INFORMATICA, 2016, Vol. 27, No. 3, 649–672

DOI: http://dx.doi.org/10.15388/Informatica.2016.104

Comparing Dissimilarity Measures: A Case of Banking Ratios

Laurynas NARUŠEVIČIUS*, Alfredas RAČKAUSKAS
Faculty of Mathematics and Informatics, Vilnius University
Naugarduko st. 24, LT-03225, Vilnius, Lithuania
e-mail: laurynas.narusevicius@mif.vu.lt

Received: January 2016; accepted: July 2016

Abstract. The aim of this paper is twofold. Firstly, to discuss a clustering of a given set of the European banks into groups based on their performance during 1999–2013. Secondly, to compare different dissimilarity measures and to determine which of them suits best for clustering banking ratios. Six ratios that reveal profitability, efficiency, stability and loan portfolio quality of the banks were used. The similarity/dissimilarity between banks was estimated using measures that are based on time series or functional data properties. Two dissimilarity measures that are not commonly used in the literature are proposed and two measures are extended from univariate into multivariate case. The results of our study show that there is no dissimilarity measure which would provide the best clustering results for all ratios. However, dissimilarity measures based on functional data properties in many cases outperfomed measures based on time series properties. The choice of the number of clusters is not that clear. According to different banking ratios, it is found that banks could be grouped into 6–12 clusters.

Key words: banking ratios, dissimilarity, time series clustering, functional data clustering, clustering comparison.

1. Introduction

In recent years cluster analysis, aiming to discover group structures among a set of observations, gains much popularity in the literature. Partitioning of the time series data helps to detect characteristic patterns, to forecast future performance, etc. The methods used in the cluster analysis can be divided into three categories: the methods based on (1) similarity of raw data; (2) features extracted from raw data and (3) models build from raw data. Recall that a measure D of dissimilarity (or equivalently similarity) of objects X and Y is symmetric: $D(X, Y) = D(Y, X)$, non-negative $D(X, Y) \geqslant 0$, and such that $D(X, X) = 0$. Similarity measure can be, but not necessarily is, a metric, i.e. $D(X, Z) \leqslant D(X, Y) + D(Y, Z)$. One of our goals of this paper is to consider various dissimilarity measures and apply them to the data under investigation.

After the Global financial crisis in 2007–2008 the financial sector and especially banks gained much attention. Authorities which are responsible for safeguarding the stability of

*Corresponding author.

the banking system, realized that microprudential supervision, where focus is on an individual institution, is not enough to maintain banking sector stable. Therefore, macroprudential supervision, which focused on the banking system as a whole, began to be actively used. A large number of the macro-level instruments were introduced and those are applied to all banks. However, the banking sector is heterogeneous and some tools could be ineffective to some banks. It would be useful to find groups of banks which have similar characteristics and design or calibrate some macroprudential instruments that would become appropriate for that group. Therefore, our goal is to discuss a clustering of the banks.

We exploited distance measures based on time series as well as on functional data properties. In addition to univariate clustering, where banks are grouped into clusters according to one bank-specific ratio, we applied multivariate clustering, where banks are clustered based on their several ratios. Since in cluster analysis data are unlabelled, a related issue is to find appropriate number of clusters that are the most proper for the data. The resulting clusters should not only have good statistical properties, but also give results that are, in our case, economically explainable.

In our study we used six ratios that reflect banks' profitability (return on average assets, return on average equity, net interest margin), efficiency (cost to income), stability (capital adequacy ratio) and portfolio credit risk (loan losses over loan portfolio). We applied twelve different dissimilarity measures. Ten of these measures are commonly used. We proposed two new distance measures, based on functional data properties, that, to our knowledge, were not used in the clustering literature. Furthermore, we extended two univariate distance measures to multivariate case. The results of the univariate clustering show that there is no dissimilarity measure which would be the best to all ratios. However, in many cases clustering methods based on functional data properties outperfomed distance measures based on time series properties. Estimation also shows that simple Euclidean distance is a relatively good distance measure for clustering banking data. Average silhouette width mostly suggested small number of clusters, usually the highest value was when banks were divided into 2 groups, and sometimes into 4 clusters. Other clustering validity indices give mixed results, especially if the profitability ratios are considered. If one considers bigger number of clusters, e.g. 20 clusters, then the results of the clustering show that there are 6–12 larger groups and other clusters are formed from few banks only. Clusters with low number of banks could be treated as outliers. The results of the multivariate clustering revealed that it is important to take into account not only how close banks' ratios are, but how similarly they change over the years.

The remainder of this paper is structured as follows. Section 2 gives a brief literature review related to the time series cluster analysis. Section 3 describes data and time series dissimilarity measures. Section 4 and Section 5 introduce functional data and multivariate clustering methodology, respectively. Section 6 describes clustering algorithm and clustering validation indices that were used in this study. Section 7 gives clustering results and their interpretation. Finally, Section 8 summarizes main findings.

2. Review of the Related Literature

There has been an increased interest in time series clustering as more time dependent data in various fields became available. The results of cluster analysis depend on many choices which must be fixed during the clustering process. In general, cluster analysis consists of few basic steps (Halkidi *et al.*, 2001). The first step is to select the features of time series on which clustering is going to be performed. The features should contain all possible information related to the task of interest. The second step is to define a dissimilarity measure between time series. A dissimilarity or distance measure quantifies and compares similarities of two time series. The next step is to choose the clustering algorithm which groups data into clusters. Since the precise number of clusters is not known a priori, clustering results must be evaluated using appropriate criterion. The final step is the interpretation of results. Expert judgement is also important when drawing the conclusion of cluster analysis.

One of the key elements in cluster analysis is determining an appropriate dissimilarity/similarity measure between two time series. Since time series are of dynamic character, the concept of similarity is complex. The two most widely used dissimilarity measures work with raw data. Conventional Euclidean distance measures the distance between two time series at each point in time. According to dynamic time warping distance (Berndt and Clifford, 1994), two time series are close if there exists a mapping, expressing a time distortion by a deceleration or acceleration so that the maximum length between all coupled observations is minimized. However, these two dissimilarity measures do not take into account the growth behaviour of the time series. Chouakria and Nagabhushan (2007) proposed dissimilarity measure which accounts for both closeness of values and behaviour of time series.

Given that time series usually are high dimensional data which could be noisy, various methods are used to extract some features of data. Dissimilarity is then measured based on these features. Some distance measures take into account the properties of time series such as correlation (Golay *et al.*, 2004), autocorrelation (Bohte *et al.*, 1980) or partial autocorrelation. Other distance measures proposed in the literature transform raw data and then estimate closeness based on transformed data. Chan and Fu (1999) among others used discrete wavelet transform, Faloutsos *et al.* (1994) employed discrete Fourier transform, Keogh *et al.* (2001) proposed piecewise aggregate approximation, Lin *et al.* (2003) introduced symbolic aggregate approximation. Many other representations are also used in the literature.

A different approach, which is used in the time series clustering literature, is to assume that time series are generated from particular parametric model. For example, Piccolo (1990) defined a distance measure in the class of invertible ARIMA processes as the Euclidean distance between the AR(∞) operators approximating the corresponding ARIMA structures. For the class of invertible and stationary ARMA processes, Maharaj (1996) proposed measure based on hypothesis testing to determine whether data generating processes significantly differ between two time series. Another group of dissimilarity measures are based on comparing levels of complexity of time series.

This category of distances includes normalized compression distance, proposed by Li *et al.* (2004) and complexity-invariant dissimilarity measure introduced by Batista *et al.* (2011, 2014).

Many authors working on time series clustering or comparing different dissimilarity measures make an assumption that time is discrete (e.g. Ding *et al.*, 2008). However, there is another field of research in which dissimilarity is measured using functional data or their properties. There are several approaches which are used in the literature. Filtering approach consists of a first step in which curves are expanded into a some finite basis of functions and a second step in which clustering is performed using the basis expansion coefficients. For example, Abraham *et al.* (2003) considered B-splines and Peng and Müller (2008) used principal component scores. Adaptive methods perform simultaneously dimensionality reduction and clustering as they consider that the functional form of data depends on clusters. James and Sugar (2003) assumed that the basis expansion coefficients of the curves into a spline basis are distributed according to a mixture of Gaussian distribution with different mean for each cluster and common variance. Samé *et al.* (2011) assumed that the curves come from a mixture of regressions on a basis of polynomial functions, with possible changes in regime. Another approach considers dissimilarity or distance between curves. Examples of this method could be found in Ferraty and Vieu (2006) and Ieva *et al.* (2013). Meanwhile, Jacques and Preda (2014) provide a good survey on methods used for the functional data clustering.

