ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 286, 2013

Daniel Kosiorowski[*], *Mateusz Bocian*[**],
Anna Węgrzynkiewicz[**], *Zygmunt Zawadzki*[**]

{DEPTHPROC} PACKAGE IN MULTIVARIATE TIME SERIES MINING

Abstract. In this paper we present our novel R package {*depthproc*} which implements several multivariate statistical procedures induced by statistical depth functions and we discuss some examples and applications of the package in data mining concerning the multivariate time series.

Key words: R package, Statistical depth function, robustness, multivariate time series

I. STATISTICAL DEPTH FUNCTIONS

A **data depth** is a way to measure the "depth" or "outlyingness" of a given point with respect to a multivariate data cloud or its underlying distribution. Given a cdf F on $\mathbf{R}^d$, a depth function $D(\mathbf{x}, F)$ provides an associated center-outward ordering of points $\mathbf{x}$ in $\mathbf{R}^d$. Statistical depth function compensates for lack of a linear order in $\mathbf{R}^d$, $d \geq 2$, by orienting points to a "center". Higher depth represents greater "centrality". This ordering allows us for a quantifying the many complex multivariate features of the underlying distribution, including location, quantiles, scale, skewness and kurtosis. For a sample $\mathbf{X}^n = \{\mathbf{x}_1, ., \mathbf{x}_n\}$, an expression $D(\mathbf{x}, \mathbf{X}^n)$ denotes a sample depth where distribution F is replaced by its sample counterpart F_n calculated on base of the sample $\mathbf{X}^n$ (for an overview see Serfling, 2006).

As an example of a **statistical depth function** let us recall **a symmetric projection depth** $D(\mathbf{x}, F)$ of a point $\mathbf{x} \in \mathbf{R}^d$ being a realization of some d dimensional random vector $\mathbf{X}$ with probability distribution F, defined as

[*] Ph.D., the work was partially supported by the by Polish National Center of Science Grant DEC-011/03/B/HS4/01138.

[**] Students of Cracow University of Economics (Informatics & Econometrics and Finance).

$$D(\mathbf{x},F)=\left[1+sup_{\|\mathbf{u}\|=1}\frac{|\mathbf{u}^T\mathbf{x}-Med(\mathbf{u}^T\mathbf{X})|}{MAD(\mathbf{u}^T\mathbf{X})}\right]^{-1}, \tag{1}$$

where *Med* denotes the univariate median, $MAD(Z) = Med(|Z-Med(Z)|)$, a sample version denoted by $D(\mathbf{x},F_n)$ or $D(\mathbf{x},\mathbf{X}^n)$ is obtained by replacing distribution F by its empirical counterpart F_n calculated on base of the sample $\mathbf{X}^n$.

The projection depth function possesses among others an affine invariance property, induced location and scatter estimators have high finite sample replacement breakdown points and good properties in terms of Hampel's influence function and Huber's maximum bias (for details see Serfling, 2006 and references therein).

For a sample $\mathbf{X}^n=\{\mathbf{x}_1,.,\mathbf{x}_n\}$ a set of points $D_\alpha(\mathbf{X}^n)=\{\mathbf{x}\in\mathbf{R}^d : D(\mathbf{x},\mathbf{X}^n)\geq\alpha\}$ is called $\alpha-$**central region**. Its border could be treated as an analogue of the univariate quantile. Figures 1-2 present sample projection depth calculated for two samples drawn from bivariate normal and mixture of two bivariate normal distribution correspondingly. Figures 1-2 were prepared by means of approximate algorithm proposed by Dyckerhoff (2004) implemented within our {**depthproc**} package free available via R-Forge server. We can define depth for vectors, matrices, functions, families of sets, geometrical objects (see Zuo and Serfling, 2006). Depth functions yield nested contours of equal outlyingness. Depth functions uniquely characterize a wide range of populations (see Kong and Zuo, 2010). For a general discussion of the depth concept see Serfling (2006) and references therein.

Figure 3 shows a relation between numbers of dwellings completed divided by the number of employed persons vs. number of employed persons in Polish voivodships in 2011 year. Figure 4 shows a relation between employment in thousands vs. GDP in Polish voivodships in 2009. Blue crosses in these figures represent mean vectors, orange stars two-dimensional Tukey medians. Two dimensional medians being placed in the most central regions differ from the mean vectors due to the existence of outlying observations. Figures 3-4 were prepared by means of {aplpack} R package using halfspace depth.

