A Study of Usefulness of Selected Robust Model Based Clustering Techniques in a Digital Development Modelling

Daniel Kosiorowski[1], Przemysław Jaśko[2], Ewa Szlachtowska

Abstract

This paper critically discusses three robust model based clustering techniques in a context of their applicative usefulness in a process of specifying a two dimensional model generating spatio-temporal data related to a digital economy. We among others study TCLUST, OTRIMLE algorithms and certain algorithms, which are available in the mclust R package. Theoretical considerations are illustrated by means of empirical issues related to a preliminary analysis of spatial phenomena of a digital economy. Additionally, we present results of simulation studies involving spatial processes departing from regularity.

Keywords: *Robust Clustering, Spatio-Temporal Models, Mixture Models*
JEL Classification: *C14, C53, R10*

1. Introduction and problem formulation

Sustainable socio-economic development of modern societies belongs to hot topics in a current public and scientific debate. A motivation of this paper relates to a problem of specifying a spatial model describing "digital development" in a certain region of space, for example in a geographic or administrative region of a country. The paper concentrates on a choice of an appropriate robust model-based clustering technique, which shall provide functions, which are further used in an estimation of a functional regression model describing a digital development of the region. It should be pointed out that in robust modelling we concentrate on an influential majority of cases, having more or less formalised idea in mind, which does and which does not belong to a data generating mechanism (Kosiorowski and Zawadzki 2019). Despite of the fact that one can find several promising clustering procedures in the literature, which are described by authors as robust, an issue of robustness of a clustering procedure is still an open problem (Hennig, 2004). One cannot find a comprehensive study which compares these procedures in a context of modelling of economic spatial-temporal phenomena. In order to fill that gap we among others conducted a simulation research, in which clusters were generated by a mixture of skewed Student T distributions, a noise has "irregular" spiral shape support and additionally samples contained outliers. We aimed at modelling a situation of an existence of "asymmetric centres of gravity", e.g., cities, a net of roads and an object playing a role of a capital.

A sequence of mappings $E = \{E_n\}_{n\in\mathbb{N}}$ is called a general clustering method, if E_n maps a collection of entities $X^n = \{x_1, ..., x_n\}$ to a collection of subsets $\{C_1, ..., C_G\}$ of X^n. It is as-

[1] Daniel Kosiorowski: Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-510 Cracow, Poland, daniel.kosiorowski@uek.krakow.pl.

[2] Przemysław Jaśko: Cracow University of Economics, Department of Computational Systems, 27 Rakowicka St., 31-510 Cracow, Poland, jaskop@uek.krakow.pl.

sumed that entities with different indexes can be distinguished. For a disjoint clustering method (DCM) $C_i \cap C_j = \emptyset$ for $i \neq j \leqslant G$. Most popular DCM yield partitions $\bigcup_{j=1}^{G} C_j = X^n$.

We consider a following general spatio-temporal model of a digital development

$$m\colon \mathbb{R}^d \times T \ni (z,t) \underset{m}{\rightarrow} P(z,t) \in \mathcal{F} \tag{1}$$

where $z \in \mathbb{R}^d$ denotes spatial coordinates, $T = [0, t_*)$ denotes time, $\mathcal{F}$ denotes a certain family of probability density functions defined on $\mathbb{R}^d$ and $P(z, t)$ denotes an element of the family indexed by a space point and time point t (for a detailed presentation of general issues related to a specification of (1) see chapter nine of Kokoszka and Reimherr, 2017).

Assume, we have a collection of samples from the above model $X^n_{z1,t1}, X^n_{z2,t2}, \ldots, X^n_{zG,tG}$, indexed by points in the space and time. For each sample we perform robust model based clustering, $E^m_{z1,t1}, E^m_{z2,t2}, \ldots, E^m_{zG,tG}$. Taking into account, that we consider model based clusterings, the sequence of clusterings denotes a sample of estimates of mixture densities (well-defined functions) indexed by the spatial-temporal parameters. The sequence enables us for an estimation of a specific form of the model (1) using known methods of estimation (e.g. using functional principal component method of estimation of a linear model of a type "vector-function" or functional kernel regression, see Ramsay and Silveman, 2005).

