Banks’ business models in the euro area: a cluster analysis in high dimensions

Published Version:
Farne', M., Vouldis, A.T. (2021). Banks’ business models in the euro area: a cluster analysis in high dimensions. ANNALS OF OPERATIONS RESEARCH, 305(1-2 (October)), 23-57 [10.1007/s10479-021-04045-9].

Published:
DOI: http://doi.org/10.1007/s10479-021-04045-9

Banks’ business models in the Euro Area: a cluster analysis in high dimensions

Abstract

In this paper, we propose a data-driven approach for identifying the business models of the 365 largest Euro Area banks. Our methodology is suitable for very granular harmonised supervisory data. Our dataset allows us to consider the full range of the activities in which banks are involved. The proposed method combines in an optimal way data clustering, dimensionality reduction and outlier detection. We identify four business models and exclude as ‘outliers’ banks that follow idiosyncratic business models. We find out that traditional commercial banks are characterized by the lowest levels of credit risk while the loan portfolios of securities holding banks are riskier compared to the other banks.

Keywords: Bank business model, Non-performing loans, Robust clustering, Dimensionality reduction.

JEL classification codes: C63, G21, G32, L21, L25

1. Introduction

This paper applies a novel methodology to classify all systemically important Eurozone banks into business models, utilising a uniquely granular dataset. Our emphasis is on the post-financial crisis banking system of Eurozone countries. We aim to overcome the limitations of the business model clustering approaches used in the literature that are based on narrow sets of broad pre-defined variables and do not fully capture the range of activities performed by banks. Our methodology combines optimally classification, dimensionality reduction, and outlier detection at the same time and accounts for the existence of idiosyncratic mixtures of banks' activities present in banks' business models (as noted e.g. in Mergaerts and Vennet, 2016).

The concept of 'business model' is increasingly used to refer to the heterogeneous mixture of activities to which banks engage and the difference in risk-return outcomes that this entails throughout different phases of the financial cycle (e.g. Yellen, 2012; Carney, 2015; Draghi, 2016). There is a burgeoning literature which proposes methods to classify banks into business models (Roengpitya et al., 2014; Ayadi et al., 2015; Köhler, 2015; Mergaerts and Vennet, 2016; Hryckiewicz and Kozlowski, 2017; Lucas et al., 2017). However existing empirical studies usually concentrate on a limited set of pre-selected dimensions when classifying banks into business models. In contrast, our method is data-driven and instead of relying on researcher's priors utilizes a unique, harmonised dataset detailing the activities undertaken by banks with an unprecedented level of granularity.

The paper contributes to filling the gap in the literature regarding banks' business models in the following ways. First, it formulates a methodology for identifying business models using granular data. Our proposed methodology combines optimally classification with dimensionality reduction allowing us to infer the factors which primarily determine banks' business models. It also incorporates an outlier detection component, allowing to identify banks which follow especially idiosyncratic business models, and whose inclusion into the normal clusters of banks would 'contaminate' the sub-sample and affect the results of analyses on differential risk or performance across business models.

Second, a unique data set, which has been made possible by the centralisation of supervision in a subset of countries within the European Union and the collection of supervisory data using harmonised definitions, is utilised. Our granular dataset comprising in total 1039 variables allows us to avoid biased classifications due to mismeasurement which could arise when broad categories are used, as will be explained below. In addition, the emphasis on the cross-sectional granular dimension of the input dataset is justified due to the slow-changing nature of banks' business models.[1]

Our paper complements and leverages previous research on banks' business models. During the last couple of years, a number of studies appeared that derive the business model classification from a narrow set of predefined variables (Roengpitya et al., 2014; Mergaerts and Vennet, 2016; Lucas et al., 2017), usually dictated by data availability. Other studies use a classification provided by the data provider (e.g. Köhler 2015, Becchetti et al., 2016) like the Bankscope's 'specialisation' attribute or focus on bank's ownership i.e. public or private (Eichengreen and Gupta, 2013). However, it has been already noted in the literature that the concept of business models, at least in Europe, belongs to a continuum (Mergaerts and Vennet 2016) if a restricted number of dimensions is used to classify banks.

Some specific examples will be used to illustrate the need to employ a granular dataset in order to distinguish banks into business models. For example, in previous studies broad categories like 'loans' or 'deposits' are used to identify the mix of activities into which banks engage. Therefore, a bank that holds an amount of loans to real economy agents above a certain threshold would most probably be classified as a traditional bank rather than an investment bank, however one may have to look also at off-balance sheet items, like loan commitments, to identify correctly a bank's involvement in financing the real economy. For example, BNP Paribas, one of the largest bank in Europe, has off-balance sheet loans commitments and financial guarantees to the real economy representing 56% of the loans to the real economy which it holds in the balance sheet.[2] Therefore, the classification of such banks depends crucially whether one relies on broad aggregates (like the value of loans in the balance sheets) because these aggregates may not represent fully the range of the banks' activities, or whether one adopts a more granular view that permits a robust identification of the range of activities in which the bank is involved (e.g. by taking into account information on off-balance sheet financing).

The measurement of banks' involvement in real sector lending is also highly affected by credit risk conditions. As a result, traditional banks with large

[1] Our classification results remain unchanged to a level higher than 95% when a different reference date within the 2014Q4 – 2015Q3 time period is used. This is expected as in a short time frame the composition of banks' activities is not expected to change significantly. This is consistent also with the approach adopted in the literature. For example, Lucas et al. (2017) assume a fixed cluster assignment although their data set spans a larger time period while Mergaerts and Vennet (2016) find that 'between' variation (differences across banks) exceeds the 'within' variation (changes over time within banks) for a sample of banks from 30 European countries. Finally, also studies which define business models with respect to governance structures assume constancy of business models (e.g. Becchetti et al., 2016).

[2] Loans to the real economy equal EUR 712 bln while the sum of loan commitments and financial guarantees to the real economy equal EUR 404 bln, according to the bank's 2016 financial statement: https://invest.bnpparibas.com/sites/default/files/documents/etats_financiers_31.12.16_en.pdf.

volumes of non-performing loans could be mistakenly perceived as being less focused on providing loans to the real economy, because in the balance sheet statements the amount of net loans is given i.e. the amount of loans after deducting allowances. This is the variable used in the literature, however when banks increase allowances this amount will diminish as a percentage of assets. The statistics for the Greek banks illustrate clearly this point: between end-2009 and end-2016 their amounts of gross loans to the domestic real economy remained almost unchanged (from EUR 188 bln to EUR 185 bln), however net loans decreased by 20% (from EUR 177 bln to EUR 141 bln) due to much higher provisions.[3] As a result, one could think that they decreased their activities in the real economy during this period, however this results is an artefact of their higher provisions due to realised credit risk and not of a change in their business model.

In addition, information on the use of derivatives, which in existing studies is found to be a distinctive element of different business models, could easily be distortive, if one does not incorporate information on both carrying amounts and notional values, or uses information on only the asset or the liability side, or does not take into account the intended use of derivatives e.g. as hedging or trading instruments. If only one of these measures is taken into account, then a distorted view of the degree to which a bank uses derivatives will be obtained, affecting subsequently its classification into business model clusters.

Finally, classifications based e.g. on the class of a bank within a national financial system would probably also not be optimal. For example, the German 'Landesbanken' differ among themselves as regards the composition of their balance sheets e.g. their composition of funding sources. Consequently, lumping these banks into one category would distort results of an econometric analysis. For example, deposits from other banks or customers are generally an important source of funding for Landesbanken, however the percentage varies considerably[4] and consideration of additional granular information on the remaining part of the liability side is needed in order to classify these banks meaningfully to a business model.

---

[3] Source: Bank of Greece statistics on the aggregate balance sheet of Greek credit institutions.

