INFORMATION
SYSTEMS IN
MANAGEMENT

Information Systems in Management (2012) Vol. 1 (3) 197–210

DIVERSIFICATION MAPS AS A TOOL FOR MULTIDIMENSIONAL DATA PRESENTATION

PIOTR JAŁOWIECKI, EWA JAŁOWIECKA

Department of Informatics, Warsaw University of Life Sciences (SGGW)

In addition to data analysis, equally important issue, is the appropriate presentation of the analysis results during the economic, social and demographic studies. Depending on the legibility of the used form, it may make it easier or more difficult to interpret and draw conclusions. Especially difficult is the presentation of multidimensional data, according to the limitations of traditional types of graphs. The paper presents a graphical presentation of the three-dimensional data in the form of the so-called "map of differentiation" that is suitably modified spatial graph. There were presented different types of maps, layers determination ways and examples of specific applications for cigarettes consumption and diversification of logistics in the Polish food processing data.

Keywords: data presentation, multidimensional data, spatial graphs, diversification maps

1. Introduction

In scientific studies, often multi-dimensional data is used. Apart from processing and analysis of such type of data, it is very important to properly present the results. While nowadays technical possibilities and available methodology (data warehouses, OLAP methods) allow, in principle, to process and analyze data with a few, several, or even dozens of dimensions, in practice there are most frequently only two- or three-dimensional data, which are the easiest for users intuitive perception. High possibilities in the matter of multidimensional data processing and analysis are not accompanied by the corresponding capabilities for clear of such analysis results presentation. This problem is especially severe in the case of pre-processing and basic data analysis of multidimensional data in many companies with using of basic software, which is in many cases a popular spreadsheet of Microsoft Excel. Although in the market there are offered different analytical packages, which enables of multidimensional analysis results presentations in many forms, however, the prevalence of their using is rather limited, also by the price or the degree of complexity [1], [2].

High frequency of processing, analysis and presentation of multidimensional data during economic, social, demographic and many other studies is the result of frequent need to present the level or frequency of a phenomenon two differentiating factors or depending on one factor, but chronologically. Meanwhile, widely used in such cases, MS Excel software is in this area quite limited, mostly due to the simplicity of implemented multidimensional data presentation tools: pivot tables, point, surface and bubble charts. Using of these methods in the basic, not modified versions often makes difficult to interpret.

The paper presents a proposal of surface charts modification in the form of so called "Diversification Maps" similar to hypsometric maps, which are clear, intuitively comprehensible and also popular method of three-dimensional data presentation in two dimensions. Two dimensions on a plane represent the longitude and latitude, while the color is the height of the area above or below sea level. The number and width of terrain layers presented on the map can be different, but always the same color represents the area, the height of which is within the specified range of values. A proposal of transfer of hypsometric maps idea to present three-dimensional data or two-dimensional data changed in time on charts is presented with using commonly available software MS Excel. Discussed in the paper issues are presented with examples of real data analysis results from research of cigarettes consumption in Poland and diversification of logistic structure in Polish agribusiness companies conducted in the past. The idea of presented method is an old, and the first time was proposed by French economist Toussaint Loua in 1873 in form of color representation of numbers in a table [2].

2. The objective and the methods

The idea of data presentation on these type of charts is to divide the entire range of presented values in layers and marking of each of them by a separate color according to a scheme used on the hypsometric. The number and width of layers is taken as needed or may be a constant value determined earlier. Often, in addition to the same layers of color on the chart there are placed labels, which represent the variable values presented in the chart. “Diversification maps” can present different types of values, such as: raw, average or percentage values. The aim of this paper is an overview of “Diversification maps” charts as a tool of multidimensional data presentation with discussion about the most important aspects of its creation.

Charts of this type are prepared in several stages: (A) choice the method of presented data distribution between layers; (B) determination the number of layers; (C) division the data into layers; (D) definition of color scheme presented in layer; (E) determination the order of values assigned to the vertical and horizontal axis; (F) formatting individual chart elements; (G) placing additional elements on the chart, (H) correction of chart (for example, by changing the range of presented data), of course if it is necessary. Presented type of charts can be prepared both with using of basic MS Excel tools or with using of VBA (Visual Basic for Applications) programming language, which is integrated with MS Office software. Every version of Excel may be used to prepare “diversification maps”, but the most intuitive for this kind of charts is MS Excel 2003.

