Outliers in Functional Time Series – Challenges for Theory and Applications of Robust Statistics

Daniel Kosiorowski[1], Dominik Mielczarek[2], Jerzy P. Rydlewski[3]

Abstract

This paper critically discusses the most promising, known from the literature, approaches to analysis of robustness in the functional time series setup. We also propose our own method of detecting functional outliers appealing to the generalized Young inverse function. The method is dedicated for detecting a certain kind of shape outliers. In empirical example we study day and night air pollution with PM10 in Katowice in 2016.

Keywords: *Functional Outliers Detection; Functional Data Analysis; Robust Economic Analysis*
JEL Classification: C12, C13, C14
DOI: 10.14659/SEMF.2018.01.21

1 Introduction

Robust statistics studies various relationships between majorities and influential minorities of observations within data sets and within underlying statistical models. These studies involves behaviour of estimators or tests when samples come from data generating processes, which are in neighbourhoods (minority) of assumed models (majority). Having at our disposal a reasonable definition of majority as a by-product one may define an outlier as observation departing from the majority in a certain justified way (Maronna et al., 2006). Although main aim of robust statistics is to propose robust procedures, i.e., techniques having good properties at the assumed model as well as in case of a slight departure from the model, a detection of outliers may be even more important issue in a context of developing new theories explaining economic phenomena. Notice that departure from the main pattern may signal problems with an old theory or a completely new phenomenon. In practice, outliers detection procedures may be important in a context of safety systems development or e-economy monitoring.

Many economic phenomena may be treated as functional time series (FTS) – series of functions of a certain continuum representing, for example, time, temperature, interest rate, an aversion to risk, age of a credit applicant etc. Assuming a certain degree of regularity as to

[1] Corresponding author: Cracow University of Economics, Department of Statistics, ul. Rakowicka 27, 31-510 Kraków, e-mail: daniel.kosiorowski@uek.krakow.pl.
[2] AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland, e-mail: dmielcza@wms.mat.agh.edu.pl.
[3] AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland, e-mail: ry@agh.edu.pl.

data generating the phenomena, one can effectively model, conduct statistical inference and predict them within a framework offered by relatively new discipline of multidimensional statistics called functional data analysis (FDA). Unfortunately, observed empirical data sets very often manifest departure from the assumed regularity in a form of existence of outliers within data. These facts may rise doubts as to credibility of applications of functional generalizations of standard time series techniques like ARMA modelling (Horváth and Kokoszka, 2012; Górecki et al., 2017). It should be stressed however, that although there are known effective techniques of coping with outliers in one- and multidimensional cases, the situation is far from satisfactory in classical time series setting (Maronna et al., 2006; Galeano et al., 2006). Within the FDA we additionally face new challenges for development of robust statistical procedures. We have to cope with completely new classes of outliers. Unfortunately, straightforward generalizations of simple outliers detection tools, like the boxplot or the quantile-quantile plot, do not exist.

This paper critically discusses the most promising approaches, known from the literature, to outliers detection appearing in the FDA and focuses on selected challenges for the FTS framework. We also present our own proposal related to the generalized Young inverse function. The method is designed for detecting a certain kind of shape outliers. In an empirical example we study day and night air pollution with PM10 in Katowice in 2016. A reasonable outliers detecting algorithm in this context should indicate anomalies which may give us an insight into relationships between various factors and events influencing air quality and in consequence leading to a better local pro-ecological regulations. The rest of the paper is organized as follows. In Section 2 we list basic types of outliers in the FDS setup and briefly indicate the best detection procedures known from literature. We also discuss their drawbacks and challenges. In Section 3 we introduce our original procedure for detecting of certain kind of shape outliers. In Section 4 selected challenges in the FTS setup are discussed. Section 5 presents a comparison of the procedures in the empirical example. Section 6 consists of some conclusions and summary.

