DE GRUYTER OPEN

ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2018, Vol. 22, No. 1
ISSN 1507-3866; e-ISSN 2449-9994

APPLYING MARKET BASKET ANALYSIS TO OFFICIAL STATISTICAL DATA

Marcin Szymkowiak
Poznań University of Economics and Business, Poznań, Poland
Statistical Office in Poznań, Poznań, Poland
e-mail: m.szymkowiak@ue.poznan.pl

Tomasz Klimanek, Tomasz Józefowski
Statistical Office in Poznań, Poznań, Poland
e-mails: t.klimanek@stat.gov.pl; t.jozefowski@stat.gov.pl

DOI: 10.15611/eada.2018.1.03
JEL Classification: C10, C40

Abstract: Market basket analysis, which is a method of discovering co-occurrence relationships, is widely used for the purposes of marketing research and e-commerce, mainly by supermarkets and online stores. Moving beyond the traditional notion of a market basket understood as a fixed list of products, the technique can be applied for data mining in other fields of research which do not involve traditional transactions and purchases made by customers. The following article describes theoretical aspects of market basket analysis with an illustrative application based on data from the National Census of Population and Housing 2011 with respect to marital status. This is the first application of market basket analysis to census data to be conducted in Poland, in which attributes of the market basket have been replaced with respondents' demographic characteristics. This approach makes it possible to identify relationships between legal (*de jure*) marital status and actual (*de facto*) marital status, taking into account other basic socio-demographic variables available in large datasets. Using the R software to generate choropleth maps classified by province as a method of visualizing association rules, it was possible to conduct a spatial analysis of the phenomenon of interest.

Keywords: market basket analysis, Apriori algorithm, National Census of Population and Housing 2011, marital status, *arules* package, *arulesViz* package.

1. Introduction

The main purpose of market basket (association) analysis is to identify co-occurrence relationships and discover logical rules that describe associations between variables in a dataset by identifying positions where they co-occur. Market basket analysis (also known as association-rule mining) is a useful method of discovering customer purchasing patterns by extracting associations or co-occurrences from stores' transactional databases [Chen et al. 2005]. The technique originates from the analysis

of transactional data generated by retail chains. The name of its most common application, i.e. market basket analysis, refers to the customer's shopping basket. The method enables shop owners to make sales management decisions, plan production, promotion and arrange products in the most optimal places. It also helps the marketing analyst to understand the behavior of customers, e.g. which products are being bought together [Kaur, Kang 2016]. Over the years, market basket analysis has started to play an increasingly important role in the analysis of financial and insurance transactions [Roodpishi, Nashtaei 2015], in telecommunications [Jaroszewicz 2008] and in the pharmaceutical industry [Cerrito 2007; Hsieh et al. 2008]. This technique was also used on a large data set of cyclone conditions to derive hypotheses regarding which particular conditions are the best predictors of cyclones [Yang et al. 2007]. This trend reflects a shift away from the traditional notion of a market basket, understood as a collection of products, towards a wider interpretation as a bundle of services, service plans or subpage visits on a website. A market basket can now be associated with services used by customers of a bank or a mobile operator.

As a tool typically associated with marketing research [Kaur, Kang 2016] and management [Aguinis, Forcum, Joo 2013], market basket analysis has so far not been applied to analyse statistical data collected by the Central Statistical Office in Poland. If the traditional contents of a market basket is replaced with specific socio-demographic variables, the technique can also be applied to data from sampling surveys or censuses in order to discover association rules and co-occurrence relationships. For instance, using census data and information about legal (*de jure*) and actual (*de facto*) marital status, one could try to discover associations that can be used to identify groups of people for whom legal marital status is most likely not identical with actual marital status. A similar analysis could also be conducted for disability with the intention of identifying people or groups of people for whom disability in the biological sense is not identical with disability in the legal sense (and vice versa).

The main objective of this article is to present an application of market basket analysis in order to identify relationships between legal (*de jure*) marital status and actual (*de facto*) marital status, taking into account other basic socio-demographic variables available from official statistical data produced in Poland. Using data from the last National Census of Population and Housing (Census 2011), the R programming language and two R packages − *arules* and *arulesViz*, the authors show the usefulness of market basket analysis for identifying association rules in this dataset. The authors also try to find and explain the reasons behind the spatial differentiation of the detected rules between legal and actual marital status.

In addition to the introduction (the first part), the article consists of three main parts. The second part describes the theoretical foundations of market basket analysis, including the most essential concepts and definitions. It also provides information about the available software and areas of applications for this method of data exploration. We also briefly discuss how market basket analysis has been implemented

in the R software, particularly in the two main packages *arules* and *arulesViz* dedicated to this technique.

The third part presents an overview of the problems associated with surveys conducted for the purposes of official statistics, which contain information about respondents' marital status, including definitional difficulties of capturing the distinction between legal and actual marital status. The authors focus on data from the 2011 Census, in which the question of marital status and its two possible dimensions were analysed in detail.

The fourth part includes a description of the research procedure applied to real data from the 2011 Census. The results obtained for specific association rule are visualised in choropleth maps showing the spatial arrangement of identified relationships. Although market basket analysis and its potential applications are well documented in the literature on data mining, the research procedure applied to actual and legal marital status is an original implementation of this technique. It is also an example showing that many methods used in marketing can be adapted for the purposes of official statistics, which is dominated by techniques originating from estimation theory.

2. The idea of market basket analysis

Market basket analysis, also known as affinity analysis or association analysis, is one of the techniques used in data mining [Aguinis, Forcum, Joo 2013; Kaur, Kanwalpreet 2014]. The term 'data mining' itself refers to the process of exploring large data sets to discover patterns and extract meaningful information. Frawley et al. [1992] define data mining as the process of extracting hidden, previously unknown and potentially useful information from data. Hand et al. [Hand, Mannila, Smyt 2001] use the term to refer to the process of extracting information from large data sets or databases. According to Cabena et al. [1998], data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases. It means that data mining covers various groups of data exploration techniques, such as classification, clustering, regression, discrimination analysis and methods based on association rules.