The stages A and B are tightly interlinked. The simplest method of distribution the data into layers in the graph is equal distribution of ranging from the smallest to the largest value of presented variable to a fixed number of layers in accordance with the formulas (1) defining the width of a single layer z_w and (2) defining the initial value, that is the lower limit of each layer k_w.

$$z_w = \frac{\max k \quad - \min k}{n_w} \tag{1}$$

$$k_w = \begin{cases} \min k & for \qquad w = 1 \\ k_{w-1} + (w-1) \cdot z_w & for \qquad w > 1 \end{cases} \tag{2}$$

where: k – variable, which values are presented on the chart, which represents one of the differentiating factors; n_w – number of layers on the chart, w – single layer number (w = 1, …, n_w).

This method does not work very well in case where some groups have all or most of values significantly higher or significantly lower than the other groups. Then usually too many layers are taken by one or a few groups, which represent the highest values. Additionally, too many layers are empty, (there represents any values), because placed between single or a few highest groups and majority of remaining groups with lower values of variable. In such cases, other method of data distribution into layers works better [8]. In this method, the distribution of values is done on the basis of the mean and standard deviation multiplied by a factor of accuracy p in accordance with formulas (3), (4) and (5).

$$z_w = \frac{s}{p} \tag{3}$$

$$k_w = \begin{cases} \bar{k} - a \cdot z_w & for \quad w = 1 \quad oraz \quad \bar{k} - (a-1) \cdot z_w \geq \min k \\ k_{w-1} + z_w & for \quad 1 < w < \max w \\ \bar{k} + b \cdot z_w & for \quad w = \max w \quad oraz \quad \bar{k} + b \cdot z_w \leq \max k \end{cases} \tag{4}$$

$$\begin{aligned} a &= \min a \quad where \quad \bar{k} - (a-1) \cdot z_w \geq \min k_{gr} \\ b &= \min b \quad where \quad \bar{k} + b \cdot z_w \leq \max k_{gr} \end{aligned} \tag{5}$$

where: $\bar{k}$ – mean value for variable values presented on the chart; s – standard deviation for data presented on the chart; w – layer number; p – accuracy level, which determines number of layers on the chart; a, b – coefficients for low border of the first (lowest) and last (highest) layer.

With using of the second described method of layers on the chart determining, the number of layers is not pre-established constant value as in the first method. It depends on the minimum, maximum, mean, standard deviation for the data presented in the chart, and primarily on accuracy level p. In studies, which results are presented in the paper as examples, a value 0.25 of p was used. It corresponds to a single layer width z_w equal to one-quarter of standard deviation from average value for the data presented in the chart. In the smaller number of layers, differences between the two methods of division into layers are more visible. With higher numbers of layers these differences are smaller.

Division of values of the researched variable into layers then presented on charts in stage C was done with using of Excel formulas which aggregates data into two-dimensional tables, which were direct base for charts. These formulas based on standard Excel functions, like as COUNTIF, IF, SUMIF, VLOOKUP. In MS Office 2003 and earlier it was necessary, to define a special formulas to aggregate data with conditions for two or more columns in database. In newest version of Excel, there are available COUNTIFS and SUMIFS functions to do it.

In stage D, the individual layers on charts were marked sequentially from the lowest to the highest by tones of color like on hypsometric maps. Two methods of assignment colors to individual layers were used. In both methods, all layers below the reference level were marked with shades of blue, the darker the lower values of the variable represented. However, layers representing values of the variable above the reference level were marked turn shades of colors green, yellow, orange and red. The methods used to assign colors to layers on charts were differed by the method of determining the reference level. In the first method, the reference level was determined on the basis of the average of all researched variable values. In the second, however, it was determined based on averages of all values of the test variable in two-dimensional groups presented on the chart. If number of layers is higher than number of selected shades of colors, it is possible to assign one shade of color to a few layers. The reference level was an average smoking level for the whole of observations, and all layers below this level are marked by blue color shades, and all above by shades of the other colors.

The orders of values assigned to the vertical and horizontal axis were determined in stage E on the base of average values of differentiating factors represented on both axis. The exceptions are cases in which some of the differentiating factors had a natural or logical order (e.g. time, age, town size, education level). Then the value of the axis was sorted according to this order. The last two stages F and G were done by standard Excel methods of charts formatting, especially with using of legend scaling and charts joining techniques.

In some cases, especially when the part of presented values is much higher or much lower than most of the other, there is a need to correct the chart to increase its readability. Typically, it consists in limiting the range of values presented in the chart and re-designation of width and layer boundaries. This is the last stage (H) of this chart type preparation. At this stage, you should also determine how missing values will be presented on the chart, which are the result of the lack of observations assigned to some pairs of values shown on the horizontal and vertical axes. Usually, the missing values in the graph are presented as a separate layer marked by separate color, which differs significantly from the other used in the chart.

