SPAG: A NEW MEASURE OF SPATIAL AGGLOMERATION. THEORETICAL BACKGROUND AND EMPIRICAL EXAMPLES[1]

Tomasz Kossowski, Jan Hauke

Institute of Socio-Economic Geography and Spatial Management, Adam Mickiewicz University, Poznań, Poland

Manuscript received: July 18, 2018
Revised version: November 27, 2018

Kossowski T., Hauke J., 2018. *SPAG*: A new measure of spatial agglomeration. Theoretical background and empirical examples. *Quaestiones Geographicae* 37(4), Bogucki Wydawnictwo Naukowe, Poznań, pp. 33–42. 1 table, 10 figs.

Abstract: Kopczewska (2017) proposed a new empirical measure of spatial agglomeration (*SPAG*) of economic activity based on geolocations of firms. The aim of the paper is to introduce theoretical backgrounds of *SPAG*. The measure is a product of two random variables with beta and gamma distributions. The moments of the product are described and estimated for Poland with spatial centroids of LAU2 treated as geolocations of firms for empirical distribution as well as for the set of firms located in a regular region. Another approach to *SPAG* properties has its origin in a geometric probability concept. We present the research results on geometric probability, applied to *SPAG*, as distance probability distributions for a regular region.

Key words: agglomeration, concentration, specialisation, *SPAG*, clustering, distribution, geometric probability, distance distribution, regular region, economic activity

Corresponding author: Tomasz Kossowski, Institute of Socio-Economic Geography and Spatial Management, Adam Mickiewicz University, Poznań, ul. B. Krygowskiego 10, 61-680 Poznań, Poland; e-mail: tkoss@amu.edu.pl

Introduction

Arbia et al. (2015) presented a number of statistical approaches to the study of the spatial concentration and dispersion of economic activities. They noticed that in the analysis of the problem of spatial location, the following economic activities are used: “a mesoeconomic approach (looking at the distribution of the agents within geographical partitions such as administrative units, regions) or an approach based on the individual geo-localizations of firms”.

The measurement of spatial concentration or specialisation has long history and many papers addressed this problem. Generally, we have two types of measures for spatial agglomeration, concentration or specialisation: 1) cluster-based measures, and 2) distance-based measures. Among them, cluster-based measures seem to be more popular. Interesting review was delivered in a recently published book *Measuring regional specialisation. A new approach* (2017). Measures of the first type usually divide territories into finite number of regions and use aggregated data describing an economic activity (for example Gini or Location Quotient, see Marcon and Puech 2010). These measures mainly have a the Modifiable Areal Unit Problem (Arbia 2001a, Morphet 1997) because they are fragile on the shape or size of spatial units. A few papers suggest improving cluster-based indices with the use of a spatial weights matrix **W** (Arbia 2001b; Arbia, Piras 2009; Guillain, Le Gallo 2010; Sohn 2014) This matrix represents a spatial structure of regions and adds information about possible spatial autocorrelation within regions.

Ripley (1976) proposed a descriptive statistics for detecting deviations from spatial

[1] The paper is financed by the Polish National Science Centre www.ncn.gov.pl as the research project in OPUS 6 call, contract No. UMO-2013/11/B/HS4/01098.

doi: 10.2478/quageo-2018-0041
ISSN 0137-477X, eISSN 2081-6383

homogeneity in the process. He introduced so-called *K* function as an integral of g-function in respect to the distance r:

$$K(r) = 2\pi \int_0^r g(\rho)\, d\rho.$$

Ripley (1976) also delivered a rigorous foundation for the second-order analysis of stationary point processes on general spaces and introduced *K* function. The modelling of spatial patterns was continued by Ripley (1977) with the term 'model' understood as the distribution of a simple second-order point process strictly stationary under a rigid motion. For point processes in the plane, *K* function is a measure of the distribution of the inter-point distances.

The distance-based measures (Marcon, Puech 2003, 2010; Duranton, Overman 2005, 2008) use real locations of firms in geographical space, basing on the concept of *K* Ripley's function and its many modifications (Baddeley 2000; Penttinen et al. 1992; Penttinen 2006). These measures also omitted the MAUP, but their results are presented in an inconvenient form as a chart of the *K* function. Kopczewska (2017) proposed a new empirical measure of spatial agglomeration (*SPAG*) of economic activity based on the geolocation of firms. This measure is a distance-based measure and corresponds to a geometric model of spatial agglomeration (Marcon, Puech 2003, 2010; Duranton, Overman 2005, 2008; Arbia et al. 2010; Mori, Smith 2014).

The aim of our paper is to introduce theoretical backgrounds of *SPAG* because this issue is not present in the original work by Kopczewska (2017). In this paper, we provide the background for this empirical indicator, using primarily the concept of the spatial process. Then, we assume that *SPAG* can be considered as a product of two random variables possessing beta and gamma distribution. The moments of the product are described and estimated for Poland with spatial centroids of LAU2 treated as geo-localisations of firms for reference distribution. In this paper we also follow the idea of geometrical probability. With the reference to some older and more contemporary papers related to this approach, we present results obtained for a regular region analysis.

In the second section of the paper we construct the *SPAG* index from a theoretical point of view, while Section 3 analyses statistical properties of the index, both theoretical and empirical. Section 4 presents a geometrical probability approach to *SPAG* and is followed by the conclusion.

