EUROPEAN CENTRAL BANK

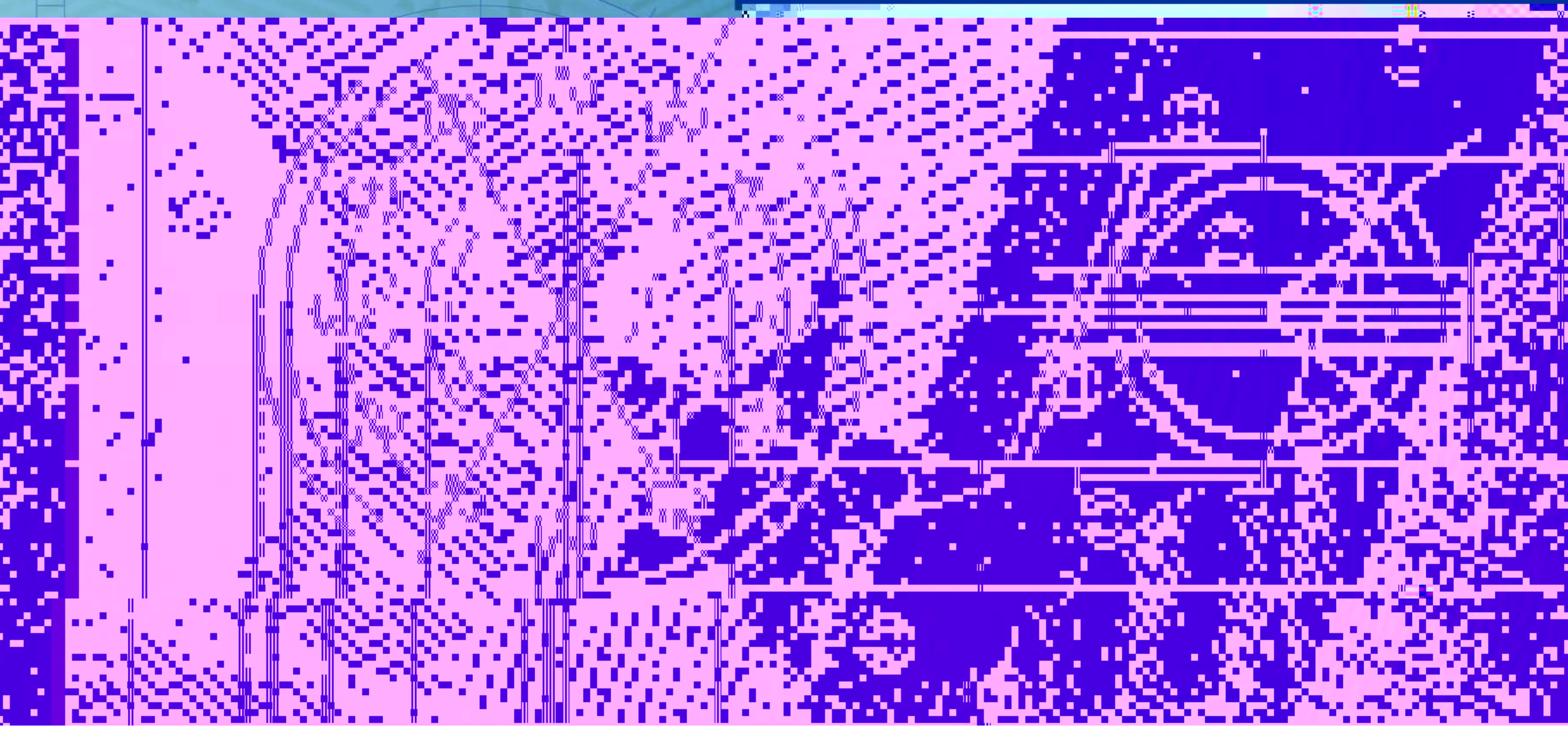
EUROSYSTEM

Working Paper Series

Matteo Farnè, Angelos Vouldis

Business models of the banks in the euro area

No 2070 / May 2017



Abstract

The paper identifies the business models followed by banks in the euro area, utilising a proprietary dataset collected in the context of the supervisory reporting of the Single Supervisory Mechanism. The concept of a 'business model' has been neglected by economic theory and is defined here with respect to the set of activities performed by banks. We adopt a clustering methodology to provide evidence for the existence of distinct business models. Clustering is combined with dimensionality reduction optimally, given the nature of our dataset which features a large number of dimensions for each bank ('fat' data). The method produces a level and a contrast factor which are intuitive in the economic sense. Four business models are identified alongside a set of 'outlier' banks that follow unique business models. The risk and performance indicators of each cluster are examined and evidence is provided that they follow distinct statistical distributions.

Keywords: Banking sector, Business models, Cluster analysis, Single Supervisory Mechanism

JEL classification codes: C63, G21, L21, L25

Non-technical summary

Unlike the textbook description of financial intermediation, whereby banks simply transform deposits to loans, the banks in reality are very much diversified. Banks are involved to varying degrees in a range of activities, both as regards the composition of their profit earning assets that they hold and their funding sources. Given the range of the aforementioned activities and the need of banks to develop knowledge and capabilities to be involved in each of them, banks follow a different models characterised by the selection of and the degree to which they are involved in various activities. It may also be assumed that different business models exhibit also differential behaviour and there may be systematic differences in their performance.

The paper proposes and applies a data-driven methodology to the business model identification problem, especially suited for high-dimensional data. The aim is to minimise the influence from the researcher's priors on the identification of business models, which is confined to the conceptual demarcation of the business model features and, consequently, to the definition of the input set. The methodology builds upon a statistical clustering algorithm enhanced by a procedure to detect 'outlier' banks i.e. banks which use idiosyncratic business models.