3. Data distribution into layers

The first group of examples presented in the paper is originated from results of studies of polish tobacco products market in Poland, which has been researching from 2006. These studies were based mainly on household budgets survey results provided by Central Statistical Office [3]. Therefore, the average relative level of cigarettes consumption in Poland is presented on charts. It was determined in many economic, social, demographic or territorial groups, and it was defined as the number of cigarettes consumed by households belong to a specific group per 1 person in the year in relation with the same level for the whole of Poland. Distribution of households into groups were made by some differentiating factors, which were represented by discrete, e.g. socio-economic group, biological type of family, voivodeship or continuous values, e.g. revenues, expenses, age. Some of the factors were represented as coefficients indicating their average value, such as level of education or number of mature persons of particular gender. These examples are intended to demonstrate the process of selecting the number of layers, the data distribution between layers and possible measures to improve the readability of the charts.

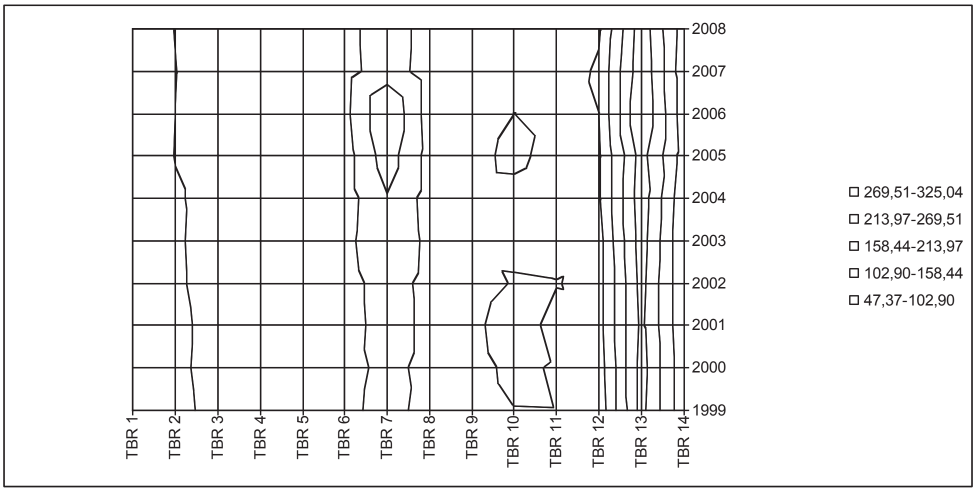

Figure 1. The result of distribution of average smoking level in different biological types of households into five layers of "diversification map" according to equal method. Source: own preparation on the base of [4]

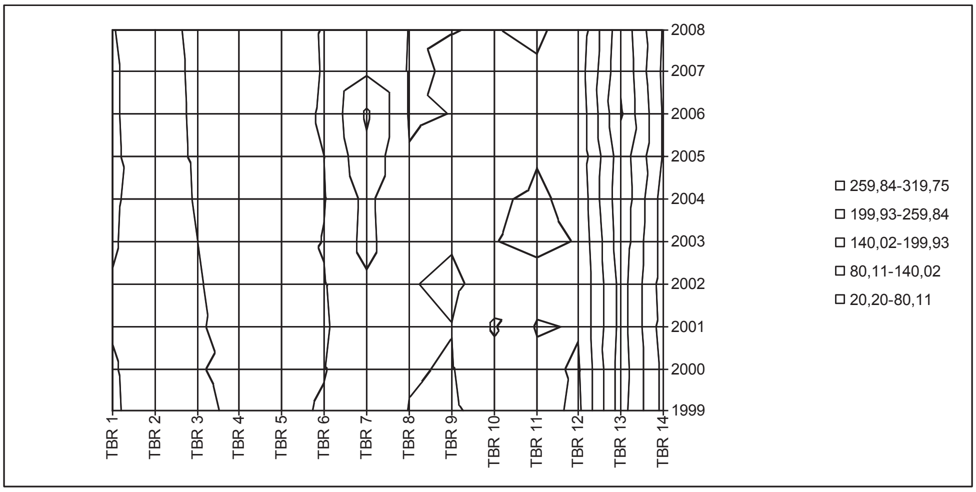

Figure 2. The result of distribution of average smoking level in different biological types of households into layers of "diversification map" according to method based on mean and standard deviation. Source: own preparation on the base of [4]