Construction of the index

The theoretical construction of *SPAG* presents an approach, which includes the density measurement of economic activity in a given area. Similarly to the other (known) distance-based measures, it starts with individually geolocated firms. Therefore, it might be applicable considering territorial divisions and the problem of zoning (MAUP) is naturally omitted. The idea differs from the concepts of Ripley's *K* function or kernel estimation (Ripley 1977). Instead of using the function, we propose the index of spatial agglomeration (*SPAG*) based on the geometrical representation of firms by circles with radii depending on the size of firms.

The index is introduced taking into account the more general framework of *spatial random processes*. Its construction is as follows:

1. We define a spatial random process, where a random variable describes some economic activity,
2. We introduce a parameter κ describing the spatial concentration/agglomeration with respect to five conditions of 'good measure'[2],
3. We formulate *SPAG* index for spatial concentration/agglomeration as an estimator of parameter κ.

Let us consider a set S = {s: s = (φ,λ)} where φ,λ are coordinates in any two-dimensional coordinate system (i.e. on the plane, on the ellipsoid WGS84, geographical coordinates, UTM coordinates, etc.). Coordinates φ,λ are restricted by some additional conditions in such a way that the set S is limited by some boundary ∂S. Let $X(s)$ will be a random variable associated with any

[2] These five conditions were described by Marcon and Puech (2010). The ideal measure of concentration: (i) compares geographic concentration results across industries, (ii) controls industrial concentration, (iii) controls the overall aggregation patterns of industries, (iv) tests the significance of the results and (v) keeps the empirical results unbiased across geographic scales.

point $s \in S$. For further simplifications we restrict random variables $X(s)$ to be independently, identically distributed with a given CDF f. Under the above conditions $\{X(s): s \in S\}$ is a spatial random process. We will interpret the points of S as possible places of economic activity.

Let us introduce a metric function d over the space S that fulfills standard (for a measure to be a distance measure) conditions:

1. $d(s_i, s_j) = 0$ if and only if $i = j$,
2. $d(s_i, s_j) = d(s_j, s_i)$,
3. $d(s_i, s_k)+d(s_k, s_j) \geq d(s_i, s_j)$.

The pair (S, d) creates a metric space. Let us

define $A \subset S$ as a field ('region') bounded by ∂S. The area of A, denoted by $|A|$, is equal to

$$\iint_S \partial S \, d\varphi d\lambda.$$

Let us assume that the parameter κ is defined over the random field $\{X(s): s \in S\}$. It is expected that it fulfills five conditions for the good measure of the spatial concentration (Marcon, Puech 2010). The parameter is unobservable and it will be constructed as a selected function of three parameters

$$\kappa = \kappa(\theta_c, \theta_d, \theta_o),$$

where $\theta_c, \theta_o \in [0,1]$ and $\theta_d \in [0,\infty)$.

We take the function, which is a product as it is the simplest function preserving monotonicity with respect to each of these three parameters

$$\kappa = \theta_c \cdot \theta_d \cdot \theta_o.$$

The aim of the study is to find a 'proper' estimator for κ. For this, we consider a finite sample from the random field $\{X(s): s \in S\}$, written as a finite length (n) random vector

$$\mathbf{Z}(s) = (X(s_1), X(s_2), \dots , X(s_n)).$$

Denote its realisation by

$$\mathbf{z}(s) = (x(s_1), x(s_2), \dots , x(s_n)),$$

where $x(s_i)$ are values of random variables associated with points $s_i = (\varphi_i,\lambda_i)$.

In the next step of the construction of the estimator, we convert vector $\mathbf{z}(s)$ to vector $\mathbf{z}(r)$ with the use a special transformation (let us call it T) related to radii r. In the beginning, we focus on the location of n economic activity units. In the empirical distribution, each point $s_i = (\varphi_i,\lambda_i)$ of n economic activity locations is covered by the circle, the area of which is proportional to any feature of the location, for example, employment. The radius i of the i-th circle might be a continuous variable for precise data on employment or discrete for interval data. The sum of the areas $|A_i|$ of n circles is equal to the area $|A|$ of the region. Radii of the circles create the "impact zones of economic activity", which are bigger in the case of larger firms. Setting circles in real business locations is to reflect upon the phenomena of a spatial agglomeration or other spatial patterns. As a consequence, we obtain vector $\mathbf{z}(r)$:

$$\mathbf{z}(r) = (r_1, r_2, \dots , r_n),$$

where r_i are calculated by

$$r_i = \sqrt{\frac{x(s_i)\,|A|}{\pi \sum_{j=1}^{n} x(s_j)}}.$$

For the discrete case (the distribution of radii is discrete) a classification function g: $\mathbf{Z}(s) \rightarrow \{1,2,\dots,k\}$, where k is a number of classes, is defined. Then, for each class i the radius with the use of the following formula

$$r_i = \sqrt{\frac{\sum_{\{j:g(x(s_j))=k\}} x(s_j)}{n_i \pi \sum_{j=1}^{n} x(s_j)}},$$

where n_i is a number of elements in class i, is calculated.

Assuming that radii are equal for all the places representing a location of economic activity, the following formula is used to calculate the radii r_i

$$r_i = \sqrt{\frac{|A|}{n\pi}}.$$

After the T transformation, a circle with the center in point s_i and radius r_i represents each element of vector $\mathbf{z}(s)$. The metric space (S, d) is invariant with respect to the T transformation.

