0.706
11	0.188	5	0.558	16	0.874
8	0.269	10	0.565	3	0.897
13	0.386	4	0.597	–	–
15	0.509	2	0.618	–	–

Source: own work.

The values of Mahalanobis observation depth measures in the sample PR_{16}^{3} (Table 4) were used to prepare a statistical map. Five depth measure intensity ranges were adopted on the map, for which corresponding grey scale shades were used. The presence of two coherent areas can be observed in the north-eastern and west-central parts of the country, which include the voivodeships with depth measure values from 0.4 to 0.6.

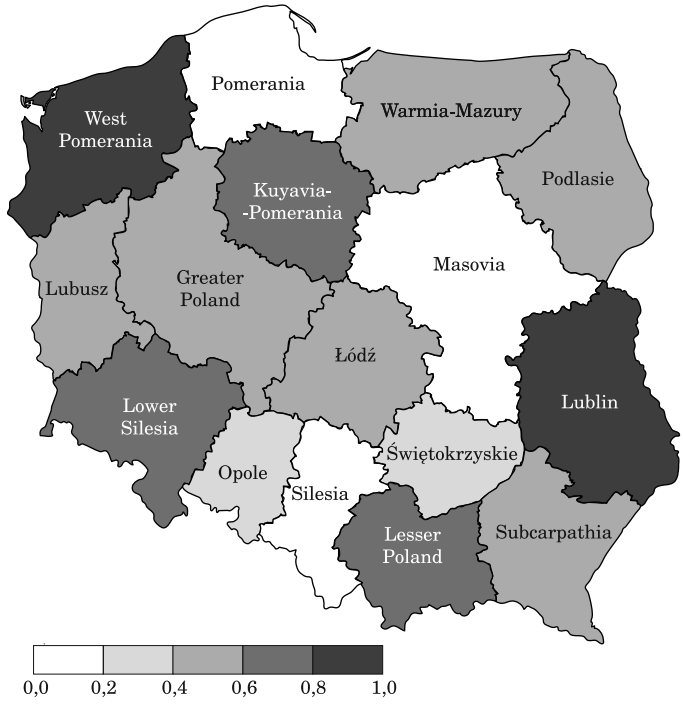

Fig. 3. Statistical map of the voivodeships for the depth measures
Source: own work.

Summary

The study used an observation depth measure in a sample for multivariate object classification, based on the example of voivodeships with respect to selected features concerning the property market in 2011. This allowed groups of objects similarly distant from the central cluster to be determined with respect to the studied diagnostic features. As a result of using the presented method, voivodeships characterized by typical values of the mean dwelling unit purchase/sale transaction price, gross average monthly pay and the number of new dwellings put into use per 1000 residents were distinguished. These were the West Pomeranian (observation 16) and Lublin voivodeships (observation

3), which make up 12.50% of the studied units. For each of these voivodeships, the depth measure values (Tables 3 and 4) reached the highest values (over 0.8). 18.75% are voivodeships outlying because of high or low values of the studied variables obtained in them. The largest group (37.5%) were the voivodeships in which the depth measure value was from 0.4 to 0.6 (observations 8, 13, 15, 14, 9, 5, 10, 4). No variables reached values which would classify a voivodeship in the outlying or most centrally-located observations in the set PR_{16}^{3} in any of these voivodeships.

The year 2011 was a stable period in the sale of dwellings compared to the previous year. 29.7 thousand dwelling units were sold, over 7% more than in the previous year (*Property market in Poland in 2011*). As in the previous years, the highest level of sales of dwellings was observed in the largest housing markets in Poland, which include the cities of: Warsaw, Gdańsk, Gdynia, Sopot, Poznań and Cracow.

The proposed statistical methods based on observation depth measures in a sample can supplement more detailed analysis, which is of high importance in the study of socioeconomic phenomena. Using an observation depth measure in a sample, difficulties connected with ordering multivariate observations are overcome. Many multivariate statistical methods require the fulfilment of some assumptions, e.g. conformity of the analysed variables with normal distribution. The fulfilment of this assumption is not necessary in the presented classification method.

Translated by JOANNA JENSEN

Accepted for print 31.12.2013

References

DONOHO D.L., GASKO M. 1992. *Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness*. The Annals of Statistics, 20: 1803–1827.

GATNAR E., WALESIAK M. 2004. *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu.

GRABIŃSKI T., WYSYMUS S., ZELIAŚ A. 1989. *Metody taksonomii numerycznej w badaniach zjawisk społeczno-gospodarczych*. PWN, Warsaw.

HE X., WANG G. 1997. *Convergence of Depth Contours for Multivariate Datasets*. The Annals of Statistics, 25: 495–504.

KOBYLIŃSKA M., WAGNER W. 2010. *Klasyfikacja obiektów wielocechowych z wykorzystaniem miary zanurzania obserwacji w próbie*. Monografie i Opracowania 570. Oficyna Wydawnicza SGH in Warsaw.

LIU R.Y., PARELIUS J.M., SINGH K. 1999. *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference*. The Annals of Statistics, 27: 783–858.

ROUSSEEUW P.J., RUTS I. 1996. *Bivariate Location Depth*. Applied Statistics, 45: 516-526.

Rynek nieruchomości w Polsce w 2011 roku. No. 15. BRE Property Partner.

TUKEY J. W. 1975. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

WASILEWSKA E. 2009. *Statystyka opisowa od podstaw. Podręcznik z zadaniami*. Wydawnictwo SGGW, Warszawa.