Our free available R package {depthproc} offers among other following procedures:

1. depthContour(X,method = "Projection",plot.title = paste(method,"depth"),...) – **2d sample depth contour plot**.

2. depthPersp(X,method = "Projection",plot.method = "rgl",xlim = extendrange(X[,1],f=0.1),...) – **3d sample depth perspective plot**.

3. *ddPlot(x, y=NULL, distribution = c("mvnorm", "t", "smvnorm", "st"), method = "Projection", scale = FALSE,...)* – **generalization of the quantile–quantile plot**.

4. *(ddmvnorm(x, method = "Projection", robust=FALSE,...)* – descriptive normality inspection.

5. *scalecurve(X, Y = NULL, alpha = seq(0,1,0.01), method = "Projection", draw = TRUE, nameX = "X", nameY)* - **nonparametric method for measuring a multivariate dispersion**.

6. asymmetrycurve<-function(X,Y = NULL, alpha = seq(0,1,0.01), method = "Projection", moving median = FALSE, draw = TRUE, nameX = "X", nameY = "Y",...) - **nonparametric method of measuring multivariate asymmetry.**

7. deepreg2d(...) – **robust regression**

8. trimprojreg2d(...) – **robust regression**

Our package depends on the following R packages {rgl}, {geometry}, {ggplot2}, {lattice}, {MNM}, {sn}, {MASS}, {robustbase}

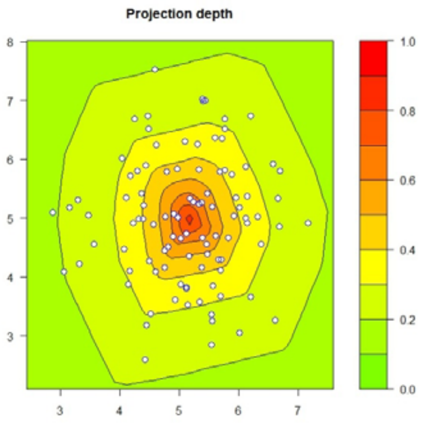

Fig. 1. Projection depth contour plot – 100 observations form 2d normal distribution

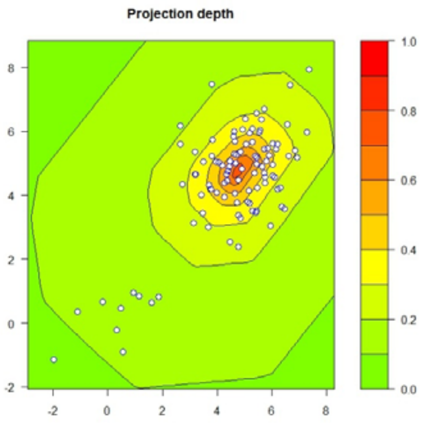

Fig. 2. Projection depth contour plot – 100 observations form a mixture of two 2d normal distributions

Source: Our own calculations – {depthproc 1.0} R package.

II. APPROXIMATE DEPTH CALCULATION IN {DEPTHPROC}

Direct calculation of the statistical depth function is generally a very challenging computational issue. Within the {depthproc} package we use approximate algorithm proposed by Dyckerhoff (2004) to calculation of a certain class of location depth functions (depths possessing so called strong projection property), direct algorithm proposed by Rousseeuw i Hubert (1998) for deepest regression estimator calculation and direct algorithm for Lopez-Pintado i Romo (2009) depth for functional data.

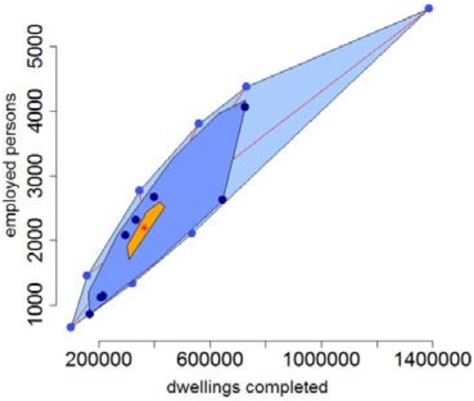

Fig. 3. Dwellings completed vs. the number of employed persons in Polish voividships in 2011 – 2D boxplot