Now, let us introduce SPAG as an estimator for the κ parameter. It is defined as

$$SPAG = \theta_c \widehat{\cdot\, \theta_d \cdot}\, \theta_o .$$

We assume that

$$\theta_c \widehat{\cdot\, \theta_d \cdot}\, \theta_o \approx \widehat{\theta}_c \widehat{\theta}_d \widehat{\theta}_o ,$$

which gives us the possibility to estimate each factor in the product separately.

The first factor of the above product is an estimator of parameter θ_c. For its specification, we use subvector $\mathbf{z}^*(r) = (r_1^*, r_2^*, ..., r_l^*)$ of the vector $\mathbf{z}(r) = (r_1, r_2, ... , r_n)$, where $l \leq n$. Then, the estimator of θ_c is defined by

$$\widehat{\theta}_c = \frac{\pi \cdot \sum_{i=1}^{l} (r_i^*)^2}{|A|} .$$

The range of the values of the estimator is the interval [0,1].

For the specification of the estimator θ_d a metric (distance) function d over the space S will be given. Then, the estimator of the index exceeding parameter θ_d is as follows

$$\widehat{\theta}_d = \frac{\sum_{i=1}^{l} \sum_{j=1}^{l} d(i,j)}{l^2 \cdot \overline{d}} ,$$

where $\overline{d}$ is a mean distance for the uniform spatial distribution of l locations of economic activities. The values of this estimator belong to the interval $[0,\infty)$.

Finally, the estimator of parameter θ_o is specified by the formula

$$\widehat{\theta}_o = \frac{|\bigcup_{i=1}^{l} A_i|}{\pi \cdot \sum_{i=1}^{l} (r_i^*)^2} ,$$

where A_i is a circle with the centre at point $s_i = (\varphi_i, \lambda_i)$ and radius r_i. The range of values of the estimator $\widehat{\theta}_o$ is interval [0,1].

Following Kopczewska (2017), we will denote estimators $\widehat{\theta}_c$, $\widehat{\theta}_d$, $\widehat{\theta}_o$ as indices I_c, I_d, I_o respectively. We will call these indices: an index of coverage, an index of distance and, an index of overlap. All components have their economic interpretation.

According to Kopczewska (2017), the *SPAG* index as a measure of the degree of spatial agglomeration allows: 1) comparing the regions or sectors over time, 2) comparing sectors inside a region and between regions. The *SPAG* index uses combination of information on the area, size and sectors of companies. Thus, the application of SPAG by policy makers, as well as the comparability between regions and/or sectors over time is easy and powerful. *SPAG* has the range values in interval [0, M], where

$$M = SPAG_{max} ,$$

the maximum value of *SPAG* estimated, with some small bias, as a ratio

$$SPAG_{max} = \frac{d_{max}}{\mathbf{E}(d)} ,$$

with d_{max} being a maximum distance within the region, and $\mathbf{E}(d)$ being the expected value of the distribution of distances within a given region.

The bias of $SPAG_{max}$ is related to the fact, that $\mathbf{E}(d)$ is slightly higher than the mean of distances between the locations of firms in a referenced spatial distribution, and in consequence, is usually underestimated $SPAG_{max}$. Values of the index exceeding 3 were not reported in empirical studies.

Statistical properties of *SPAG*

In this section, we briefly discuss some of the statistical properties of *SPAG*. For this purpose we will introduce some results reported in a paper by Nadarajah and Kotz (2005), and then, we will show how these results can be used in the *SPAG* case.

Beta and Gamma distribution case

Let us observe that *SPAG* is in fact a product of two independent following factors

$$SPAG = \frac{|\bigcup_{i=1}^{l} A_i|}{|A|} \widehat{\theta}_d . \tag{1}$$

For further simplifications, we will denote

$$\widehat{\theta_{co}} = \frac{|\bigcup_{i=1}^{l} A_i|}{|A|} .$$

Let's assume, that $\widehat{\theta_{co}}$ has beta distribution with the probability density function

$$f_{\widehat{\theta_{co}}}(\widehat{\theta_{co}}) = \frac{\widehat{\theta_{co}}^{a-1}(1-\widehat{\theta_{co}})^{b-1}}{B(a,b)}, \qquad (2)$$

for $\widehat{\theta_{co}} \in (0,1)$, $a > 0$, $b > 0$. The expected value and variance of $\widehat{\theta_{co}}$ are

$$\mathbf{E}(\widehat{\theta_{co}}) = \frac{a}{a+b}, \quad \sigma^2_{\widehat{\theta_{co}}} = \frac{ab}{(a+b)^2(a+b+1)}.$$

For $\widehat{\theta_d}$ let us assume that it has gamma distribution with the probability density function

$$f_{\widehat{\theta_d}}(\widehat{\theta_d}) = \frac{\lambda^{\beta}\widehat{\theta_d}^{\beta-1}e^{-\lambda\widehat{\theta_d}}}{\Gamma(\beta)}, \qquad (3)$$

for $\widehat{\theta_d} > 0$, $\beta > 0$, $\lambda > 0$. The expected value and variance of $\widehat{\theta_d}$ are

$$\mathbf{E}(\widehat{\theta_d}) = \frac{\beta}{\lambda}, \quad \sigma^2_{\widehat{\theta_d}} = \frac{\beta}{\lambda^2}.$$

