Spatial statistics: Recognition of Spatial Relationships of Companies

Ilona Berková, Tomáš Mrkvička

Abstract: The paper solves the methodology for recognition of spatial relationships of bankruptcy and creditability indicators in Zapadočesky, Jihocesky and Vysocina regions. The main aim of our research is revealing whether companies are spatially dependent (behaviour of indices is same or different) or whether companies are spatially independent. Spatial statistics will be focused on relationships within the sector as well as between sectors. The classification model IN 95 is used in the analysis and it will be modelled by marked point process. The main tool will be the analysis of Ripley’s K-function or other second order characteristics. Sectors and company sizes will be used as marks of points and they will be incorporated in the studied characteristics. For introduction of the methodology we applied these methods on a sample of trees location in Redwood.

Key words: IN 95 · spatial statistics · bankruptcy and credibility models · market point proces · Ripley’s K-function

JEL Classification: C31

1 Introduction

Point process statistics analyses geometrical structures of patterns formed by objects which are randomly distributed in space and can be used for classification and identification of structural changes. Objects are represented by points or marks where points describe their locations and marks provide additional information, thus characterising the objects further, e.g. size, shape, type and so on. Point pattern is a collection of points in an area or set and is interpreted as a sample from a point process. Point processes are denoted by N and it means a random set of points \( x_1, x_2, \ldots \), i.e. \( N = \{x_1, x_2, \ldots \} \). The set N can be finite or infinite. The main aim of point process statistics is to understand and describe interactions among points that explain the mutual positions of points. Spatial statistics is applied in economy where you can reveal economical interactions which typically lead to clustering and repulsion among points (Illian et al., 2008).

It is possible to apply statistical models on point patterns’ data. These models can be used for data summarization and making predictions. Also they can be simulated, it means, random pattern can be generated by model. The importance of statistical modelling can be identified as the best and the most effective form of data analysis. An analysis without statistical modelling leads to less informative results. Statistical model of point pattern is denoted as point process (Baddaley, 2010).

Lots of empirical studies have tried to develop suitable indices and statistical tests which can measure the degree of spatial behaviour of points. The measurement can be performed by traditional methods by indices, for example by Aggregation Clark-Evens index (Dixon, 2002). However, more popular are distance methods (J-function, D-function and g-function) and nearest neighbours methods (L-function and K-function). These methods identify interaction of points through summary characteristics of point pattern (Illian et al., 2008). An important issue of point process statistic is simulation of point processes, where recent research of Myllymäki et al. (2015) introduced new methods for comparing a summary characteristics estimated from data and estimations from its simulations.

Data collection methods depend on objects which are represented by the points, the objectives of the study and the available resources. During data collection methods it is important to optimise unbiasedness, representativeness and control sampling errors. Among points it is possible to observe three main interaction, i.e. clustering, repulsion and random interactions. For exact qualification and distinction in particular types of spatial behaviour it is necessary to choose suitable statistical methods which provide detailed information about spatial structures. Each method view random sets from different angle.

The main aim of the paper is to introduce statistical methodology, which can be applied to point pattern of corporations. Results will be used for monitoring of aggregated behaviour of companies, i.e. if companies make clusters or not.

Ing. Ilona Berková, University of South Bohemia in České Budějovice, Faculty of Economics, Department of Applied Mathematics a Informatics, Studentská 13, 370 05 České Budějovice, Czech Republic, e-mail: berkoi00@ef.jcu.cz
doc. RNDr. Tomáš Mrkvička, PhD., University of South Bohemia in České Budějovice, Faculty of Economics, Department of Applied Mathematics a Informatics, Studentská 13, 370 05 České Budějovice, Czech Republic, e-mail: mrkvicka.toma@gmail.com
Spatial concentration of companies is studied in long term with many authors and is solved especially by K-function and its modifications. As important authors in the field of application of spatial statistics on company data may be consider for example Giuseppe Arbia, Giuseppe Espa a Diefo Giuliani, who applied K-function to find spatial interaction of companies in Italy in the important article “Clusters of firms in an inhomogeneous space: The high-tech industries in Milan” (Arbia et al., 2012). Authors modified K-function by extension of time in the article “Detecting the existence of space-time clustering of firms” (Arbia et. al, 2010).

2 Methods

The research will focus on spatial dependences of corporations settled in Zapadocesky, Jihocesky and Vysocina region. These regions were chosen because of comparatively homogeny conditions of regions and their position makes the suitable observation window. It contains more than 11 000 companies. As we haven’t collected data of corporations yet, we will introduce basic principles of spatial analysis by using data which describe a location of seedlings of trees in Redwood.

The most popular tradition method of spatial statistic is the Aggregation Clark-Evans Index (CE index). The CE index is based on summary characteristics which compares mean of nearest-neighbour distance. Values of the CE index greater than 1 indicate that the pattern has a tendency towards regularity, while the CE<1 indicates clustering. The advantage of this index is its easy application. On the other hand it leads to losing of information through reflection of point processes as simple statistics.

Ripley’s K-function is a tool to analyse complete data of spatial point processes. In the analysis there is considered two dimensional space. Data contain locations of each point in defined observation window. This function can be used to summary point patterns, test hypotheses, estimate parameters and fit model. K-function provides more information and more sensitive analysis than traditional methods. Sometimes this function is considered as the second order moment cumulative function which is related to second order intensity function reference. Ripley’s K function is the mean number of points other than the typical point in a ball of radius r centred at the typical point. The aim is testing deviations from complete spatial randomness (CSR). Ripley’s K-function is defined as:

\[ K(h) = \frac{1}{\lambda}E(N_h), \]

Let \( N_h \) be the number of points within distance \( h \) of random window and \( \lambda \) be intensity of process. For more information see Illian et al. (2008) page 215.