Market basket analysis is one of the methods of discovering association rules, i.e. co-occurrence relationships between specific values of categorical variables, usually in large data sets [Raorane, Kulkarni, Jitkar 2012]. Its main objective is to find out what kind of products are usually purchased together, which ones are not purchased at all or are purchased rarely and what is the likelihood that a customer who has bought a given product will also buy another one. Providing answers to such questions was the original purpose of the method, which, as already mentioned in the introduction, was created to analyse transactional data from store chains [Russell, Petersen 2000]. The range of applications for market basket analysis is not limited to

shop transactions, which is the conventional application of this technique in data mining. It is also used in e-commerce, direct marketing, merchandising, logistics and the pharmaceutical industry, as mentioned in the introduction. Aguinis et al. [Aguinis, Forcum, Joo 2013] give a wide overview of market basket analysis use in bioinformatics, nuclear science, pharmacoepidemiology, immunology, geophysics and other fields. Market basket analysis can also be very useful in discovering spatial association rules. As noted by Łapczyński [2009], it was used to discover a link between the characteristics of districts in Helsinki and the number of car accidents that occurred in them. The authors of that study took into consideration such district characteristics as the presence of bars, restaurants, parks and cemeteries, main and local roads, etc. They analysed association rules such as "if a bar or restaurant, then a car accident". Thanks to market basket analysis, it was possible to identify districts with certain characteristics where car accidents were very likely.

Market basket analysis has been applied in the United Kingdom to census data in order to support decisions related to the transportation policy of the city of Manchester based on spatial association rules [Łapczyński 2009]. The application described in the present article is another example of discovering spatial association rules, in particular detecting association rules between categories of legal and actual marital status depending on certain demographic variables (sex, age, place of residence, etc.). This is the first application of market basket analysis to data collected by official statistics in Poland, i.e. to data from the last census conducted in 2011.

Market basket analysis focuses on association rules in the form of if-then statements, such as $A \rightarrow B$, where A and B are sets of the so-called attributes. In an association rule, A is an antecedent and B is a consequent. For association rules like $A \rightarrow B$, it is possible to define three important measures of significance and interest: support, confidence and lift [Berry, Linoff 2004; Larose 2005; Zhang, Zhang 2002]. These measures can be expressed in terms of probability as follows:

- Support of a rule expresses the probability of the co-occurrence of A and B; it indicates the proportion of transactions in the dataset of all transactions containing A and B:

$$\frac{n(A \cap B)}{N} = P(A \cap B). \tag{1}$$

- Confidence of a rule defines the conditional probability $P(B \mid A)$, i.e. the probability of event A occurring given that event B has occurred and indicates the proportion of transactions containing A, which also contain B:

$$\frac{n(A \cap B)}{n(A)} = P(B \mid A). \tag{2}$$

- Lift is defined as the ratio of the confidence to the unconditional probability of the consequent B; when the lift of a rule is greater than 1, it means that the purchase of A increases the probability of purchasing B:

$$\frac{\text{confidence}}{P(B)} = \frac{P(B \mid A)}{P(B)}, \tag{3}$$

where N denotes the number of all transactions, and $n(x)$ is the number of transactions containing x.

Association rules concerning transactions made by customers of shops can be formulated as:

- shirt → tie,
- cigarettes → lighter,
- diapers → beer,
- car and satnav → current map of Europe.

Association rules can include so-called trivial rules, which describe classic patterns of customer behaviour. This is exemplified by the rule shirt → tie, since the purchase of a new shirt is often accompanied by the purchase of a matching tie. Much more interesting, however, are the so-called useful rules, which have practical relevance and reflect unknown behaviour patterns. The rule diapers → beer is an example of a useful rule because the purchase of diapers does not necessarily determine the purchase of beer and the rule is not intuitive. As it turns out, many men who are shopping for diapers also buy beer. The discovery of such a non-trivial rule can affect the shop's promotional strategy on the sales floor and the placement of products on the shelves. Association rules can also include inexplicable rules, which are difficult to interpret and apply in the daily practice of running a shop.

Applied to other fields not related to the conventional market basket, association rules can take different forms depending on the topic of interest. For instance, with respect to marital status, which is the topic of this article, we can construct the following association rule: male and legal marital status (married) → actual marital status (partner). We can be interested in discovering rules which will help us to determine for which groups, defined by certain demographic characteristics, legal marital status can differ from actual marital status.

The search for association rules is extremely time-consuming owing to the large number of products that can be bought and the large number of transactions. It is therefore necessary to construct algorithms that optimise the search for association rules. There are many algorithms which are available for association rule mining. Existing algorithms find association rules on the basis of various metrics, such as support, confidence or lift [Kaur, Kang 2016].

A major contribution in the field of association algorithms was made by Agrawal and Srikant [1994], who proposed two new algorithms: Apriori and its extension AprioriTID. The Apriori algorithm is an iterative algorithm which looks for so-called frequent itemsets, which are representatives of sets of items that occur together in transactions. It is also assumed that the support of a frequent itemset is equal to or greater than a certain minimum support. Frequent itemsets are used to create association rules whose confidence is greater than or equal to a predefined minimum

value. The algorithm is described in detail in the article by Agrawal and Srikant [1994].

Market basket analysis, as a method of discovering association rules, has been implemented in many commercial and free statistical packages for data analysis. The list of commercial software includes SAS Enterprise Miner, IBM SPSS Modeler, Azmy SuperQuery, LPA Data Mining Toolkit, Magnum Opus and Wizsoft. Free packages include Apriori, ARtool, DM-II system, FIMI and R packages. The R software and its two packages – *arules* and *arulesViz* – are particularly effective tools for discovering and visualizing association rules. The first one – *arules* – is designed for discovering association rules. It can be used to identify all kinds of rules, set minimum values of support, confidence and lift, define antecedents and consequents and sort rules in ascending or descending order. The *arulesViz* package is particularly suitable for visualizing association rules using a wide range of charts. Detailed information about ways of visualizing association rules with examples can be found in the article by Hahsler and Chelluboin [2011].

Both packages were used to identify association rules concerning marital status on the basis of the 2011 census data. Although the R packages provide a number of chart types to visualize results of market basket analysis, we have decided to use choropleth maps in order to analyse the spatial distribution of association rules for marital status. This, however, does not limit the visualisation possibilities offered by the *arulesViz* package.

3. Marital status in selected statistical surveys

Marital status is a key demographic variable, which determines people's behaviour and attitudes in the socio-economic context. Categories of marital status can be treated as a formally or legally defined status or as indications of the actual state of affairs. In Poland, censuses and microcensuses are among the main sources of information about the population structure by age and marital status. Until 2002, censuses only provided information about the actual marital status, and not according to the legal definition. Information on this topic was based on respondents' declarations that did not have to be supported by any documents [Kędelski, Paradysz 2013].

During the census of 2002, for the first time information was collected not only about the actual marital status but, following international recommendations, also about respondents' legal marital status, which was defined in accordance with Polish law [GUS 2003].