Following the paper by Nadarajah and Kotz (2005: 437), the probability density function of $SPAG = \widehat{\theta_{co}} \cdot \widehat{\theta_d}$ is

$$f_{SPAG}(SPAG) = \frac{\lambda^{\beta}\Gamma(b)}{\Gamma(\beta)B(a,b)} \cdot SPAG^{\beta-1}e^{-\lambda \cdot SPAG} \cdot \Psi(b,\, 1+\beta-a;\, \lambda \cdot SPAG),$$

for $SPAG > 0$, where Ψ is the Kummer function defined by

$$\Psi(a,\, b;\, x) = \frac{1}{\Gamma(a)}\int_0^{\infty} e^{-xt}t^{a-t}(1+t)^{b-a-1}dt.$$

The above specification of the distribution allows calculating moments of *SPAG*. Its expected value is

$$\mathbf{E}(SPAG) = \frac{a\beta}{(a+b)\lambda}, \qquad (4)$$

and variance has a form of

$$\boldsymbol{\sigma}^2(SPAG) = \sigma^2_{\widehat{\theta_{co}}}\sigma^2_{\widehat{\theta_d}} + \sigma^2_{\widehat{\theta_{co}}}\mathbf{E}(\widehat{\theta_d^2}) + \sigma^2_{\widehat{\theta_d}}\mathbf{E}(\widehat{\theta_{co}^2}).$$

Taking into consideration that

$$\mathbf{E}(\widehat{\theta_{co}^2}) = \frac{a(a+1)}{(a+b)(a+b+1)}, \text{ and } \mathbf{E}(\widehat{\theta_d^2}) = \frac{\beta(\beta+1)}{\lambda^2},$$

after some calculations we obtain

$$\boldsymbol{\sigma}^2(SPAG) = \frac{\beta a(3b+\beta b+a^2+ab+a)}{(a+b)^2(a+b+1)\lambda^2}. \qquad (5)$$

Using formulas from the paper by Nadarajah and Kotz (2005), percentage points z_p associated with cdf function of SPAG are available. These points are obtainable from the following equation (using numerical procedure)

$$\frac{1}{\Gamma(\beta)B(a,\, b)}\left[\frac{(\lambda z_p)^{\beta}}{\beta}B(b{,}a-\beta)H(1-a-b+\beta,\, \beta;\, \beta+1,\, 1-a+\beta; -\lambda z_p) + \frac{(\lambda z_p)^{a}}{a}\Gamma(\beta-a)H(a,\, 1-b;\, a-\beta+1{,}1+a;\, -\lambda z_p)\right] = p,$$

where $H(\cdot)$ is a hypergeometric function.

Empirical case – the distribution of *SPAG f*or Poland

In this section we analyse a specific empirical case – looking for the distribution of *SPAG* for Poland with the use of formulas (4) and (5). In this case, we have to find parameters of distributions formulated in (2) and (3). For the beta distribution, we will use the following estimators for parameters:

$$\hat{a} = \frac{\hat{\mu}\hat{b}}{1-\hat{\mu}},$$

$$\hat{b} = \frac{(1-\mu)^2\hat{\mu}}{(1-\hat{\mu}-\hat{\mu}^2)^2\hat{\sigma}^2} + \hat{\mu}-1, \qquad (6)$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are estimators of the expected value and variance of a random variable X with the B(a,b) distribution respectively.

Parameters of Γ distribution are obtained from its estimators, described by the formulas

$$\hat{\beta} = \frac{\hat{\mu}^2}{\hat{\sigma}^2},$$

$$\hat{\lambda} = \frac{\hat{\mu}}{\hat{\sigma}^2}, \qquad (7)$$

where $\hat{\mu}$ is the estimator of an expected value and $\hat{\sigma}^2$ is the variance estimator of a random variable Y with the Γ(β,λ) distribution.

The estimation procedure is based on the following datasets for Poland:

Table 1. Classes of the employment size of firms in Poland.

Class of employment	Number of firms
1–9	3,938,654
10–49	146,926
50–249	29,610
250–999	3,706
1000 and more	775
	4,119,671

Source: Statistics Poland.

1) georeferenced set of LAU2 spatial units in Poland (communes),
2) the distribution of firms in different employment size classes,
3) the total number of firms located in every spatial unit from the georeferenced set mentioned in 1).

The georeferenced set of spatial units is written in a commonly used shapefile format and has information about 2,479 spatial units. The data considered in 2) and 3) were obtained from Statistics Poland in Warsaw.

The maximum number of employers was restricted to 5,000. Using data from Table 1 the estimation procedure led to the following *X* distribution: B(0.027916752,17.30099551). The distribution is asymmetric and its PDF is presented in Fig. 1.

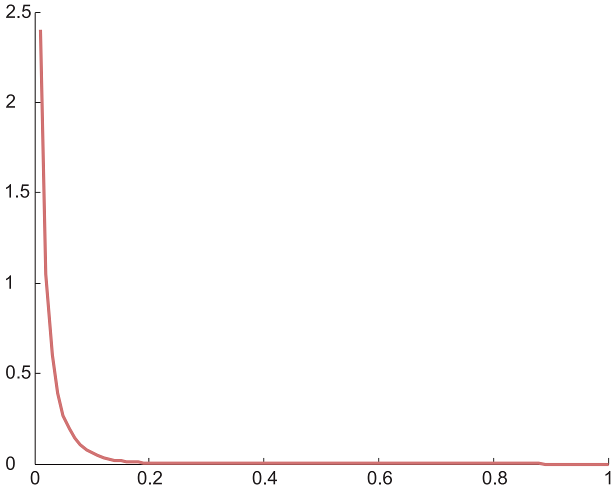

Fig. 1. PDF for B(0.027916752,17.30099551) distribution.
Source: own study.