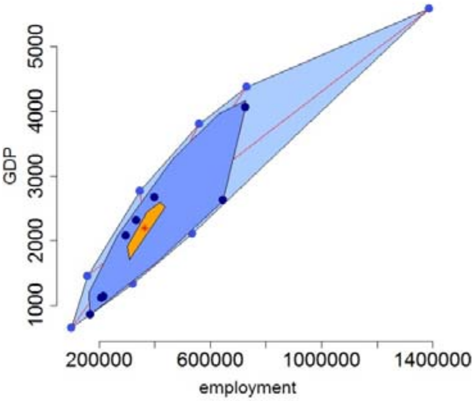

Fig. 4. Employment in thousands vs. GDP in Polish voivodships – 2D boxplot

Source: Our own calculations, data GUS, {aplpack} R package.

Let $D(\mathbf{x},\mathbf{Z}^n)$ denote sample depth in a point $\mathbf{x}\in\mathbf{R}^d$, $d\geq 2$, $\mathbf{Z}^n=\{\mathbf{z}_1,...,\mathbf{z}_n\}\subset\mathbf{R}^d$. Following Dyckerhoff (2004) we assume depth of the point $\mathbf{x}\in\mathbf{R}^d$ with respect to $\mathbf{Z}^n=\{\mathbf{z}_1,.,\mathbf{z}_n\}\subset\mathbf{R}^d$ equals minimum of a one-dimensional projection $\mathbf{u}^T\mathbf{x}$ with respect to $\mathbf{u}^T\mathbf{Z}^n=\{\mathbf{u}^T\mathbf{z}_1,.,\mathbf{u}^T\mathbf{z}_n\}\subset\mathbf{R}$, $\mathbf{u}^T\in\mathbf{R}^d$, $\|\mathbf{u}\|=1$, i.e. $D(\mathbf{x},\mathbf{X}^n)=\min\limits_{\|\mathbf{u}\|=1}D^1(\mathbf{u}^T\mathbf{x},\mathbf{u}^T\mathbf{X}^n)$. Let $D^1(y,Y^n)$ denote the one dimensional depth, $y\in\mathbf{R}$, $Y^n=\{y_1,...,y_n\}$, let $Q(\alpha)$ be a quantile of amount α. The above idea brings us to the following one dimensional depths leading to multidimensional depths using Dyckerhoff ideas:

1. *Simplicial depth*: $D^1(y,Y^n)=F_n(y)(1-F_n(y))$, where F_n- denotes sample cdf.

2. *Half space depth*: $D^1(y,Y^n)=\min\{F_n(y),1-F_n(y)\}$, where F_n- denotes sample cdf.

3. *Projection depth*:

$$D^1(y,Z^n)=\min_{\alpha}\left\{\alpha : y\in\left[Med(Z^n)-\frac{1-\alpha}{\alpha}MAD(Z^n),Med(Z^n)+\frac{1-\alpha}{\alpha}MAD(Z^n)\right]\right\}$$

where Med denotes median, MAD denotes median of absolute deviations from the median.

4. *Zonoid depth*: $D^1(y,Z^n)=\min_{\alpha}\left\{\alpha : y\in\left[\frac{1}{\beta}\sum_{\beta\le\alpha}Q(\beta),\frac{1}{\beta}\sum_{\beta\le\alpha}Q(1-\beta)\right]\right\}$.

For two probability distributions F and G, both in $\mathbf{R}^d$, we can **define depth vs. depth** plot being very useful generalization of the one dimensional quantile-quantile plot (see Li and Liu 2004:

$$DD(F,G)=\left\{\left(D(\mathbf{z},F),D(\mathbf{z},G)\right),\mathbf{z}\in\mathbf{R}^d\right\} \tag{2}$$

Its sample counterpart calculated for two samples $\mathbf{X}^n=\{X_1,.,X_n\}$ from F, and $\mathbf{Y}^m=\{Y_1,.,Y_m\}$ from G is defined as

$$DD(F_n,G_m)=\ \left\{\left(D(\mathbf{z},F_n),D(\mathbf{z},G_m)\right),\mathbf{z}\in\ \{\mathbf{X}^n\cup\mathbf{Y}^m\}\right\} \tag{3}$$

