K-local median algorithm for functional data in empirical analysis of air pollution data

Daniel Kosiorowski[1], Ewa Szlachtowska[2]

Abstract

Novel tools offered by functional data analysis enable economists for enlarging a range of considered issues as well as for obtaining new insights into classical areas of empirical economic research. This paper presents a novel k–local functional median algorithm for functional data. Its statistical properties as well as its usefulness in analysis of real data set concerning air pollution monitoring in Malopolskie voivodeship in Poland in 2016 are shown. An implementation of the algorithm within a free R package DepthProc is indicated and its comparison with selected alternatives presented in the literature is performed.

Keywords: *air pollution monitoring, k-local functional median clustering, functional k-means clustering, functional data analysis*

JEL Classification: C14, C53, C55

1 Introduction

Pollination in Malopolska is an extremely important social problem. It influences on the social costs (such as the average length and quality of life) and economic (e.g. medical expenses, loss of tourist values). Cluster analysis for functional data representing daily trajectories of concentrations of hazardous substances (including nitrogen oxides) allows for an optimization of the environmental policy of the region.

2 K-local functional median algorithm

Main aim of the paper is to propose a novel statistical method of conducting preliminary analysis of the problem of air pollution in Cracow. For this purpose, we used two clustering algorithms for functional data: k-means algorithm and k-local functional median algorithm. The k-local functional median algorithm is our original proposal, in which we used an idea of local depth proposed in Paindaveine and Van Bever (2013) for the modified band depth proposed by Lopez-Pintado and Romo (2006) and intensively studied in Nieto-Reyes and Battey (2016) .

[1] Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, e-mail: daniel.kosiorowski@uek.krakow.pl.

[2] Corresponding author: AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, e-mail: szlachto@agh.edu.pl.

2.1 Local depth

Depths express centrality of objects with respect to samples or probability disstributions, see for example Nagy et. al. (2016). Depths describe global property of data cloud or the underlying distribution - a degree of outlyingness of a point from a center - the median. In many situations however local properties of data set are of prime importance. To these situations belong clustering issues, where multimodality of data set has to be stressed. In this context several local extensions of depths have been proposed. Our proposal base on one of them presented in Paindaveine and Van Bever (2013) for multivariate data case. For more examples on local depths we refer to Lopez-Pintado et. al. (2007).

Let x1(t),...,xn(t) denotes a set of real functions, for simplicity let us assume that they belong to C[0,1] continuous functions defined on an interval [0,1]. A graph of a function x is a subset of $\Re^2$ defined as

$$G(x) = \{(t, t(x)) : t \in [0,1]\}\,. \tag{1}$$

A band in $\Re^2$ determined by k functions from a sample x1,...,xn is defined:

$$\begin{aligned} V(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) &= \left\{(t, y) : t \in [0,1],\ \min_{r=1,\ldots,k} x_{i_r}(t) \le y \le \max_{r=1,\ldots,k} x_{i_r}(t)\right\} \\ &= \left\{(t, y) : t \in [0,1], y = \alpha_t \min_{r=1,\ldots,k} x_{i_r}(t) + (1-\alpha_t) \max_{r=1,\ldots,k} x_{i_r}(t), \alpha_t \in [0,1]\right\}. \end{aligned} \tag{2}$$

For any function x and set of functions {x1,...,xn} an index of j functions

$$S_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} I\left\{G(x) \subset V(x_{i_1}, x_{i_2}, \ldots, x_{i_j})\right\}, \tag{3}$$

$j \ge 2$, expresses a fractions of bands $V(x_{i_1}, x_{i_2}, \ldots, x_{i_j})$ determined by *j* different functions $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$, covering a graph of *x*.