[4] Specifically, some Landesbanken rely almost exclusively on deposits for their funding, either from other banks or from customers while for some other this is not the case e.g. the Oldenburgische Landesbank had according to its 2016 financial statement more than 85% of its funding via deposits, while this percentage is less than 50% for Landesbank Baden-Württemberg – again as shown in the bank's 2016 financial statement.

All the examples presented above show why using a narrow set of variables to classify banks into business models could be problematic and potentially misleading. Consequently, there is a high value added of using a granular data set as the multidimensional concept of a business model requires detailed information to be captured and an identification methodology which can handle such granular information. A business model identification method which uses granular input data needs to perform not only clustering, but also dimensionality reduction and the identification of factors, in addition to including outlier detection in the sense of identifying banks following very idiosyncratic business models.

Besides granularity, other features of our dataset render it ideal for the purpose of identifying banks' business models. First, the dataset includes all large banks in Europe, with harmonised data consolidated at the prudential perimeter. This means that it provides the most reliable information on banks of systemic importance operating in Europe without exclusions which may bias the results e.g. inclusion of only listed banks. Furthermore, the prudential perimeter of consolidation as opposed to accounting consolidation or use of stand-alone data is optimal when considering banks' business models from a financial stability perspective because it provides simultaneously a consolidated view of banks' activities while abstracting from non-banking activities which possess very different risk characteristics than banking. Finally, it is important to note that all banks in our sample operate under the same regulatory environment, based on the transposition of Basel III in Europe via the CRD IV package, therefore the effect of regulation on the composition of the portfolio of their activities is similar.

The paper is structured as follows. Section 2 presents a review of the literature on the concept of the business model and on relevant empirical studies on banking. Section 3 describes the input set and presents the clustering methodology. Section 4 presents the results and provides a discussion about the identified business models. Finally Section 5 concludes.

2. Review of the literature

The literature of business models in banking had been until recently driven mainly by the concept of 'strategic groups', introduced by Hunt (1972). The concomitant concept of 'mobility barriers' (Caves and Porter 1977) had been introduced to explain persistent performance differentials between firms within one industry and also applied in banking (Amel and Rhoades 1988; DeSarbo and Grewal 2008, Halaj and Zochowski 2009; Mehra 1996;

Reger and Huff 1993; Tywoniak et al. 2007). Clustering methods are applied to identify strategic groups and consequently performance indicators are examined to assess whether performance differences exist. Data constraints are dictating the choice of the dimensions along which clustering is performed while the focus has been always in national or regional banking systems. Expert judgment is used extensively in the selection of the input set.

2.1. Empirical analyses of banks' business models

In the last years there is an expanding number of empirical studies of banking systems based on the business model concept. Roengpitya et al. (2014) (RTT henceforth) provide a clustering method to distinguish an international sample of banks according to their 'business models', based on the Ward's algorithm (1963), by using a selection of asset and liability variables (*choice* variables). Specifically, RTT define business models based on eight balance sheet ratios (loans, securities, trading book, interbank lending, customer deposits, wholesale debt, stable funding and interbank borrowing) which are interpreted as "reflecting strategic management choices" that leverage on the strengths of each organisation. They test their model on 1299 data points from 222 banks operating in 34 countries across the period 2005-2013, identifying three main business profiles: the Retail-funded, the Whole-funded and the Trading one. Finally, they provide a description of the bank performance for each business cluster by using a selection of key balance sheet ratios (*outcome* variables).

Ayadi and de Groen (2014) and Ayadi et al. (2015) also define banks' business models based on their activities. They examine a set of European banks (covering 80% of all banking assets in EEA in Ayadi and de Groen (2014)) and select a small number of dimensions (specifically, loans, trading assets, liabilities to other banks, customer deposits, debt liabilities and derivative exposures[5]) to perform hierarchical clustering. Both RTT and Ayadi et al. clearly distinguish between "activities", which determine the business model, and "outcomes", the latter measured by profitability and

[5] In Ayadi et al. (2015), liabilities to other banks and customer deposits were substituted by customer loans, because the expansion of the dataset compared to Ayadi and de Groen (2014) imposed more constraining data limitations.

performance indicators.

Ayadi and de Groen (2014) find that retail banks, specifically the two business models which they label “diversified retail” and “focused retail”, exhibit lower leverage compared to the business models of “investment” and “wholesale” banks which depart from the traditional intermediation function. However, this lower leverage is not reflected in their risk-adjusted ratios which are similar across business models, and can be interpreted as a symptom of “risk optimisation” on the part of sophisticated large banks. The results regarding performance are not clear-cut, also because of the volatility in the time dimension, however, it seems that “diversified retail” banks performed overall better than the other business models when taking into account both the pre-crisis and the crisis periods. Ayadi et al. (2015) expand the sample compared to Ayadi and de Groen (2014) and cover 95% of all banking assets in EEA. They also find that retail banks are less risky than wholesale and investment banks when using market measures of risk. RTT identify three business models and also find that their “retail-funded” banks perform better than “wholesale-funded” and “trading” banks while “trading” banks hold the higher levels of capital. Other studies which follow this line of research include Köhler (2015), Mergaerts and Vennet (2016), Hryckiewicz and Kozlowski (2017) and Lucas et al. (2017).

The important distinction between “choice” and “outcome” variables, aims to differentiate the set of variables reflecting strategic choices from the differential performance which is investigated ex post. The empirical strategies followed to classify banks into strategic groups usually focus on balance sheet “choice” variables (Amel and Rhoades 1988; DeSarbo and Grewal 2008; Mehra 1996). Halaj and Zochowski (2009), include additionally income and cost components, however this expansion of the type of variables is justified as a proxy for the unavailability of granular balance sheet breakdowns. Finally, Tywoniak et al. 2007 use also customer satisfaction ratings, although this seems to be better suited as a performance variable which could be investigated ex post.[6]

2.2. Identification methodologies

[6] Reger and Huff (1993) should be considered separately in this strand of the literature as they focus on the cognitive dimension of the managers and utilises data originating from interviews with bankers. As regards, the criteria used to determine the differences among strategic groups, DeSarbo and Grewal (2008) include performance, efficiency and size in the outcome set. Halaj and Zochowski (2009) also incorporate risk indicators (‘irregular loans’) arguing that this allows to position banks in a risk-return space, an idea which is especially relevant for the banking sector.

The identification of banks' business models requires the use of clustering methods that are known to belong to the class of unsupervised learning methods (Hastie et al. 2009). Agglomerative hierarchical methods like the Ward's clustering method (Ward 1963), which minimises the variance within clusters, rely on expert judgment. These methods are not suitable in a high dimensional context as it is not so easy to characterize clusters based on a large number of variables. This method is employed by RTT who select a priori subsets of eight variables representing bank assets and liabilities, excluding highly correlated variables.[7]

In the direction of classifying large dimensional objects, clustering methods which incorporate a dimension reduction process have also been proposed. The dimensionality reduction component is critical for the set-up where the input set is granular and relatively large compared to the number of entities to be classified. The most obvious way by which dimension reduction issues can be incorporated into a clustering methodology could be through applying a principal component analysis (Hotelling 1933) or a classical factor analysis before conducting the clustering. Consequently, a standard unsupervised clustering algorithm like the Ward's one on the obtained principal components or factors can be applied. This approach is called *tandem analysis* (Arabie and Hubert 1994).

However, as pointed out in De Soete and Carrol (1994) and De Sarbo et al. (1990), this approach may not be the most efficient for classification. The dimensions identified by the principal components or the factor analysis are not necessarily the ones that maximise the distance among the latent clusters identified by the second step. Performing the dimensionality reduction in a separate, initial step may mask or obscure the true cluster structure of the data, since it classifies the objects according to directions which are not optimal for discriminatory purposes.