Unlike earlier ones, the census of 2011 was a mixed mode census, where data were obtained from administrative registers and also collected from respondents in a sample survey and in a full enumeration survey.

In keeping with international recommendations, information was collected about legal and actual marital status. Taking into account the regulations existing in

different countries, marital status was to be established for people aged 15 and over. Legal (official) marital status was defined as the legal status of a given person according to the law of a given country.

According to Polish law, there are four categories of legal marital status:

- single – persons who have never been in a legal marriage;
- married – persons whose marriage has been contracted according to civil law;
- widower, widow – persons whose legal marriage has ceased to exist following the spouse's death;
- divorced – persons whose marriage has been dissolved by a court decision.

It is worth noting that married couples for whom courts have issued separation orders, from the legal point of view, remain legally married.

The person's actual marital status was defined on the basis of the legal marital status and the nature of the relationship in which the person currently lives, i.e. based on information about relationships and mutual bonds between members of a given household. Taking into account this information, the following categories of actual marital status can be distinguished:

- single – persons who have never been in a legal marriage and at the time of the census are not in a cohabiting union;
- married – persons who have contracted a legal marriage and actually live as a married couple. The declaration is valid regardless of whether both spouses are enumerated together (in one dwelling) or separately (as a result of absence due to study, work or lack of a common dwelling);
- partner – persons living in cohabiting unions in the same household, regardless of the legal marital status of cohabiting partners;
- widower, widow – persons whose legal marriage has ceased to exist following the spouse's death and at the time of the census are not living in a cohabiting union with another person;
- divorced – persons whose marriage has been dissolved by a court decision and at the time of the census are not living in a cohabiting union with another person;
- separated.

The last category refers to persons who at the time of the census:

- are in a state of legal separation and are not in a cohabiting union with another person;
- are legally married but no longer live as a married couple and are not in a cohabiting union with another person [GUS 2013].

It is worth noting that legally married persons who no longer live as a married couple as a result of a decision made by one or both spouses, but the dissolution of their marriage has not been sanctioned by a court decision (divorce or separation), are not treated as married, and their actual marital status is recorded according to the respondent's declaration, either as separation or as being in a cohabiting union with another person.

Information about marital status can also be found in repeated surveys conducted by the Central Statistical Office, such as the Household Budget Survey, the EU-SILC survey, or the Labour Force Survey. All of these surveys, however, only collect information about legal marital status and are based on samples that are too small to support reliable estimates at lower levels of spatial aggregation. It should be noted that the questionnaires for the Household Budget Survey and the EU-SILC survey (BR-01a and EU-SILC-G, EU-SILC-I, respectively), apart from the question about marital status, also contain a question in which the respondent is asked whether they are currently in a relationship with a person living in the same household. Although possible answer options for this question provide additional information about the relationship status (formal, informal, no relationship), it is not sufficient to obtain such detailed classification of actual marital status as that provided by the 2011 census; there is also no information about relationships with persons not living in the same household.

Finally, a person's legal marital status can also be determined on the basis of information stored in administrative data sources, such as the PESEL register (Universal Electronic System of Population Register Number), the National System of Monitoring Family Benefits, the National System of Monitoring the Alimony Fund or the Central Statistical Application (with respect to childcare for children up to the age of 3).

4. Application of MBA to data from Census 2011 with respect to marital status

4.1. Description of the research procedure

The purpose of the research procedure described in this article is to test the possibility of applying market basket analysis to data from the 2011 census. The following preliminary research hypothesis has been formulated: "The spatial distribution of characteristics that describe the identified association rules across provinces is not uniform, which could imply that the territorial factor affects relationships between different categories of legal and actual marital status."

In order to verify the above hypothesis and achieve the main research objective, we used data from the 2011 census, restricting the population of interest to persons aged 15 and over[1]. Also, given the fact that it was the first mixed-mode census in the history of Polish statistics, supplied with data from administrative registers and from a survey (e.g. on the actual marital status) involving a 20% sample of all dwellings in Poland, the dataset was limited to those respondents that could be unambiguously

[1] Since different countries have different regulations concerning the marriageable age, it has been decided that in censuses, the marital status should be determined for people aged 15 and over [GUS 2013].

identified in terms of legal and actual marital status[2]. The final dataset contained information about 7,676,815 persons and, in addition to the target variables, i.e. legal and actual legal status, included the following demographic variables[3]:

- sex (male, female),
- place of actual residence (urban, rural),
- actual marital status (single, married, partner, widower/widow, divorced, separated),
- legal marital status (single, married, widower/widow, divorced),
- age (under 18, 18-24, 25-34, 35-44, 45-54, 55-64, 65 and over),
- province[4].

The next step consisted in converting data into binary format (see Table 1), where different levels of variables are encoded as binary variables (0/1) and the dataset contains an additional variable indicating the transaction number (in our case it is the number of the person in the dataset)[5]. This stage of data processing was performed by means of a 4GL code implemented in SAS. Table 1 shows a fragment of the resulting dataset.

Table 1. Data from Census 2011 converted into binary format for market basket analysis

TRANS	MAN	WOMAN	CITY	RURAL	SS_F	MM_F
1	0	1	1	0	0	1
2	0	1	1	0	0	1
3	0	1	1	0	0	0
4	0	1	0	1	0	1
5	0	1	1	0	0	1
6	0	1	1	0	0	1
7	0	1	1	0	0	0
8	1	0	1	0	0	0
9	0	1	1	0	0	1
10	1	0	0	1	1	0
11	0	1	1	0	1	0
12	0	1	1	0	0	0
13	0	1	1	0	0	0
14	1	0	1	0	1	0
15	1	0	1	0	1	0

Source: own elaboration.

[2] We removed records with missing data or those containing the category "unspecified", e.g. with respect to the area of residence (urban/rural).

[3] Information in brackets refers to variable levels used in the analysis.

[4] The codes used by the Central Statistical Office, 02 – Dolnośląskie, 04 – Kujawsko-Pomorskie, …, 32 – Zachodniopomorskie.

[5] For example, MAN is a variable created by converting the SEX variable: it is 1 when the person is male, and 0 when the person is female. The other variables have been converted in the same way (except the variable TRANS, which indicates the number of transaction – person).

After conversion, the dataset contained 7,676,815, "transactions" and 38 columns. Owing to the technical difficulties of processing such a large dataset in R (about 2.3 GB), it was split into 16 parts, each containing data about one province. After converting provincial datasets into .csv files, they were imported into the R workspace and processed using the *arules* and *arulesViz* package.