The function has a simple form in the Poisson process case (CSR): \( K(h)=\pi h^2 \), for \( h \geq 0 \). Usually L-function is defined instead

\[ L(h) = \sqrt{\frac{K(h)}{\pi}}, \]

in CSR \( L(h)=h \). Sometimes L-function is called Ripley’s L-function because it is just standardized version of Ripley’s K-function. If \( K(h)>\pi h^2 \) or \( L(h)>h \) for small \( h \), in the pattern there are clusters. If we can define \( K(h)<\pi h^2 \) or \( L(h)<h \) for small \( h \), points in the pattern are regular. Results of L-function are more conclusive due to better interpretation.

For a comparison of empirical functions and their counterpart, which is simulated from zero hypothesis, should be applied Envelope tests. Myllymäki et al. (2015) introduced new global Envelope tests which provide graphic illustration and exact p-value. Tests present distances where behaviour of function leads to rejection or confirmation of zero hypothesis. It is used for better understanding and suggestion more suitable models. Especially Rank Envelope test is based on Envelope test and it is recommended in case of high number of simulations. In our research it is considered 2 499 simulations. The value of significance level is 5 %.

3 Research results

In the research there were observed 62 seedlings of tree in Redwood. The observed window is square with 1 m² area. The average intensity is 62 seedlings of tree per m². Location of seedlings is displayed in the Figure 1.

In the Figure 1 we can see that seedlings tend to formation of clusters. For objective evaluation statistics methods must be used to confirm the statement.

The CE index is considered as the most exploited tradition method. The test statistics of the CE index is \( R=0.6187 \). It is ratio of average neighbour distance in observed window (W) to expected distance for CSR. The alternative hypothesis is clustering. P-value of Bootstrap test that the CE index is equal to 1 is \( 4\times10^{-4} \). Thus data generate clusters.
The main tool for point processes analysis is Ripley’s K-function which is illustrated in Figure 2. The Figure shows that empirical K-function (solid line) is different from expected theoretical value (dashed line), which shows value of CSR and responds $\pi h^2$ for distance.

**Figure 2** Ripley’s K-function for Redwood data

In the similar way it is possible to construct L-function which can be seen in the Figure 3. L-function was used due to its linear behaviour in case of Poisson model and its better interpretation.

It is obvious that estimations of $\hat{K}(h)$ and $\hat{L}(h)$ are above the CSR curve. It means that the pattern generates clusters. It is necessary to find out if clustering is significant. This significance is identified by simulation of global envelope. Global envelope for clusters of these functions are illustrated in Figures 4. In this case we considered 2499 simulations. L-function is often adjusted to horizontal graphic illustration for better understanding. We used this adjusting for another modelling of function too.
The CE index, K-function and L-function refer that dates tend to generating clusters. In first case it was considered homogenous Poisson process so the dates were tested for CSR. Rank Envelope test for this model is illustrated in the Figure 4. From the results mentioned above it is necessary to reject hypothesis which considers that the pattern is Poisson model on base of L-function. The graphical illustration is supported by p-value which is in interval from 0 to 0.011, so we can reject the zero hypotheses of CSR.

**Figure 4** Rank envelope test of Poisson process

The results in Figure 4 show that point pattern tends to generate clusters in distance of points from 0.04 metres to 0.20 metres. The statement can be confirmed by application of Cox processes which represent clustering processes. Neyman-Scott process is particular Poisson cluster process which considers that parents of clusters follow Poisson process and daughters are spread around parents with dispersion density function. The most famous Neyman-Scott process is Modified Thomas process and Matérn cluster process. In the first we considered Modified Thomas process which testing is illustrated in the Figure 5 (on the left side) and then we considered Matérn cluster process which testing is illustrated in the Figure 5 (on the right side).
**Figure 5** Rank envelope test modified Thomas process (left) a Matérn cluster process (right)

Source: Authors

Modified Thomas process is the specific version of Neyman-Scott process where distribution of daughter point around points of parents has symmetric standard distribution. In Matérn cluster process, which is type of Neyman-Scott process, there are points spread uniformly and independently in the circle with ratio $r$ around the centre of clusters (Illian, J. et al., 2008).

Result shows that pattern follows clustering models and we cannot reject hypothesis which claims that pattern is Neyman-Scott process on base of L-function. The result is confirmed by p-value which is in interval from 0.091 to 0.093 for Modified Thomas process and in interval from 0.073 to 0.079 for Matérn cluster process. The hypothesis of generation of clusters is confirmed, so it is not necessary to model Gibbs processes which are used in case of mutual repulsion of points.

4 Conclusions

The aim of this paper was to introduce basic principles of spatial statistics, chosen methods and their application. These principles will be used for deeper monitoring of spatial relation of companies. In the analysis of point processes there were found out that seedlings of tree in Redwood generate clusters. It was found out due to modelling of processes on base of L-function. First of all there was modelled Poisson process which shows deviation from complete spatial randomness and showed clusters generation. Then there were used cluster processes as Modified Thomas process and Matérn cluster process which confirmed the hypothesis about clustering.

This issue will be applied in next research which main aim will be identification if companies are spatial independent or if there exists some spatial dependence. In case that companies will be spatially dependent we will solve if this dependency is the same (it means if companies develop in same way) or if dependency is different (it means if companies develop in different way). Analysis in this paper could help with deeper recognition of spatial relations of corporations where marks describing the health of companies will be added to the point process.

References