2. A comparison of robust model based clustering methods

The term "model-based cluster analysis" was coined by Banfield and Raftery (1993) for clustering based on finite mixtures of Gaussian distributions and related methods. In multidimensional case the standard Gaussian mixture model is to assume that data $X^n = \{x_1, ..., x_n\}$ are modelled as drawn i. i. d. from a distribution with density

$$f(x;\theta) = \sum_{j=1}^{G} \pi_j \phi(x;\mu_j,\Sigma_j) \tag{2}$$

where $\phi(\cdot, \mu_j, \Sigma_j)$ is the density of a Gaussian distribution with mean vector μ_j and covariance matrix Σ_j, π_j is the proportion of the j-th mixture component, $\sum_{j=1}^{G} \pi_j = 1$. The parameter vector θ contains all proportions, means and covariances.

The most popular estimator of θ is the maximum likelihood estimator (ML). It leads to a natural clustering rule: *classify an observation to the mixture component maximizing its posterior probability for a class membership*.

3. Selected robust model based clustering methods

MCLUST. The MCLUST (Scrucca et al. 2016), uses a Gaussian mixture model which has a single term representing a noise, and is given by

$$\prod_{i=1}^{n} \left[\pi_0 \frac{\mathbf{I}\{x_i \in S\}}{V} + \sum_{j=1}^{G} \pi_j \phi(x_i;\mu_j,\Sigma_j) \right], \tag{3}$$

where V is a volume of a convex hull S of the sample X^n and $\mathbf{I}\{\cdot\}$ denotes an indicator function, x_i represents an observation, G denotes number of components, π_j is the probability that an observation belongs to the j-th component ($\pi_j \geqslant 0, \sum_{j=1}^{G} \pi_j = 1$), $\phi(\cdot; \mu_j, \Sigma_j)$ denotes density of the p-dimensional normal distribution with parameters (μ_j, Σ_j).

Model selection is obtained via maximization of Bayesian Information Criterion (BIC)

$$BIC \equiv \ell_{\mathcal{M}}(X^n, \theta^*) p_{\mathcal{M}} \log(n), \tag{4}$$

where $\ell_{\mathcal{M}}(X^n, \theta^*)$ is the maximized loglikelihood for the model and the data X^n, $p_{\mathcal{M}}$ is the number of parameters to be estimated in the model $\mathcal{M}$, and n is number of observations.

TCLUST. The second algorithm we consider is the TCLUST (Fritz et al. 2012). Various aspects of its empirical usefulness were studied in Szlachtowska et al. (2016).

OTRIMLE. The third algorithm we consider is the RIMLE (Coretto, Hennig 2013), which maximizes a pseudo-likelihood, based on the improper pseudo-density of the form

$$\psi_\delta(x; \theta) = \pi_0 \delta + \sum_{j=1}^{G} \pi_j \phi(x; \mu_j, \Sigma_j), \tag{5}$$

where $\phi(\cdot; \mu_j, \Sigma_j)$ is the p-dimensional Gaussian density with mean $\mu_j \in \mathbb{R}^\rho$ and covariance matrix Σ_j, $\pi_0, \pi_j \in [0,1]$ for $j = 1, 2, \ldots, G$, $\pi_0 + \sum_{j=1}^{G} \pi_j = 1$, and $\delta > 0$ is the improper uniform density representing outliers.

This improper uniform density, which is not spanned on the predefined support set (unlike in the MCLUST, where support set S is selected to cover the data sample X^n, is not aimed to model the noise component, but it's rather treated as a technical tool to account for the points, which are in low density areas for Gaussian components. In contrast to the MCLUST model, extreme points in the data sample, won't have impact on the uniform density, which in the RIMLE model takes the improper form.

The parameter vector θ contains all Gaussian components parameters and each of the proportion parameters, including π_0. In the RIMLE model the δ and the number of Gaussian components G are treated as fixed.

Given the sample X^n the improper pseudo-log-likelihood function takes the form

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \psi_\delta(x_i; \theta), \tag{6}$$

and for prespecified value of δ the RIMLE estimator is defined as

$$\hat{\theta}_n^{RIMLE}(\delta) = \arg\max_{\theta \in \Theta_n} \ell_n(\theta), \tag{7}$$

where Θ_n is a constrained parameter space defined as

$$\Theta_n = \left\{ \theta : \pi_j \geqslant 0 \forall j \geqslant 1, \pi_0 + \sum_{j=1}^{G} \pi_j = 1;\ \frac{\lambda_{\max}(\Sigma^\theta)}{\lambda_{\min}(\Sigma^\theta)} \leqslant \gamma;\ \frac{1}{n} \sum_{i=1}^{n} \tau_0(x_i; \theta) \leqslant \pi_{\max} \right\}. \tag{8}$$

The parameter space Θ_n is defined by (Correto and Hennig, 2013) to ensure existence of the RIMLE estimator $\hat{\theta}_n^{RIMLE}(\delta)$. To obtain boundedness of the pseudo-log-likelihood criterion function, along with constraints on the proportion parameters and the eigenratio constraint, additional constraint called the "noise proportion" is needed in the RIMLE case.