Once the initial distance matrix is computed, a clustering algorithm can be used to divide data into clusters. There are many different clustering algorithms that are used to cluster time series. The literature provides several categories of algorithms and methods in each category. The clustering is crisp if each element belongs to only one cluster, or the partition is fuzzy if one element could be in more than one cluster to a different degree. A popular category of crisp clustering are partitional algorithms. This category includes methods like *k-means* (MacQueen, 1967), where mean of the elements in the cluster represents each cluster, and partitioning around metoids (*PAM*) (Ng and Han, 1994), where the most centrally located element in a cluster represents each cluster. The similar methods of fuzzy clustering are the fuzzy *c-means* (Bezdek *et al.*, 1984), modified fuzzy *c-means* (Hemanth *et al.*, 2016) and fuzzy *c-metoids* (Krishnapuram *et al.*, 1993). A second commonly used category is hierarchical clustering. There are two types of hierarchical algorithms: agglomerative where each element is placed in its own cluster and then elements are merged to form larger clusters until there is one cluster, and divisive method, which works in opposite direction. Other category of clustering algorithms is density-based algorithms, which include methods like *DBSCAN* (Ester *et al.*, 1996). In this clustering algorithm cluster is extended as long as the density (number of elements) in the neighbourhood exceeds some threshold. The main idea of grid-based clustering algorithms (methods like *STING*, Wang *et al.*, 1997) is to quantize the element space into a finite number of cells that form a grid structure on which clustering operations are performed.

Since clustering algorithms divide unlabelled data into significant groups, it is important to evaluate clustering results and find partitioning that fits data the best. There are three basic criteria on which clustering evaluation is usually performed. First is the compactness of a cluster which should be minimized, i.e. the members of each cluster should

be as close to each other as possible. The second is the connectedness of a cluster, i.e. to what extent elements are placed in the same cluster as their nearest neighbours. The third is the separation of clusters which should be maximized, i.e. the clusters should be widely spaced. There are many validity assessment methods provided in the literarure, which combine measures mentioned above. An example of validity indices are the Dunn index (Dunn, 1974), Davies and Bouldin Index (Davies and Bouldin, 1979) or Silhouette Width (Rousseeuw, 1986; Kaufman and Rousseeuw, 1990).

Time series clustering problems arise in a wide range of fields, including business and economics, physics, medicine, meteorology and many others. Liao (2005) provided a good survey on time series clustering which includes many references. Meanwhile, an interesting overview on recent time series data mining methods and algorithms can be seen in Fu (2011). As was pointed out by Liao (2005), there are not so many studies comparing different time series dissimilarity measures. Few examples of papers that compared several distance measures are works by Ding *et al.* (2008) and Díaz and Vilar (2010).

To summarize, existing literature provides many choices of methods or algorithms in each step of the cluster analysis. Some methods may produce good clustering results in one instance, while in other cases a different method would do better. So the choice of a particular method may influence final results significantly. Therefore, expert judgement is also an important step in conducting the clustering of a given dataset.

3. Time Series Clustering Methodology

In this section we present time series data under consideration and describe some measures of dissimilarities of time series that are used to cluster the data.

3.1. *Data*

Six bank-specific variables were taken in our clustering exercise. We included three profitability measures: return on average assets (ROAA), return on average equity (ROAE), net interest margin (NIM); operational efficiency measure – cost to income ratio (CIR); credit quality measure – loan loss provisions over total gross loans (LLP) and bank riskiness measure – total capital ratio (CAR). These bank-specific measures are the main variables, which describe situation in the banking sector.

In this study we used annual unconsolidated bank accounts data covering time period from 1999 to 2013. Dataset is obtained from Bureau van Dijk *Bankscope* database and includes all commercial, savings and cooperative banks from European Union countries. These institutional bank types are mainly focused on financial intermediation. Therefore, we do not include data from investment banks or other bank types as their business model is essentially different from commercial, savings and cooperative banks. The preliminary sample consists of six bank-specific variables from 2800 banks.

We needed to edit our data in the following ways. First of all we excluded all banks with missing data entries, i.e. we left only those banks which had complete data for the years 1999–2013. Secondly, we excluded banks that had extreme values or large unexplained

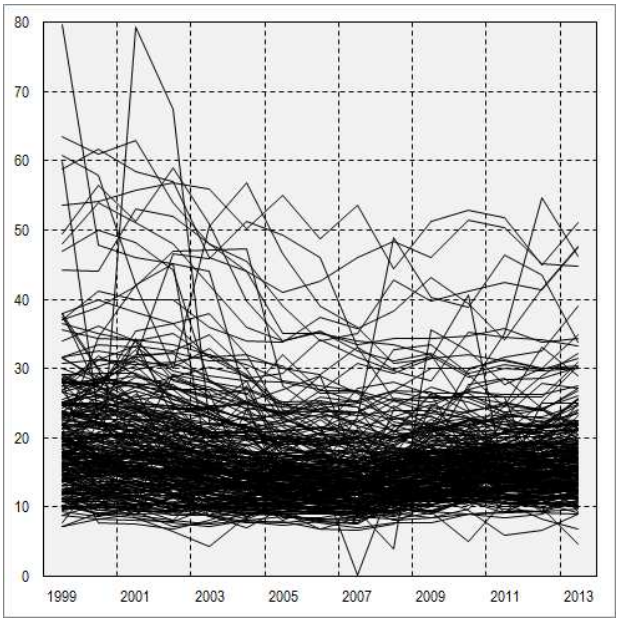

Fig. 1. Capital adequacy ratio.

shifts in the values of variables. The final dataset varied from 260 banks for capital adequacy ratio to 1332 banks for ROAA variable. For example, Fig. 1 shows our sample of capital adequacy ratio.

The return on average assets usually is used as the main bank profitability variable. The ROAA is calculated as the ratio of net profits over average total assets. This ratio shows bank's ability to generate profits from all activities related to their assets. Average assets are used to calculate ratio, because they help to capture any changes in assets that occured during the fiscal year. Golin (2001) describes ROAA as the key measure to evaluate bank's profitability.

The second measure of profitability is the return on average equity. ROAE is the net profits expressed as percentage of average equity. This ratio gives information about the return to shareholders on the equity. Banks usually report both ROAA and ROAE ratios to indicate their profitability. The main difference between these two ratios is that ROAE does not take into account the risk that is associated with higher leverage. Thus banks with a higher equity (lower leverage ratio) generally report a lower ROAE but a higher ROAA.

The third measure of profitability is the net interest margin. The NIM ratio is defined as interest income minus interest expense divided by the average of interest-bearing assets. This ratio is narrower than the other two profitability measures as it only focuses on the profit earned on interest rate related activities.

In our study we also include cost to income ratio, which shows the efficiency of the bank performance. The CIR is calculated as operating costs divided by total generated revenue. This ratio is a measure of bank's ability to turn resources into revenue. Changes

Table 1
Descriptive statistics.

	Mean	Median	Standard deviation	Min	Max
ROAA	0.41	0.29	0.76	−18.8	11.21
ROAE	5.10	4.46	6.77	−91.70	95.91
Net interest margin	2.74	2.65	0.98	−0.70	15.68
Cost to income ratio	67.80	68.25	12.43	5.26	186.36
Loan loss provisions/gross loans	0.71	0.60	1.09	−7.69	17.62
Capital adequacy ratio	17.62	15.67	7.88	0.13	79.60

in CIR can highlight potential problems: if costs are rising at a higher rate than income, CIR will rise from one period to the next.