Estimation of parameters for distribution is more complicated. Let's observe that

$$\mathbf{E}(Y) = \frac{1}{d}\,\mathbf{E}\left(\frac{\sum_{i=1}^{l}\sum_{j=1}^{l} d(i,j)}{l^2}\right),$$

where (i, j) are coordinates of firms. We assume, that all firms in each spatial unit are located at the centroid of spatial unit. So, we related the total number of firms in every spatial unit to its centroid. Then, we used p = 9,999 permutations of numbers of firms in the analysed spatial units. In this way, we estimated both a mean value of distances from one firm to another, and also its variance. The mean distance was equal to 289 km.

The second problem was to gain the value of *d*, which is necessary for estimation of parameters *β* and *λ*. Firstly, using some procedures in R, we simulated a regular grid distribution of firms over Polish territory (n = 4,118,671). Then, we found *d*, which is equal to the mean distance in this grid. It should be underlined that the procedure of calculation of $n(n\text{-}1)/2$ distances is time consuming (the number of distances is larger than $8 \cdot 10^{12}$ and needs an extremely large size of memory for holding it). So, we decided to use the permutation procedure for the estimation of *d*.

In the procedure we permuted p = 10,000 times rows of a two-column matrix of points coordinates on a reference-distribution grid and we calculated 10,000 mean distances between original and permuted matrices of coordinates. The mean value of those estimations was found, and was treated as the estimate of *d*. The dispersion between the minimum and maximum distance estimated was acceptable being no larger than 0.7 km. The obtained estimate of parameter *d* was 292.0782 km.

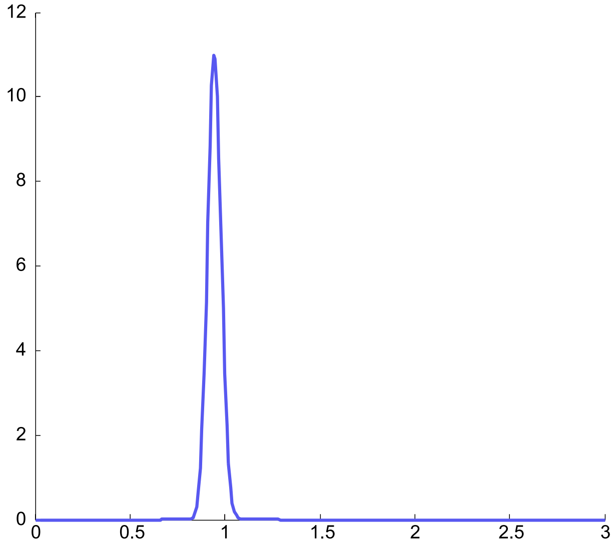

Fig. 2. PDF for Γ(686.0985,725.829) distribution.
Source: own study.

Finally, we calculated values of β and λ for Γ distribution, and they were equal to 686.0985 and 725.829 respectively. The estimated PDF distribution is presented in Fig. 2. The last part is a calculation of *SPAG* distribution parameters and moments using formulas (4) and (5). The result is as follows:

$$f_{SPAG}(SPAG) = \frac{725.829^{685.0985}\Gamma\,(17.30099551)}{\Gamma\,(686.0985)\mathrm{B}(0.027916752{,}17.30099551)}\,SPAG^{686.0985}$$
$$\cdot \exp(-725.829 \cdot SPAG) \cdot \Psi(17.30099551{,}687.070583;\,725.829 \cdot SPAG)$$
$$\mathbf{E}(SPAG) = 0.00152281$$
$$\hat{\sigma}^2(SPAG) = 0.0000787543$$

Let us notice that in the analysed empirical case, values of the expected value and variance of *SPAG* are relatively small.

Geometric probability approach to *SPAG*

Another approach to *SPAG* properties has its origin in a geometric probability concept. Although the idea of geometric probability is quite old, it has not received much attention in recent studies. There exist some traditions in applications of the geometric probability to spatial analysis. One of the crucial problems related to the use of a geometric probability concept is the question of the distribution of distances between two random points within a given region. To make it simple, it is often assumed that a given region has a regular shape in one of the geometrical forms: triangle, square, parallelogram, rhombus, hexagon or finally, a form of circle. It is also worth mentioning, that this problem is from time to time rediscovered (Alagar 1976). Some bases of the geometric probability are described in the early book written by Kendall and Moran (1962). Additional notes related to the concept of geometric probability were presented by Moran (1966, 1969).

The results of the research on geometric probability, applied to our work, i.e. distance probability distributions for a regular region were published in the mid-1970s by Alagar (1976). This work was extended by two researchers: Zhuang and Pan (2011, 2017). They calculated PDF and CDF functions as well as an expected value and variance for distances between two random points in a hexagon. The theoretical results of their work were confirmed by a series of numerical simulations.

A case of a regular region with a fixed location of firms

Let us consider a case of a regular region, when the shape of a region is a hexagon. In this region, a set of n = 4,910 firms is located. The set of firms with their locations was taken from the Statistics Poland database. The borders of the region were created artificially by their mathematical definition on the map (with respect to a map projection). The location of firms is shown in Fig. 3. In this research, our question is what the distribution of *SPAG* is.