To establish the eigenratio constraint, we need to introduce some notation. Given $\lambda_{j,k}$ is the k-th eigenvalue of the j-th component covariance matrix Σ_j, we define the set of eigenvalues $\Lambda(\Sigma^\theta) = \{\lambda_{j,k}: j = 1,2,\ldots, G; k = 1,2,\ldots, p\}$, from which we select respectively minimal and maximal element: $\lambda_{\min}(\Sigma^\theta) = \min_{j,k} \Lambda(\Sigma^\theta)$, $\lambda_{\max}(\Sigma^\theta) = \max_{j,k} \Lambda(\Sigma^\theta)$. Eigenratio constraint is aimed at preventing from the degeneracy of model "regular" components distributions, in other words, from having some components with Gaussian distributions, concentrated in the vector subspace of $\mathbb{R}^p$, resulting in the singular covariance matrix parameters for them, which in turn translates into the infinite value of the pseudo-likelihood function. In the RIMLE case, eigenratio constraint alone, cannot preclude forming "regular" components with distributions concentrated on single points, and all other points fitted by the improper uniform component. In order to mitigate such kind of situations, the "noise proportion" constraint is added to the definition of RIMLE parameter space. Considered constraint is defined using pseudo posterior probabilities of the noise component for consecutive sample observations.

Under the RIMLE model with parameter θ, noise component pseudo posterior probability $\tau_0(x_i; \theta)$, conditional on the observation $x_i \in X^n$, is given by

$$\tau_0(x_i;\theta) = \frac{\pi_0 \delta}{\psi_\delta(x_i;\theta)} = \frac{\pi_0 \delta}{\pi_0 \delta + \sum_{j=1}^{G} \pi_j \phi(x_i;\mu_j,\Sigma_j)}. \tag{9}$$

Using ergodic arguments, X^n sample mean of the pseudo posterior probabilities $\frac{1}{n}\sum_{i=1}^{n} \tau_0(x_i;\theta)$, can be treated as an approximation to the expected proportion of noise points, under model parameter θ. So, constraint $\frac{1}{n}\sum_{i=1}^{n} \tau_0(x_i;\theta) \leqslant \pi_{\max}$, imposes that expected noise proportion should be no higher than $\pi_{\max} \in (0,1)$.

As can be seen the "noise proportion" constraint is sample dependent, which in turn translates into the dependency of the parameter space Θ_n on the sample X^n. Pseudo posterior probabilities $\tau_j(x_i; \theta)$, $j = 1,2,\ldots, G$ for the Gaussian components are defined analogously. The observation x_i is assigned to the cluster whose index corresponds to the highest pseudo posterior $J(x_i; \theta) = \arg\max_{j\in\{1,2,\ldots,G\}} \tau_j(x_i; \theta)$. To fix the value of δ for the noise component improper uniform density, on which estimator $\hat{\theta}_n^{RIMLE}(\delta)$ depends, (Correto, Hennig, 2013) proposed the OTRIMLE (Optimally Tuned RIMLE) procedure. Under the OTRIMLE approach, improper density level δ is selected according to the formal criterion, which allow to make trade-off between conformity of components' empirical distributions with Gaussian distribution and a proportion of outliers. Optimal level for the improper density is given by

$$\delta_n = \arg\min_{\delta \in [0,\delta_{\max}]} D(\delta) + \beta\pi_{0,n}, \tag{10}$$