Credit portfolio quality is an important aspect of overall bank performance. Therefore we analyse loan loss provisions over total gross loans ratio. The loan loss provisions are taken from a bank's income statement. A higher LLP ratio indicates problems in credit portfolio and also potential problems for bank's stability.

Capital adequacy ratio is defined as bank's total capital expressed as a percentage of its risk-weighted assets. CAR determines the capacity of the bank to meet potential losses from credit risk, market risk, operational risk and others. This ratio ensures that the banks do not expand their business without having adequate capital. Capital adequacy ratio helps to measure the riskiness on the banking sector. A higher CAR implies a more stable banking system.

Descriptive statistics of the data are presented in Table 1. The descriptive statistics indicates the need to cluster banks as data show high standard deviations compared to mean values, e.g. ROAA, ROAE. Cost to income ratio has min and max values greater than 3σ, which also motivates to separate banks into several groups.

The data under investigation have the form:

$$\boldsymbol{x}_t^{(i)} = \left(x_{1,t}^{(i)}, \ldots, x_{d,t}^{(i)}\right), \quad t = 1, \ldots, T;\ i = 1, \ldots, N.$$

Here index $i = 1, \ldots, N$ corresponds to a bank whereas index t corresponds to time (years in our case) and d corresponds to a bank-specific ratio. Since not all banks in our data set have all six ratios, we will mostly consider univariate clustering, i.e. we will cluster banks according to each ratio separately. In addition, we will take profitability and efficiency ratios and consider multivariate clustering based on four ratios.

3.2. *Dissimilarity Based Time Series Clustering*

As was pointed out by Liao (2005) and Batista *et al.* (2014), dissimilarity measure between two time series is one key choice in clustering that has to be made. The choice of distance measure is more important than the choice of clustering algorithm. In this section, we review six dissimilarity measures used in time series clustering studies.

Euclidean distance

In general, any metric of the finite dimensional Euclidean space could be used as a measure of dissimilarities of two time series. In this research we used conventional Euclidean

distance based measure as a starting method for clustering. Ding *et al.* (2008) showed that simple Euclidean distance could outperform other dissimilarity measures in many cases.

Suppose that $x = (x_1, \ldots, x_T)$ represents values of some ratio of the bank i and $y = (y_1, \ldots, y_T)$ represents values of some ratio of the bank j ($i, j = 1, \ldots, N$ and $i \neq j$). In our study $t = 1, \ldots, 15$. Euclidean distance is then described as:

$$D_{\lambda,\mathrm{EUCL}}(x, y) = \left(\sum_{t=1}^{T} \big((x_t - \lambda x_{t-1}) - (y_t - \lambda y_{t-1}) \big)^2 \right)^{1/2}, \tag{1}$$

where λ is weighting parameter. A classical approach is to take $\lambda = 0$. Then the proximity depends on the closeness of the values at the corresponding point of time. However, the distance $D_{\mathrm{EUCL}}(x, y) := D_{0,\mathrm{EUCL}}(x, y)$ does not take into account the growth rates of the vectors x and y. Therefore, we also considered dissimilarity with $\lambda = 1$. To be more precise, we applied $D_{\Delta,\mathrm{EUCL}} = D_{0,\mathrm{EUCL}}(x, y) + D_{1,\mathrm{EUCL}}(x, y)$ in this study.

Adaptive dissimilarity index

Chouakria and Nagabhushan (2007) introduced dissimilarity index, which is based on an adaptive tuning function and addressed to cover both behaviour and values proximity measures. They used first order temporal correlation coefficient to evaluate the proximity between the dynamic behaviour of the series. This coefficient is defined as follows:

$$CORT(x, y) = \frac{\sum_{t=1}^{T-1} (x_{t+1} - x_t)(y_{t+1} - y_t)}{\sqrt{\sum_{t=1}^{T-1} (x_{t+1} - x_t)^2} \sqrt{\sum_{t=1}^{T-1} (y_{t+1} - y_t)^2}}.$$

Temporal correlation coefficient belongs to the interval $[-1, 1]$. The value $CORT(x, y) = 1$ means that the series x and y at any time point show a similar dynamic behaviour, i.e. series decrease or increase with a similar growth rate and direction (similar behaviour). The value $CORT(x, y) = -1$ means that both series have a similar growth rate but direction is opposite (opposite behaviour). The value $CORT(x, y) = 0$ implies that growth rates are stochastically linearly independent and there is no monotonicity between series x and y (different behaviour). The proximity of the values of two time series $D_{\mathrm{EUCL}}(x, y)$ is estimated using Euclidean distance.

Dissimilarity index proposed by Chouakria and Nagabhushan (2007) automatically modulates the proximity of the values according to the proximity of the behaviour. This index is defined by:

$$D_{\mathrm{CORT}}(x, y) = \phi_k\big[CORT(x, y)\big] \cdot D_{\mathrm{EUCL}}(x, y), \tag{2}$$

where $\phi(u)$ is exponential adaptive tuning function:

$$\phi_k(u) = \frac{2}{1 + e^{ku}}, \quad k \geqslant 0.$$

An adaptive tuning function decreases the weight of the proximity between values when the temporal correlation increases from 0 to 1. And it works viceversa when correlation decreases from 0 to -1. In case of $CORT(x, y) = 0$, i.e. time series show different

behaviour, the dissimilarity index is approximately equal to the value of $D_{\mathrm{EUCL}}(x, y)$. The parameter k modulates the contribution of the temporal correlation and Euclidean distance to the dissimilarity index $D_{\mathrm{CORT}}(x, y)$.

A complexity-invariant distance measure

Batista *et al.* (2011, 2014) proposed dissimilarity index, which use information about complexity difference between time series x and y. Authors argued that many dissimilarity measures tend to place more complex pairs of time series further apart than pairs of simple series. A complexity-invariant dissimilarity measure $D_{\mathrm{CID}}(x, y)$ is defined as follows:

$$D_{\mathrm{CID}}(x, y) = CF(x, y) \cdot D_{\mathrm{EUCL}}(x, y), \tag{3}$$

where $CF(x, y)$ is complexity correction factor:

$$CF(x, y) = \frac{\max\{CE(x), CE(y)\}}{\min\{CE(x), CE(y)\}}, \quad CE(x) = \sqrt{\sum_{t=1}^{T-1}(x_{t+1} - x_t)^2}.$$

The complexity correction factor increases the distance between two time series if there is complexity difference between them. Furthermore, if time series have similar complexity, then the distance is approximately equal to $D_{\mathrm{EUCL}}(x, y)$.

The main idea of Batista *et al.* (2011, 2014) is that if a time series is stretched to become a straight line, then more complex time series would result in a longer line. Dissimilarity index $D_{\mathrm{CID}}(x, y)$ is parameter-free, simple and increased accuracy of clustering in several experiments accomplished by Batista *et al.* (2011).

Autocorrelation based distance

Bohte *et al.* (1980), Geleano and Peña (2000) and several other authors used estimated autocorrelation function to measure the distance between two time series. Suppose that $\hat{\rho}_x = (\hat{\rho}_{1,x}, \ldots, \hat{\rho}_{L,x})'$ and $\hat{\rho}_y = (\hat{\rho}_{1,y}, \ldots, \hat{\rho}_{L,y})'$ are the estimated autocorrelation vectors of x and y. Here L is such that $\hat{\rho}_{i,x} \approx 0$ and $\hat{\rho}_{i,y} \approx 0$ when $i > L$. A dissimilarity between two univariate time series can be measured by:

$$D_{\mathrm{ACF}}(x, y) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)' \Omega (\hat{\rho}_x - \hat{\rho}_y)},$$

where Ω is a weighting matrix.