2 Outliers in functional data analysis setup

Within the FDA, functional data are considered as sample observations of random functions, i.e., random elements of certain function space. In general, such spaces are typically considered to be infinite dimensional, real and separable Banach or Hilbert spaces (Horváth and Kokoszka, 2012). In empirical studies we observe discrete noisy data, often appearing in unequally spaced time points. A first step in the analysis is to transform these data into

regular, smooth functions (see Horváth and Kokoszka, 2012, and references therein). Another important task that usually occurs during the pre-processing stage is the analysis of variability and splitting it into what is called *phase variability* and *amplitude variability*. The distinction among these two kinds of variability is a significant feature of FDA, as it does not have any counterpart in univariate or multivariate analysis (see Tarabelloni, 2017, and references therein). Referring to these two kinds of variability, one defines two types of outliers namely *shape outliers* and *magnitude outliers*, having in general different effect on the FDA procedures. The first taxonomy of functional outliers was proposed in Hubert et al. (2015). Although a formal and commonly accepted definition of functional outlier is still missing in the literature, a discrimination between amplitude and phase variability inspired the main and widely accepted distinction between the two kinds of outliers, i.e. magnitude and shape outliers. The first are related to amplitude, and are a direct analogue of the outlyingness concept in the multivariate context, while the second are related to phase variability, hence they are completely new and do not have a direct counterpart in classical statistics. The different nature of such outliers motivates researchers to propose various tools to detect and handle with them.

The most popular method designed for shape outliers detection is based on a concept of outliergram proposed in Arribas-Gil and Romo (2014) and intensively studied, for example, in Tarabelloni (2017). The method is rooted in a certain interesting property of the modified band depth of functional observations (López-Pintado and Romo, 2009). The best method of magnitude outliers detection is based on the functional boxplot proposed in Sun and Genton (2011), and further improved in Tarabelloni (2017). It is worth stressing that in practice a dataset may be affected by both magnitude and shape outliers, and a first step of an analysis is to separate them.

The boxplot, a well-known one-dimensional visualization technique, is used for detecting outliers when assuming normality. Under normality, within an interval [Q1-F*IQR;Q3+F*IQR], where Q1,Q3,IQR are correspondingly the first quartile, the third quartile and the interquartile range, and for F=1.5 we have about 0.99 probability mass. Points outside that interval are treated as outliers. Clearly, atypical observations can be either genuine but rare outcomes of the random process generating data, or can be corrupted data, due to a possible contamination of the sample. In a relation to this idea in a functional case, Sun and Genton (2011) proposed to assume a Gaussian process and choose F so that only fraction of 1% of the functions was flagged as outliers. Unfortunately, as in the functional case an analytic expression for F cannot be directly derived, a relevant resampling procedure

must be used. Since Gaussian functional data are far more complex than standard normal random variables, the procedure should be designed to take into account the first and the second moment of the dataset, i.e., the mean and covariance functions. Therefore, the adjustment process must be data-driven and has to be repeated for each dataset. Tarabelloni (2017) has recently proposed to use certain robust estimators and then to use bootstrap method for estimating appropriate quantiles and value for F. His method has been supported with a free roahd R package.

3 Our proposal using boxplot for Young inverse functions

Although the outliergram indicates well outliers belonging to a rich family of shape outliers, it lacks a sufficient precision of identification of type of departure of observation from a majority of observations, as it has been pointed out in Nagy et al. (2017). The sensitivity for types of outlyigness is especially important for developing new economic theories and applications, i.e., the outlying trajectory of a country development may be treated as incentive for developing a more general development theory that explains the root-cause of that trajectory. This drawback of the outliergram motivates us to propose an original method underlying a certain kind of shape "outlyingness in variation", frequently studied in economics (business cycle analysis). We aim at proposing a method, which is less computationally sophisticated than methods proposed in Nagy et al. (2017).

Let $\phi:[a,b]\to[c,d]$ be any real function (called a parent function). A function $\Phi:[c,d]\to[a,b]$ defined for any number $y\in[c,d]$ with a formula

$$\Phi(y)=\min\{x\in[a,b]:\phi(x)\geq y\}$$

we call generalized Young inverse function of function ϕ.