where $\delta_{\max} > 0$ is a prespecified value, and $\beta \geqslant 0$ is the penalty parameter for increasing the proportion of outliers component, while $D(\delta)$ measures departures from Gaussianity of "regular component" empirical data clusters, resulting under the model improper constant density level of δ. With the observation assignment rule for J the RIMLE estimator $\hat{\theta}_n^{RIMLE}(\delta)$, value of the $D(\delta)$ is based on the Kolmogorov distances between the components' empirical distributions of square Mahalanobis distances and the chi-square distribution with degrees of freedom, which is expected for them under Gaussianity of components. Under the OTRIMLE procedure, to find the optimal value of δ_n, the RIMLE estimator $\hat{\theta}_n^{RIMLE}(\delta)$ value is computed over the candidate set, within certain "golden section search algorithm". To compute values of the RIMLE estimators the ECM-algorithm (Expectation Conditional Maximization algorithm) is used, for which pseudocode is presented in (Correto and Hennig, 2013). This kind of algorithm imposes fulfillment of the RIMLE parameter space constraints in each of its iteration.

4. A comparison of computational algorithms for considered robust estimators

Table 1. A summary comparison of the clustering algorithms

Aspect \ Algorithm	MCLUST	TCLUST	OTRIMLE
An approach to take noise into account	Uniform distribution with a predefined support dependent on a convex hull of a sample X^n.	A noise is jointly modelled with outliers.	Using an improper uniform distribution.
Conditions for the identification of the estimator	Different kinds of parametrisations for cluster dispersion matrices (imposing constraints of cluster distributions).	Conditions on ratios of cluster dispersion matrices eigenvalues.	Conditions on ratios of cluster dispersion matrices eigenvalues, and on proportion of noise.
Consistency of the estimator	By default ML estimator enhanced with BIC is used.	By default ML estimator based on the MCD is used.	
Affine equivariance of the estimator	Maximised objective function is affine equivariant, but initialization steps are not.	Depending on the predefined constraints on scatter matrices.	Not, for the basic form of the estimator. Yes, for the modified RIMLE.
Algorithm for finding an optimal value of the objective function	Expectation-Maximization (EM)	Classification EM	Expectation-Conditional Maximization (ECM)

Aspect \ Algorithm	MCLUST	TCLUST	OTRIMLE
Breakdown point	Problems with outliers at extreme positions.	Depends on a configuration of data in a sample.	Depends on a configuration of data in a sample.
Implementation in R	`mclust`	`tclust`	`otrimle`

OTRIMLE criterion for the δ constant, allows for a trade-off between a departure of components of empirical distributions from normality and proportion of outliers in a sample. Thanks to it, OTRIMLE is less prone to existence of the extreme points in the sample data than the MCLUST, where support for the noise component uniform density is a set including all the sample points (most often convex hull or hyperrectangle).

Simulation studies. In order to investigate small and moderate sample properties of the algorithms in a context of exploring spatial phenomena of digital economy we conducted extensive simulation studies. We among others generated samples from 2D mixtures of three component skewed Student T distributions "noised" by distribution with two-spiral support. The samples consisted of 3330 points, from which 3000 were "clean" points, 300 noise points and 30 outlying observations.

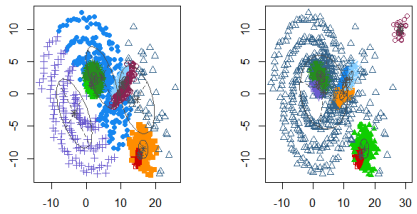

Fig. 1. Results of clustering using MCLUST

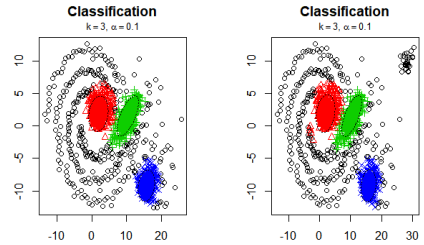

Fig. 2. Results of clustering using TCLUST

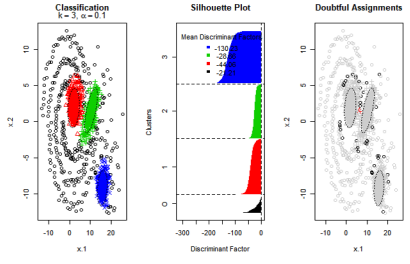

Fig. 3. Properties of TCLUST clustering for dataset with 10% spiral noise

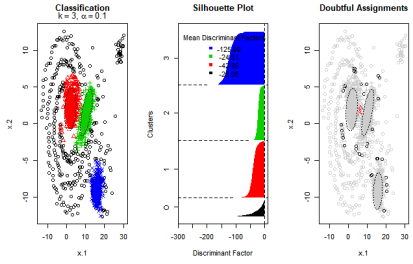

Fig. 4. Properties of TCLUST clustering for dataset with 10% spiral noise with 1% outliers

Figure 1 presents results of the clustering using MCLUST for "clean data with noise" (left panel) and "noised data with outliers" (right panel). Figures 2–4 present analogous situation

for the TCLUST algorithm and figures 5–6 for the OTRIMLE algorithm. The figures may be treated as an illustration of a one experiment from a collection of 1000 experiments.

