ANALYSIS OF SIMILARITIES IN CONTEXT OF ENTERPRISE INNOVATIONS

Zdena Dobešová, Vít Pászto, Karel Macků

Palacký University in Olomouc, Faculty of Science, Department of Geoinformatics

Abstract: *Innovations are important activities in enterprise development. The presented research detects similarities in enterprises innovation strategies. In this paper, the authors used dataset provided by Community Innovation Survey 2010, describing various innovation activities of surveyed enterprises in the Czechia. The enterprise innovations are reported in two groups of indicators: technical (product and process) innovations and non-technical (marketing and organisational) innovations. As the information about innovation is represented by binary values, coefficients of association were used for analysis of relations between objects. Detection of similarities by a coefficient of similarity is one of basic data mining methods. Moreover, the cluster analysis was used for result values of the similarity coefficients. Finally, the evaluation of similarity is presented for the selected Czech district.*

Keywords: *Innovations, association, enterprises, data mining, cluster analysis.*

JEL classification: *O3, O32, Q5*

Grant affiliation: *This paper is supported by the Spationomy project (2016-1-CZ01-KA203-024040) funded by the European Union and by the student project IGA_PrF_2017_024 of the Palacký University Olomouc.*

1. Introduction

Innovations are crucial for economic growth and (regional) development and are considered as a key to growth and long-term success, especially in "knowledge-driven-economy". According to (Sternberg, 2000), seeing as innovation requires information and knowledge, continual development and production of innovative products function as decisive elements for successful regional development. Generally, in entrepreneurship context, innovations are treated as something new with added value to company's performance helping to increase its competitiveness on the (regional) market. In developed countries, innovations are widely seen as the basis of their competitive economies (Porter and Ketels, 2003). An enterprise that tries to improve its position on the market should implement or adopt an appropriate innovation policy for sustaining competitive advantage. Due to innovation activities, managers are empowered to influence and cultivate their environment actively. Therefore, it is important to analyze innovation activities and potential of the (regional) market in the managerial decision-making process. In the literature, a term regional innovation network or system is used. The regional innovation network or system is stimulating growth and innovation from an individual business perspective and a regional perspective (Sternberg, 2000; Cooke, 1992). Then, it is a question, how to describe, evaluate, and analyse a (regional) innovation performance. One of the most used methods for acquiring desired data about innovations is questionnaire survey (e.g. Kirton, 1976; Maillat, Quévit and Senn, 1993, Sternberg, 2000). In this paper, a questionnaire-based survey called Community Innovation Survey conducted by the European Union was used in order to analyse the similarity of enterprises' innovation activities. The main objective of this study was to explore (dis)similarities in selected LAU 1 (Local Administrative Units) regions in Czechia with the use of the coefficient of similarity and consequent clustering.

2. Data

Fifth period of Community Innovation Survey (CIS) covering years 2008 to 2010 was used (hereafter as CIS 2010). This survey is carried out by all EU member states and uses harmonised questionnaire (EUROSTAT, 2012). Data collection for CIS 2010 were organised in 2011 by a questionnaire a survey focusing on all enterprises with ten or more employees, stratified by size and economic activity. In total 5,151 responses, representing 21 % of the total statistical population, were received with 83 % return rate of useful answers (ČSÚ, 2013; Vaculík et al., 2017).

The enterprise innovations are reported in two groups of indicators: technical (product and process) innovations and non-technical (marketing and organisational) innovations. The technical innovations consist of two product indicators concerning innovation of final products or service. Next three technical process indicators refer to an improvement of production, supply and distribution of products, and change of accounting and information systems. The non-technical innovations are comprise four marketing indicators (new design of packaging of products, advertisements, licensing and franchising), and by three organisational indicators (internal business practices for process organization, internal workflow changes, and external change of relationships). A total number of twelve innovation indicators represents source data for the analysis of similarities and clustering of enterprises in the Czechia.

3. Methodology

This chapter describes the main methodological steps of data processing. Firstly, the calculation of similarity coefficients is mentioned. Secondly, the possibilities of use of clustering according to the values of dissimilarity coefficients are presented.

3.1. Coefficient of similarity

The coefficient of similarity is one possible measure of the enterprise innovation similarity. Binary similarity coefficients are used when only presence-absence data are available. The basic data for calculation of binary similarity coefficients is a 2x2 frequency table (TAB.1), i.e. for two entities. That frequency table is calculated from input binary data that describes types of innovations.