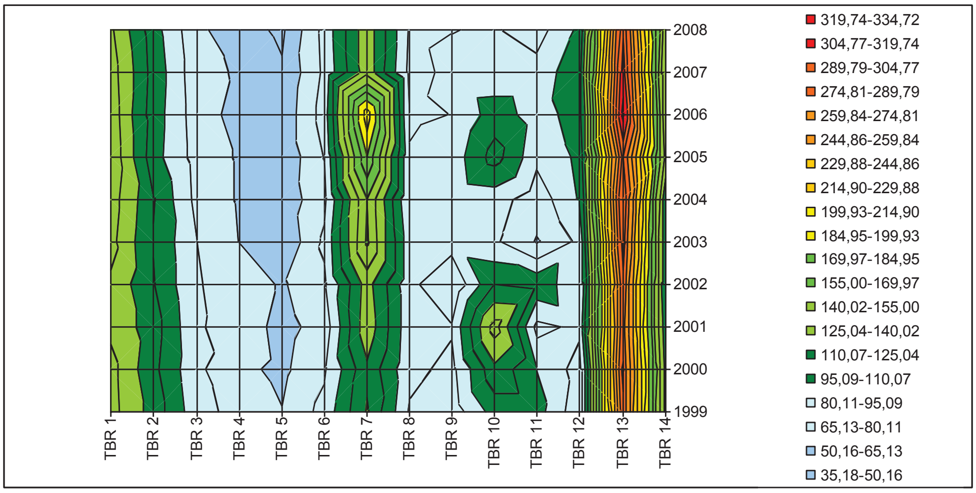

Figure 3. The result of distribution of average smoking level in different biological types of households into layers of "diversification map" according to method based on mean and standard deviation with using of $p = 4$ accuracy level. Source: own preparation on the base of [4]

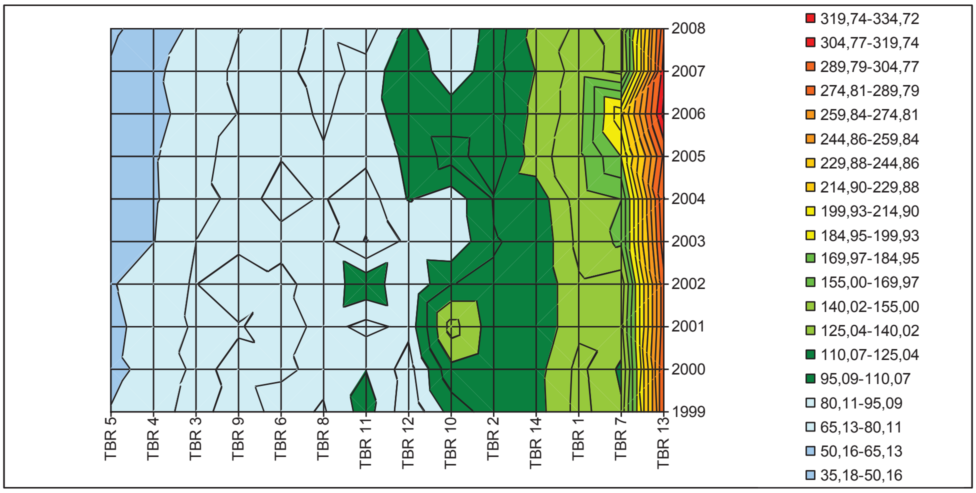

Figure 4. The result of distribution of average smoking level in different biological types of households into layers of "diversification map" according to method based on mean and standard deviation with using of $p = 4$ accuracy level and biological types sorted ascending by mean. Source: own preparation on the base of [4]

On Figures 1 and 2 the results of using of two described methods of data distribution into layers are presented. Codes from TBR1 to TBR14 represents different biological types of households. Detailed explanation of these codes is presented in the paper [4]. The second method allows an easier identification and evaluation of characteristic points on the graph, for example periodic depreciation or observation groups significantly different from the neighboring. Distribution into layers presented in Figure 2 was made at the level of accuracy $p = 1$. It means that a single layer has a width equal to the standard deviation for smoking level. In order to increase the accuracy of data presentation, the level of accuracy can be increased. In practice of studies, accuracy level p was increased to 2 and more often to 4. In described studies of cigarettes consumption level, a uniform color scale was used for all differentiating factors. Due to using of $p = 4$ accuracy level (the width of single layer equal to quarter of standard deviation), which resulting large number of layers, a few (usually two) layers was marked by the same shade on charts (Figure 3 and 4).