Definition 1: For functions x1,...,xn the band depth of a function f equals

$$S_{n,J}(x) = \sum_{j=2}^{J} S_n^{(j)}(x), \quad J \ge 2. \tag{4}$$

In case, when X1,...,Xn are independent copies of stochastic process X, which generates x1,...,xn, population versions of depth indices are defined:

$$S^{(j)}(x) = P\left\{G(x) \subset V(X_{i_1}, X_{i_2}, \ldots, X_{i_j})\right\}, \tag{5}$$

$$S_J(x) = \sum_{j=2}^{J} S^{(j)}(x) = \sum_{j=2}^{J} P\left\{G(x) \subset V(X_{i_1}, X_{i_2}, X_{i_j})\right\}. \tag{6}$$

A function being a sample median with respect to a sample $\hat{m}_{n,J}$ is a curve which maximizes the sample depth:

$$\hat{m}_{n,J} = \underset{x\in\{x_1,\ldots,x_n\}}{\arg\max} S_{n,J}(x). \tag{7}$$

In a population case as the median we take m$_J$ in C[0,1] which maximize S$_J$(). Unfortunatelly there are great difficulties in applications of the above concept of functional depth in case of economic time series. Trajectories of economic objects are crossing for many times what makes the band depth rather useless. Lopez-Pintado and Romo (2006) proposed much better concept of functional depth for economic applications. For any function x from a sample {x$_1$,...,x$_n$}, $j \geq 2$ let

$$A_j(x) \equiv A(x; x_{i_1}, \ldots, x_{i_j}) \equiv \left\{ t \in [0,1] : \min_{r=i_1,\ldots,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,\ldots,i_j} x_r(t) \right\}, \tag{8}$$

denotes a set of points in the interval [0,1], for which a function x is inside a band determined by $x_{i_1}, x_{i_2}, \ldots, x_{i_j}$.

If λ is the Lebesque's measure on the interval [0,1], λ(A$_j$(x)) is a fraction of time, in which the function x is inside a band.

Definition 2: A generalized band depth of a curve x is

$$GS_n^{(j)}(x) = \binom{n}{j}^{-1} \sum_{1\leq i_1<i_2<\cdots<i_j\leq n} \lambda(A(x; x_{i_1}, x_{i_2}, \ldots, x_{i_j})), \quad j \geq 2. \tag{9}$$

A function being a sample median is defined as:

$$\hat{m}_{n,J} = \underset{x\in\{x_1,\ldots,x_n\}}{\arg\max} S_{n,J}(x) . \tag{10}$$

In a population case the median m$_J$ maximizes S$_J$(·) in C[0,1].

For further generalization of the band depth and their theoretical properties see Nieto-Reyes and Battey (2016). For any depth function D(x,P), the depth regions, $R_\alpha(P) = \{x \in L^2([0,T]) : D(x,P) \geq \alpha\}$ are of paramount importance as they reveal very various characteristics of probability distribution *P*: location, scatter, dependency structure (clearly these regions are nested and inner regions contain larger depth). When defining local depth it will be more appropriate to index the family $R_\alpha(P)$ by means of probability contents. Consequently, for any $\beta \in (0,1]$ we define the smallest depth region with P-probability equal or larger than β as

$$R^\beta(P) = \bigcap_{\alpha\in A(\beta)} R_\alpha(P), \tag{11}$$

where $A(\beta)=\{\alpha\geq 0: P(R_\alpha(P))\geq\beta\}$. The depth regions $R_\alpha(P)$ or $R^\beta(P)$ provide neighborhoods of the deepest point only. However we can replace P by its symmetrized version $P_x = 1/2P^x + 1/2\,P^{2x-X}$. Let $D(\cdot,P)$ be a depth function. The corresponding *sample local depth function at the locality level* $\beta\in(0,1]$ is $LD^\beta(x,P^{(n)})=D(x,P_x^{\beta(n)})$, where $P_x^{\beta(n)}$ denotes the empirical measure with those data points that belong to $R_x^\beta(P^{(n)})$. $R_x^\beta(P^{(n)})$ is the smallest sample depth region that contains at least a proportion β of the 2n random functions $x_1,...,x_n,2x-x_1,...,2x-x_n$. Depth is always well defined – its an affine invariance originates from original depth. Notice for β=1 we obtain global depth, while for β≈0 we obtain extreme localization.