TAB. 1: Basic frequency table of number of presence-absence innovations

		Enterprise A	
		Number of innovations present (1)	Number of innovations absent (0)
Enterprise B	Number of innovations present (1)	a	b
	Number of innovations absent (0)	c	d

where:

a - number of innovations in enterprise A and enterprise B (joint occurrences)

b - number of innovations in enterprise B but not in enterprise A

c - number of innovations in enterprise A but not in enterprise B

d - number of innovations absent in both enterprises (zero-zero matches)

There are mentioned several similarity coefficients in the literature (Petr, 2014; Krebs, 2014; Šarmanová, 2012). Possible coefficients to be used are Jaccard's, Sokal-Micheren, Sorensen, Russel-Rao, Dice, Rogers-Tanimoto, Hamman, etc. Most often used similarity coefficient for binary data is Jaccard´s coefficient, which was finally applied in this study. This index does not consider the number d. Jaccard´s coefficient (SJ) calculates the similarity by this equation:

$$SJ = a / (a + b + c),$$

where a, b, c are defined above in presence-absence matrix (TAB. 1).

The value of SJ is equal to 1 for a presence of all types of innovations. Value 1 represents the maximum and indicates the maximum similarity of two enterprises. The minimum value is 0 and means total dissimilarity of two enterprises.

Jaccard's index can be modified to a coefficient of dissimilarity by taking its inverse. Dissimilarity is calculated by the equation:

$$DSJ = 1 - SJ$$

The higher value of dissimilarity index means the greater dissimilarity of two enterprises. For the evaluation of similarities, the Jaccard´s dissimilarity coefficient was chosen. The result of similarity calculation for two objects (a pair) is one coefficient (number). In case of more entities (enterprises), the result is square matrix of Jaccard's dissimilarity coefficient. Clustering of data is used as the next step of data analysis (Petr et al., 2010).

3.2. Clustering

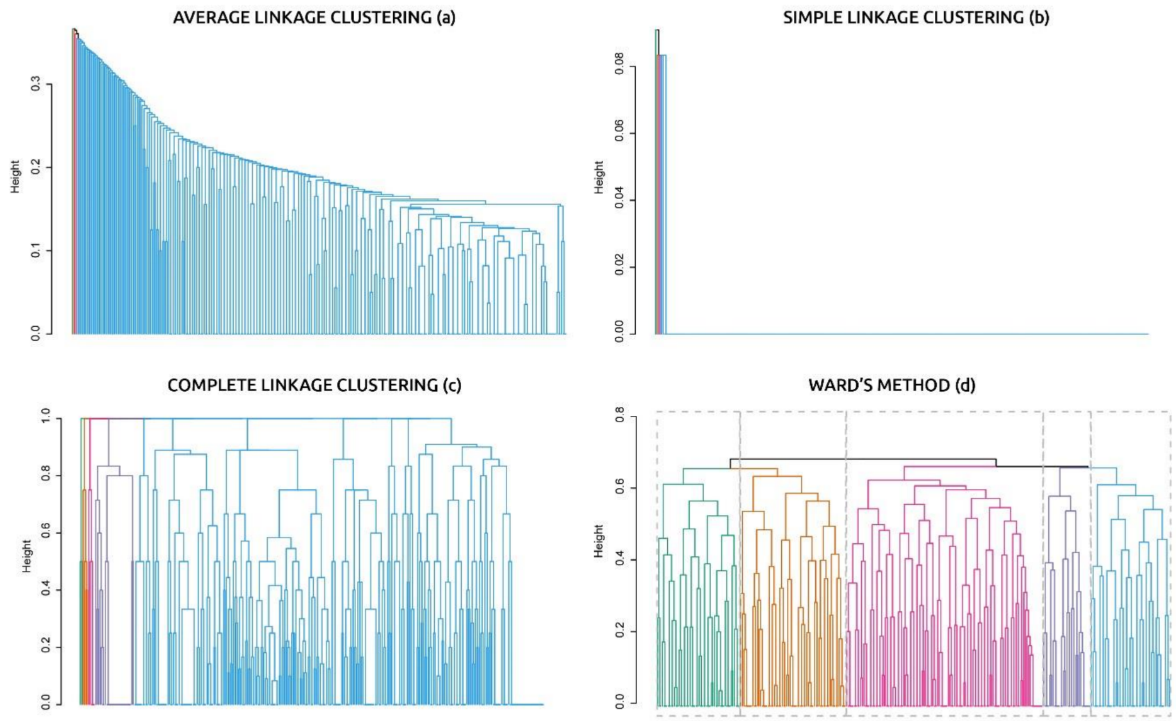
Association coefficients describe similarities between companies regarding the number of applied innovations. However, these results cannot be reasonably visualised, and with a larger number of records (more than twenty), the resulting matrix of coefficients is obfuscatory and misinterpretable. In case of large number of records (large matrix), it is for discussion if any grouping technique based on samples similarity can make the interpretation clearer.