An effective solution which incorporates dimension reduction into the class of partitional clustering techniques labelled as "k-means" (MacQueen, 1967) is provided by Vichi and Kiers (2001), who develop the factorial k-means algorithm, where a subspace is defined such that the projected data

[7] The number of clusters is chosen using the pseudo F-index, as proposed in Calinski and Harabasz (1974), which quantifies the trade-off between parsimony and ability to discriminate between clusters.

points on this subspace are closest to the centroids. As the name of the procedure suggests, it involves both factor analysis (reducing dimensionality) and k-means procedure (clustering objects and finding out their centroids in this low-dimensional subspace). We adopt an enhanced version of this clustering approach which seems to optimally combine the two essential features, dimensionality reduction and clustering. We incorporate in the clustering algorithm an intrinsic procedure to identify outliers within clusters, using the factor scores obtained by the iterative algorithm.

Before proceeding to the detailed description of the methodology used, we mention the alternative family of methods based on finite mixture models, which has also been used in the literature to identify banks' business models (Lucas et al. 2017).[8] Density-based approaches are computationally heavy in large dimensions, since they are likely to result in a large number of clusters. In addition, distribution hypotheses on economic data are potentially more distortive in the banking context than, for instance, on genetic data with pre-defined labels (see Lin et al. 2016 and Murray et al., 2014a,b). Furthermore, due to the distribution assumptions made in finite mixtures models, the outlier detection as regards the banks' business models cannot readily incorporate information from all input dimensions. This may lead to the "contamination" of the identified banks' clusters with very idiosyncratic institutions potentially distorting the results. In addition, several low rank spaces are identified instead of one in cited works, with the exception of Murray et al. (2014b). This further complicates the description of identified outliers via those methods.

Compared to the finite mixtures approach as used e.g. in Lucas et al. (2017), our enhanced clustering methodology offers the possibility to utilise a granular set of input dimensions, without the need to commit to a restricted set of inputs. The "trimmed" factorial k-means approach relies on a least squares algorithm which is effective in large datasets, because we identify only one latent space in place of several ones with a constrained distribution. In addition, a distribution-free approach lets the data speak with respect to the shape of clusters while also identifies banks lying far from the estimated clusters ("radial" outliers) without relying on parametric assumptions.

3. Methodology for identifying business models

[8] The standard reference on finite mixture models is McLachlan and Peel (2000). Recently, the literature has provided robust versions estimating mixtures of multivariate skew-normal (Lin et al. 2016) and skew-t distributions (Murray et al., 2014a) by maximum likelihood and the EM algorithm respectively. A distribution-free alternative is provided in Yang et al., 2017 using trimmed likelihood. Lucas et al. (2017) estimate via EM dynamic mixtures of normal or t distributions with time-varying means and possibly covariance matrices and find that the choice of Student's t causes clusters to be more robust to outliers due to fat tails.

3.1. Input set

We use a set of proprietary supervisory data which are collected in the context of the ECB Supervision. These data have been developed by the European Banking Authority (EBA) and employ harmonised definitions, thus representing an ideal set for a comparative analysis across countries. The availability of harmonised data across jurisdictions represents a necessary precondition when attempting to classify banks into respective business models.

We focus on Financial Reporting (FINREP) variables, providing a detailed decomposition of the balance sheet. FINREP is a standardised EU-wide framework for reporting accounting data, with a prudential scope of consolidation. Our sample consists of 365 banks residing in the 19 Eurozone countries. All systemically significant banks, as defined by the ECB Supervision (based on their absolute and within-country, relative size), are included in this sample. Our data set is cross sectional with reference date end-2014. The variables included in the input set could be interpreted as the “choice” variables, reflecting banks’ choices about the set of activities in which they are involved.

In particular, our input set contains information on the banks’ balance sheet composition under four types of breakdowns, specifically accounting portfolios, instruments (loans, securities etc), counterparties (households, non-financial corporations etc) and products (mortgage loans, credit cards etc). Appendix A presents a detailed description of the input data set.

Each of the 1039 initial variables is standardised using total assets as the scaling factor, except from the ‘total assets’ variable which is normalised using its maximum value within the sample. Therefore, a ‘size’ variable is retained in the initial data set while almost all the remaining variables lie in the interval [0,1] since they are expressed as a percentage of size.[9] Standardisation is used because we define business models with respect to the composition of banks’ activities, consistently with the literature reviewed above, and to avoid a dominance of the classification procedure by the large banks. Our results are invariant if the to-asset-ratios of any balance sheet variable and the relative size of any bank with respect to the maximum asset size in the sample are kept constant.

[9] It should be noted that there exist few variables presenting values higher than unity, like notional amounts of derivatives.

In this initial set there is a number of variables which are highly correlated and information which is redundant. Correlation is not per se an issue for the application of our clustering algorithm. However, given that we run the clustering algorithm with a number of different initialisations in order to search in the space of solutions and that we use the covariance matrix of the input set for the initialisation, the presence of nearly duplicated variables among the input data is not desirable.

Therefore, we follow a procedure to minimise the presence of very correlated variables in the input data set. The procedure consists of selecting the variables that should remain in the input set according to their 'importance', which is measured for each variable as the sum of the absolute correlations respect to all the others. This is a pre-processing step intended to avoid nearly duplicated variables by detecting pairs of variables that show a sample correlation very close to 1. Taking also into account the fact that some of the initial variables were very sparsely populated, we narrowed down the initial set of 1039 variables into a final input set of 382 variables (see Appendix B).

3.2. Clustering method

The statistical clustering problem can be defined as follows. Given a $n \times p$ data matrix X, where n is the number of banks (objects or observations) and p is the number of the variables, we would like to classify the banks into distinct clusters which contain objects which are 'close' in a statistical sense. Each cluster would represent a specific business model. The salient feature of our problem is that the dimension is relatively high compared to the number of objects: $p > n$. This feature is not common in similar classification problems; typically the objects which are to be classified are many more than the number of observed variables. Therefore, our problem belongs to the field of clustering in high dimensions. In addition, the absolute number of dimensions necessitates the use of data reduction techniques in order to compress the large initial data set into meaningful composite variables.

As explained in Section 2.2, existing studies on clustering banks do not provide a readily available suggestion on how to approach the clustering problem in a high dimensional space. Specifically, it is not clear how strong is the impact of distribution assumptions for the data in our large-dimensional setting. Consequently, density-based clustering methods, which are based on normal or Student's t mixtures, may hinder the interpretation of results. In addition, there is no clear rationale for defining ex ante the distribution shapes of variables across business models. Alternative methods, like hierarchical and partitioning (i.e. centroid-based) methods, do not provide any dimension reduction by themselves.

Let us call r the latent rank (i.e. the dimension of the reduced space) and c the number of clusters. In formal terms, the model involves the minimization of a measure of the following matrix

$$\boldsymbol{XA} - \boldsymbol{U\bar{Y}} \quad (1)$$

where $\boldsymbol{A}$ is a p × r column-wise orthonormal matrix (coefficients matrix), U is a n × c membership (or grouping) matrix such that $u_{ij} = 1$, if and only if $o_i \in Pj$, where $o_i, i = 1, .., n$, is the i-th observation and $Pj,\ j = 1, ..., c$ is the j-th cluster. The c × r matrix $\boldsymbol{\bar{Y}}$ contains the centroids of the clusters in the low rank space. The left term of this expression represents the projections into the factor space of the original objects (the transformed variable space or low rank or reduced space), while the second term represents the centroids of the clusters.