Let a functional data ϕ_j be defined on a finite number of knots, i.e.,

$$\phi_j:\{x_1,\ldots,x_{j_i}\}\to\{y_1,\ldots,y_{j_m}\}.$$

Therefore, an empirical generalized Young inverse function ϕ_j, namely, $\Phi_j:\left[\min\{y_1,\ldots,y_{j_m}\},\max\{y_1,\ldots,y_{j_m}\}\right]\to\{x_1,\ldots,x_{j_i}\}$ can be effectively defined for any real number $y\in\left[\min\{y_1,\ldots,y_{j_m}\},\max\{y_1,\ldots,y_{j_m}\}\right]$ with a formula:

$$\Phi(y)=\min\left\{x\in\{x_1,\ldots,x_{j_i}\}:\phi_j(x)\geq y\right\}.$$

Furthermore, we have a functional sample $Y^n=\{y_1,...,y_n\}$. We would like to show that a mapping of the following form

$$D(y \mid y_1,...,y_n)(z) = \{z \in [c,d] : \min_{i=1,...,n} y_i^{-1} \le y^{-1}(z) \le \max_{i=1,...,n} y_i^{-1}\},$$

where y^{-1} denotes generalized Young inverse function of function y, is a functional depth of a function y with respect to the sample Y^n.

An illustration of Young inverse functions is presented in Fig. 1-2.

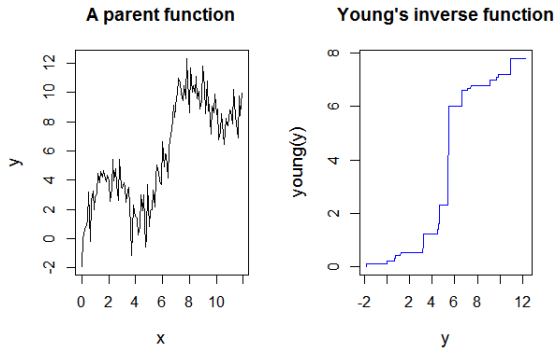

Fig. 1. An illustration of Young inverse function.

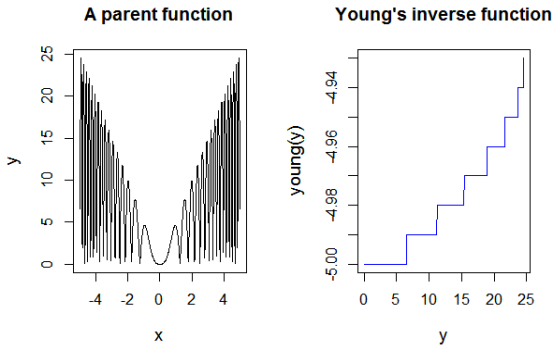

Fig. 2. An illustration of Young inverse function.

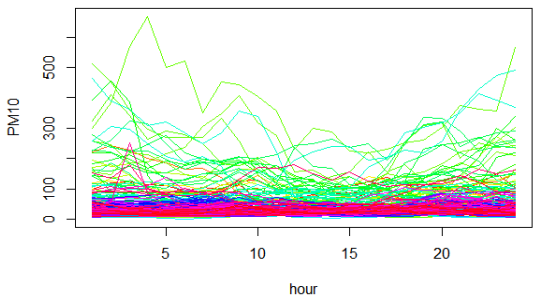

Fig. 3. Raw data on day and night air pollution with PM10 in Katowice in 2016.

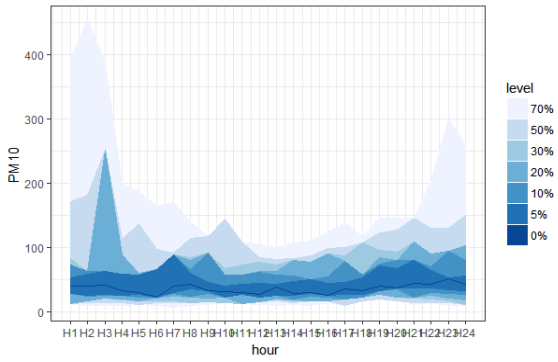

Fig. 4. Functional boxplot for air pollution data with PM10 in Katowice in 2016.