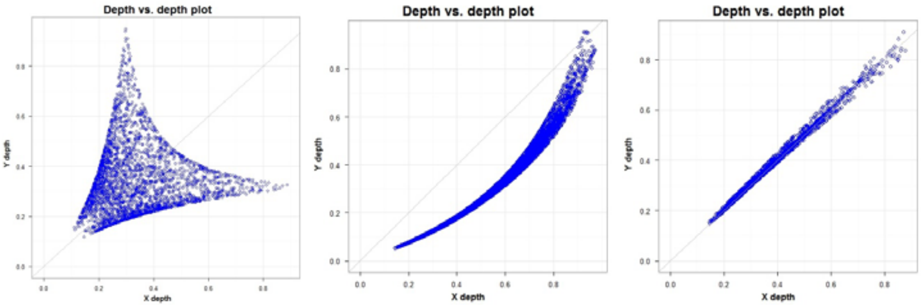

Fig. 5. Sample depth vs. depth plot for a difference in location (left), a difference in scale (middle) and the same distribution (right)
Source: Our own calculations – {depthproc 1.0} R package.

For sample depth function $D(\mathbf{x},\mathbf{Z}^n)$, $\mathbf{x}\in\mathbf{R}^d$, $d\geq 2$, $\mathbf{Z}^n=\{\mathbf{z}_1,.,\mathbf{z}_n\}\subset\mathbf{R}^d$, $D_\alpha(\mathbf{Z}^n)=\{\mathbf{z}\in\mathbf{R}^d : D(\mathbf{z},\mathbf{Z}^n)\geq\alpha\}$ $\alpha -$ central region, for $\alpha\in[0,1]$ we can define a **scale curve**

$$SC(\alpha)=\left(\alpha, vol(D_\alpha(\mathbf{Z}^n)\right)\subset\mathbf{R}^2, \tag{5}$$

and **asymmetry curve**

$$AC(\alpha)=\left(\alpha,\left\|c^{-1}(\{\overline{\mathbf{z}}-med\,|\,D_\alpha(\mathbf{Z}^n)\})\right\|\right)\subset\mathbf{R}^2, \tag{6}$$

being nonparametric scale and asymmetry functional correspondingly, where $c-$denotes constant, $\overline{\mathbf{z}}-$denotes mean vector, denotes multivariate median induced by depth function and $vol-$ denotes a volume.

III. EXAMPLES OF APPLICATIONS OF {DEPTHPROC} PACKAGE

Our {depthproc} package offers a variety of possibilities for a preliminary analysis of multivariate time series. We can among others prepare ***robust scatter diagrams*** time series value in a while *t* vs. its value in whiles *(t-1), (t-2)*,… - what can help us in a correct model specification. We can prepare ***moving depth vs. depth plot*** and monitor multivariate location, scatter, skewness of the considered process. We can prepare ***moving scale or asymmetry curve***. We can ***predict future values*** of the analyzed process by means of ***deepest regression*** applied to the moving window from the time series. We can consider robust filters, smoothing and by means of depth functions provided by {depthproc}.

In order to show usefulness of the selected statistical procedures offered by {depthproc} we simulated 3500 observations from a certain regular two dimensional vector autoregressive model VAR(1). We assumed the simulated data consist up to 5% of additive outliers generated from i.i.d normal distribution. Observations from number 1401 to 2450 are shifted with respect to assumed VAR(1) model. We considered inference process conducted on base of window from the series of length 500 observations. Observations from number 1 to 500 were treated as a reference sample. Figure 6 presents the results.

We considered two dimensional empirical data set consisted of opening and closing points values for WIG20 index from 01.01.2009 to 30.06.2012 (861 observations). We compared consecutive six month periods. We treated first period as a reference sample. Figure 7 presents the calculated depth vs. depth plots. The plots indicate significant differences in locations of the half year windows.