Equation (1) lies in the low-dimensional space spanned by the columns of the column-wise orthonormal matrix $\boldsymbol{A}$. Consequently, the model can be specified as follows

$$\boldsymbol{XAA'} = \boldsymbol{U\bar{Y}A'} + \boldsymbol{E} \quad (2)$$

where E is a residual matrix. Equation (2) describes the partition in the original space. The optimal partition therefore is sought by minimizing the function

$$\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{\bar{Y}}) = \left||\boldsymbol{XAA'} - \boldsymbol{U\bar{Y}A'}|\right|^2 = ||\boldsymbol{XA} - \boldsymbol{U\bar{Y}}||^2 \quad (3)$$

which can be equivalently expressed as

$$\boldsymbol{F}(\boldsymbol{A}, \boldsymbol{U}) = ||\ \mathbf{XA} - \mathbf{U}(\boldsymbol{U'U})^{-1}\boldsymbol{U'XA}\ ||^2 \quad (4)$$

since $\boldsymbol{\bar{Y}} = (\boldsymbol{U'U})^{-1}\boldsymbol{U'XA}$. This minimisation is performed under the constraints that $\boldsymbol{A'A}$=$\boldsymbol{I_r}$ and $\boldsymbol{U}$ is binary with only one non-zero element per row. In geometrical terms, we seek for the orthogonal linear combinations of the variables (factors) which best partition the objects by minimising the least-squares criterion (Eq. 4) in this reduced space.

We follow a robust approach which belongs to the class of Alternated Least Squares (ALS) algorithms (Vichi and Kiers, 2001) which is explained in detail in Appendix C. A discrete clustering model and a continuous factorial model are specified *simultaneously* for our data set. So, we perform at the same time data reduction (i.e. data synthesis) and variable selection by a single cluster analysis method, thus identifying the composite variables which most contribute to the classification of objects.

The strong consistency in statistical sense of the factorial k-means procedure is proved in Terada (2015). Underlying assumptions only require that the p-dimensional data vectors X_i , $i = 1,..,n$, are IID with a common distribution P. The only constraint is that the random space spanned by the p components of X_i is not isomorphic to any random space of dimension r, where r is the chosen latent rank. In high dimensions, it is clear that this is very unlikely to occur as r is extremely small with respect to p. Therefore, consistency is ensured if $n \to \infty$. Note that the condition $p \geq n$ is not ruled out as long as the described constraint is satisfied.

The selection of the latent rank r and the number of clusters c is not straightforward. It has to be noted that these two parameters depend on each other. Specifically, the number of components r cannot be larger than $c - 1$. The reason is that $Rank((\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{X}\boldsymbol{A})) = \min(c - 1, r)$, therefore describing the low-dimensional space of the clusters using more dimensions than necessary does not seem to make sense. The process for selecting these parameters will be described in Section 4.1 since it combines statistical criteria and the aim of obtaining interpretable results.

3.3. Robustified clustering with simultaneous radial outlier detection

It is known a priori that some institutions in our dataset follow unique business models, e.g. functioning as central clearing counterparties, focusing exclusively on refinancing public sector loans etc. Therefore, there is a clear rationale for excluding these outliers from the clusters to avoid distortions of the final results. We want to make sure that the presence of such cases does not distort the classification of banks. Therefore, we present a robustified version of the Vichi and Kiers (2001) procedure, which identifies clusters taking iteratively into account the presence of radial outliers. Our method is specifically intended to identify the so-called radial outliers, that is, observations deviating so much from assigned centroids to be considered external to assigned clusters. Therefore, our approach is robust both in the sense of measuring in a robust way the composition of the banks' activities, due to the granularity of the used dataset, and in avoiding the 'contamination' of the identified clusters with banks following particularly idiosyncratic business models.

As well established in the literature (see e.g. Rousseeuw and Leroy, 2003), the most used method for detecting multivariate outliers is via the Mahalanobis distance, $D = \sqrt{(x-\overline{x})'\boldsymbol{S}^{-1}(x-\overline{x})}$, with D^2 being asymptotically a chi-squared with p degrees of freedom under the assumption of normality for x (S is the unbiased sample covariance matrix). Under the normality assumption, Hotelling (1933) showed that $t^2 = n(x-\overline{x})'\boldsymbol{S}^{-1}(x-\overline{x})$, called Hotelling's T-squared, is proportional to $F_{p,\,n-p}$, where F is the Fisher's F. However, it is easy to see in our context this approximation cannot be used, because normality is not respected and the degrees of freedom $n-p$ would be negative, since $p > n$.

The trimmed k-means algorithm proposed in Cuesta-Albatos et al. (1997) could be used to detect anomalous data simultaneously with clustering. However, this method may be computationally intractable when both p and n are large while it does not offer a clear interpretation and visualisation of the identified clusters in large dimensions. In contrast, we would like to identify both partitions and outliers in a reduced space rather than in a p-dimensional space.

For this reason, a method is developed here to find the partition of the $100 \text{ x } (1-\alpha)\%$ most concentrated objects with respect to the scores in the low-dimensional space. Specifically, in the absence of any distribution assumptions we compare Mahalanobis distances across banks in order to exclude radial outliers i.e. banks clearly different from the rest. Mahalanobis distances are based upon $\boldsymbol{C_F}$, the unbiased covariance matrix of factor scores estimated over the entire sample, because the heuristics based on Equation (4) do not explicitly address the possibility of significantly different covariance matrices across clusters.

In more detail, our problem may be stated as follows. Our task is to minimise $\boldsymbol{F}(\boldsymbol{A},\boldsymbol{U},\overline{\boldsymbol{Y}}) = ||\boldsymbol{X}\boldsymbol{A} - \boldsymbol{U}\overline{\boldsymbol{Y}}||^2$ under the constraints $\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij} = [(1-\alpha)n]$, $\sum_{j=1}^{c} U_{ij} \leq 1$ for each $i = 1, \dots, n$, where the trimming proportion is $\alpha \in [0,0.5]$. This trimmed problem may be practically solved under the framework of Rousseuw and Van Driessen (2000). As a subset selection step (H-step), we set 100 initialisers and compute the initial estimates of loadings, centroids and cluster memberships by the Alternated Least Squares algorithm of Vichi and Kiers (2001). As a concentration step (C-step), we compute the Mahalanobis distance at each observation and we exclude the $[\alpha n]$ observations with the largest ones, since they are the observations that contribute the most to $\mathbf{F}(\mathbf{A}, \mathbf{U}, \overline{\mathbf{Y}})$, which is the variance within clusters. In this way, it is ensured that $\mathbf{F}(\mathbf{A}, \mathbf{U}, \overline{\mathbf{Y}})$ is decreasing at each step, such that the overall optimum is found out over the entire range of initializers which approximate the parameter space. The details of the outlier detection part of the algorithm are also described in Appendix C.

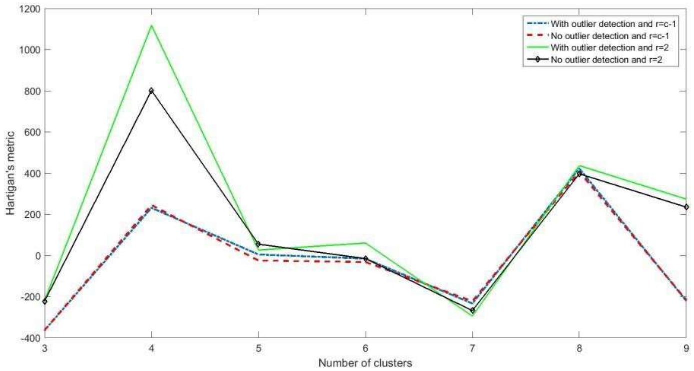
This robustified version of factorial k-means algorithm has two major advantages. First, the clusters are optimally shaped, given that distortions arising from the dimensionality reduction stage are avoided. Second, the outliers are automatically identified during the procedure by the clustering algorithm, with no need of applying any subsequent procedure.