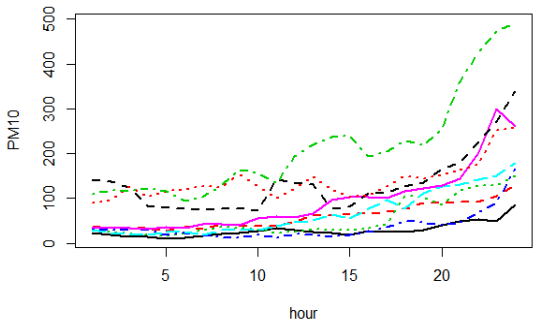

Fig. 5. Outlying day and night PM10 pollution curves indicated by Young inverse function method.

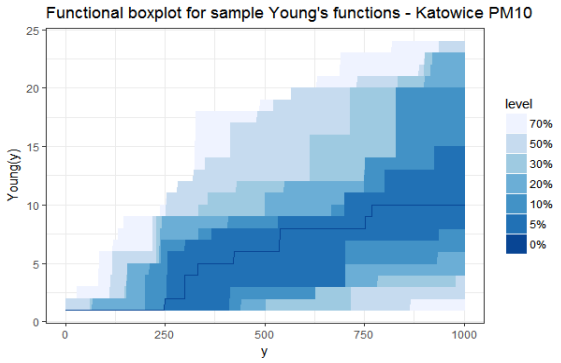

Fig. 6. Functional boxplot for Young inverse functions of air pollution curves in Katowice in 2016.

PROPOSAL: Let *MBD* denote a sample modified band depth and let *FM* denote a sample Frainman and Muniz depth (see López-Pintado and Romo, 2009). For a sample of functions $Y^n = \{y_1,...,y_n\}$ take the following steps (we fix thresholds of a_{MBD}, a_{Young}):

1. Verify the null hypothesis that Y^n has been generated by Gaussian law using statistic T_4 of Jarque-Berra type test for functional data proposed in Górecki et al. (2017). If the null hypothesis is rejected, then go to step 2, if it is not rejected, then go to step 3.

2. Calculate Tarabelloni (2017) version of the adjusted boxplot, and place outliers into a set O^T, indicating amplitude outliers.

3. Calculate the functional boxplot using the *MBD*, and place functions with $MBD < a_{MBD}$ into a set O^{MBD}, indicating amplitude outliers.

4. Calculate the outliergram, and place outliers into a set O^{AG}, indicating ”roughly” shape outliers.

5. Calculate the generalized Young inverse functions for the sample, i.e., $\left(Y^n\right)^{-1}$ and then calculate the functional boxplot for $\left(Y^n\right)^{-1}$. Place observations with *FM* depth $< a_{Young}$ into a set O^{Young}, indicating “outliers in variation”.

The final outliers set is $(Y^n \setminus (O^T \check{C} O^{MBD})) \text{Ç} O^{AG} \text{Ç} O^{Young}$.

Notice that one of the difficulties, when using a functional boxplot for outlier detection, is to indicate a number *c* such that the following formula is fulfilled: $P(MBD(y,Y^n) < c) = \alpha_{MBD}$,

where P denotes probability. In order to indicate a number c, we have to know P, which is a sampling distribution of functional depth and it is generally an unknown object.