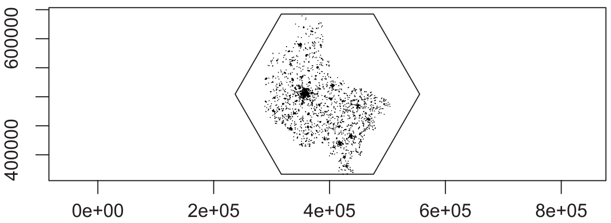

Fig. 3. Firms distribution in a hexagonal region.
Source: own study.

In the beginning, a possible range of *SPAG* in a hexagonal region was found. For the case studied the distance probability function $f_{\mathrm{D}_{\mathrm{H_I}}}(d)$, where 0 < d < 2 in a unit hexagon, was applied using formula (3) from Zhuang and Pan (2011). The expected value of d is equal to

$$\mathbf{E}(d) = \int_0^2 x f_{\mathrm{D}_{\mathrm{H_I}}}(x)dx = 0.826. \tag{8}$$

In our case, the side length of the hexagon (as shown in Fig. 3) is s = 159.286 km. Then, the expected value of distances between two random points in this hexagon is equal to 131.570 km. Thus, the estimate of the maximum value of *SPAG* is

$$SPAG_{max} = \frac{2s}{\mathbf{E}(d)} = \frac{2}{0.826} = 2.422.$$

The size of firms was sampled from exponential distribution with parameter 1/2, and then rounded down to the nearest integer number. The parameter of exponential distribution with the value of ½ was selected due to the fact that the expected value of the distribution is 2.57, i.e.

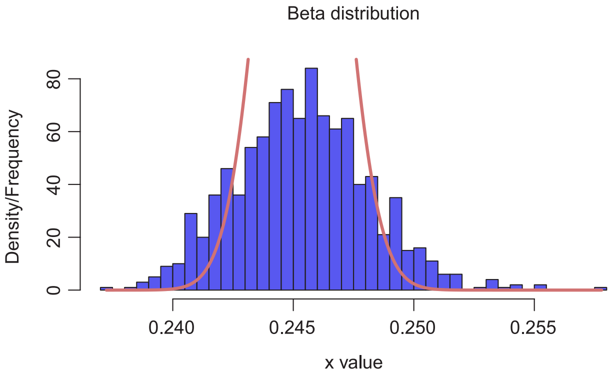

Fig. 4. PDF for B(20078, 61748) distribution.
Source: own study.

the mean of the real size of employment distribution in firms. For the known size of employment, circles located around firms and their union set was obtained. The empirical distribution of $\widehat{\theta_{co}}$ was achieved with the use of p = 9,999 permutations of the set of circle sizes. Then, the estimated parameters of B distribution for $\widehat{\theta_{co}}$ applying formula (6) were gained. The distribution is presented in Fig. 4.

Repeating the procedure described above, a theoretical average distance $\overline{d}_{grid}$ between firms located on a referenced hexagonal grid inside the region, and $\widehat{\theta_d}$ distribution were estimated. Comparing to previous analyses, the number of permutations was reduced to l = 999 (due to time wasting calculations). Finally, we obtained $\overline{d}_{grid}$ = 131.036 km, while a real average distance between the locations of firms is $\overline{d}_{real}$ = 7.807 km. Let us observe that the value of $\overline{d}_{grid}$ is slightly smaller than **E**(d) for this hexagon. Associated Γ distribution and its parameters are drawn in Fig. 5.

Finally, the empirical distribution of *SPAG* was found. It is shown in Fig. 6. The moments of the distribution, were obtained from (4) and (5) formulas. The expected value is equal to 0.146, and variance is $3.55 \cdot 10^{-6}$.

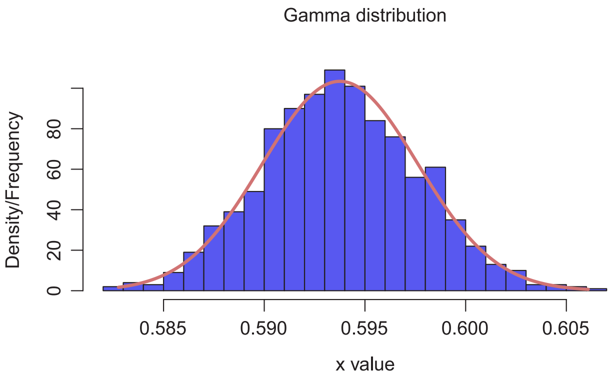

Fig. 5. Histogram and PDF for Γ(23692,39900) distribution.
Source: own study.

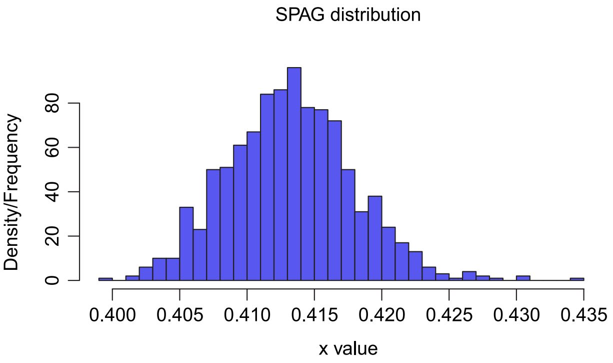

Fig. 6. Histogram for *SPAG*.
Source: own study.