4. Business models and their characteristics

4.1. Clusters and factors identified

The selection of the number of clusters and factors follows the methodology described in Appendix C. Figure 1 presents the Hartigan's statistic for different number of clusters and for two different strategies of selecting the number of factors which represent the lower and upper bounds, respectively. The first strategy is to keep the number of factors fixed and equal to two (2). The second is to use the maximum number of factors, $r = c - 1$. In addition, in both cases we plot the results both for the case where outlier detection is performed and for the case where no outliers are excluded. A common feature of all these lines is that Hartigan's condition is satisfied for $c = 4$ and therefore, we select 4 clusters.

Figure 1: Hartigan's statistic for different number of clusters and factors.



The maximum number of factors that could be used is $\mathrm{c}-1=3$, however when examining the singular values of the n x r matrix $(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{X}\boldsymbol{A}$ (see Vichi and Kiers 2001), it is clear that clusters nearly fall in a subspace with dimensionality lower than 3.[10] Therefore we select $\mathrm{r}=2$, a decision which is further reinforced by our aim to obtain interpretable results.

In addition, we set $\alpha=0.1$, i.e. the 10% of the banks which are identified as outliers represent a separate group. The selection of the quantile value was chosen based on the examination of the set of banks which were selected in the outlier set and on the visual examination of results in the low dimensional space. Specifically, the chosen parameter values lead to a set of outlier banks that contains institutions with idiosyncratic features which are also distinctively far from the clusters' centroids in the low dimensional space, as it will be elaborated later. On the other hand, our classification results are not sensitive to this assumption in the sense that the cluster membership of all the remaining banks is not affected.

The two factors produced by the factorial k-means model consist of a 'level' factor and a 'contrast' (slope) factor. Intuitively, the first factor represents a measure of the presence of standard elements in banks' balance sheets, like loans, deposits, derivatives and issued debts, excluding trading assets. The second factor represents the contrast between loans and 'standard liabilities' (which include deposits and issued debt) and therefore discriminates banks with respect to the imbalance of these standard items on the two sides of their balance sheet.

[10] Specifically, the singular values for three factors were 17.3, 3.7 and 0.4 while for the two-factor case they were equal to 16.0 and 3.

4.2. The business models identified

Figures 2 and 3 present the "median" balance sheet composition of banks in each cluster allowing a better understanding of the composition of activities which characterise the identified business models. To provide intuition we name the business models as follows:

1. *Wholesale funded banks* are generally large banks, their asset side consists mostly of loans (second only to traditional commercial banks, see below), they rely much more than other types of banks on debt for their funding and less on household deposits (see Figure 3). These banks are characterised by far the largest use of derivatives, both for hedging and trading. This cluster contains the lowest number of banks: in total, 58 banks belong to this category.

2. *Securities holding banks* hold a relatively large securities portfolio and cash buffer, fund themselves with deposits and do not use derivatives much. This business model holds the higher amount of cash, mainly to be able to carry out its trading activities. This business model grants the lowest amount of loans (see Figure 2). The liability side of the securities holding banks looks pretty 'traditional' with a significant amount of deposits.[11] They are usually small, but this cluster is the most heterogeneous one as regards their size. The number of banks which follow this model is 86.

3. *Traditional commercial banks* are medium-sized, have loans on their asset side more than all other banks (see Figure 2). These banks fund themselves more with deposits compared to all other business models (see Figure 3) and use derivatives primarily for hedging. They represent the textbook prototype of banks as financial intermediaries. The number of banks contained in this cluster is 77.

4. *Complex commercial banks* are medium sized, possess a significant percentage of loans on their asset side but lower compared to traditional commercial banks because they also own securities to a larger extent, fund themselves mostly with deposits (but less than traditional commercial

[11] Therefore, it is clear that the two types which sometimes are lumped together as 'investment banks', namely the securities holding banks and the wholesale funded ones should be distinguished because their activities differ substantially.

banks) and use derivatives mostly for trading purposes. This is a hybrid category, between traditional commercial and wholesale funded banks. It is the largest cluster and includes 108 banks.[12]

The numbers of banks that are classified in the various categories can be compared with those of RTT and Ayadi et al. (2015). RTT classify most out of the 67 European banks contained in their sample as "retail" with "wholesale funded" following and with "trading" banks representing the lowest number. Our results are in accordance with those of RTT when it comes to retail banks representing the majority of the banking population. There is however a discrepancy with respect to the relative numbers of "wholesale funded" and "securities holding" ("trading") banks, given that in our case the securities holding banks are more than the wholesale funded. This result could be driven by our extra category, namely the complex commercial banks which may include some banks that in RTT could have been labelled as wholesale funded. It may also be due to our larger sample that contains smaller banks that follow the "securities holding" business model. Our results are consistent with those of Ayadi et al. (2015) where retail banks are the majority followed by "investment" banks and with "wholesale" banks representing the minority. Given that the sample of Ayadi et al. is the largest among those compared here (with 2,542 banks from the EEA and Switzerland), it seems plausible that our results are somewhere between those of RTT who consider a small sample and Ayadi et al. in the sense that the importance of securities holding banks seems to increase as the sample gradually becomes larger. There is also a clear correspondence of our business models to those identified by Hryckiewicz and Kozlowski (2017).

[12] We have preferred to label this business model as "complex" rather than "diversified", as the latter label would imply that they are safer against risks. On the other hand, both names refer to the variety of the activities in which these banks are engaged to. The characterisation "universal" could also be fitting for this class of banks.

Figure 2. "Median" asset composition per business model.

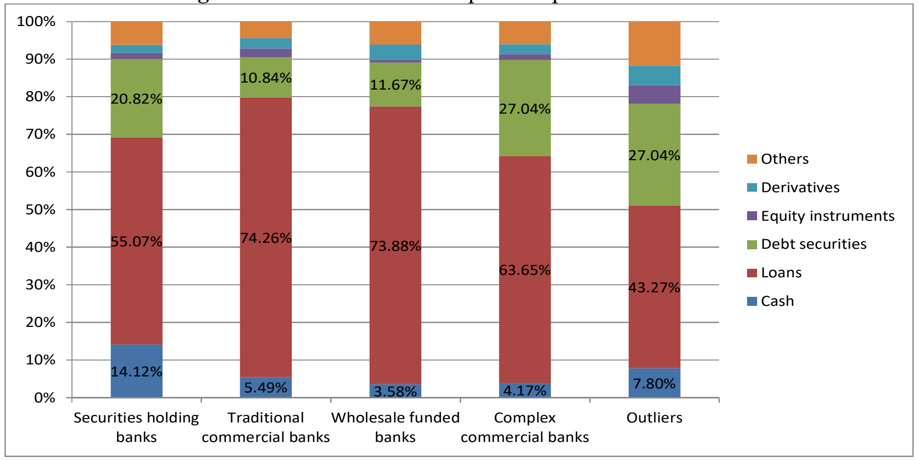


Figure 3. "Median" liabilities composition per business model.

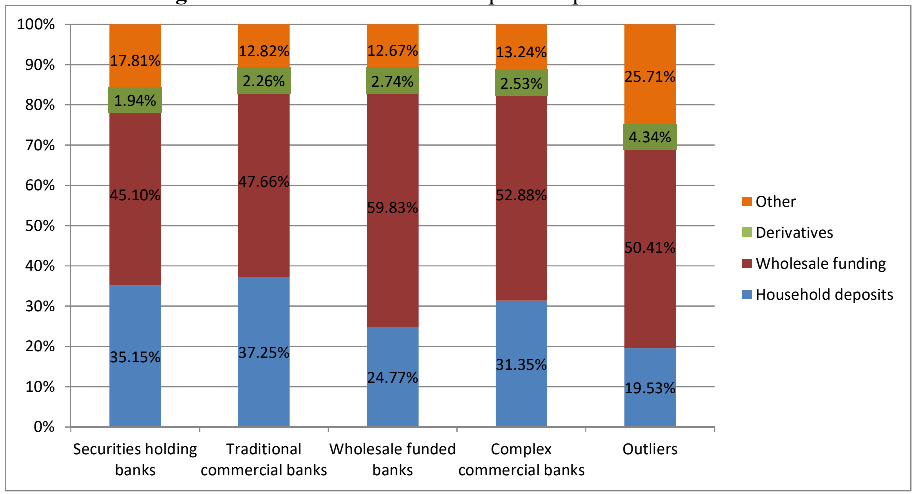


Figure 4 presents a graphical illustration of the positions of all banks and clusters in the factor space while Figure 5 also includes the outlier category in order to show the position of outliers compared to the other classified banks. In addition, we report in Table 1 the centroids of the various clusters in the factor space.