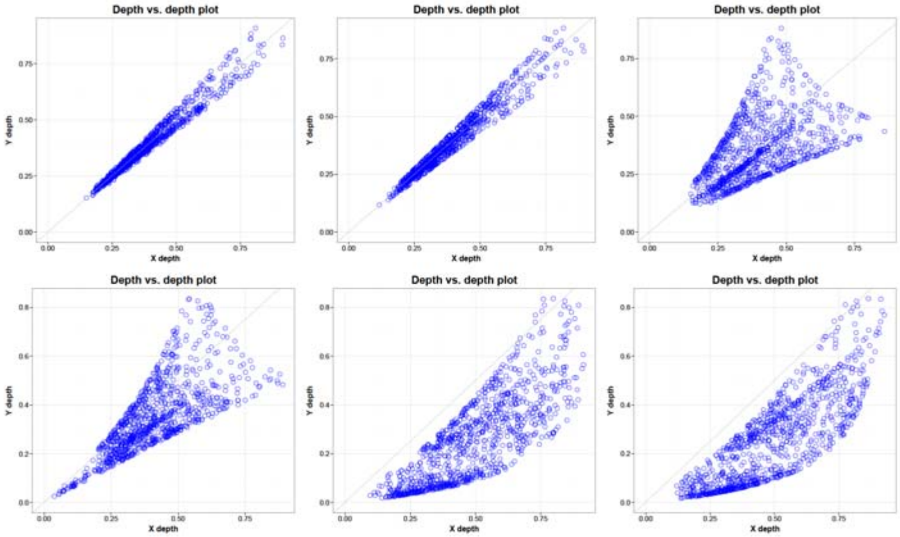

Fig. 6. Sample depth vs. depth plot for the simulated data from VAR(1) process. Plots were prepared on base of windows of length 500 observations from the series
Source: Our own calculations – {depthproc 1.0} R package.

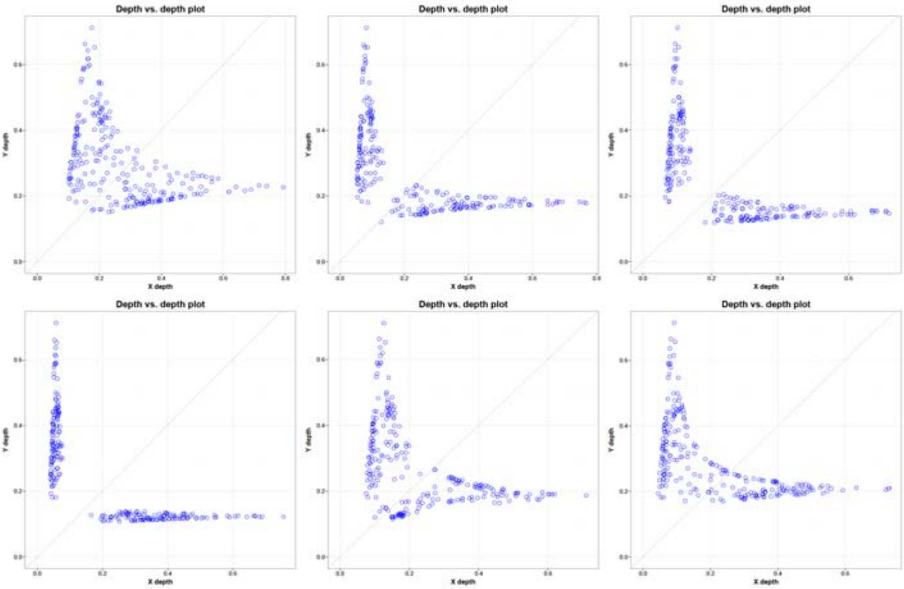

Fig. 7. Depth vs. depth plots for the WIG 20 index considered wrt opening and closing values from 01.01.2009 to 30.06.2012 year (861 observations). We compared consecutive six month periods
Source: Our own calculations – {depthproc 1.0} R package.

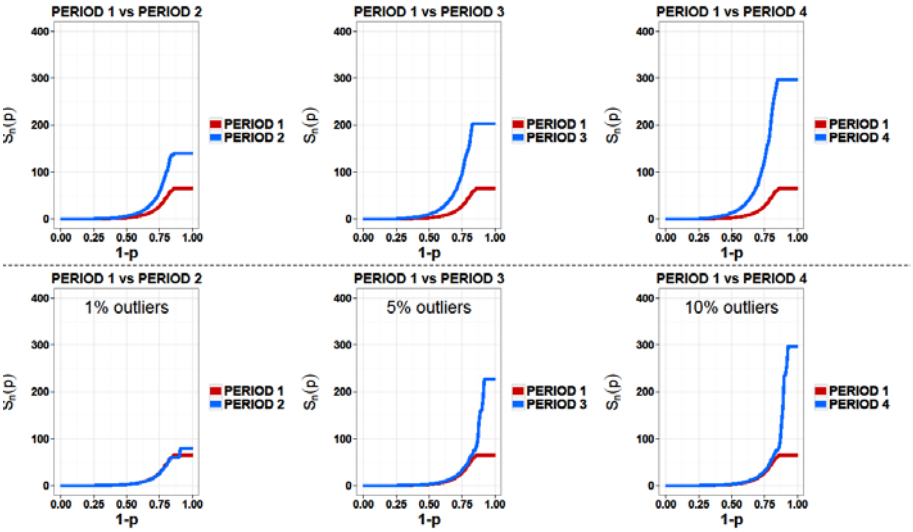

Fig. 8. Sample scale curves for two dimensional time series simulated from VAR(1) process. Samples without outliers and samples with up to 1%, 5% and 10% of additive outliers
Source: Our own calculations – {depthproc 1.0} R package.