Better clarity of the chart can be obtained by sorting the values presented on the axes in ascending or descending order (Figure 3 and 4). But not always sorting makes sense due to the interpretation of some of the data, the values of which have a logical sequence, e.g. time, age, education. Effect of Items of the graph axes sorting is even more evident, with a smaller number of layers.

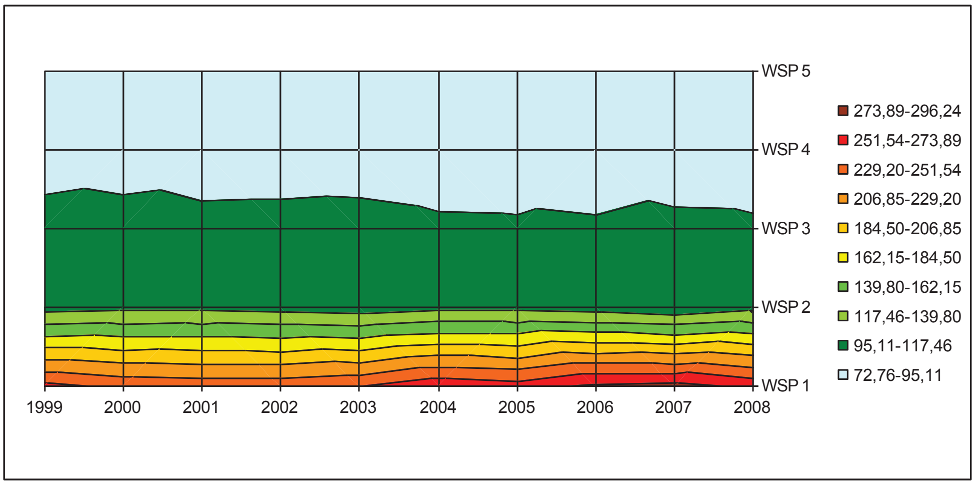

Figure 5. The result of distribution of average smoking level in different categories of proportion between adults both genders in household into layers of "diversification map" according to method based on mean and standard deviation with using of $p = 2$ accuracy level. Source: own preparation on the base of [5]

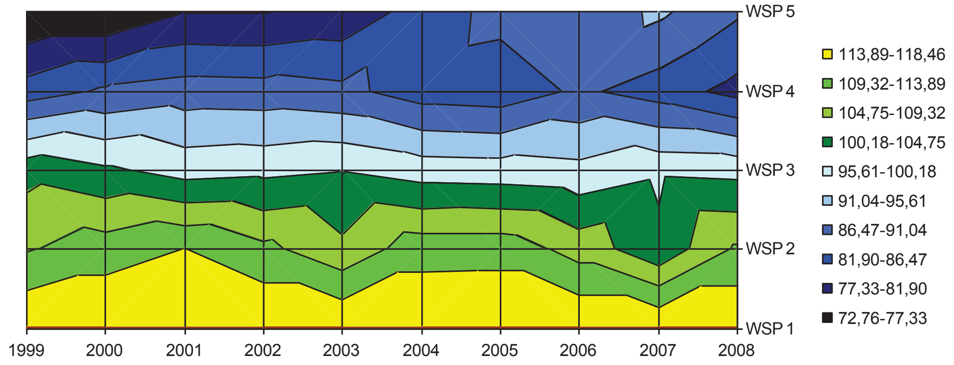

Figure 6. The result of distribution of average smoking level in different categories of proportion between adults both genders in household into layers of "diversification map" according to method based on mean and standard deviation with using of p = 2 accuracy level after removing values of WSP1 layer characterized by the highest level of cigarette consumption. Source: own preparation on the base of [5].

In cases where one or more groups of values is much higher or lower than the other, increasing the number of layers does not increase the readability and accuracy of data presented on chart (Figure 5). The best solution of this problem is to limit the range of presented data by removing outliers and redistribution data into layers once again (Figure 6). In Figures 5 and 6, WSP1 to WSP5 codes mean different proportions between adults of both genders. WSP1 is largely outnumbered by men, while WSP5 is largely outnumbered by women. In WSP1 group average smoking level is more than 2.5 times higher than in other groups.