Looking at the relative positions in the x-axis (level factor), the wholesale funded banks are clearly located leftwards compared to all other categories. Both types of commercial banks occupy approximately the same range across the x-axis while the securities holding banks are located on the right of all other types. This relative positioning conforms to the composition of the level factor as explained above. In particular, while both commercial banks and wholesale funded contain similar amounts of loans, deposits and issued debt, they differ with respect to the use of derivatives (higher for wholesale funded banks)[13] and this places the latter at the left end. On the other hand, the securities holding banks are at the right end of the x-axis given the large presence of trading assets and the relatively low presence of loans which lead to low absolute values for the Level factor.

[13] Specifically, the median carrying amount of hedge accounting derivatives on the asset side of the "wholesale-funded" cluster is 1.05% of the total assets while this number is less than 0.25% for the remaining clusters. The carrying amount of derivatives in the "Held for trading" portfolio is 0.96% for the "wholesale-funded" cluster while it is less than 0.57% for the remaining clusters.

Figure 4. Location of banks and clusters in the two-dimensional factor space. In the table below the graph, the coordinates of the centroids position in the factor space is presented.

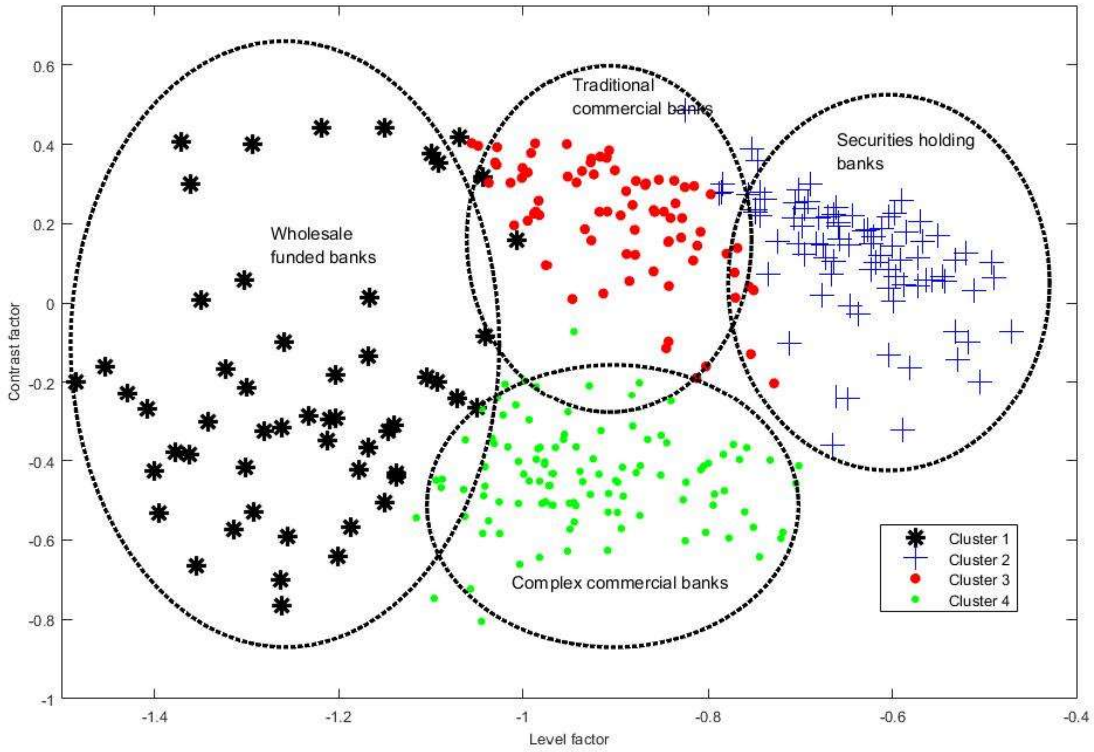


Table 1. Centroids of clusters in the factor space

	Level factor	Contrast factor
1. Whol. funded	-1.07	-0.13
2. Sec. holding	-0.64	-0.04
3. Trad. comm.	-0.82	-0.08
4. Complex comm.	-0.96	-0.16

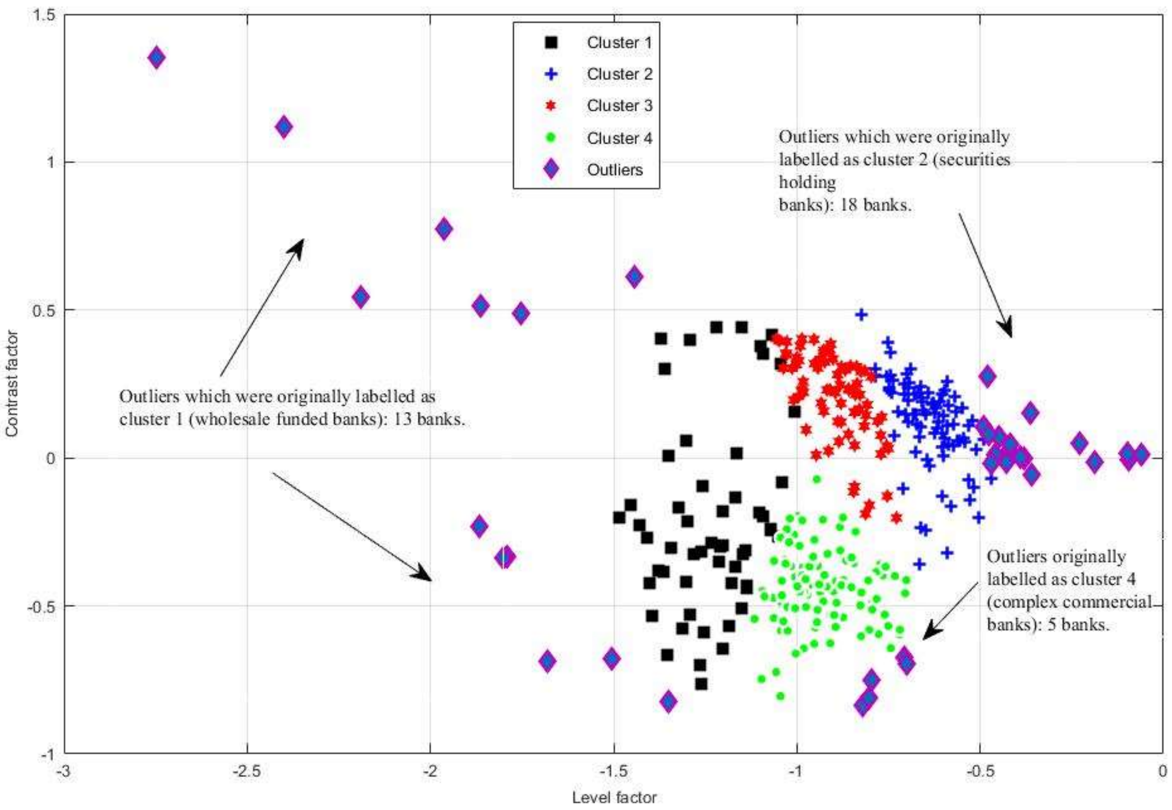
With regard to the relative positions in the y-axis (contrast factor), it is interesting to note that the traditional commercial banks are located higher than the complex commercial ones, due to the more pronounced presence of loans on their asset side (loans enter with a positive sign in the contrast factor). On the other hand, wholesale funded banks occupy a wide range of positions with respect to the y-axis, reflecting varying contrasts between loans and standard liabilities. A seemingly counter-intuitive observation is the high contrast values exhibited by the securities holding banks, despite their relatively small percentage of loans (due to higher levels of trading assets and cash). This is explained when one considers their liability side which also includes a lower level of 'standard' liabilities compared to other types of banks. Specifically, the category "other liabilities" (besides deposits, debt and derivatives) is noticeably higher for securities holding banks. The trading activities are reflected in the high percentage values of this item comprising e.g. amounts payable in respect of future settlements of transactions in securities or foreign exchange transactions.[14]