Figure 8 presents sample scale curves calculated on base of data windows from the simulated VAR(1) process with scale shifts and with up to 10% of the additive outliers. Figure 9 presents sample asymmetry curves calculated on base of two dimensional data sets simulated from two-dimensional normal and two-dimensional T distributions with several parameters of skewness.

IV. SUMMARY AND CONCLUSIONS

We presented selected functions of R package {depthproc} which is freely available with detailed description under the address: https://r-forge.r-project.org/projects/depthproc/

Our package is still developing and in our opinion in the future will find several interesting applications in the robust economic analysis.

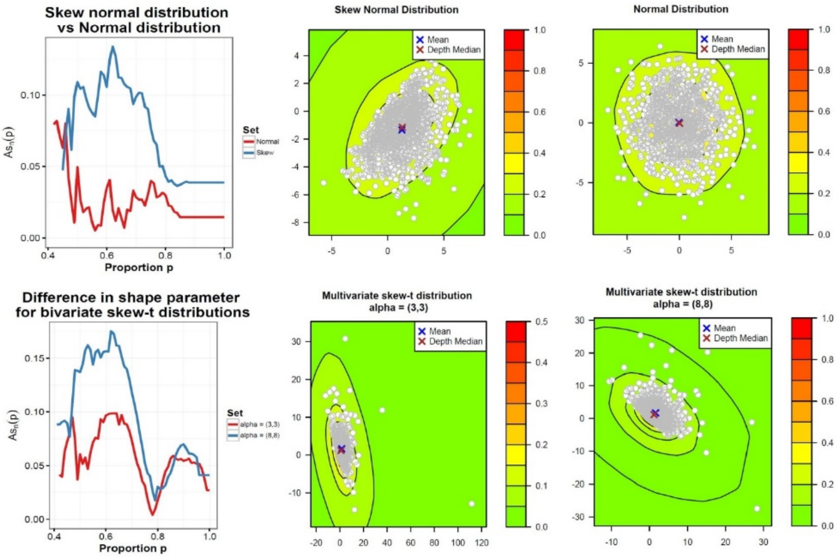

Fig. 9. Sample asymmetry curves for two dimensional data sets simulated from two-dimensional normal and two-dimensional T distributions with several parameters of skewness
Source: Our own calculations – {depthproc 1.0} R package.

REFERENCES

Dyckerhoff, R. (2004), Data depths satisfying the projection property. *Allgemeines Statistisches Archiv*. 88, 163-190.

Li, J., Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, bf 19(4), 686-696.

Kosiorowski, D. (2012), Student depth in robust economic data stream analysis, Colubi A.(Ed.) Proceedings of COMPSTAT'2012, The International Statistical Institute/International Association for Statistical Computing.

Maronna, R. A., Martin, R. D., Yohai, V. J. (2006), Robust statistics - theory and methods. Chichester: John Wiley & Sons.

Rousseeuw, P. J., Hubert, M. (1999), Regression depth, *Journal of the American Statistical Association*, 94, 388-433.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference, In: Liu R.Y., Serfling R., Souvaine D. L. (Eds.): *Series in Discrete Mathematics and Theoretical Computer Science*, AMS, 72, 1-15.

Daniel Kosiorowski, *Mateusz Bocian*, *Anna Węgrzynkiewicz· Zygmunt Zawadzki*

PAKIET {DEPTHPROC} W EKSPLORACYJNEJ ANALIZIE WIELOWYMIAROWEGO SZEREGU CZASOWEGO

W artykule przedstawiamy pakiet środowiska R naszego autorstwa o nazwie {*DepthProc*}. Pakiet zawiera implementacje kilku wielowymiarowych procedur statystycznych indukowanych przez statystyczne funkcje głębi. Przedstawiamy przykłady zastosowań pakietu w eksploracyjnej analizie wielowymiarowego szeregu czasowego.