4. Data presentation

The source of the second group of examples presented in the paper, are results of the logistics state in Polish agribusiness companies studies, which has been provided from 2009. These studies based on survey performed among food and beverages companies in 2010-2011 years. The first studies was a chronological character, thus on the first group of charts, one dimension always are a time. Whereas, the second group of charts with logistic origin presents typical three-dimensional data on two-dimensional charts in form "diversification maps". Data on these examples are described by codes for number of employees category: MK – microcompanies (up to 9 employees), ML – small companies (10 to 49 employees), SR – middle companies (50 to 249 employees) and large companies (250 and more employees) and by codes for food production company branch: MS – meat, OW – fruits and vegetables, OT – fat and oils, ML – milk, ZS – cereal and starch, WP – bakeries, WS – other food products, PS – animal feed, and NP – beverages.

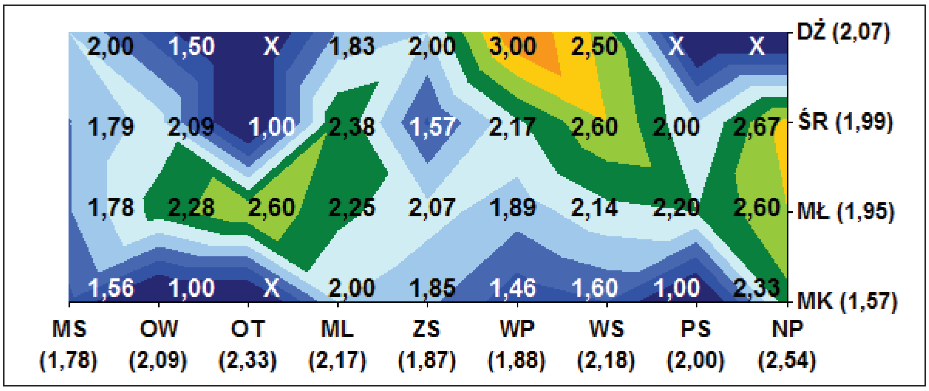

Figure 7. The "Diversification map", which presents the average values of logistic costs level coefficient according to food production branch (horizontal axe) and number of employees category (vertical axe). Source: own preparation on the base of [7]

In presented earlier graphs, the level of the investigated phenomenon was presented in the form of a number of layers of equal width, which are marked in accordance with the adopted color scheme, similar to hypsometric maps. In such cases, the identification of values for a particular point on the chart is only possible with an accuracy of a range of values assigned to a particular layer. The other method of identification of concrete values presented on chart is data labels adding to the chart. Due to the fact that the surface charts in Excel do not have the data labels, it must be done in a separate chart and integrate it with the main chart. The best choice is the three-dimensional column chart rotated and formatted appropriately. The integration of the two graphs can be handled by joining tool or with using of VBA. You should remember to have a 3D column chart transparent background, because it is the outer layer (Figure 7).

In Figure 7, in addition to data labels, next to values placed on both axes plus, averages for the whole of data category (industry or size of company) were placed. There are so called boundary means. Another, but more complicated way to add them to the chart is a creation four auxiliary charts one one-dimensional surface chart and one 3D column chart with labels for each axis. It may be joined and integrated with the main charts in form presented on Figure 8.

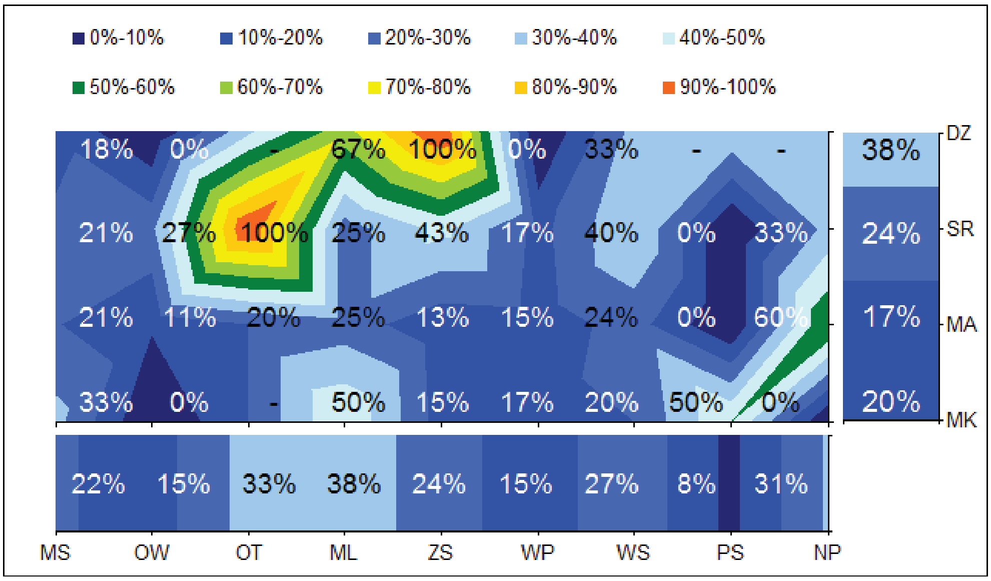

Figure 8. The “Diversification map”, which presents the percentage of companies declared using of separate information systems to logistic activities support according to food production branch (horizontal axe) and number of employees category (vertical axe). Source: own preparation on the base of [6]