Figure 5 provides a validation for the outlier component by showing the position of the outliers in the two-factor space. It is clear that the large majority of detected outliers are located distinctively apart from the other classified banks. Further insight into the composition of the outlier set can be gained by examining the initial classification of the banks which end up in the outlier set, before the outlier detection algorithm is applied. Specifically, the outlier banks set is composed of 18 banks which were initially characterised as *securities holding banks*, 13 banks which were initially characterised as *wholesale funded banks*, and 5 banks which were initially characterised as *complex commercial banks*. No bank from the *traditional commercial banks* category was reallocated as an outlier. Therefore, consistently with the qualitative observations above, mainly banks which depart from the standard model of a commercial bank were reclassified into the outlier category.

A closer examination of the set of 'outlier' banks reveals that it includes primarily small investment banks and specialised lenders. For example, we find in this set some local government funding agencies, specialising in providing financing for (semi-)publicly owned organizations and institutions refinancing loans to local public sector entities. Also included are some specialised subsidiaries of larger groups, a bank in a run-down mode and central clearing counterparties. Overall, included in this group are banks following clearly idiosyncratic business models.

[14] Fair valued financial commitments and guarantees are also included under this item – according to anecdotal evidence, such "other liabilities" is relatively more important for the other categories of banks, however this further decomposition is not readily available.

Figure 5. Location of banks and clusters in the two-dimensional factor space, including outliers.



5. Conclusion

We present the first study that makes use of an exceptionally granular data set on European banks in order to infer the types of existing business models. We have adopted a data driven clustering approach which combines optimally the classification of banks with data reduction, enhanced with an 'outlier' banks detection component in order to avoid the 'contamination' of the derived clusters with very specialised institutions. Our approach minimises the impact of the researcher's priors on the results while taking into account the full range of banks' activities and avoiding misclassifications due to the use of a narrow set of balance sheet ratios.

The results provide an anatomy of the Euro area banking sector and indicate the co-existence of distinct business models. We label the four business models identified by our method as "traditional commercial", "complex commercial", "wholesale funded" and "securities holding" banks. Specialised institutions such as state-owned entities aimed at refinancing loans to semi-public and public entities, central clearing counterparties or banks in a run-down mode, are identified as outliers by the clustering algorithm.

The statistical analysis identifies two main factors as the most efficient composite variables to discriminate banks: a level factor representing the presence of "standard" asset and liability items, with the notable exception of trading assets, and a contrast factor which represents the imbalance in the presence of loans on the asset side compared to "standard liabilities" with the latter including deposits and issued debt. These factors represent banks' activities and are robust to the presence of the various forms through which these activities are reflected in the bank's balance sheet e.g. whether a bank provides credit to non-financial corporations via regular loans or credit lines.

Our results reflect the history of macroeconomic developments and policy decisions which have taken place during the crisis period while dynamic effects may also be important e.g. like migration of banks towards different business models, and should be further investigated. For example, both the ECB Financial Stability Review (ECB 2015) and the IMF Global Financial Stability Report (IMF 2015) point to structural business model changes in the aftermath of the global financial crisis, in a context characterised by low inflation, weak profitability and large, albeit heterogeneously distributed, non-performing loans.

These results are highly relevant for monetary policy, micro-prudential banking supervision and the design of macro-prudential policy. Specifically, asymmetries in the monetary policy transmission mechanism across countries, as identified for example by Barigozzi et al. (2014), could be also partly explained by the prevalence of different banks' business models in single countries. Empirical research points to the significance that should be attached to the heterogeneity of banks' business models when designing micro-prudential supervision (Blundell-Wignall et al., 2014) and this is already reflected in the supervision process in many jurisdictions (e.g. see ECB, 2016). It would be of great interest to investigate a possible link between business models and measures of systemic risk that would inform also the calibration of structural macro-prudential capital tools e.g. the systemic risk buffer (SRB) or the other systemically important institutions (O-SII) buffer. Finally, it would be of great interest to investigate whether the business models aspect could shed light on the relationship between competition and risk-taking in banking (Boyd and De Nicolò, 2005) e.g. by influencing the correlation structure of losses (as for example Hakenes and Schnabel, 2011).

References

Albertazzi, U., Gambacorta, L., 2009. Bank profitability and the business cycle. Journal of Financial Stability 5, 393-409.

Altunbas, Y., Binici, M., Gambacorta, L., 2018. Macroprudential policy and bank risk. Journal of International Money and Finance 81, 203-220.

Amel, D. F., Rhoades, S. A., 1988. Strategic groups in Banking. Review of Economics and Statistics 70, no. 4, 685-689.

Anastasiou, D., Louri, H., Tsionas, M., 2016. Determinants of non-performing loans: Evidence from Euro-area countries. Finance Research Letters 18, 116-119.

Angrist, J., Pischke, J., 2009. Mostly harmless econometrics: An empiricist's companion. New Jersey: Princeton University Press.

Ayadi, R., De Groen, W., 2014. Banking business models monitor 2014. Centre for European Policy Studies and International Observatory on Financial Services Cooperatives.

Ayadi, R., De Groen W., Sassi, I., Mathlouthi, W., Rey, H., Aubry, O., 2015. Banking business models monitor 2015. Alphonse and Dorimène Desjardins International Institute for Coopeatives & International Research Centre on Cooperative Finance.

Barigozzi, M., Conti, A., Luciani, M. 2014. Do euro area countries respond asymmetrically to the common monetary policy? Oxford Bulletin of Economics and Statistics 76, no. 5, 693-712.

Beechetti, L., Ciciretti, R., Paolantonio, A., 2016. The cooperative bank difference before and after the global financial crisis. Journal of International Money and Finance 69, 224-246.

Beltratti, A., Stulz, R.M., 2012. The crecit crisis around the globe: Why did some banks perform better? Journal of Financial Economics 105, 1-17.

Berger, A. N., DeYoung, R., 1997. Problem loans and cost efficiency in commercial banks. Journal of Banking and Finance 21, 849-870.

Berger, A. N., Mester, L. J., 2003. Explaining the dramatic changes in performance of US banks: technological change, deregulation, and dynamic changes in competition. Journal of Financial Intermediation 12(1), 57-95.

Berger, A. N., Bouwman, C. H. S., 2013. How does capital affect bank performance during financial crises? Journal of Financial Economics 109, 146-176.

Blundell-Wignall, A., Atkinson, P., Roulet, C., 2014. Bank business models and the Basel system: Complexity and interconnectedness. OECD Journal: Financial Market Trends 2014, vol. 2013/2.