4.2. Construction of association rules

To construct association rules in the form of A→B, i.e. if A then B, where A is the antecedent, and B is the consequent of a transaction, binary variables representing different levels of legal marital status and sex, age group, place of residence were chosen as antecedents, and binary variables representing different levels of actual marital status were used as consequents.

The following R code implements the search for association rules[6]:

```
asocjacje2 <- apriori(trans,
parameter = list(support=0.000001,conf=0.000001,minlen=2),
appearance   = list(rhs=c('SS_F','MM_F','PP_F','WW_F','DD_F','SP_F','NN_F'),
              lhs=c('RURAL','CITY', 'AGE1','AGE2','AGE3','AGE4',
                    'AGE5', 'AGE6','AGE7','MAN','WOMAN', 'SS_J',
                    'MM_J','WW_J','DD_J','NN_J')))
```

One aspect that requires a brief explanation are the settings for the measures of support (support = 0.000001) and confidence (conf = 0.000001). The default values are 0.1 and 0.8, respectively. The values used in the above code are dictated by the specific nature of market basket analysis. Whereas in the classic application of this technique[7] one typically looks for frequent rules, the search for rules that characterise relatively rare phenomena (e.g. actual marital status = "widow/widower") requires a change of focus. With the default setting, i.e. high values of support and confidence, it would be impossible to find such rules.

The parameter "minlen" is set to 2 in order to prevent the creation of empty rules, such as: {}=>PP_F, which would be the case if it was set to 1. The parameters "lhs" and "rhs" indicate which binary variables should appear on the left or the right side of the rule.

---

[6] Strings in the R code denote names of variables used in the rules detection procedure. Therefore all variable names ending in "_F" refer to de facto marital status. All variable names ending in "_J" are related to de jure marital status. Characters preceding the underscore in the variable names denote different categories of marital status: SS – male single/female single, MM – male married/female married, WW – widower/widow, DD – male divorced/female divorced, SP – male separated/female separated, PP – male partner/female partner, NN – unknown marital status. The AGE variable followed by a number denotes different age groups. For example, AGE2 denotes people aged 18-24.

[7] This application is associated with sales management, production planning, promotion, optimal product placement in the store, but is also used in the analysis of financial and insurance transactions, telecommunications, logistics and the pharmaceutical industry.

By applying the Apriori algorithm proposed by Agrawal and Srikant [1994] and implemented in the *arules* package, a total of 21,325 rules were identified (for all provinces). The next section presents selected results of rules detection together with their spatial variation.

4.3. Selected results of market basket analysis

From the set of rules describing associations between legal and actual marital status we have selected eight for closer analysis. The selection of the rules was partly arbitrary and partly data driven. We focused on two subpopulations: widows/widowers and divorced people and the question whether their de facto status involves 'living alone' or in a kind of informal relationships like a partnership. Having such a subset of all the rules, we selected those with the highest support, taking into account that the rules should not be trivial but they should contain at least information about de jure marital status, age, sex and place or residence. In addition to visualizing the spatial variation in two measures of the association rules (Figures 1-8) – confidence and lift, our aim is to investigate how other demographic variables (sex, age and place of residence) affect actual marital status, which is obviously directly associated with legal marital status.

The legal marital status of nearly all widows aged the 55-64 and living in urban areas, is consistent with their actual marital status. The confidence of this type of rules ranges from 97% for the provinces situated along the western border (Zachodniopomorskie, Lubuskie and Dolnośląskie) to over 99% in the provinces of south-eastern Poland (Świętokrzyskie, Małopolskie and Podkarpackie). Cohabitation is not common for this group of women, and there is an evident difference between western provinces – confidence of about 3%, and the provinces of south-eastern Poland – confidence of below 1% (Figure 1). Additional information is provided by the analysis of lift, which indicates how many times greater the conditional probability is than the unconditional probability (Figure 2). It turns out that the probability of cohabitation for this group of women is the biggest in Opolskie and Lubelskie (the lift is 1.61 and 1.73, respectively), while in Lubuskie, Łódzkie and Pomorskie the lift is around 1, which means that in these provinces for women aged 55-64 and living in urban areas the status of legal widowhood does not increase the likelihood of cohabitation.

The legal marital status of nearly all widowers aged 55-64 and living in urban areas is consistent with the de facto status. Rules of this type have the confidence ranging from 95% for western provinces (Zachodniopomorskie and Lubuskie) to over 98% in south-eastern provinces (Lubelskie and Podkarpackie). Interestingly, this group also includes Wielkopolskie province – confidence of 98% (Figure 3). The probability of entering into partnership unions in this group of widowers is slightly higher than in the group of widows described previously. Once again, there is an evident difference between the western provinces – confidence exceeding 5%, and the south-eastern provinces – confidence under 2%. The analysis of lift, especially with

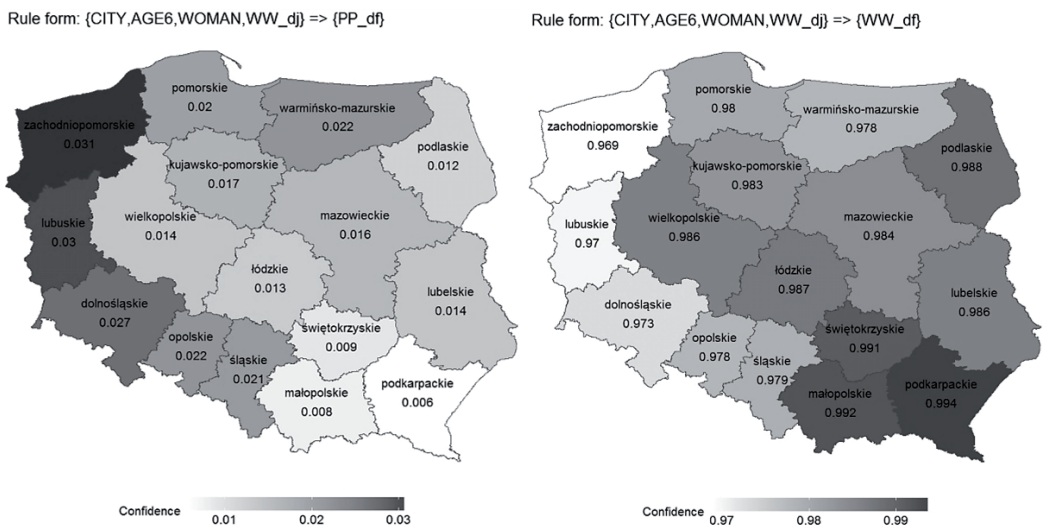

Figure 1. Spatial distribution of confidence for selected association rules for widows (de jure marital status – left panel) and for female partners (de facto status – right panel) aged 55-64, living in urban areas

Source: own elaboration.