On “Diversification map” charts different data categories may be presented, e.g absolute, average, percentage values of the studied phenomena or different coefficients defined by user. The example is Figure 8, on which a percentage of companies declared using a separate information system to logistic activities support is presented.

Often, in comparative objectives, on the several “Diversification map” charts, especially in percentage version, the same values and color scale is used, regardless of the different value ranges presented on single charts. For example in Figure 9, which presents percentage of companies, which declared presence of separated: logistic, transport, supply management, warehousing, packaging and information management departments presence. By using an uniform scale for all charts, for example you can see easy, a much lower frequency of separate packaging and information management departments presence in the vast majority of company categories.

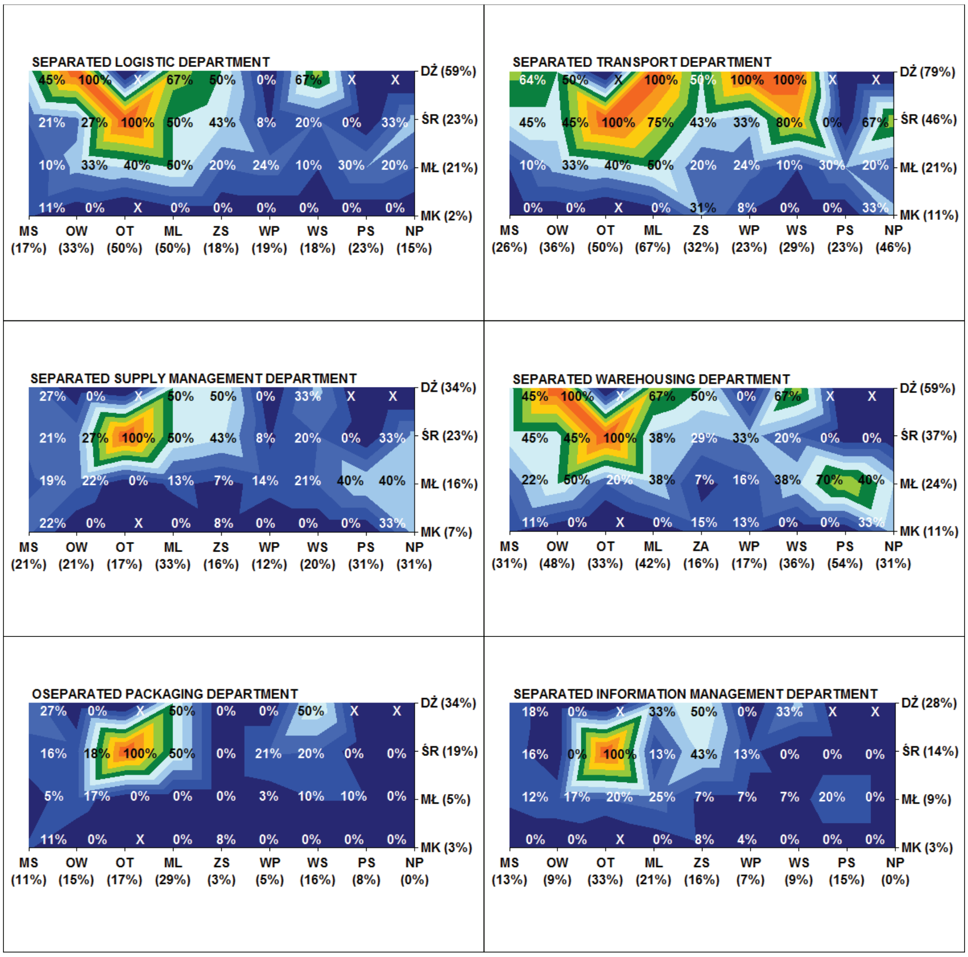

Figure 9. "Diversification map" charts comparison for declarations of different separate departments presence in companies according to food production branch (horizontal axe) and number of employees category (vertical axe).
Source: own preparation on the base of [6]

Sometimes, mainly for technical reasons (printing technique) "Diversification map" charts must be performed out only in shades of gray. Then, in addition to the various shades of gray colors, dotted surfaces are often used. Different kinds of dotting is using usually only for white and light gray colors (Figure 10).