4 Challenges in functional time series setup

Despite of many scientists efforts, robust statistics in area of time series analysis has much more gaps in comparison to uni- and multivariate statistics. The eighth chapter of the influential book by Maronna et al. (2006) may be treated as an introduction in this context. Outliers may cause bias in the model parameter estimates, and then, distort the size and power of statistical tests based on biased estimates. Secondly, outliers may increase the confidence intervals for the model parameters. Thirdly, as a consequence of the previous points, outliers strongly influence predictions. The best developed theory is available for ARIMA models. However, recent developments involve advances in coping with outliers when the series is generated by a general nonlinear models including as particular cases the bilinear model, the self-exciting threshold autoregressive (SETAR) model, the exponential autoregressive model, and the generalized autoregressive conditional heteroscedasticity (GARCH) models. Outliers in multivariate time series have been much less analysed than in the univariate case. For more recent researches related to our studies, let us only indicate that multivariate outliers were introduced in Tsay et al. (2000). Galeano et al. (2006) used projection pursuit methods to develop a procedure for detecting outliers, showing in particular, that testing for outliers in certain projection directions can be more powerful than testing the multivariate series directly. In view of these findings, an iterative procedure to detect and handle multiple outliers (based on a univariate search in these optimal directions) was proposed. The main advantage of this procedure is identification of outliers without prespecifying a vector ARMA model for the data. These ideas are in a close relation to the typical FDA framework, where due to infinite dimensional nature of the objects and richness of possible models, investigation of appropriate projections of data is a natural research tool. Note only, that even in case of outliers in an ARCH(1) model, we encounter two different scenarios, because an outlier can affect the level of a process as well as the volatility of the process. The case of a functional ARCH(1) model introduces the next level of complications.

In an analogy to the one dimensional setup (Maronna et al., 2006), one may introduce functional additive outliers (FAO). The FAO may have adverse influence on all the steps of the time series analysis. Functional innovative outliers (FIO) affect a single innovation, and correspond to an internal change of a single innovation of the time series and are often associated with isolated incidents. The effects of an FIO on a series are different depending on

whether the series is "stationary" or not (stationarity in the FTS setup is still an open issue). Maronna et al. (2006) introduced also functional level shift (FLS) which is a change in the level of the series, and the functional temporary change (FTC) which is an exponentially decreasing change in the level of the series. This shows a great number of possible functional outliers definitions.

The research challenge here is that the effect of a functional outlier not only depends on the model and the outlier size, as in the univariate case, on the interaction between the model and size as in multivariate case, but also on the outlier type. Additionally, as in the classical time series setup, the masking and swamping effects are still present, if a sequence of outlier patches is present in the functional time series (Galeano et al., 2006), i.e., a few outliers of the same magnitude appearing one after another (and forming a path).

5 Empirical study

The first application of functional outliers detection tools to air quality monitoring may be found in Febrero et al. (2008). In our empirical studies we, among others, have studied day and night air pollution with PM10 in Katowice in a period 01.09.2016 to 28.02.2017, using the data from the automatic measurement station located on Kossuth's street in Katowice and available from WIOŚ (http://powietrze.katowice.wios.gov.pl). The dataset consisted of 181 curves. Fig. 3 presents raw data on day and night air pollution with PM10 in Katowice in 2016. The same dataset has been exploited in the paper by Kosiorowski et al. (2018), where the authors have analyzed PM10 concentration in the air for five measurement stations in Silesian Region. The air pollution in the Silesian Region has been treated as a realization of a hierarchical functional random variable. Double functional median method has been used to compute a robust aggregate representing air pollution in the Silesian Region.

Fig. 4 presents the functional boxplot for air pollution data with PM10 in Katowice in 2016. Fig. 5 presents the outlying day and night PM10 pollution curves indicated by the generalized Young inverse function method and Fig. 6 shows the functional boxplot for the generalized Young inverse functions of air pollution curves in Katowice in 2016. Fig. 7 presents an adjusted functional boxplot, and air pollution data. Fig. 8 presents outliergram, and shape outliers indicated by it in the air pollution data. The adjusted functional boxplots and the outliergram were prepared using roahd R package (Tarabelloni, 2017), whereas functional boxplots for raw data and the generalized Young curves using DepthProc R package (Kosiorowski and Zawadzki, 2017).

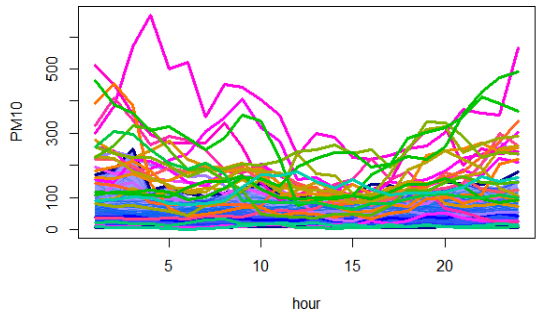

Fig. 7. Adjusted functional boxplot, air pollution data.