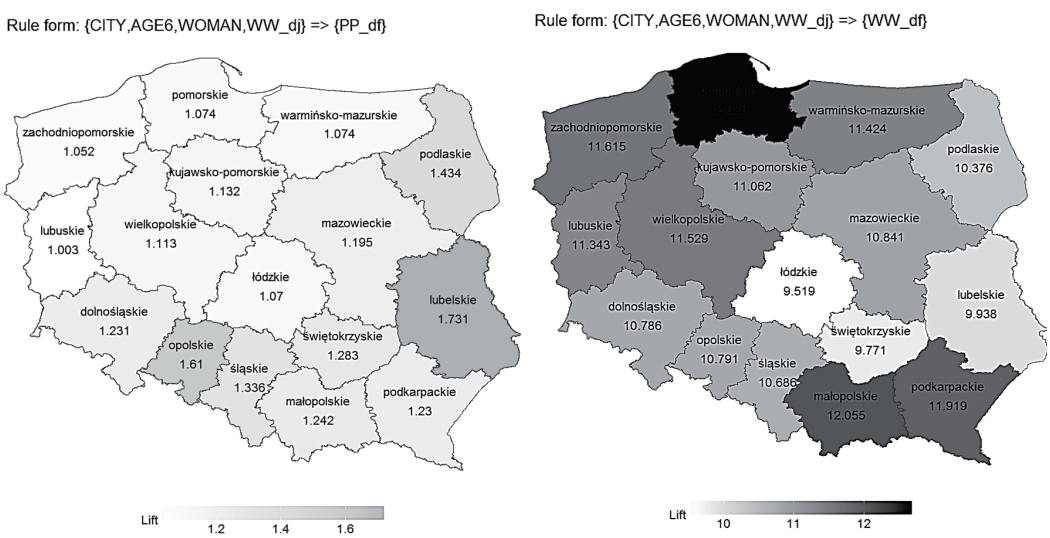

Figure 2. Spatial distribution of lift for selected association rules for widows (de jure marital status – left panel) and for female partners (de facto status – right panel) aged 55-64 living in urban areas

Source: own elaboration.

respect to the group of widowers actually living in partnership unions, indicates that the probability of entering into such unions is higher in Małopolskie and Podkarpackie province (lift is equal to 3.90 and 3.67 respectively), i.e. in these provinces the fact

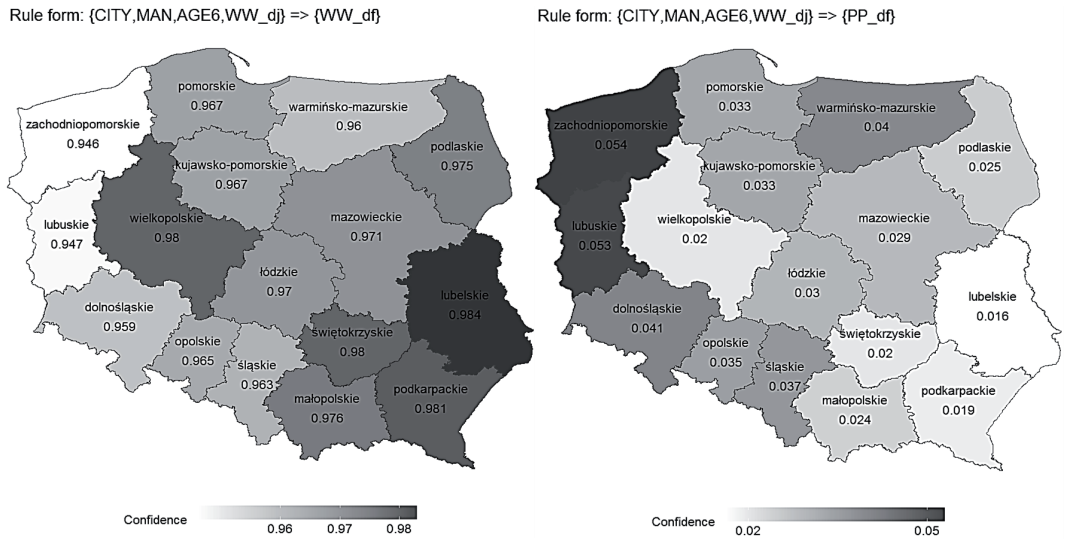

Figure 3. Spatial distribution of confidence for selected association rules for widowers (de jure marital status – left panel) and for male partners (de facto status − right panel) aged 55-64 living in urban areas

Source: own elaboration.

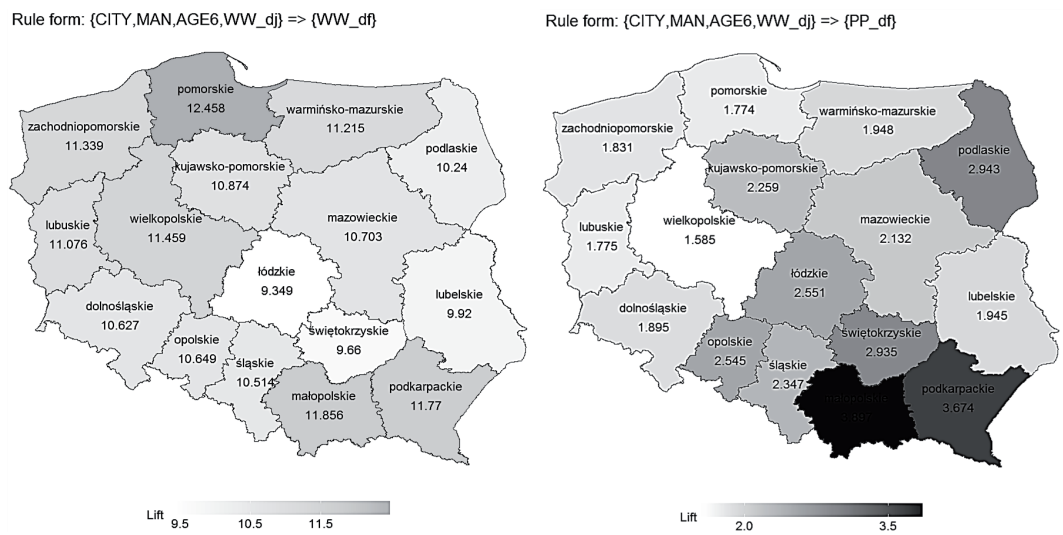

Figure 4. Spatial distribution of lift for selected association rules for widowers (de jure marital status – left panel) and for male partners (de facto status − right panel) aged 55-64, living in urban areas

Source: own elaboration.

of being a widower (in the legal sense) aged 55-64 living in an urban area, is associated with a four-fold increase in the probability of entering into partnership unions. This correlation is less likely in Wielkopolskie province, where lift is equal to 1.59 (Figure 4).