Boyd, J., Gertler, M., 1994. The role of large banks in the recent US banking crisis. Federal Reserve Bank of Minneapolis Quarterly Review 18, 1-21.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis, Communications in Statistics 3, 1-27

Carlson, M., Shatto, M., Warusawitharana, M., 2017. Matching banks by business model, geography and size: A dataset. FEDS Notes, August 8 2017.

Carney, M., 2015. Breaking the tragedy of the horizon – climate change and financial stability. Speech at Lloyd's, London, 29 September.

Caves, R. E., Porter, M. E., 1977. From entry barriers to mobility barriers: Conjenctural decisions and contrived deterrence to new competition. Quarterly Journal of Economics 91, no. 2, 241-262.

Cuesta-Albertos, J.A., Gordaliza, A., Matrán C., 1997. Trimmed k-means: An attempt to robustify quantizers. Annals of Statistics 25, 553-576.

Demirgüç-Kunt, A., and Huizinga, H., 1999. Determinants of commercial bank interest margins and profitability: Some international evidence. The World Bank Economic Review 13, 379-408.

DeSarbo, W.S., Jedidi, K., Cool, K., Schendel, D., 1990. Simultaneous multidimensional unfolding and cluster analysis: an investigation of strategic groups. Marketing Letters 2, 129-146.

De Soete G. and Caroll J. D, 1994. K-means clustering in a low-dimensional Euclidean space, in E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy (Eds.). New approaches in classification and data analysis. Berlin, Springer-Verlag.

Draghi, M., 2016. Introductory statement at the European Parliament's Economic and Monetary Affairs Committee. Brussels, 15 February.

Eichengreen, B., Gupta, P., 2013. The financial crisis and Indian banks: Survival of the fittest? Journal of International Money and Finance 39, 138-152.

Elsas, R., Hackethal, A., Holzhäuser, M., 2010. The anatomy of bank diversification. Journal of Banking and Finance 34, 1274-1287.

Ennis, H., Malek, H., 2005. Bank risk of failure and the too-big-to-fail policy. Federal Reserve Bank of Richmond Economic Quarterly 91(2), 21-44.

European Central Bank, 2015. Financial Stability Review. November 2015. European Central

Bank, 2016. SSM SREP: Methodology booklet. 2016 edition.

Freixas, X., Rochet, J., 2013. Taming systemically important financial institutions. Journal of Money, Credit and Banking 45, 37-58.

Ghosh, A., 2015. Banking-industry specific and regional economic determinants of non-performing loans: Evidence from US states. Journal of Financial Stability 20, 93-2014.

Hakenes, H., and Schnabel, I., 2011. Capital regulation, bank competition, and financial stability. Economic Letters 113, no. 3, 256-258.

Halaj, G., Zochowski, D., 2009. Strategic groups and bank's performance. Financial Theory and Practice 33, 153-186.

Hartigan, J.A., 1975. Clustering Algorithms. New York: J. Wiley & Sons.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data mining, Inference, and Prediction. New York: Springer.

Ho, T., Saunders A., 1981. The determinants of bank interest margins: Theory and empirical evidence. The Journal of Financial and Quantitative Analysis 16, 581-600.

Homar, T., Kick, H., Salleo, C., 2016. Making sense of the EU wide stress test: A comparison with the SRISK approach. ECB Working Paper Series No 1920.

Hotelling, H., 1931. The generalization of Student's ratio. Annals of Mathematical Statistics 2 (3), 360-378.

Hotelling, H., 1933. Analysis of a Complex of Statistical Variables Into Principal Components. Journal of Educational Psychology 24, 417-441 and 498-520.

Hryckiewicz, A., Kozlowski, L. 2017. Banking business models and the nature of financial crisis. Journal of International Money and Finance 71, 1-24.

Hunt, M. S., 1972. Competition in the major home appliance industry, 1960-1970. Unpublished PhD dissertation, Business Economics Committee, Harvard University.

IMF 2015. Global Financial Stability Report. April 2015.

Judd, C., McClelland, G., 2008. Data analysis: A model comparison approach. Routledge, New York.

Klein, N., 2013. Non-performing loans in CESEE: Determinants and impact on macroeconomic performance. IMF Working Paper Series 13/72.

Köhler, M. 2015. Which banks are more risky? The impact of business models on bank stability. Journal of Financial Stability 16, 195-212.

KPMG 2016. Peer analysis: Predicting supervisory challenges. May 2016.

Laeven L., Levine, R., 2007. Is there a diversification discount in financial conglomerates? Journal of Financial Economics 85, 331-367.

Lin, T.-I., McLachlan, G. J., and Lee, S. X., 2016. Extending mixtures of factor models using the restricted multivariate skew-normal distribution. Journal of Multivariate Analysis 143, 398-413.

Louzis, D. P., Vouldis, A. T., and Metaxas, V. L., 2012. Macroeconomic and bank- specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. Journal of Banking and Finance 36, 1012-1027.

Lucas, A., Schaumburg, J., Schwaab, B. 2017. Bank business models at zero interest rates. Journal of Business and Economic Statistics (forthcoming)

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations Proc. Fifth Berkeley Symp. on Math. Statist. and Prob. Vol. 1 (Univ. of Calif. Press, 1967), 281-297.

McLachlan, G.J.; Peel, D. (2000). Finite Mixture Models. Wiley.

Mergaerts, F., Vennet, R. D., 2016. Business models and bank performance: A long- term perspective. Journal of Financial Stability 22, 57-75.

Murray, P. M., Browne, R. P, and McNicholas, P. D., 2014a. Mixtures of skew-t factor analyzers. Computational Statistics & Data Analysis 77, 326-335.

Murray, P. M., McNicholas, P. D. and Browne, R. P., 2014b. A mixture of common skew-t factor analysers. Stat 3, 68-82.

Nkusu, M., 2011. Nonperforming loans and macrofinancial vulnerabilities in advanced economies. IMF Working Paper 11/161.

Porter, M. E., 1979. The structure within industries and companies' performance. Review of Economics and Statistics, 61, 214-227.

Quagliariello, M., 2007. Banks' riskiness over the business cycle: A panel analysis on Italian intermediaries. Applied Financial Economics 17, 119-138.

Reger, R. K., Huff, A. S. (1993). Strategic groups: A cognitive perspective. Strategic Management Journal, 14, 103-123.

Roengpitya, R, Tarashev, N, and Tsatsaronis, K, 2014. Bank business models. BIS Quarterly Review. December 2014, 59-65.

Rousseeuw, P. J., Leroy, A., 2003. Robust Regression and Outlier Detection. John Wiley, New Jersey.

Rousseeuw, P. J., Van Driessen, K., 2000. An algorithm for positive-breakdown regression based on concentration steps. Data Analysis, 335-346.

SNL, 2013. Fundamentals of Peer Analysis, January 2013.

Terada, Y., 2015. Strong consistency of factorial k-means clustering. Annals of the Institute of Statistical Mathematics, 67(2), 335–357.

Tywoniak, S., Galvin, P., Davies, J., 2007. New Institutional Economics' Contribution to Strategic Groups Analysis. Managerial and Decision Economics 28, 213-228.

Vichi, M., Kiers, H.A.L., 2001. Factorial k-means analysis for two-way data. Computational

Statistics and Data analysis 37, 49-64.

Ward, J.H. 1963. Hierarchical Grouping to optimize an objective function. Journal of the American Statistical Association 58, 236-244.

Wheelock, D.C., Wilson, P.W., 2012. Do large banks have lower costs? New estimates of returns to scale for U.S. banks. Journal of Money, Credit and Banking 44, 171-199.

Yang, L., Sijia X., and Weixin Y., 2017. Robust fitting of mixtures of factor analyzers using the trimmed likelihood estimator. Communications in Statistics-Simulation and Computation 46(2), 1280-1291.

Yellen, J. (2012). Testimony before the Committee on Financial Services, U.S. House of Representatives, Washington, 4 November.