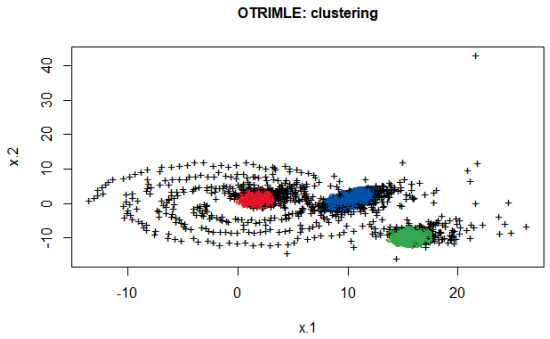

Fig. 5. Results of clustering using OTRIMLE

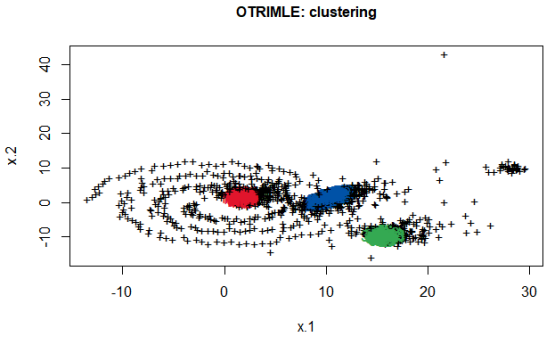

Fig. 6. Results of clustering using OTRIMLE

Empirical example. We used considered algorithms in the image segmentation task for the 24-bit RGB digital raster image, with the resolution of 200 × 200 pixels, representing night lights satellite image of Poland. We aimed at establishing spatial clusters, with similar level of activity, measured by the night light intensity. We assumed that the light intensity correspond to a degree of development. To extract data from the image we used R package magick. Data were downloaded from the NOAA Database (https://www.ngdc.noaa.gov/eog/download.html). Figures 7 and 8 present results of the clustering obtained via MCLUST and TCLUST algorithms correspondingly. For the algorithms we assumed data sample partition into six clusters (indicated by the BIC for MCLUST model) and a set of outliers.

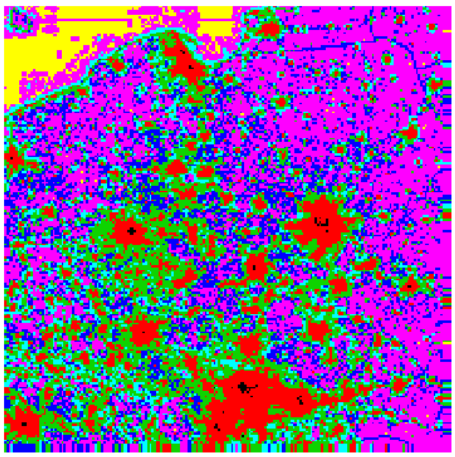

Fig. 7. Night light intensity–results of MCLUST

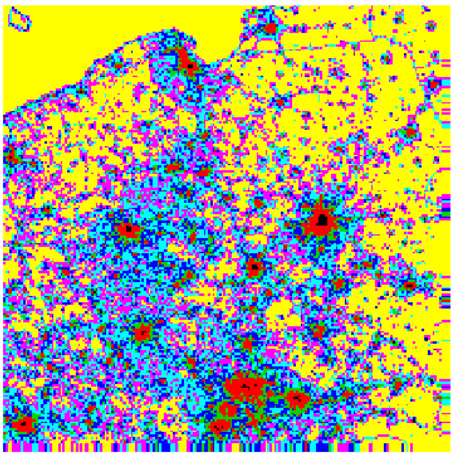

Fig. 8. Night light intensity–results of TCLUST

For both methods results, cluster numbers are sorted according to the decreasing order of its mean nightlight brightness. Set grouping outlying observations, associated with the brightest pixels is numbered as zero. First we describe results of the MCLUST algorithm.