Analysis of association rules concerning city-dwelling divorced women aged 35-44 reveal a number of patterns (Figure 5). First of all, once again there is a clear division between the western, more liberal part of the country, where this group of women is more likely to enter into partnership unions (confidence for Zachodniopomorskie and Lubuskie provinces fluctuate around 12%) and the eastern part, characterised by lower values of confidence (below 7% for Podkarpackie, Świętokrzyskie and Podlaskie provinces). Śląskie province, with confidence of 12%, is an interesting exception in this respect. Taking into account the interpretation of the rule, this means that the population of the province can also be described as more liberal. Another characteristic of the provinces of eastern Poland is the high likelihood of consistency between the legal and actual marital status of women in this group: confidence for this type of association rules exceeds 93% in the case of Podkarpackie, Świętokrzyskie and Podlaskie.

Analysis of lift, especially regarding the group of divorced women living in partnership unions (Figure 6) indicates that the probability of such unions increases the most in Małopolskie province (lift equal to 13.73), which means that the fact of

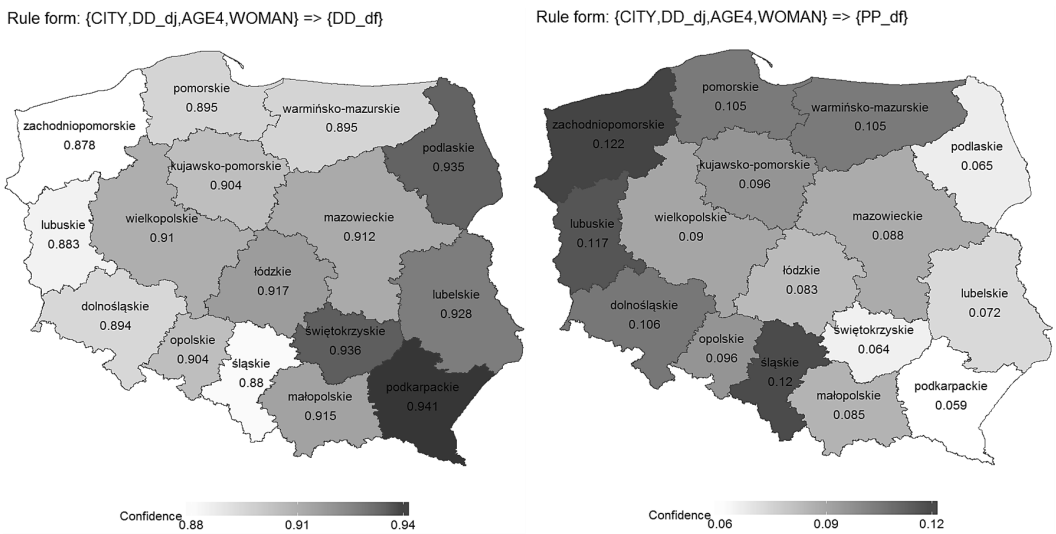

Figure 5. Spatial distribution of confidence for selected association rules for divorced women (de jure marital status – left panel) and for female partners (de facto status – right panel) aged 35-44, living in urban areas

Source: own elaboration.

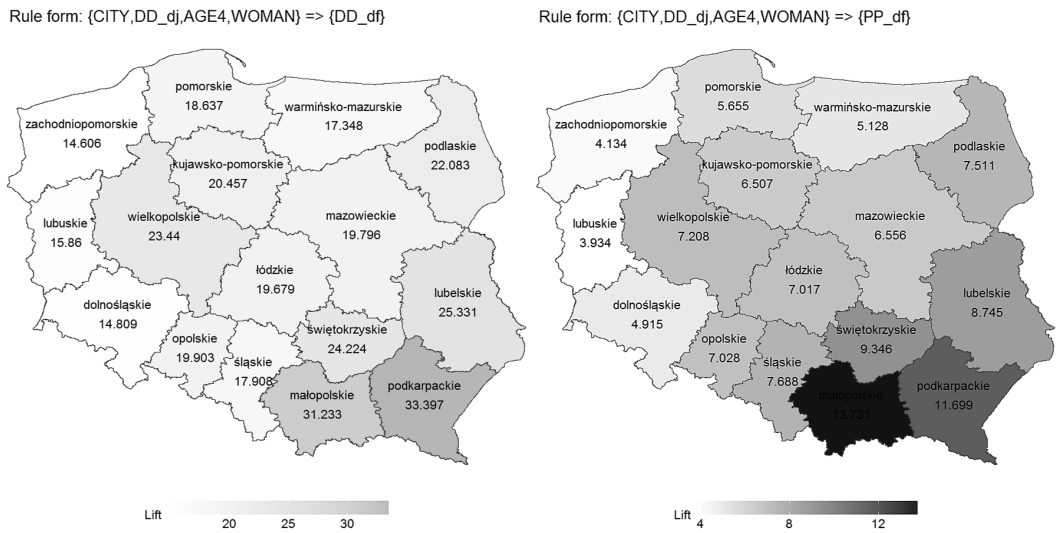

Figure 6. Spatial distribution of lift for selected association rules for divorced women (de jure marital status – left panel) and for female partners (de facto status – right panel) aged 35-44, living in urban areas

Source: own elaboration.

being a legally divorced female city-dweller aged 35-44 is associated with a 14-fold higher likelihood of entering into partnership unions.

The marital status of widowers aged 35-44 living in urban areas is generally consistent with the actual status (rules of this kind have a confidence of 85% for Lubuskie province and 94% in the case of Świętokrzyskie province – Figure 7). The probability of entering into partnership unions in this group of widowers is similar to that identified for the above mentioned group of widows. In most provinces confidence for this type of rules is higher for widowers than for widows, with the exception of the provinces of Śląskie and Łódzkie, where widows are slightly more likely to enter into partnership unions than widowers. Once again we can observe the disparity between the western and eastern part of the country: for Zachodniopomorskie and Lubuskie confidence exceeds 13.5%, compared to under 9% for Świętokrzyskie, Małopolskie and Podkarpackie. Analysis of lift for the group of widowers actually living in partnership unions indicates that the probability of entering such unions increases the most in the provinces of Małopolskie and Podkarpackie (lift equal to 11.81 and 16.67 respectively), which implies that in these provinces the fact of being a city-dwelling widower (in the legal sense) aged 35-44 is associated with an 11-fold and 16-fold higher probability of entering into partnership unions. In the province of Zachodniopomorskie this probability is the lowest – lift of 4.62 (Figure 8).