Comparing theoretical results for the moments of *SPAG* distribution with empirical moments, it should be noticed that in this case theoretical distribution of *SPAG* doesn't describe empirical distribution with sufficient accuracy. The histogram in Fig. 6 shows that the empirical mean is equal to 0.413, and is far away from theoretical point 0.146. The theoretical variance is more than 10 times larger than the empirical value which is equal to $2.2 \cdot 10^{-5}$.

A case of a regular region with a simulated-random location of firms

One more approach to a regular region case (for example triangles, squares or hexagons) is the modification of the analysis conducted in the previous subsection. As above, we assume a regular hexagonal region with n = 4,910 firms randomly located within it. The size of employment in firms was randomly assigned using distribution as in the previous case. In Fig. 7, the random distribution of firms within a regular hexagonal region is presented.

Then, permuting p = 999 times the set of firm sizes, the empirical distribution for $\widehat{\theta_{co}}$ was determined. The distribution and its parameters

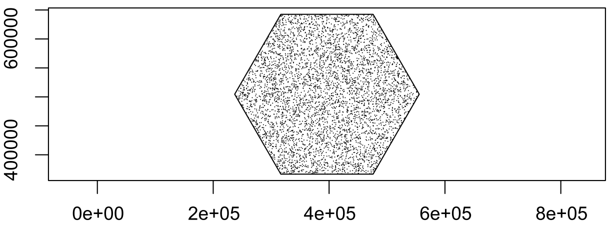

Fig. 7. Firms distribution in a hexagonal region.
Source: own study.

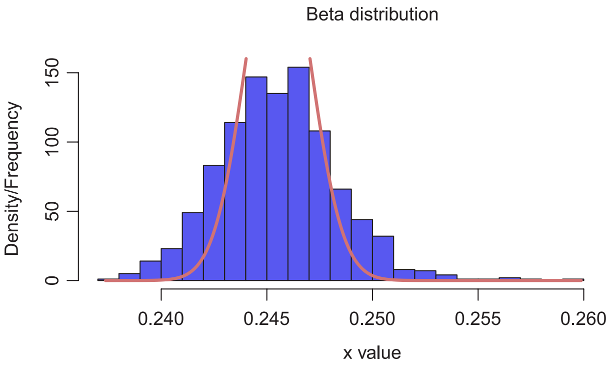

Fig. 8. PDF for B(19797, 60830) distribution.
Source: own study.

are shown in Fig. 8. Next, the average distance between real locations of firms, and a mean distance between firms in a uniform hexagonal distribution across the region were estimated. We found $\overline{d}_{grid}$ = 131.036 km and, $\overline{d}_{real}$ = 131.955 km.

Consequently, the empirical distribution of $\widehat{\theta}_d$ was delivered. As was mentioned earlier, the $\widehat{\theta}_d$ distribution has the Γ probability density function. The empirical distribution of $\widehat{\theta}_d$ is drawn (with its parameters) in Fig. 9.

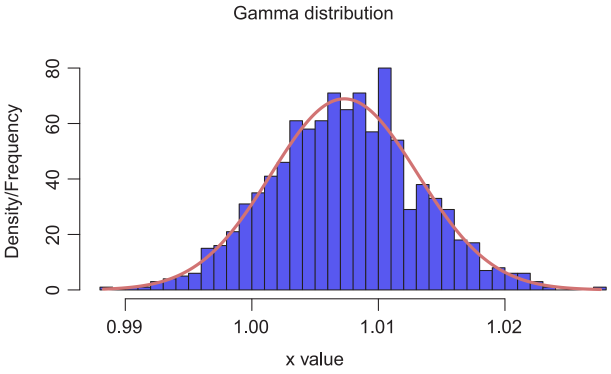

Fig. 9. Histogram and PDF for Γ(30270,30049) distribution.
Source: own study.

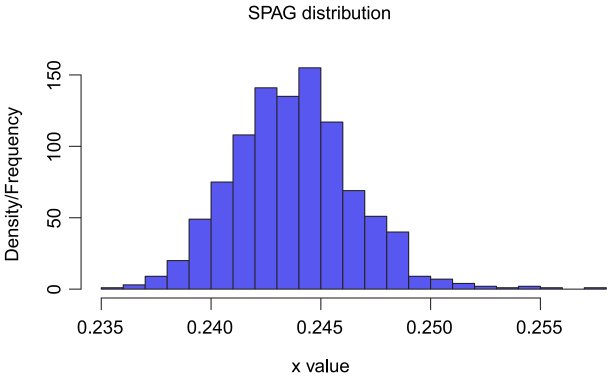

Fig. 10. Histogram of *SPAG*.
Source: own study.

In the last step, we calculated the distribution for *SPAG*. Its expected value is equal to 0.247, while variance is $3.55 \cdot 10^{-6}$. The histogram in Fig. 10 presents the empirical distribution of *SPAG*. The moments of the empirical distribution are similar to the theoretical one. The expected value is equal to 0.243, so it is only slightly different from the theoretical value. The variance of the empirical distribution is and, it is also very close to the theoretical result.

Conclusion

In the study of socio-economic processes, many indexes are used, but only some of them have the appropriate theoretical framework allowing to apply inferential statistical methods in the analysis. The theoretical construction of *SPAG* with the use of the assumed distribution of its components makes it possible to specify the distribution and moments of *SPAG* (an approach based on the properties of beta and gamma distribution). This allows constructing exact statistical significance tests useful for the comparison of empirically obtained values. However, the obtained moments of theoretical distribution were not very close to the empirical one. It suggests that this approach should be adjusted by additional research.