Table 2. Pixels (area) partition between clusters by MCLUST and TCLUST

Cluster\ Fraction of pixels	0 (black)	1 (red)	2 (green)	3 (blue)	4 (cyan)	5 (magenta)	6 (yellow)
MCLUST	0.18	13.96	22.46	16.86	9.87	32.76	3.91
TCLUST	0.20	4.13	6.23	10.95	19.03	19.65	39.81

The brightest pixels associated with the main parts of Polish biggest cities (Upper Silesian conurbation, Warsaw, Poznan, Cracow, Gdansk) were assigned as outlying observations (these pixels are marked by the black colour on the maps). The first cluster is formed by the brightest pixels (marked on the map by the red colour) not assigned as outliers, which are concentrated in the areas near the Polish biggest cities, especially area between Upper Silesia and Krakow and Warsaw suburbs are distinguished. Subsequent three clusters with decreasing mean brightness level, are continuously formed by pixels associated with places with increasing distance from the main cities (pixels from clusters 2 to 4 are marked on the map respectively by the colours: green blue and cyan). Cluster 5 groups pixels representing terrestrial areas, which are least illuminated, which are large parts of NE and NW Poland. The sixth cluster, groups pixels representing areas on the Baltic Sea, distant from the shores, which are not artificially enlightened. For the TCLUST (with assumed fraction of outliers equal 0.2%) clustering results seems to be much better than that for the previous method. Most of the clusters include spatially concentrated group of pixels. Pixels assigned as outliers (black), form areas which highly resemble administrative territories of the biggest Polish cities. First cluster contains pixels (in red) associated with strict metropolitan areas of the mentioned cities (the radius for this areas are much smaller than in MCLUST case). The cluster 2 contains comparatively to the MCLUST, tiny fraction of pixels (in green) concentrated mainly in the area of Upper Silesia and western part of Lesser Poland. Cluster 4 contains pixels (cyan) which form fairly vast, continuous areas outside the metropolitan areas of biggest cities. Cluster 5 embraces highly scattered pixels (in magenta), mainly in the western part of the country. Cluster 6 contains the largest fraction of pixels, representing the least illuminated terrestrial areas (north-eastern and eastern part of Poland and north-west) and the full analysed area of the Baltic Sea (MCLUST assigned to the sixth cluster only distant parts on the sea). To sum up, in the case of the MCLUST and TCLUST respectively, pixels associated with each of the six clusters (0 stands for "outlying" pixels group), constitute following fractions of the pixels, representing all of the analysed area (which is presented on the map).

Conclusions and further studies

We have critically studied three high quality model based clustering algorithms in the context of their applications in modelling of spatial phenomena. We cannot indicate a "total winner", which uniformly maximizes all criteria of evaluation.

Acknowledgements

DK and PJ thanks for financial support from the Ministry of Science and Higher Education within "Regional Initiative of Excellence" Programme for 2019–2022. Project no.: 021/RID/2018/19. Total financing: 11 897 131,40 PLN. DK thanks for the support related to CUE grant for the research resources preservation 2019.

References

Banfield, J.D. and Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics,* 49, 803–821.

Fritz H., García-Escudero L. A., Mayo-Iscar A., (2012), tclust: An R Package for a Trimming Approach to Cluster Analysis, *Journal of Statistical Software*, 47(12), 1–26.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location – scale mixtures. *The Annals of Statistics*, 32(4), 1313–1340.

Coretto, P., & Hennig, C. (2013). Finding approximately Gaussian clusters via robust improper maximum likelihood. *arXiv preprint arXiv:1309.6895*.

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6), 1154–1176.

Kokoszka, Reimherr, M. (2017). Introduction to Functional Data Analysis, CRC, London.

Kosiorowski, D., & Zawadzki, Z. (2019). DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena, *Journal of Statistical Software*, forthcoming.

Ramsay J., & Silverman, B. (2005). *Functional Data Analysis*. Springer.

Scrucca, L., Fop, M., Murphy, T.B., & Raftery, A.E. (2016). mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1), 289.

Szlachtowska E., Kosiorowski D., Mielczarek D., (2016). Ocena jakości aplikacyjnej odpornego algorytmu analizy skupień TCLUST na przykładzie zbioru danych dotyczących jakości powietrza w Krakowie, *Przegląd Statystyczny*, R. 63(1), 67–80.