There is no limitation of clustering methods usage to process the values of association matrix. Cluster analysis is a general logical process formulated as a procedure by which individuals are clustered objectively into groups based on their similarity and difference (Tyron, 1939). It helps with investigating of similarity between multidimensional objects and their classification into clusters. Singh and Rajamani (1996) presents possibilities of clustering on binary data using single linkage, complete linkage, and average linkage clustering algorithm. These methods are examples of hierarchical agglomerative clustering, when the objects are classified in the "from the bottom" order, i.e. firstly, clusters are created from individual entities, and in next iterations, these clusters are aggregated together based on their similarities.

Once the (dis)similarity coefficients are determined for pairs of enterprises, clustering algorithm evaluates the similarity between two groups of companies (or a company and an existing cluster), and consequently, the highest similarity is grouped. Final results of clustering can be visualised by a dendrogram, from which the clusters can be read.

Several hierarchical clustering methods were tested: simple linkage clustering (SLC), complete linkage clustering (CLC), average linkage clustering (ALC) and Ward's method. By CLC (FIG. 1c), the dendrogram tends to merge records at dissimilarity value 1. This is because the similarity of each cluster to the others is defined by the least similar pairs among the two, which is often complete dissimilarity. Analogically, SLC (FIG. 1b) merges many records at value 0. For ALC (FIG. 1a) is typical increasing chaining in clusters. The best result is provided by Ward's method with five clusters (FIG. 1d). The number of clusters was defined by the total within-cluster sum of square (WSS), which measures the compactness of the clustering, and it should be as small as possible.

FIG. 1: Dendrograms for different clustering methods



4. Results

In total, CIS 2010 data contains 5,151 records about various enterprises. More than half of the enterprises (2,938) reported any type of innovation activity. Only 21 enterprises innovated in all 12 indicators. They could be considered as the top innovative enterprises. For investigation of similarities and consequence clustering, the sample of data was selected, based on the sector of industry C – Manufacturing (TAB. 2) with the use of the classification of economic activities in the European Community (NACE) (Eurostat, 2008). The selected codes are from code 11 (Manufacture of beverages) to 15 (Manufacture of leather and related products). The total number of enterprises is 631, from which 334 of them have gone through any innovation activity.

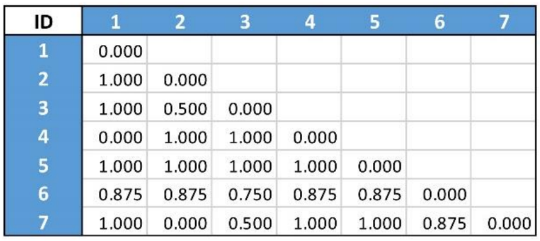
TAB. 2: Statistical overview about enterprise innovations - selected sectors

Section C – MANUFACTURING INDUSTRY		
Code of sector (according to NACE)	Manufacturing Sector	Number of innovating enterprises
11	beverages	81
12	tobacco products	2

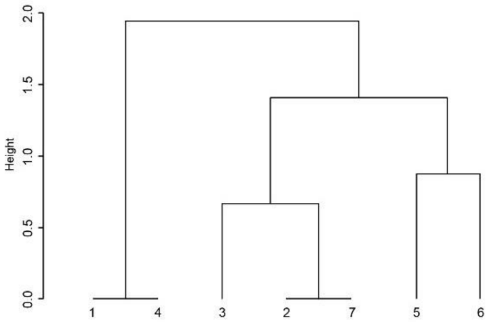
13	textiles	127
14	clothing	135
15	leather and related products	71

Unfortunately, the clustering of hundreds of records did not bring relevant results. The problem is following – two companies are similar to each other and have a certain value in the association. Another two different companies may be similar, with the same Jaccard's coefficient, but in other innovation types. Cluster analysis merges all of these records into one, which results in erroneous interpretations. In the extreme case, companies that innovate in all categories, along with companies that have only one innovation, can appear in one cluster. This phenomenon is due to the large variance of the association coefficient values. For comparison, smaller samples were tested – Olomouc district (69 records), Prostějov district (19 records), and Jeseník disctrict (7 records, see FIG. 2 for Jaccard's coefficient results and the final clustering dendrogram). In this smaller datasets, clustering produced better (more relevant and interpretative) results.