As in the population case, our sample local depth will require considering, for any $x\in L^2$, the symmetrized distribution P_x^n which is empirical distribution associated with $x_1,...,x_n,2x-x_1,...,2x-x_n$. Sample properties of the local versions of depths result from general findings presented in Paindaveine and Van Bever (2013). Implementations of local versions of several depths including projection depth, Student, simplicial, L_p depth, regression depth and modified band depth can be found in free R package *DepthProc* (see Kosiorowski and Zawadzki, 2014). For choosing the locality parameter β we recommend using cross validation related to an optimalization a certain merit criterion (the resolution being appropriate for comparing phenomena in terms of their aggregated local shape differences, which relies on our knowledge on the considered phenomena).

2.2 K–local functional median algorithm – our proposal

We first choose k, where k is user-specified parameter, namely, the number of cluster desired. The second parameter chosen by the researcher is the value of the parameter β (default value of β is 0.2). With the parameter β we may consider the issue at different levels, i.e. changing the value of the parameter β we can control the accuracy of the partition. Received values of local depth for all points are helpful in choosing the amount of clusters.

We consider data whose proximity measure is depth. For our objective function, which measures the quality of a clustering, we use

$$F=\sum_{i=1}^{k}\sum_{j\in C_i} LD^\beta(f_j,P_{n,i}), \tag{12}$$

where $P_{n,i}$ is empirical distribution of i-th cluster.

In first step we calculate the local depth for all points of the data set with respect to chosen values of parameters β and k. Then we are looking for k centroids $c_1,...,c_k$ satisfying

$$\sum_{i=1}^{k} LD^{\beta}(c_i, P^{(n)}) \to \max \tag{13}$$

$$\sum_{i,j=1}^{k} \| c_i - c_j \|_2 \to \max \tag{14}$$

In the second step we create new clusters in such a way that for every point we count L_2 distance from all centroids and assign the point to the closest centroid, that is,

$$Dist(f,P) = \max\{d(f,c_1), d(f,c_2), \ldots, d(f,c_k)\}, \tag{15}$$

where d denotes L_2 distance.

If there are two distances with the same values, then we assign a point to clusters with a lower number. In third step for the newly formed cluster we compute the functional local median with respect to the empirical distribution of the cluster. We repeat second and third steps, until centroids do not change or until only 1% of the points change clusters. Note, robustness of the clustering procedure may be evaluated using well known measures of clustering results quality (see Walesiak and Dudek, 2015). For example small changes of input data should lead to small changes of the silhouette plot characteristics in case of robust clustering procedure. More dilemmas of robust analysis of economic data streams can be found in Kosiorowski (2016).

2.3 Trimmed k-local functional median algorithm

The user chooses the parameter γ. We calculate a measure of degree of affiliation to cluster for each observation, i.e. $d(f,c_{l(f)})=d(f,c_{(1)})$. We set obtained values descending

$$d(f_{(1)}, c_{l(f_{(1)})}) \geq d(f_{(2)}, c_{l(f_{(2)})}) \geq \cdots \geq d(f_{(n)}, c_{l(f_{(n)})}). \tag{16}$$

Then we reject a proportion γ of the observation of the highest values of measure of affiliation. For this method discriminant factors can be obtained for every observation (trimmed and non trimmed) in the data set (see Fitz et. al., 2015). The quality of the assignment decision of a non trimmed observation f_i to the cluster j with $d(f_{(i)},c_{l(f(i))})$ can be evaluated by comparing its degree of affiliation with cluster j to the best second possible assignment. That is,

$$DF(f_i) = \log\left(\frac{d(f_i, c_{(2)})}{d(f_i, c_{(1)})} \right) \tag{17}$$

for f_i not trimmed. Let $f_{(1)}, \dots ,f_{(n)}$ be the observations in the sample after being sorted according to their $d(f_{(i)},c_{l(f(i))})$ values. It is not difficult to see that $f_{(1)}, \dots, f_{(\lceil\gamma n\rceil)}$ are the trimmed observations which are not assigned to any cluster. Nevertheless, it is possible to compute the degree of affiliation $d(f_{(i)},c_{l(f(i))})$ of a trimmed observation f_i to its nearest cluster. Thus, the quality of the trimming decision on this observation can be evaluated by comparing $d(f_{(i)}, cl_{f(i)})$ to $d\left(f_{(\lceil\gamma n\rceil+1)}, c_{l\left(f_{(\lceil\gamma n\rceil+1)}\right)}\right)$ with $f_{(\lceil\gamma n\rceil+1)}$ being the non-trimmed observation with smallest value of $d(.,c_{l(.)})$. That is