The second approach, inspired by papers on geometric distribution and papers on distance distributions in regular regions allowed finding the range of *SPAG* and verifying the average distance between firms in referenced regular distribution. The calculated empirical distributions of *SPAG* for the set of firms located in a hexagonal region, as well as theoretical and empirical moments of these distributions turn out to be promising in the characterisation of *SPAG* distribution.

From an empirical point of view values of the *SPAG* index are interpreted as follows: when all firms are located in one place within a given region, then the index value is equal to zero. The value of *SPAG* is equal to one when firms are located according to uniform spatial distribution, and the size of employment in all firms is the same. It is also possible, for some specific cases, that *SPAG* exceeds one. It is usually in the situation, when there are several clusters of firms within a given region, with large distances between them. It can be named as a border-cluster distribution of firms.

The further research of *SPAG* properties could consider its extension and include heterogeneity into the measurement and analysis of robustness of *SPAG* measurement results with respect to different methods of centroids location.

The application of *SPAG* by policy makers, as well as the comparability between regions, sectors over time is easy and powerful. According to Kopczewska (2017), the *SPAG* index can be applied for measuring the degree of spatial agglomeration (concentration, clustering). In this situation, the index gives an opportunity to follow the agglomeration process. *SPAG* also allows making comparisons between regions regarding agglomeration or repulsion processes, and makes it possible to know how advanced these processes in different regions are. As a consequence, policy makers can decide about supporting agglomeration processes in a region or not. And finally, policy makers are able to detect whether co-located sectors of the economy built spatial clusters.

References

Alagar V.S., 1976. The distribution of the distance between random points. *Journal of Applied Probability* 13(3): 558–566.

Arbia G., 2001a. Modelling the geography of economic activities on a continuous space. *Papers in Regional Science* 80: 411–424.

Arbia G., 2001b. The role of spatial effects in the empirical analysis of regional concentration. *Journal of Geographical Systems* 3: 271–281.

Arbia G., Piras G., 2009. A new class of spatial concentration measures. *Computational Statistics and Data Analysis* 53: 4471–4481.

Arbia G., Espa G., Giuliani D., Mazzitelli A., 2010. Detecting the existence of space-time clustering of firms. *Regional Science and Urban Economics* 40(5): 311–323.

Arbia G., Espa G., Giuliani D., 2015. Analysis of spatial concentration and dispersion. In: Karlsson C., Anderson M., Norman T. (eds), *Handbook of Research Methods and Applications in Economic Geography*. Elgar: 135–157.

Baddeley A.D., 2000. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences* 4: 417–423.

Duranton G., Overman H.G., 2005. Testing for localization using micro- geographic data. *The Review of Economic Studies* 72(4): 1077–1106.

Duranton G., Overman H.G., 2008. Exploring the detailed location patterns of UK manufacturing industries using microgeographic data. *Journal of Regional Science* 48: 213–243.

Guillain R., Le Gallo J., 2010. Agglomeration and dispersion of economic activities in and around Paris: An exploratory spatial data analysis. *Environment and Planning B*, 37: 961–81.

Kendall M.G., Moran P.A.P., 1962. *Geometric probability*. Griffin Statistical Monographs. Griffin, London.

Kopczewska K., 2017. SPAG – index of spatial agglomeration. In: Kopczewska K., Churski P., Ochojski A., Polko A. (eds), *Measuring regional specialisation. A new approach*. Palgrave Macmillan, Cham, Switzerland: 189–216.

Marcon E., Puech F., 2003. Evaluating the geo-graphic concentration of industries using distance-based methods. *Journal of Economic Geography* 3(4): 409–428.

Marcon E., Puech F., 2010. Measures of the geographic concentration of industries: Improving distance-based methods. *Journal of Economic Geography* 10(5): 745–762.

Moran P.A.P., 1966. A note on recent research in geometrical probability. *Journal of Applied Probability* 3: 453–463.

Moran P.A.P., 1969. A second note on recent research in geometrical probability. *Advances in Applied Probability* 1: 73–89.

Mori T., Smith T., 2014. A probabilistic modeling approach to the detection of industrial agglomeration. *Journal of Economic Geography* 14(3): 547–588.

Morphet C.S., 1997. A statistical method for the identification of spatial clusters. *Environment and Planning A*, 29: 1039–1055.

Nadarajah S., Kotz S., 2005. On the product and ratio of gamma and beta random variables. *Allgemeines Statistisches Archiv* 89(4): 435–449.

Penttinen A., Stoyan D., Henttonen H., 1992. Marked point processes in forest statistics. *Forest Science* 38: 806–824.

Penttinen A., 2006. Statistics for marked point patterns. In: *Yearbook of the Finnish Statistical Society 2006*: 70–91.

Ripley B.D., 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* 13: 255–266.

Ripley B.D., 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2): 172–212.

Sohn J., 2014. Industry classification considering spatial distribution of manufacturing activities. *Area* 46.1: 101–110.

Zhuang Y., Pan J., 2011. *Random distances associated with hexagons*. Working paper.

Zhuang Y., Pan J., 2017. *A geometrical probability approach to location-critical network*. Working paper.