FIG. 2: Example of the Jeseník district



ID	1	2	3	4	5	6	7
1	0.000						
2	1.000	0.000					
3	1.000	0.500	0.000				
4	0.000	1.000	1.000	0.000			
5	1.000	1.000	1.000	1.000	0.000		
6	0.875	0.875	0.750	0.875	0.875	0.000	
7	1.000	0.000	0.500	1.000	1.000	0.875	0.000



5. Conclusion and discussion

The calculations of the associations' coefficients have proven to be an appropriate tool for detecting similarities in binary data. Due to a large number of available coefficients, some expertise is needed to select a suitable one. The resulting matrix of similarities allows to compare individual pairs of records easily, but it does not offer added synthetic information that would help in classifying records into smaller generalised groups. For this reason, cluster analysis methods have been tested. Calculations show that for more records, there is no easy-to-interpret pattern in the final clusters, and therefore clustering is not very appropriate for this type of data. Cluster analysis has proved good results for small number of records (up to 20), where a dendrogram can be used as a visualisation tool for detection of (dis)similarities. Enterprises with the same type of innovations can be revealed this way. However, with the rising value of dissimilarity (FIG. 2, y-axis), correctness is decreasing. For future research, it would be useful to use other methods that would take into account not only the number of innovations but also their order (type) in the dataset.

Literature:

Cooke, P. (1992). *Regional innovation systems: competitive regulation in the new Europe.* Geoforum, 23(3), p. 365-382.

ČSÚ. (2013). *Statistické šetření o inovacích – metodický přehled*. Retrieved from https://www.czso.cz/documents/10180/20542669/21300314m.pdf/.

EUROSTAT. (2008). *RAMON – Reference And Management Of Nomenclatures*. Retrieved from http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2.

EUROSTAT. (2012). *The Community Innovation Survey, The harmonised survey questionnaire.* Retrieved from http://ec.europa.eu/eurostat/documents/203647/203701/Harmonised+survey+questionnaire+2012.

Kirton, M. (1976). Adaptors and innovators: A description and measure. *Journal of applied psychology*, 61(5), p. 622.

Krebs, C.J. (2014). *Ecological Methodology, Chapter 12 Similarity Coefficients and Cluster Analysis*, retrieved from http://www.zoology.ubc.ca/~krebs/downloads/krebs_chapter_12_2014.pdf.

Maillat, D., Quévit, M., & Senn, L. (1993). *Réseaux d'innovation et milieux innovateurs*. Réseaux dinnovation et milieu innovateurs: un pari pour le développement regional. Paris: GREMI/EDES.

Petr, P. (2014). *Metody Data Miningu*, part I. University of Pardubice, Pardubice.

Petr, P., Krupka, J., & Provaznikova, R. (2010). *Statistical Approach to Analysis of the Regions*. In H. Fujita & J. Sasaki (Eds.), Selected Topics in Applied Computer Science. Athens: World Scientific and Engineering Acad and Soc.

Porter, M. E., & Ketels, C. H. (2003). *UK competitiveness: moving to the next stage*, (UK) Department of Trade and Industry economics paper no. 3.

Singh, N. and D. Rajamani (1996) *Similarity coefficient-based clustering: methods for cell formation.* Cellular Manufacturing Systems. Boston, MA: Springer US, p. 70. DOI: 10.1007/978-1-4613-1187-4_4. ISBN 978-1-4612-8504-5.

Sternberg, R. (2000). *Innovation networks and regional development—evidence from the European Regional Innovation Survey (ERIS): theoretical concepts, methodological approach, empirical basis and introduction to the theme issue*. European Planning Studies, 8(4), p. 389-407.

Šarmanová, J. (2012). *Metody analýzy dat*. VŠB – Technická univerzita, Ostrava

Tryon, R.C. (1939). *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers, Ann Arbor.

Vaculík, M., Pászto, V., Švarcová, B. (2017): Spatial distribution of innovation activities in Czech Republic, 2010-2012, *Journal of International Studies* 10(1), p. 123-134, DOI: 10.14254/2071-8330.2017/10-1/8.

Contact:

doc. Ing. Zdena Dobešová, Ph.D.
Department of geoinformatics
17. listopadu 50
771 46 Olomouc
Czech Republic
E-mail: zdena.dobesova@upol.cz
www.geoinformatics.upol.cz

Brief information about the author:

Doc. Ing. Zdena Dobešová, Ph.D. is currently an associate professor at the Department of Geoinformatics, Palacký University Olomouc, Czechia. Her research interests are GIS, digital cartography, the visual programming language in GIS, scripting in Python for ArcGIS, spatial databases and data mining. She is an author of 8 books and more than 70 articles from journals and conferences.

Mgr. Vít Pászto Ph.D. is currently an assistant professor at the Department of Geoinformatics, Palacký University Olomouc, Czechia. His scientific interests cover topics such as geocomputation, spatial analysis of economic data, and GIS based human geography. For the last year, he has been leading a project "Spationomy" focusing on the spatial exploration of economic data and methods for interdisciplinary analytics.

Mgr. Karel Macků is a Ph.D. student at the Department of Geoinformatics, Palacký University Olomouc, Czechia. His main research is focused on the evaluation of the quality of life using quantitative methods and spatial/statistical data. His other professional interests are data processing, statistics, and visualization.