$$DF(f_i) = \log\left(\frac{d(f_{(i)}, c_{l\left(f_{(i)}\right)})}{d(f_{(\lceil\gamma n\rceil+1)}, c_{l\left(f_{(\lceil\gamma n\rceil+1)}\right)})}\right) \tag{18}$$

for f_i trimmed. Hence, discriminant factors $DF(f_i)$ are obtained for every observation in the data set, whether trimmed or not. Observations with small $DF(f_i)$ values (that is, values close to zero) indicate doubtful assignments or trimming decisions. Further properties and theoretical properties of the proposals may be found in Kosiorowski et. al. (2017).

3 Air pollution in Cracow

Air pollution is a very important problem in Malopolska. We use clustering algorithms for functional data to analyze air pollution in Cracow. We choose air quality data in the period from 1 to 31 December 2016 in Cracow, station Avenue Krasinski. All the considered data was taken from *Malopolskie, System monitoringu jakosci powietrza* http://monitoring.krakow.pios.gov.pl/ In the study we used the techniques used in functional data analysis. For more details we refer to Horvath and Kokoszka (2012), Ramsay et. al. (2009). We used the following packages fda.usc (see Febrero-Bande and de la Fuente, 2012) and DepthProc (see Kosiorowski and Zawadzki, 2014).

We observe that the biggest concentration of PM10 occurs in the afternoon, between the hours of 3 p.m. and 8 p.m. In contrast, the height of the concentration of nitrogen dioxide is more varied. It depends not only on time, but also depends on the day of the week.

First, the analysis was conducted for nitrogen dioxide pollution by using k-means algorithm. If we divide the observations into three clusters by k-means algorithm, then first cluster includes 1, 2, 6, 8, 9, 14, 15, 19, 22, 23, 29 December. Second cluster includes 3, 10, 11, 18, 24, 25, 26, 27, 28 December. Third cluster includes 4, 5, 7, 12, 13, 16, 17, 20, 21, 30, 31 December. In the second cluster there are weekends, holidays and the period after Christmas. Examining the first and third cluster it is difficult to find a relationship between the days of the week and the amount of pollination. It is worth noting that in the k-means algorithm, an important step is the selection of centroid. For parameter k=5 at subsequent iterations sometimes we get empty fifth cluster. The greatest concentration of suspended dust was 5, 17, 30 and 31 December, just before New Year's Eve. From the recorded data difficult to see the correctness, in which days of the week is the greatest concentration of dust. However, the lowest concentration is in the Christmas period, i.e. 23-29 of December.

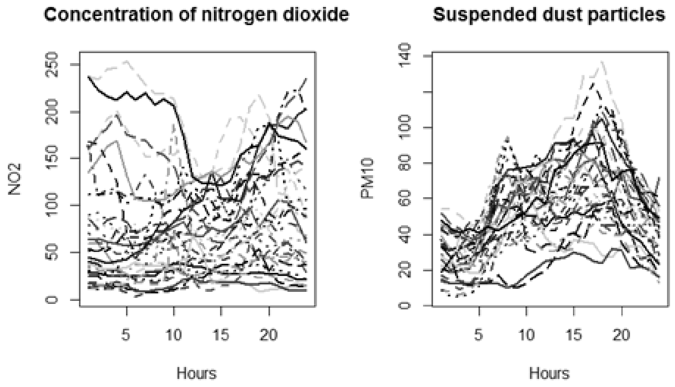

Fig. 1. The amount of nitrogen dioxide and suspended dust particles emitted into the atmosphere as air pollution in Cracow, December 2016.