If we take $\Omega = I$, i.e. uniform weights, then $D_{\mathrm{ACF}}(x, y)$ is Euclidean distance between the estimated autocorrelation functions:

$$D_{\mathrm{ACFE}}(x, y) = \sqrt{\sum_{i=1}^{L}(\hat{\rho}_{i,x} - \hat{\rho}_{i,y})^2}. \tag{4}$$

Dynamic time warping distance

Berndt and Clifford (1994) proposed dynamic time warping (DTW) to find patterns in time series. This distance measure is popular and widely used in time series clustering

literature. Let N be the set of all possible sequences of n pairs preserving the observations order in the form:

$$r = \big((x_{a_1}, y_{b_1}), \ldots, (x_{a_n}, y_{b_n})\big), \tag{5}$$

where $a_i, b_j \in \{1, \ldots, T\}$ such that $a_1 = b_1 = 1$ and $a_n = b_n = T$, and $a_{i+1} = a_i$ or $a_i + 1$ and $b_{i+1} = b_i$ or $b_i + 1$, for $i \in \{1, \ldots, n-1\}$. Then dynamic time warping distance is defined by:

$$D_{\mathrm{DTW}}(x, y) = \min_{r \in N} \left(\sum_{i=1,\ldots,n} |x_{a_i} - y_{b_i}| \right). \tag{6}$$

Dynamic time warping dissimilarity measure allows time series to be stretched or compressed to recognize similar shapes.

4. Functional Data Clustering Methodology

In this section the data are considered as observations of curves, i.e. the random variables underlying data are countinuous time stochastic processes. To cluster the curves we apply a non-parametric method which uses a specific distance or dissimilarities between functions. Besides widely used in the literature measures of dissimilarities of functions such as Hausdorff distance, L_2-distance or distance based on functional principal components, we consider also a class of Hölder distances that take into account a certain type of growth rates of curves. It is shown that this type of distances in some cases performs better compared with others.

4.1. *Functional Data*

We assume that the data under investigation $\boldsymbol{x}_t^{(i)} = (x_{1,t}^{(i)}, \ldots, x_{d,t}^{(i)})$, $t = 1, \ldots, T$, $i = 1, \ldots, N$, constitute observations of random curves:

$$\boldsymbol{X}^{(i)}(t) = \big(x_1^{(i)}(t), \ldots, x_d^{(i)}(t)\big), \quad t \in [0, 1],\ i = 1, \ldots, N.$$

Moreover, we assume that the sampled curves are observed at discrete instants of time. Hence we have:

$$\boldsymbol{x}_j^{(i)} = \boldsymbol{X}^{(i)}(j/T) + \varepsilon^i(j/T), \quad j = 1, \ldots, T.$$

We reconstruct functions $\boldsymbol{x}^{(i)}(t)$, $t \in [0, 1]$ by smoothing techniques (see e.g. Ramsey and Silverman, 2005), thus obtaining functional data

$$\widehat{\boldsymbol{x}}^{(i)}(t), \quad t \in [0, 1],\ i = 1, \ldots, N,$$

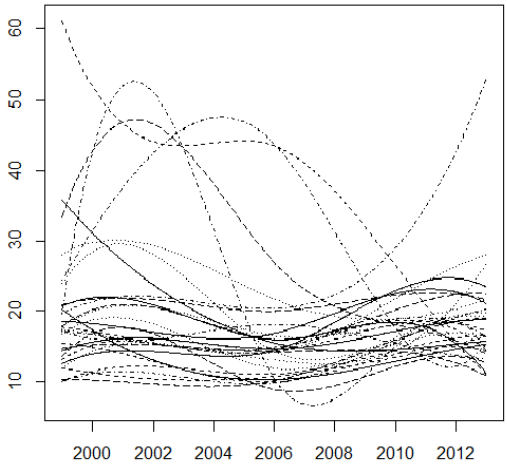

Fig. 2. Capital adequacy ratio (smoothed using B-spline approximation).

which are a subject for functional clustering analysis. It is worth to mention that each function $\widehat{\boldsymbol{x}}^{(i)}$ is d-dimensional. In Fig. 2 we present an example of 1-dimensional functional data under consideration.

As in time series clustering, we will apply clustering methodology on 1-dimensional curves, i.e. we will cluster banks according to each ratio separately.

4.2. *Functional Data Dissimilarity Measures*

In this section, we review six dissimilarity measures used in functional data clustering.

Hausdorff distance between two curves

A distance between two curves can be measured by Hausdorff distance. This distance measures maximum distance from a point in one curve to the nearest point in the other curve. Suppose that $G(x) = \{(t, x(t)) : t \in [0, 1]\} \subset R^2$ and $G(y) = \{(t, y(t)) : t \in [0, 1]\} \subset R^2$ are graphs of the curves x and y, respectively. Hausdorff distance $D_{\text{Hausdorff}}(x, y)$ is defined by:

$$D_{\text{Hausdorff}}(x, y) = \max\left\{ \sup_{x \in G(x)} \inf_{y \in G(y)} D_{\mathcal{L}_2}(x, y), \sup_{y \in G(y)} \inf_{x \in G(x)} D_{\mathcal{L}_2}(x, y) \right\}, \tag{7}$$

where $D_{\mathcal{L}_2}$ is Euclidean distance.

A formal definition of the Euclidean distance between functional data is:

$$D_{\mathcal{L}_2}(x, y) = \sqrt{\left(\frac{1}{\int_a^b w(t)\mathrm{d}t} \int_a^b \left| x(t) - y(t) \right|^2 \cdot w(t)\mathrm{d}t \right)},$$

where $w(t)$ is a weighting function.

Distance based on B-spline approximation

Ferraty and Vieu (2006) proposed a two-stage approach for functional data clustering. Proximity between two curves x and y could be estimated using:

$$D^q(x, y) = \sqrt{\frac{1}{T}\int_T \left(x^{(q)}(t) - y^{(q)}(t)\right)^2 \mathrm{d}t},$$

where $x^{(q)}(t)$ is the q-th derivative of x. In the first stage Ferraty and Vieu (2006) used B-spline to approximate functional data.

Consider a B-spline basis as a set of functions $B = \{b_1, \ldots, b_N\}$. Then derivatives of the approximated curves by n elements of B-spline are expressed: $\hat{x}^{(q)} = \sum_{n=1}^{N} c_n B_n^{(q)}$. The second stage is proximity measure, which is expressed as:

$$D_{\mathrm{B}}(x, y) = \sqrt{\frac{1}{T}\int_T \left(\hat{x}^{(q)}(t) - \hat{y}^{(q)}(t)\right)^2 \mathrm{d}t}. \tag{8}$$

In our analysis we considered two cases, i.e. we estimated distance with $q = 0$ (D_{BASIS}) and $q = 1$ (D_{DERIV}).

Distance based on continuity properties of curves

In this paper we introduce two dissimilarity measures. One of them is based on Hölderian property of a function. Dissimilarity measure is constructed from two parts. The first part shows how close functions are to each other. In this part we calculate supremum between two curves. The second part shows how similar curves change together. Hölder distance measure is defined by:

$$D_{\mathrm{Hölder}}(x, y) = \sup_t \left|x(t) - y(t)\right| + \sup_{t \neq s} \frac{\left|(x(t) - y(t)) - (x(s) - y(s))\right|}{|t - s|^{\alpha}}, \tag{9}$$

where the number $\alpha \in (0, 1]$ is called Hölder exponent.

The second dissimilarity measure, which we propose, is also constructed from two parts. The first part shows how close B-spline approximations are close together. The second part uses q-th derivative to capture how close is the change of curves. The dissimilarity measure is estimated as follows:

$$D_{\mathrm{SUP}}(x, y) = \sup_t \left|x(t) - y(t)\right| + \sup_t \left|x^{(q)}(t) - y^{(q)}(t)\right|. \tag{10}$$

This distance measure takes into account both the closeness and the behaviour of the data.

Distance based on functional principal components

Functional principal components give another tool to reduce dimension of a functional data. This distance measure is also considered as a two-stage approach. The functional data can be decomposed in a finite orthonormal basis: $\hat{x}_i(t) = \sum_{i=1}^{K} f_{ik}\xi_k(t)$, where f_{ik}

is the score of the principal component $\xi_k(t)$. In this case the distance between two curves is calculated as:

$$D_{\mathrm{FPCA}}(x, y) = \sqrt{\sum_{k=1}^{K} \left(\sum_{j=1}^{T} \big(f_x(t_j) - f_y(t_j) \big) \right)^2}, \tag{11}$$

where f_x and f_y are scores of the principal component of curves x and y, respectively.