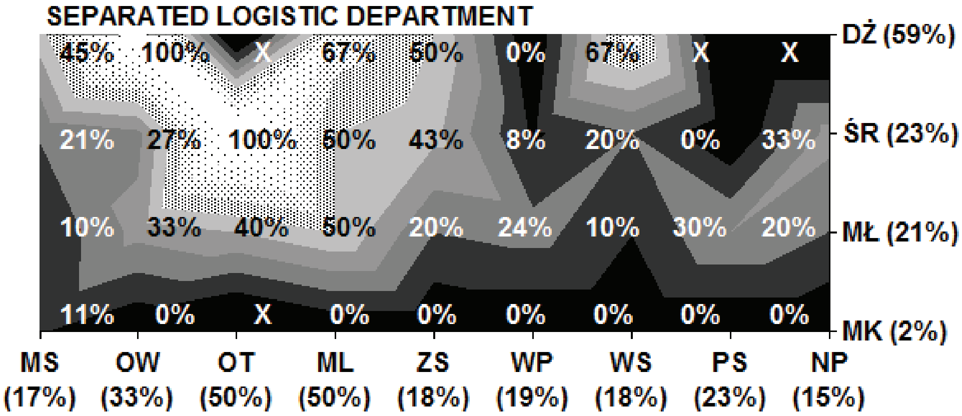

Figure 10. The "Diversification map", which presents the percentage of companies declared using of separate logistic department according to food production branch (horizontal axe) and number of employees category (vertical axe).
Source: own preparation on the base of [6]

5. Conclusion

In the paper, a proposition of simple way of three-dimensional or two-dimensional, changing in time data presentation is described. Presented "Diversification map" charts based on surface charts can be prepared with using of popular and commonly available software – Microsoft Excel. All presented charts may be prepared without using of VBA programming language, but accordingly written program can facilitate and automate the process of this type of charts creation. Discussion about assumptions, structure and implementation of such programs is outside of the thematic scope of this paper. "Diversification maps" are intuitive for users, primarily due to using of the popular scheme used to determine the values on hypsometric maps. Presented methods of graphical presentation of multidimensional data have been using for a few years in studies provided by the authors, and all examples in this paper originate from the results of these studies. Nowadays, there are provided works intended to increase the automation of the process of this type of chart creation using both VBA, and standard Excel tools as well.

REFERENCES

[1] N.Chen (1999) Extension *of Generalized Association Plots (GAP)*. Proceedings of the Statistical Graphics Section American Statistical Association, pp. 111–116.

[2] N.Gehlenborg, B.Wong (2012) Point *of view: Heats map.* Nature Methods, Vol. 9, Issue 3, pp. 213.

[3] Główny Urząd Statystyczny (1999) *Metodyka i organizacja badań budżetów gospodarstw* domowych. Zeszyty Metodyczne i Klasyfikacje. Departament Warunków Życia, Główny Urząd Statystyczny, Warszawa (in Polish).

[4] E.Jałowiecka, P.Jałowiecki, A.Orłowski (2010) Konsumpcja papierosów w gospodarstwach domowych w Polsce w latach 1999-2006 w zależności od typu biologicznego gospodarstwa. Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu, Warszawa-Poznań-Lublin, Tom XII, Zeszyt 4, 118-124, 2010 (in Polish).

[5] E.Jałowiecka, P.Jałowiecki, A.Orłowski (2012) Demograficzne zróżnicowanie konsumpcji papierosów w Polsce w latach 1999-2008. Roczniki Naukowe Stowarzyszenia Ekonomistów Rolnictwa i Agrobiznesu, 2012 (in Polish, in printing).

[6] P.Jałowiecki (2012) Informatyczne wspomaganie logistyki w przedsiębiorstwach przetwórstwa rolno-spożywczego w Polsce. Logistyka, 2012 (in Polish, in printing).

[7] P.Jałowiecki, E.Jałowiecka (2012) Struktura i koszty logistyki w wybranych branżach sektora rolno-spożywczego. Logistyka, 2012 (in Polish, in printing).

[8] E. Nowak (1990) *Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych.* Państwowe Wydawnictwo Ekonomiczne, Warszawa (in Polish).