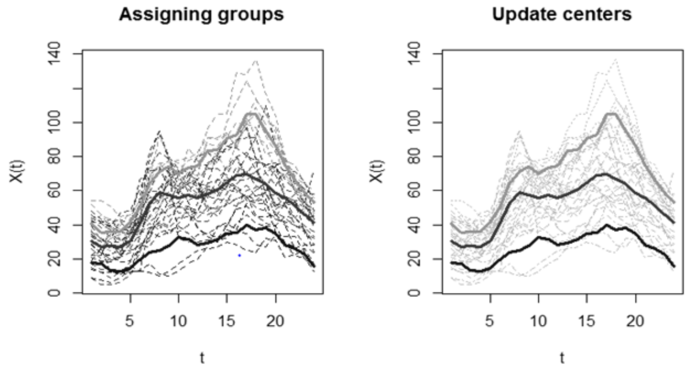

Fig. 2. Functional means for individual clusters and assigning groups for nitrogen dioxide pollution emitted into the atmosphere as air pollution in Cracow, December 2016. Method – functional k-means algorithm, k=3.

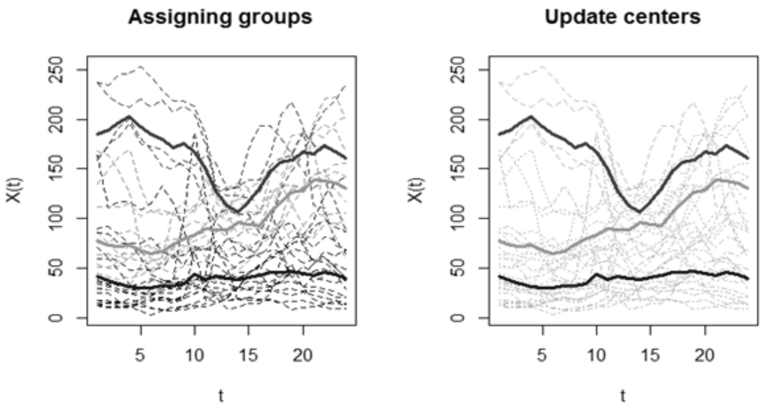

Fig. 3. Functional means for individual clusters and assigning groups for particulate matter pollution in Cracow, December 2016. Method – functional k-means algorithm, k=3.

In the next step of the analysis of concentrations of hazardous substances we used the k-local functional median. By changing input parameter k, we received the optimum division into two clusters. If we divide the observations into three clusters by k-local functional means algorithm, then first cluster includes 1, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 29, 30, 31 December. Second cluster includes 2, 10, 11, 18, 24, 25, 26, 27, 28 December. As in previous analysis in the second cluster there are weekends, holidays and the period after Christmas. While analyzing the concentration of particulate matter we obtained two clusters. First cluster includes 1, 2, 3, 9, 10, 11, 12, 15, 24, 25, 26, 27, 28, 29 December. The greatest concentration of particulate matter occurred in the period before the Christmas, as well as just before New Year's Eve.

Conclusions

In our study we faced an issue of missing data. The missing data may arise through equipment failure. They were supplemented by using the median of the observed cases on the variable in each hour. The next issue was to compare the results obtained with two clustering algorithms. They were very similar. However, because of the computational complexity of the local functional median calculation our algorithm randomly selects a sample, for which it determines centroids. Generating several times the algorithm helps us to verify the choice of number of clusters by comparing the received groups and variation in these groups.

In summary we obtain that the smallest concentration of nitrogen dioxide occurred on holidays (i.e. the first group). The largest concentration on weekdays (i.e. the second group). Just before Christmas, during the departure of the holidays nitrogen dioxide concentration remained at a medium level. Therefore it can be hypothesized that the concentration of nitrogen dioxide depends on the volume of traffic on the road. For example knowledge of the phenomenon of variability of air pollution can help in the environmental policy of the city, in the planning of free communication or traffic restrictions.