5. Multivariate Clustering

In the previous two sections we considered data as univariate time series or 1-dimensional curves. In this section we interpret the data as N observations of d-dimensional time series or d-dimensional curves.

Multivariate Euclidean distance
The Euclidean distance of univariate time series can be easily expended to multivariate case. This dissimilarity measure is expressed as:

$$D_{\lambda,\mathrm{EUCL}}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{d} \sum_{t=1}^{T} \big((x_{jt} - \lambda x_{j,t-1}) - (y_{jt} - \lambda y_{j,t-1}) \big)^2 \right)^{1/2}. \tag{12}$$

The next two measures of dissimilarity of time series are obtained by introducing certain correction of Euclidean distance. This addresses to adaptive dissimilarity index introduced by Chouakria and Nagabhushan (2007) and to complexity invariant distance measure introduced in Batista *et al.* (2011, 2014). We define the analogues for d-dimensional time series.

Multivariate adaptive dissimilarity index
Extended adaptive dissimilarity index for the d-dimensional time series $\boldsymbol{x}$ and $\boldsymbol{y}$, is defined as:

$$CORT(\boldsymbol{x}, \boldsymbol{y}) = Q_{\boldsymbol{x}}^{-1/2} Q_{\boldsymbol{x},\boldsymbol{y}} Q_{\boldsymbol{x}}^{-1/2},$$

where

$$Q_{\boldsymbol{x}} = \sum_{t=1}^{T-1} (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)^{\tau} (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t), \qquad Q_{\boldsymbol{x},\boldsymbol{y}} = \sum_{t=1}^{T-1} (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)^{\tau} (\boldsymbol{y}_{t+1} - \boldsymbol{y}_t).$$

Let $\lambda_{\max}(CORT(\boldsymbol{x}, \boldsymbol{y}))$ denote the largest eigenvalue of the matrix $CORT(\boldsymbol{x}, \boldsymbol{y})$. Multivariate addaptive dissimilarity index is then expressed as:

$$D_{\mathrm{CORT}}(\boldsymbol{x}, \boldsymbol{y}) = \phi_k \big[\lambda_{\max}\big(CORT(\boldsymbol{x}, \boldsymbol{y}) \big) \big] \cdot D_{\mathrm{EUCL}}(\boldsymbol{x}, \boldsymbol{y}), \tag{13}$$

where $\phi(u)$ is exponential adaptive tuning function.

Multivariate complexity-invariant distance measure
Consider that complexity estimate $CE(\boldsymbol{x})$ for multivariate case can be written as follows:

$$CE(\boldsymbol{x}) = \left(\sum_{j=1}^{d} \sum_{t=1}^{T-1} (x_{jt+1} - x_{jt})^2 \right)^{1/2}.$$

Then a complexity-invariant dissimilarity measure is expressed as:

$$D_{\mathrm{CID}}(\boldsymbol{x}, \boldsymbol{y}) = CF(\boldsymbol{x}, \boldsymbol{y}) \cdot D_{\mathrm{EUCL}}(\boldsymbol{x}, \boldsymbol{y}), \tag{14}$$

where $CF(\boldsymbol{x}, \boldsymbol{y})$ is complexity correction factor.

Multivariate case of other dissimilarity measures
Other dissimilarity measures (D_{ACF}, D_{DTW}, $D_{\mathrm{Hausdorff}}$, $D_{\mathrm{Hölder}}$, D_{BASIS}, D_{DERIV}, D_{SUP}, D_{FPCA}) for the multivariate case are calculated using the following expression:

$$D_{\mathrm{M}}(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{d} \left[D_j(x, y) \right]^2 \right)^{1/2}, \tag{15}$$

where $D_j(x, y)$ is a coordinate-wise dissimilarity measure.

6. Clustering Algorithm and Validity Assessment

In this section, we present clustering algorithm and clustering validity indices used in the paper.

6.1. *Clustering Algorithm*

In this study we used conventional agglomerative hierarchical clustering algorithm. This method works by clustering time series into a tree of clusters (dendrogram). At the beginning each observation is assigned to its own cluster. Afterwards, clustering algorithm works iteratively, at each step joining the two most similar clusters into larger and larger ones. This process continues until a single cluster is formed or until certain termination conditions are satisfied. The complete linkage algorithm, which was applied in this study, measures the similarity between two clusters as the similarity between farthest pair of data belonging to different clusters.

The iterative procedure of the complete linkage algorithm can be written as follows:

1. At the start each element is assigned to its own cluster. The level of dendrogram is set to $L(0) = 0$ and sequence number $n = 0$.
2. One finds a pair of clusters, say C_i and C_j, with lowest dissimilarity ($D(C_i, C_j)$). Set the sequence number to $n = n + 1$. These two clusters are then joined at a level $L(n) = D(C_i, C_j)$.

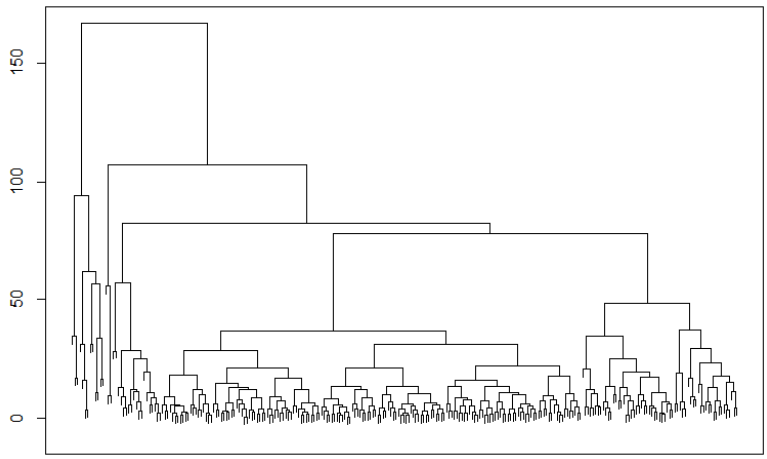

Fig. 3. Dendrogram using D_{FPCA} distance measure: capital adequacy ratio.

3. Dissimilarity matrix is updated by reducing its order by one. In complete linkage clustering the distance between clusters is the distance between the farthest pair of points, i.e. $D(C_i \cup C_j, C_k) = \max_{i,j \in C_i \cup C_j, k \in C_k}(D(i,k), D(j,k))$.
4. Steps 2 and 3 are repeated until single cluster is obtained ($N - 1$ times).

Figure 3 gives an example of dendrogram from the hierarchical clustering algorithm. We have chosen hierarchical clustering algorithm because it is more efficient in dealing with outliers than partitional algorithms.

6.2. *Cluster Validity Assessment*

The next step is to choose the optimal number of clusters. We calculated three different measures which are used validating the results of clustering analysis in clustering literature. Dunn index and Caliński and Harabasz index are used choosing number of clusters. Meanwhile, average silhouette width helps to choose number of clusters and to compare different distance measures.

Average silhouette width

Rousseeuw (1986), Kaufman and Rousseeuw (1990) introduced average silhouette width as a measure to evaluate clustering. Let $\mathcal{C} = \{C, , \ldots, C_K\}$ be particular clustering partition of the N observations into K disjoint clusters. The silhouette value measures the degree of confidence in the clustering of an observation. For observation i, value is defined by:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)}, \tag{16}$$

where a_i is the average dissimilarity of i to all other objects of C_i (cluster containing observation i) and b_i is a minimum of the average dissimilarity between i and the elements

of the other cluster which is different from C_i. Thus:

$$a_i = \frac{1}{n(C_i)} \sum_{j \in C_i} D(i, j), \qquad b_i = \min_{C_m \in \mathcal{C} \setminus C_X} \sum_{j \in C_m} \frac{D(i, j)}{n(C_m)}$$

where $D(i, j)$ is dissimilarity measure and $n(C)$ is the cardinality of the cluster C. The average silhouette value is in the interval $[-1, 1]$. A value close to 1 means that particular clustering partition is well classified, and value close to -1 means that observations are misclassified.