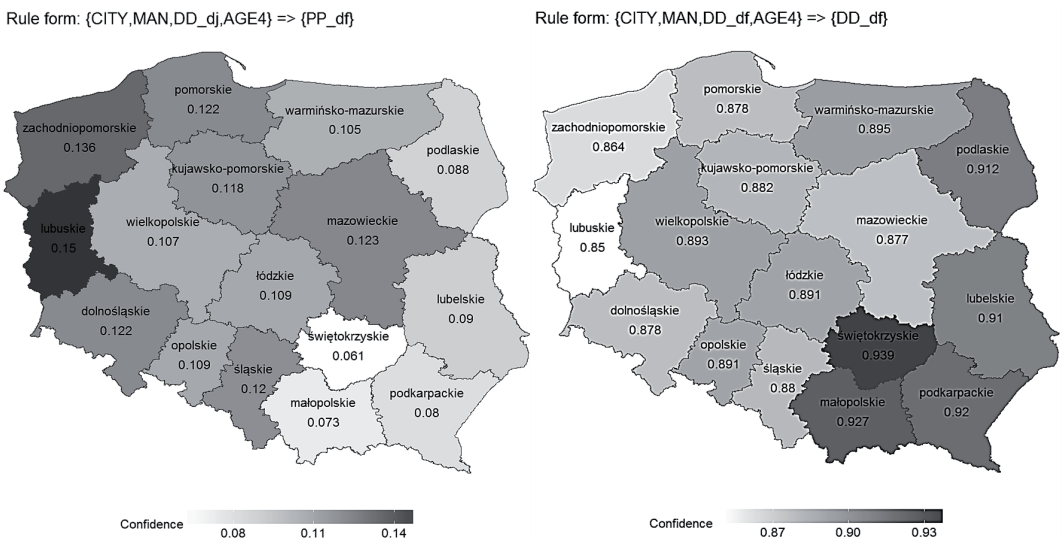

Figure 7. Spatial distribution of confidence for selected association rules for divorced men (de jure marital status – left panel) and for male partners (de facto status – right panel) aged 35-44, living in urban areas

Source: own elaboration.

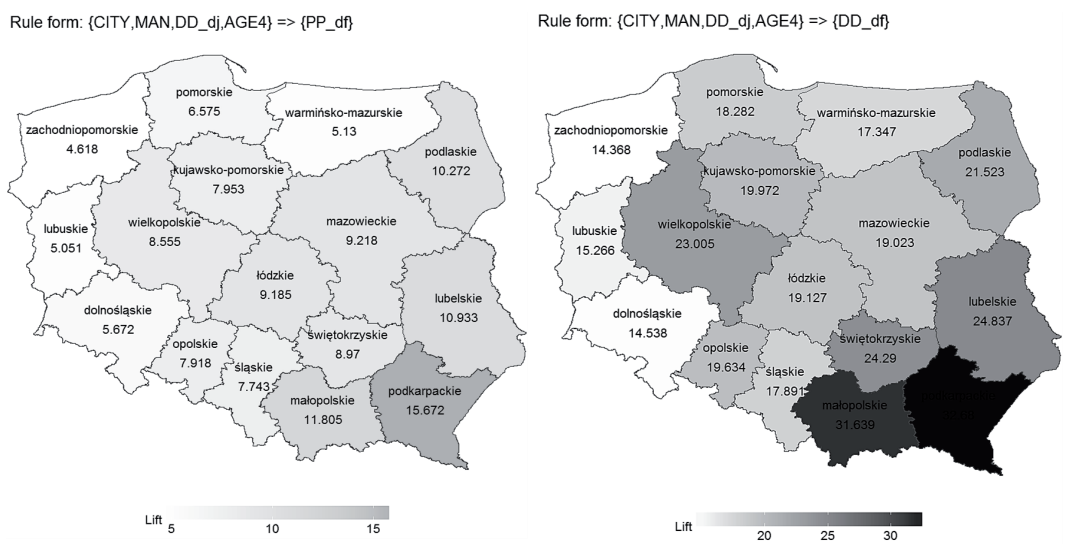

Figure 8. Spatial distribution of lift for selected association rules for divorced men (de jure marital status – left panel) and for male partners (de facto status – right panel) aged 35-44, living in urban areas

Source: own elaboration.

5. Summary

The procedure described in the article is an example of extending the range of applications for market basket analysis. In our opinion we have managed to demonstrate that market basket analysis, which is typically applied in marketing studies, can be applied to identify rules describing associations between *de jure* and *de facto* marital status, taking into account other socio-demographic variables available in large data sets. Also, the results obtained in the study confirm the possibility of identifying spatial patterns governing associations between legal and actual marital status. These patterns were analysed across provinces, which do not have to coincide with the regional distribution of similar social attitudes generally described as more liberal or more conservative. Higher values of lift for women and men creating partnership relationships in the provinces located in the eastern part of Poland: Podkarpackie, Małopolskie, Świętokrzyskie, Lubelskie and Podlaskie can be explained by the lowest probabilities of cohabitation in these regions; of course, additional sociological analysis is necessary to explain the phenomenon in more detail.

It is worth pointing out that the use of market basket analysis to detect association rules for small subpopulations requires a completely different hierarchy of measures known as support, confidence and lift. In classical MBA applications, a lot more importance is attached to high values of support. This is, of course, the result of a desire to generate the largest possible profit, e.g. in retail chains – in other words, the focus is on detecting the most frequent rules. In contrast, the study described in the article focused on rules for subpopulations of widows and widowers, divorced persons or people living in partnership unions, which are relatively rare in Polish conditions. For this reason, in applications involving the detection of rules for very specific, small subpopulations, more importance needs to be attached to confidence, and, above of all, to lift.

We realise that human behaviour related to marital status is not only affected by sex, age and place of residence and a more in-depth analysis of determinants of this phenomenon would have to include other socio-economic variables. However, it seems that market basket analysis can be used as a quick diagnostic tool to detect such differences.

It should be noted that certain choices made at the design stage of the study, which were aimed at accelerating calculations (e.g. the use of separate data sets for each province instead of one data set for the whole country), should be modified in the future. The objective would be to use territorial division (by province) as a variable to be taken into account when constructing association rules. It would also be necessary to analyse classification rules not only in terms of the rural/urban distinction but also taking into account a more detailed classification of towns in terms of population size, e.g. over 500,000, 200-500,000, etc.