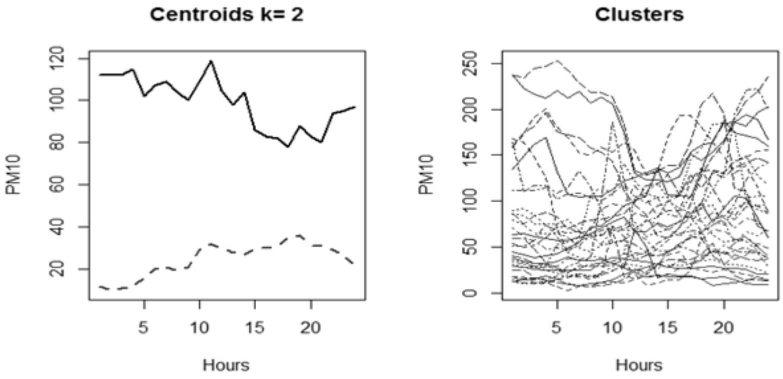

Fig. 4. Functional medians for individual clusters and assigning groups for nitrogen dioxide pollution emitted into the atmosphere as air pollution in Cracow, December 2016. Method – k-local functional median algorithm, k=2.

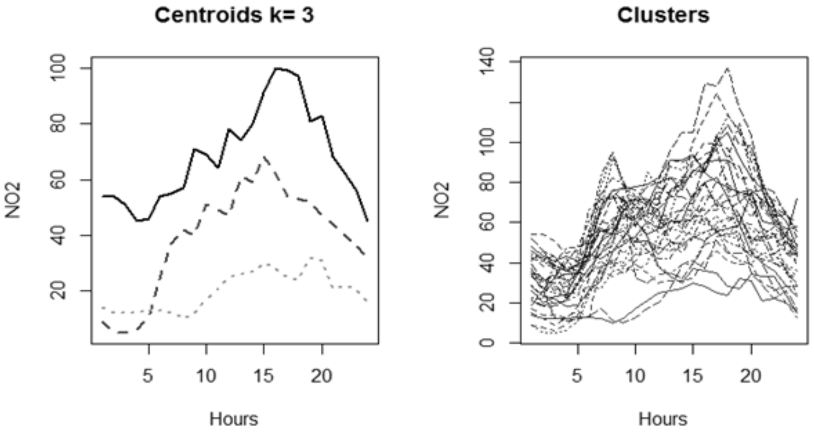

Fig. 5. Functional medians for individual clusters and assigning groups for particulate matter pollution in Cracow, December 2016. Method – k-local functional median algorithm, k=3.

Acknowledgements

DK thanks for the support related to CUE grant for the research resources preservation 2017.

References

Febrero-Bande, M., & de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software 51(4),* 1-28.

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.*

Horvath, L., Kokoszka, P. (2012). Inference for functional data with applications. *Springer.*

Kosiorowski, D., Mielczarek, D., & Szlachtowska, E. (2015). Clustering of functional objects in energy load prediction issues.*:IX International scientific conference in honour of Professor Aleksander Zelias : Zakopane, Poland, 12–15 may 2015.*

Kosiorowski, D., Szlachtowska, E., & Zawadzki Z. (2017). Selected clustering algorithms for functional data in specification of functional time series. *Submitted*

Kosiorowski, D., & Zawadzki, Z. (2014). DepthProc: An R package for robust exploration of multidimensional economic phenomena. *http://arxiv.org/pdf/1408.4542.pdf.*

Kosiorowski, D. (2016). Dilemmas of robust analysis of economic data streams. *Journal of Mathematical Sciences 1(2),* 59-72.

Lopez-Pintado, S., & Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics & Data Analysis 51*, 4957-4968.

Nagy, S., Hlubinka, D., & Gijbels, I. (2016). Integrated depth for functional data: Statistical properties and consistency. *ESIAM Probability and Statistics.*

Nieto-Reyes, A., & Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science 31(1),* 61-79.

Paindaveine, D., & Van Bever, G. (2013). From depth to local depth: a focus on centrality. *Journal of the American Statistical Association 108(503),* 1105-1119.

Ramsay, J., Hooker, G., & Graves, S. (2009). Functional data analysis with R and Matlab. *Springer.*

Walesiak, M., Dudek, & A. (2015). Searching for Optimal clustering procedure for a data set: the ClusterSim R package, https://CRAN.R-project.org/package=clusterSim.