Dunn index

Another cluster validity index was proposed by Dunn (1974). This index tries to identify compact and well separated clusters. The Dunn index is calculated as a ratio of the smallest dissimilarity between observations not in the same cluster to the largest intra-cluster dissimilarity:

$$Dunn(\mathcal{C}) = \frac{\min_{C_m, C_l \in \mathcal{C}, C_m \neq C_l} (\min_{i \in C_m, j \in C_l} D(i, j))}{\max_{C_n \in \mathcal{C}} diam(C_n)} \tag{17}$$

where $diam(C_n)$ is diameter of a cluster, i.e. maximum distance between observations in cluster C_n.

If observations are in the compact and well separated clusters, then the value of Dunn index should be large, because the dissimilarity between the clusters is expected to be large and the diameter of the clusters is expected to be small.

Caliński and Harabasz index

Caliński and Harabasz (1974) introduced a criterion, which can be used to determine the number of clusters in cluster analysis. Milligan and Cooper (1985) showed that $CH(k)$ index works in many cases.

Caliński and Harabasz index is defined by the following expression:

$$CH(k) = \frac{\mathbf{B}_k (N - k)}{\mathbf{W}_k (k - 1)}, \tag{18}$$

where $\mathbf{W}_k$ is the overall within-cluster variance:

$$\mathbf{W}_k = \sum_{h=1}^{k} \frac{1}{|C_h|} \sum_{i, j \in C_h} D(i, j)^2$$

and $\mathbf{B}_k$ is the overall between-cluster variance:

$$\mathbf{B}_k = \frac{1}{N} \sum_{i, j=1}^{N} D(i, j)^2 - \mathbf{W}_k.$$

Well separated clusters have large $\mathbf{B}_k$ and small $\mathbf{W}_k$. Therefore, larger value of $CH(k)$ indicates better data partition.

7. Clustering Results

In this section, we present results from the banking data clustering exercise. We performed the clustering experiment on 6 bank performance ratios. We used 12 dissimilarity measures for each ratio to assess the closeness of the banks. Then agglomerative hierarchical clustering algorithm was perfomed to group banks into clusters. Since the true number of clusters is unknown, we divided banks into 2, 4, 6, 8, 10 and 20 clusters. Finally, cluster validity indices were calculated, which are presented in Appendix.

In a univariate case based on the average silhouette width we can conclude that there is no dissimilarity measure which could be the best one for all ratios. For example, for some ratios (ROAA, ROAE) dissimilarity measured between first derivatives gives high average silhouette width (see Appendix, Figs. 5 and 6). However, this distance measure performs poorly for capital adequacy ratio, especially if we consider more than two clusters (Appendix, Fig. 10). Similarly, distance measure based on supremum between two curves gives good clustering results for loan loss provisions over gross loan portfolio ratio (Appendix, Fig. 9), but it is not suitable to cluster banks if we use ROAE. Our proposed distance measures ($D_{\text{Hölder}}$, D_{SUP}) performed well in this study. Dissimilarity measure based on L_∞ norm between B-spline approximations and their first derivatives showed the highest average silhouette width for ROAA. Meanwhile, distance measure based on Hölder's exponent provided the best results if we take CAR ratio. One thing that could be noted is that dissimilarity measures which are based on functional data properties ($D_{\text{Hölder}}$, D_{SUP}, D_{BASIS}, D_{DERIV}, D_{FPCA}) performed better than measures which use time series properties. In our case distance measure based on autocorrelation and *CID* gives the lowest average silhouette width values (see e.g. Appendix, Fig. 7). It could be also noted that simple Euclidean distance performed rather good for clustering banking data. This result is consistent with Ding *et al.* (2008), who also found that Euclidean distance provides relatively good clustering outcome. Thus, we can conclude that it is useful to use dissimilarity measures which employ functional data properties.

From these results we also see that average silhouette width usually is the highest if we take two clusters. Only in some cases the higher value is with 4 clusters, for example, ROAA, if we measure dissimilarity with D_{BASIS} distance or functional principal components and CIR if we use Euclidean distance (see Appendix, Fig. 8). In many cases the average silhouette width drops significantly if we consider more than four clusters. However, other clustering validity indices give mixed results. In some cases, for instance, if we take CAR ratio or LLP, Dunn index mostly shows that we should consider two ar four bank clusters (see Appendix, Table 2). Caliński and Harabasz index also suggests mostly 2 or 4 clusters. Similar results are with LLP ratio where both indices indicate that the best option is to choose two or four clusters (see Appendix, Table 3). An interesting case is with CIR ratio. While Dunn index suggests that we should consider even 20 clusters, Caliński and Harabasz index gives the opposite result, i.e. the index is highest if we take 2 or 4 clusters. Such discrepancy may arise due to the fact that CH index calculates average variance, meanwhile Dunn index takes maximum distance between observations. As our data are noisy and clusters are not well-separated, few observations may have strong impact on Dunn index values.

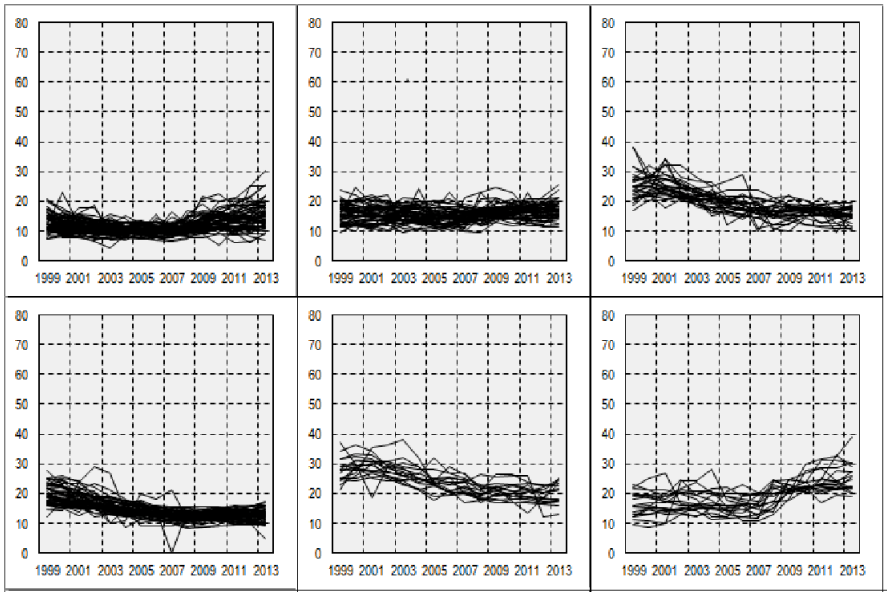

Fig. 4. Clustering results using D_{FPCA} distance measure: capital adequacy ratio.

Analysis of the data grouped into 20 clusters revealed that in many cases there are few clusters formed by larger number of banks and other clusters are formed by only few banks. For example, Fig. 4 shows 6 clusters that are formed from bigger number of banks using D_{FPCA} distance measure. From the figure we can distinguish few patterns in the development of capital adequacy ratio: one group of banks kept their CAR ratio more or less at the same level, second group showed decreasing trend, and third group increased CAR significantly after the Global financial crisis in 2009. Other clusters included only 1 or 3 banks. Clustering results show that we can extract 6 larger clusters if we take ROAA or ROAE. Taking capital adequacy ratio results in 6 clusters, whereas LLP in 8 clusters. Banks could be clustered into 10–11 groups if we take NIM and into 12 larger clusters if we take CIR. Clusters that are formed by few banks could be considered as outliers. Therefore, only the larger clusters could be further examined in the development of the macroprudential policy instruments.