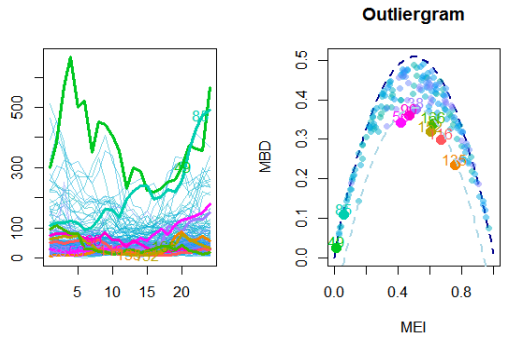

Fig. 8. Outliergram and outliers in air pollution data.

Conclusions

Reliable algorithms of functional outliers detection are desirable in a context of analysing various anomalies in economic phenomena described by functions. In the paper we have proposed a new functional outliers method and incorporated it into a decision procedure, which uses a known outlier detection apparatus i.e., the outliergram, the functional boxplot and the adjusted functional boxplot. Our approach is recommended for "outliers in variation" detection. Our future studies involve developing a theory for the proposed procedure as well as applications in optimization of a local pro-ecological policy.

Acknowledgements

DK thanks for the support related to CUE grant for the research resources preservation 2018. JPR and DM work was partially supported by the Faculty of Applied Mathematics AGH UST statutory tasks within subsidy of Ministry of Science and Higher Education, grant no. 11.11.420.004.

References

Arribas-Gil, A. & Romo J. (2014). Shape Outlier Detection and Visualization for Functional Data: the Outliergram. *Biostatistics*, *15*(4), 603-619.

Cuevas, A., Febrero, M. & Fraiman, R. (2006). On the Use of the Bootstrap for Estimating Functions with Functional Data. *Computational Statistics & Data Analysis*, *51*(2), 1063-1074.

Febrero, M., Galeano, P. & González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, *19*: 331-345.

Galeano, P., Peña, D. & Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, *101*(474), 654-669.

Górecki, T., Hörmann, S., Horváth, L. & Kokoszka, P. (2017). Testing Normality of Functional Time Series. *Journal of Time Series Analysis*. Doi:10.1111/jtsa.12281.

Horváth, L. & Kokoszka P. (2012). *Inference for Functional Data with Applications*. Springer Verlag, New York.

Hubert, M., Rousseeuw, P. & Segaert, P. (2015). Multivariate Functional Outlier Detection. *Statistical Methods and Applications*, *24*(2), 177-202.

Kosiorowski, D. & Zawadzki, Z. (2017). DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena, arXiv:1408.4542.

Kosiorowski, D., Mielczarek, D. & Rydlewski, J. P. (2018). Forecasting of a Hierarchical Functional Time Series on Example of Macromodel for the Day and Night Air Pollution in Silesia Region - A Critical Overview. *Central European Journal of Economic Modelling and Econometrics*, *10*(1), 26-46.

López-Pintado, S. & Romo J. (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, *104*(486), 718-734.

Maronna, R.A., Martin, R. D. & Yohai, V. J. (2006). *Robust Statistics - Theory and Methods.* John Wiley & Sons, Chichester.

Nagy, S., Gijbels I. & Hlubinka D. (2017). Depth-Based Recognition of Shape Outlying Functions. *Journal of Computational and Graphical Statistics*, *26*(4), 883-893.

Sun, Y. & Genton M. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316-334.

Tarabelloni, N., (2017). *Robust Statistical Methods in Functional Data Analysis*. Phd thesis and roahd R package, Politecnico di Milano.

Tsay, R.S., Peña, D. & Pankratz A.E. (2000). Outliers in Multivariate Time Series. *Biometrika*, *87*(4), 789-804.