The results of applying market basket analysis to detect rules describing associations between legal and actual marital status are encouraging and provide the motivation to continue testing the usefulness of MBA in the analysis of other socio-economic variables provided by official statistics in Poland, e.g. disability (in the legal and biological sense), or unemployment (based on registration or the ILO definition used in the LFS).

Bibliography

Agrawal R., Srikant R., 1994, *Fast Algorithms for Mining Association Rules*, Proceedings of the 20th VLDB Conference Santiago, Chile.

Aguinis H., Forcum L.E., Joo H., 2013, *using market basket analysis in management research*, Journal of Management, vol. 39, no. 7, pp. 1799-1824.

Berry M.J.A., Linoff G.S., 2004, *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management (2nd Ed.)*, Wiley, Indianapolis.

Cabena P., Hadjinian P., Stadler R., Verhees J., Zanasi S., 1998, *Discovering Data Mining: From Concept to Implementation*, Prentice Hall, Upper Saddle River, NJ.

Cerrito P.B., 2007, *Choice of antibiotic in open heart surgery*, Intelligent Decision Technologies, 1, pp. 63-69.

Chen Y-L., Kwei T., Shen R-J., Hu Y-H., 2005, *Market basket analysis in a multiple store environment*, Decision Support Systems, vol. 40, issue 2, pp. 339-354.

Frawley W., Piatesky-Shapiro G., Matheus C., 1992, *Knowledge Discovery in Databases: An Overview*, Al Magazine.

GUS, 2003, *Ludność. Stan i struktura demograficzno-społeczna*, NSP 2002, Warszawa.

GUS, 2013, *Ludność. Stan i struktura demograficzno-społeczna*, Narodowy Spis Powszechny Ludności i Mieszkań, Warszawa, http://stat.gov.pl/download/cps/rde/xbcr/gus/LUD_ludnosc_stan_str_dem_spo_NSP2011.pdf, dostęp: 30.06.2017.

Hand D., Mannila H., Smyt P., 2001, *Principles of Data Mining*, MIT Press, Cambridge, MA.

Hahsler M., Chelluboina S., 2011, *Visualizing Association Rules: Introduction to the R-extension Package arulesViz*.

Hsieh S-C., Lai J.-N., Lee C.-F., Hu F.-C., Tseng W.-L., Wang J.-D., 2008, *The prescribing of Chinese herbal products in Taiwan: A cross-sectional analysis of the national health insurance reimbursement database*, Pharmacoepidemiology and Drug Safety, 17, pp. 609-619.

Jaroszewicz Sz., 2008, *Cross-selling models for telecommunication services*, Journal of Telecommunications and Information Technology, vol. 3, pp. 52-59.

Kaur M., Kang S., 2016, *Market Basket Analysis: Identify the changing trends of market data using association rule mining*, Procedia Computer Science, 85, pp. 78-85.

Kaur P., Kanwalpreet S.A., 2014, *Data Mining: Review*, International Journal of Computer Science and Information Technologies, 5(5), pp. 6225-6228.

Kędelski M., Paradysz J., 2013, *Demografia*, Poznań.

Larose D.T., 2005, *Discovering Knowledge in Data: An Introduction to Data Mining*, Hoboken, Wiley Interscience.

Lasek M., Pęczkowski M., 2013, *Enterprise Miner. Wykorzystanie narzędzi Data Mining w systemie SAS*, Wydawnictwo Uniwersytetu Warszawskiego, Warszawa.

Łapczyński M., 2009, *Analiza koszykowa i analiza sekwencji – wielki brat czuwa*, StatSoft Polska.

Package 'arules', 2017, https://cran.r-project.org/web/packages/arules/arules.pdf, dostęp: 9.02.2017, dokumentacja pakietu program R.

Package ‘arulesViz’, 2017, https://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf, dostęp: 9.02.2017, dokumentacja pakietu program R.

Raorane A.A., Kulkarni R.V., Jitkar B.D., 2012, *Association rule – extracting knowledge using market basket analysis*, Research Journal of Recent Sciences, 1(2), pp. 19-27.

Roodpishi M.V., Nashtaei R.A., 2015, *Market basket analysis in insurance industry*, Management Science Letters 5, pp. 393-400.

Russell G.J., Petersen A., 2000, *Analysis of cross category dependence in market basket selection*, Journal of Retailing, 76, pp. 367-392.

Yang R., Tang J., Kafatos M., 2007, *Improved associated conditions in rapid intensifications of tropical cyclones*, Geophysical Research Letters, 34, pp. 1-5.

Zhang C., Zhang S., 2002, *Association Rule Mining: Models and Algorithms*, Springer, Berlin.

ANALIZA KOSZYKOWA I JEJ ZASTOSOWANIA W STATYSTYCE PUBLICZNEJ

Streszczenie: Analiza koszykowa jako metoda poszukiwania odpowiednich reguł asocjacyjnych jest szeroko wykorzystywana w badaniach marketingowych i w handlu elektronicznym, głównie przez supermarkety czy sklepy prowadzące sprzedaż on-line. Odchodząc od tradycyjnego rozumienia koszyka i zawartych w nim produktów, można zastosować również tę technikę Data Mining w innych obszarach badawczych, w których nie mamy do czynienia z tradycyjnym rozumieniem pojęcia transakcji i produktów nabywanych przez klientów. W artykule przedstawiono teoretyczne aspekty analizy koszykowej i jej egzemplifikację na danych pochodzących z Narodowego Spisu Powszechnego Ludności i Mieszkań 2011 w odniesieniu do stanu cywilnego. Jest to pierwsza tego typu aplikacja dla danych spisowych w Polsce, w których atrybuty koszyka zakupów zastąpiono odpowiednimi cechami demograficznymi osób. Dzięki takiemu podejściu możliwa była identyfikacja reguł opisujących związki między stanem cywilnym prawnym a stanem cywilnym faktycznym przy uwzględnieniu innych podstawowych zmiennych społeczno-demograficznych w dużych zbiorach danych. Wizualizacja uzyskanych reguł asocjacyjnych w programie R na odpowiednich kartogramach w układzie wojewódzkim umożliwiła ponadto przestrzenną analizę zróżnicowania badanego zjawiska.

Słowa kluczowe: analiza koszykowa, algorytm Apriori, Narodowy Spis Powszechny Ludności i Mieszkań 2011, stan cywilny, pakiet arules, pakiet arulesViz.