In a multivariate case we take three profitability measures and efficiency ratio (CIR) to form d-dimensional time series. Furthermore, we normalized data of each ratio to take into account differences between values of each ratio. We take these four ratios because most of the banks in our sample had data on them. Based on the average silhouette width we see that it is reasonable to cluster banks into groups based on few ratios at the same time. Most of the values of the ASV index are comparable with the univariate cases. Another finding is that in multivariate case it is important to take into account both closseness and behaviour of time series, as $D_{\mathrm{Hölder}}$, D_{SUP} and $D_{\Delta,\mathrm{EUCL}}$ give better clustering results. Differently from univariate case where $D_{\Delta,\mathrm{EUCL}}$ does not improve results of Euclidean distance (D_{EUCL}), in a multivariate case the change in ratios improves clustering results. In a multivariate case ASV mostly suggests two clusters, but based on Dunn index

and CH index we should take a larger number of clusters. As multivariate clustering revealed, it is possible to find homogeneous groups of banks taking into account all ratios. Of course, if we analyse separate ratios of the clustered banks, we see that some banks would not be grouped into the same cluster in a univariate case. For this result there are also economical reasons as banks might reach similar ROAA ratio having different share of equity and/or performance efficiency. Nevertheless, multivariate dissimilarity measures, proposed in this paper, might be useful in other cases.

8. Conclusions

There are two main purposes of this paper. The first one is to compare various dissimilarity measures which are used to cluster time series data. We considered dissimilarity measures based on raw time series data and measures which take into account some properties of time series (e.g. autocorrelation). Another group of dissimilarity measures is based on functional data properties. Furthermore, we analysed clustering based on multivariate data. Second purpose is to consider clustering of the banks according to their performance ratios and to find a proper number of clusters. We took 6 ratios that are commonly used to compare performance of the banks. Three ratios measure profitability: return on average assets, return on average equity, net interest margin. Cost to income shows efficiency, capital adequacy ratio shows how much risk bank is taking and loan loss provisions show the quality of loan portfolio.

Cluster analysis results show that the choice of dissimilarity measure may change the way banks are grouped significantly. The same could be addressed to the choice of number of clusters which depends on the clustering validation method. As pointed out by Batista *et al.* (2014), dissimilarity measure is the key component in clustering. Therefore, it is good option to take few distance measures and compare the results. Furthermore, based on the average silhouette width we may conclude that no dissimilarity measure worked best for all ratios. In some cases dissimilarity measured between first derivatives or dissimilarity measure based on Hölder condition gives high average silhouette width, in other cases distance measure based on functional principal components gives better clustering results. However, clustering methods based on functional data properties mostly outperfomed distance measures based on time series properties. In our study D_{ACF} and D_{CID} provided relatively poor clustering results for many ratios. Another conclusion could be that simple Euclidean distance is relatively good distance measure for clustering banking data. Third conclusion, based on average silhouette width both proposed measures, $D_{\mathrm{Hölder}}$ and D_{SUP}, were among the best for clustering banking data.

The choice of number of clusters is not that clear also. However, for some banking ratio clustering validation indices suggest low number of clusters. If we consider CAR, LLP or CIR, then optimal number of clusters would be 2 or 4. But if we take profitability ratios, then the results are mixed and the number of clusters could be chosen by the expert judgement.

Division of banks into 20 clusters revealed that there are few larger clusters and other clusters are formed by small number of banks. According to different banking ratios, there

are from 6 to 12 clusters. Those larger clusters could be further analysed and used to develop some new macroprudential tools.

Multivariate clustering revealed that it is reasonable to group banks into clusters according to profitability and efficiency ratios. Average silhouette width is comparable with univariate cases. Of course, if we analyse separate ratios of the clustered banks, we see that some banks would not be grouped in a univariate case. Nevertheless, in some data samples multivariate clustering might be useful as it divides time series based on few features.

Appendix

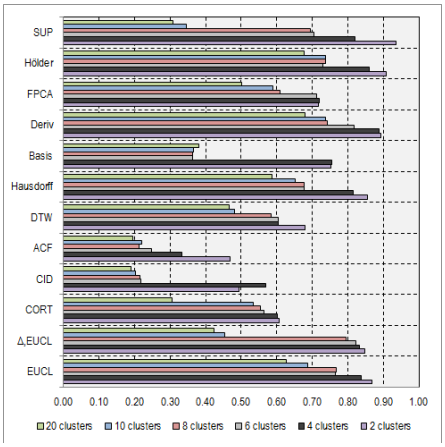

Fig. 5. Average silhouette width (ROAA).

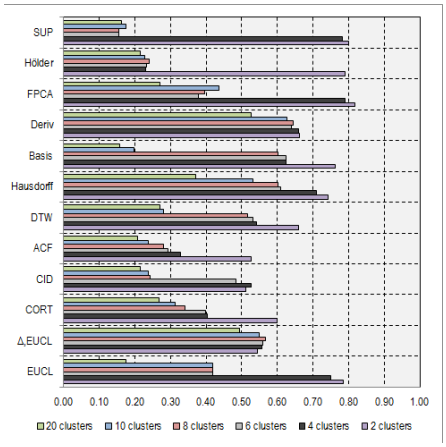

Fig. 6. Average silhouette width (ROAE).

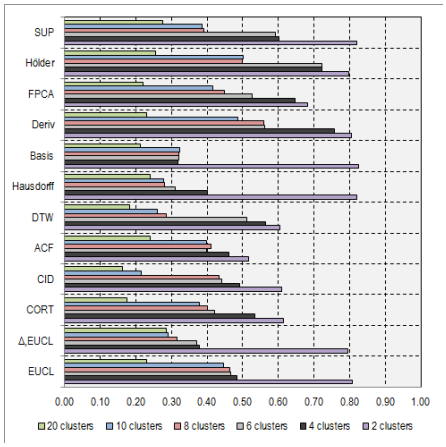

Fig. 7. Average silhouette width (NIM).

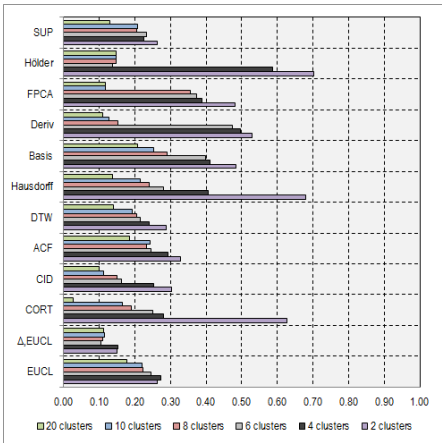

Fig. 8. Average silhouette width (CIR).

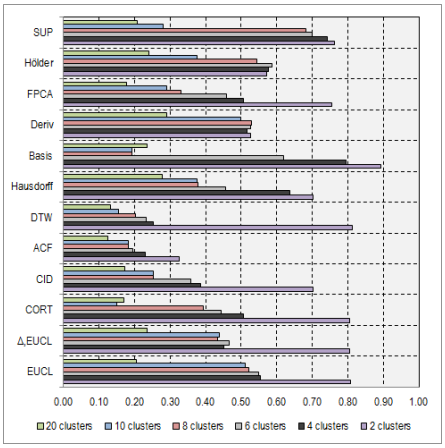

Fig. 9. Average silhouette width (LLP).

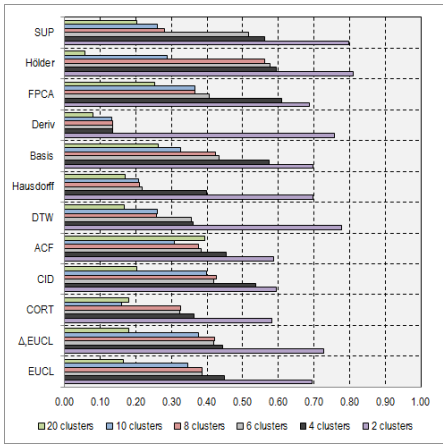

Fig. 10. Average silhouette width (CAR).

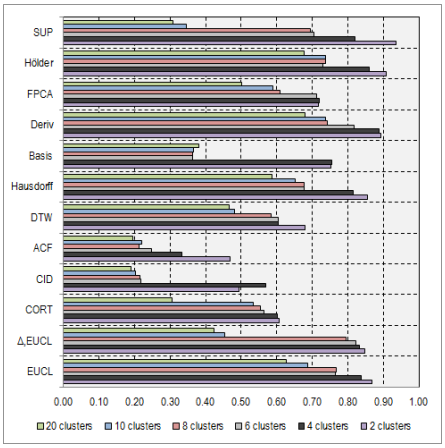

Fig. 11. Average silhouette width (multivariate case).

Table 2
Number of clusters suggested by Dunn index.

	EUCL	Δ, EUCL	CORT	CID	ACF	DTW	Hausdorff	Basis	Deriv	FPCA	Hölder	SUP
ROAA	2	8	4	20	10	10	4	4	4	6	2	2
ROAE	4	8	8	6	20	20	10	2	20	4	2	4
NIM	2	20	4	20	10	20	2	2	4	4	2	2
CIR	20	20	2	20	20	2	20	20	6	8	2	20
LLP	2	2	2	20	20	2	4	2	10	2	8	8
CAR	2	2	20	10	20	4	2	4	2	4	2	2
Multivariate	2	2	10	8	20	2	2	10	2	2	2	4

Table 3
Number of clusters suggested by Caliński and Harabasz index.

	EUCL	Δ, EUCL	CORT	CID	ACF	DTW	Hausdorff	Basis	Deriv	FPCA	Hölder	SUP
ROAA	10	2	20	6	2	2	2	6	6	20	2	2
ROAE	6	2	6	6	2	10	2	4	10	10	2	4
NIM	6	4	2	2	8	20	4	10	2	2	2	4
CIR	4	2	2	2	4	20	2	2	4	2	8	4
LLP	4	2	20	2	2	20	2	2	2	4	20	2
CAR	2	4	2	2	20	20	2	2	2	2	4	20
Multivariate	4	8	2	8	4	6	2	2	4	4	2	2

References

Abraham, C., Cornillon, P.A., Matzner-Løber, E., Molinari N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, 30(3), 581–595.

Batista, G., Wang, X., Keogh, E.J. (2011). A complexity-invariant distance measure for time series. In: *Proceedings of the Eleventh SIAM International Conference on Data Mining*, pp. 699–710.

Batista, G., Keogh, E.J., Tataw, O.M., de Souza, V. (2014). CID: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3), 634–669.

Berndt, D.J., Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In: *AAA1-94 Workshop on Knowledge Discovery in Databases*, pp. 359–370.

Bezdek, J.C., Ehrlich, R., Full, W. (1984). FCM: fuzzy c-means algorithm. *Computers and Geosciences*, 10(2–3), 191–203.

Bohte, Z., Cepar, D., Kosmelij, K. (1980). Clustering of time series. In: *Proceedings of 1980 Computational Statistics Meeting COMPSTAT 80*, pp. 587–593.

Caliński, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), pp. 1–27.

Chan, K.-P., Fu, W.-C. (1999). Efficient time series matching by wavelets. In: *Proceedings of 15th International Conference on Data Engineering (ICDE '99)*, pp. 126–133.

Chouakria, A.D., Nagabhushan, P.N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1), 5–21.

Davies, D.L., Bouldin, D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

Díaz, S.P., Vilar, J.A. (2010). Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. *Journal of Classification*, 27, 333–362.

Ding, H., Trajcevski, G., Wang, X., Keogh, E.J. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2), 1542–1552.

Dunn, J.C. (1974). Well separated clusters and fuzzy partitions. *Journal of Cybernetics*, 1(24), 95–104.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of 2nd International Conference On Knowledge Discovery and Data Mining*, pp. 226–231.

Faloutsos, C., Ranganathan, M., Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. In: *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, pp. 419–429.

Ferraty, F., Vieu, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice*. *Springer Series in Statistics*. Springer, New York, pp. 125–147.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24, 164–181.

Geleano, G., Peña, D. (2000). Multivariate analysis in vector time series. *Statistics and Econometrics. Series 15 WP*, 1(24).

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., Boesiger, P. (2004). A new correlation-based fuzzy logic clustering algorithm for fMRI. *Magnetic Resonance in Medicine*, 40(2), 249–260.

Golin, J. (2001). *The Bank Credit Analysis Handbook: A Guide for Analysts, Bankers and Investors*. *Wiley Finance Series*. Wiley, New York, pp. 261–336.

Halkidi, M., Batistakis, Y., Vizirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2–3), 107–145.

Hemanth, D.J., Anitha, J., Balas, V.E. (2016). Performance improved modified fuzzy c-means algorithm for image segmentation applications. *Informatica*, 26(4), 635–648.

Ieva, F., Paganoni, A.M., Pigoli, D., Viteli, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3), 401–418.

Jacques, J., Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3), 231–255.

James, G.M., Sugar, C.A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462), 397–408.

Kaufman, L., Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. *Wiley Series in Probability and Statistics*. Wiley, New York, pp. 68–125.

Keogh, E.J, Chakrabarti, K., Pazzani, M., Mehrotra, S. (2001). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3), 263–286.

Krishnapuram, R., Frigui, H., Nasraoui, O. (1993). Quadratic shell clustering algorithms and the detection of second-degree curves. *Pattern Recognition Letters*, 14(7), 545–552.

Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.B. (2004). The similarity metric. *IEEE Transactions on Information Theory*, 50(12), 3250–3264.

Liao, W.T. (2005). Clustering of time series data – a survey. *Pattern Recognition*, 38, 1857–1874.

Lin, J., Keogh, E.J., Lonardi, S., Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD'03*, pp. 2–11.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297.

Maharaj, E.A. (1996). A significance test for classifying ARMA models. *Journal of Statistical Computation and Simulation*, 54(4), 305–331.

Milligan, G.W., Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159–179.

Ng, R.T., Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In: *Proceedings of the 20th VLDB Conference*, pp. 144–155.

Peng, B.J., Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic process, with application to online auctions. *The Annals of Applied Statistics*, 2(3), 1056–1077.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11(2), 153–164.

Ramsey, J., Silverman, B.W. (2005). *Functional Data Analysis*. *Springer Series in Statistics*, second edition. Springer, New York.

Rousseeuw, P.J. (1986). Silhouettes: a graphical aid to interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Samé, A., Chamroukhi, F., Govaert, G., Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, 5(4), 301–321.

Wang, W., Yang, J., Muntz, R. (1997). STING: a statistical information grid approach to spatial data mining. In: *Proceedings of 23rd VLDB Conference*, pp. 186–195.

L. Naruševičius is a doctoral student at Vilnius University, Faculty of Mathematics and Informatics. His research interests include the problems of financial stability, macroprudential policy, banking sector, stress testing.

A. Račkauskas is a professor at Vilnius University, Faculty of Mathematics and Informatics. He is the head of the Department of Econometric Analysis. His research interests include the functional data analysis, limit theorems of the probability theory and their applications in statistics and econometrics, bootstrap methods in statistics and econometrics.

Panašumo vertinimo metodų lyginimas pagal bankų veiklos rodiklius

Laurynas NARUŠEVIČIUS, Alfredas RAČKAUSKAS

Šiame straipsnyje nagrinėjama, kaip būtų galima suskirstyti Eupopoje veikiančių bankų imtį į klasterius pagal jų veiklos rodiklius 1999–2013 metais. Antrasis šio straipsnio tikslas – palyginti panašumo vertinimo metodus ir nustatyti, kurie yra tinkamiausi klasterizuoti bankams. Darbe naudojami šeši rodikliai, kurie parodo bankų pelningumą, veiklos efektyvumą, stabilumą ir paskolų portfelio kokybę. Panašumas tarp bankų vertintas metodais, kurie remiasi laiko eilučių ypatybėmis arba funkcinių duomenų ypatybėmis. Darbe taip pat pasiūlyti du panašumo vertinimo metodai, nedažnai taikomi klasterizavimo literatūroje. Be to, išplėsti keli metodai iš vienmačių į daugiamačius atvejus. Skaičiavimų rezultatai rodo, kad nėra vieno panašumo vertinimo metodo, kuris duotų geriausius klasterizavimo rezultatus pagal visus rodiklius. Vis dėlto panašumo vertinimo metodai, kurie remiasi funkcinių duomenų ypatybėmis duoda geresnius rezultatus nei metodai, paremti laiko eilučių ypatybėmis. Rezultatai rodo, kad sunku tiksliai parinkti tinkamą bankų klasterių skaičių. Remiantis skirtingais veiklos rodikliais, bankus būtų galima suskirstyti į 6–12 klasterius.