Using this methodology the paper identifies the business models followed by the banks in the euro area, utilising a proprietary dataset which is collected in the context of the supervisory reporting of the Single Supervisory Mechanism.

The statistical method produces two composite variables which enable the classification of banks into clusters, representing discrete business models, and are intuitive in the economic sense. Four business models (wholesale funded, traditional commercial, complex commercial and securities holding) are identified alongside a set of 'outlier' banks that follow unique business models. The ensuing classification is validated by examining the risk and performance indicators of each cluster and providing evidence that they follow distinct statistical distributions.

Regarding the characterisation of the business models in the risk-return plane, the securities holding banks exhibit relatively higher returns while also holding relatively high capital buffers (therefore, they excel on RoA outcome) and relatively risky assets.

Wholesale funded banks also hold a risky portfolio, on average, while exhibiting high returns (especially with respect to RoE, given their relatively low capital). On the other hand, the traditional commercial banks hold, on average, the safest assets and they outperform the complex commercial banks although their returns are lower compared to both wholesale funded and securities holding banks. The complex commercial banks seem to present a non-optimal risk-performance combination.

1. Introduction

Unlike the textbook description of financial intermediation, whereby banks transform deposits to loans (as presented e.g. in Ho and Saunders, 1981), the 'ecology' of operating banks is very much diversified. Banks are involved to varying degrees in a range of activities, both as regards the composition of their profit earning assets that they hold and their funding sources.

Specifically, in the textbook business model, banks maximise profits derived from the interest spread between lending and borrowing rates which are charged and paid to 'entrepreneurs' and 'savers' respectively. In reality, however, the source of banks' profits is more diversified including fees and commissions charged for the provision of financial services and also trading. Moreover, the counterparts of these transactions are much more diverse involving consumers, SMEs, large non-financial corporations, other banks, central banks etc. Finally, banks are also involved in a number of other activities, besides loan granting and deposit taking, such as securitisation and hedging, using mainly derivatives. Given the range of the aforementioned activities and the need of banks to develop knowledge and capabilities to be involved in each of them, banks follow a different models characterised by the selection of and the degree to which they are involved in various activities. It may also be assumed that different business models exhibit also differential behaviour and there may be systematic differences in their performance.

Policy makers have used the concept of business models to describe developments in the banking sector. Mark Carney (2015), Governor of the Bank of England, has referred to the need to adapt supervisory practices to the different subsets of banks: "Our supervision is forward-looking and judgement-based. It is risk-based and proportionate – tailored to different business models around the sector". In a similar vein, Janet Yellen (2012), Chair of the Board of Governors of the Federal Reserve System, has used the concept referring to its cross-sectional aspect and noted that "when it comes to bank regulation and supervision, one size does not fit all … rules and supervisory approaches should be tailored to different types of institutions". Mario Draghi (2016), President of the European Central Bank, has used the concept in a dynamic context to refer to the need for banks to change their range of activities in the face of a new macroeconomic environment: “banks may have to do more to adjust their business models to the lower growth/lower interest-rate environment and to the strengthened international regulatory framework that has been put in place since the crisis” (Brussels, 15 February 2016). Furthermore, the Supervisory Review and Evaluation Process (SREP) of the ECB Banking Supervision refers to the ‘business model’ concept in order to assess the overall banks’ risk, referring specifically to its viability and sustainability (ECB 2015). Therefore, policy makers have used the concept to refer to differences in the sets of activities performed by the banks, both at the cross-sectional and the time dimension.

Despite the awareness of academics, market participants and policy makers that distinct business models are observed in reality and that the banking sector analysis should incorporate such differences, a stylised depiction of the banking sector is usually preferred where this heterogeneity is assumed away. The present paper undertakes a statistical analysis providing evidence on the existence of the different business model in the European banking sector and identifying their salient features.

The availability of harmonised data across jurisdictions represents a significant precondition when attempting to classify banks into respective business models. In this direction, the harmonisation of supervisory reporting templates which has been gradually achieved to a large extent in the European Union, also within the general context of harmonising supervisory practices across the world, represents the opportunity to inform this type of analysis with a large, comparable data set, 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 proposes a data-driven methodology to the business model identification problem, especially suited for high-dimensional data. The aim is to minimise the influence from the researcher’s priors, which is confined to the conceptual demarcation of the business model features and, consequently, to the definition of the input set. The methodology builds upon the clustering algorithm proposed by Vichi and Kiers (2001) enhanced by a procedure to detect ‘outlier’ banks i.e. banks which use idiosyncratic business models e.g. state-owned banks refinancing state entities.

Second, a unique data set, which has been made possible by the centralisation of supervision in the European Union and the collection of supervisory data using harmonised definitions, is utilised. The data set presents breakdowns of the banks' balance sheets into four dimensions i.e. accounting portfolios (following the IFRS categories), instruments (loans, debt securities etc.), counterparties (households, non-financial corporations etc.) and products (project finance, credit cards, types of derivatives).

Third, we identify the distinctive characteristics pertaining to each business model as regards risk and performance dimensions. We find that in some cases the distributions of the respective indicators differ substantially across the subsets of banks defined by the business models. These results prove the potential for enhancing the current, aggregate approaches from incorporating the business model aspect into the analysis of the banking sector.

Our aim therefore is to develop a statistical tool which would enable the classification of the banks reporting to the Single Supervisory Mechanism (SSM) into groups ("clusters"), whereby each group contains only banks which are "similar", whereby similarity is defined according to statistical criteria, and taking into account a large number of data for each bank. We design the clustering methodology in a way that these groups should represent banks following distinct business models.

With respect to existing methodologies, our aim is to 'let the data speak', i.e. to minimise the importance of expert judgement in the choice of clustering variables and method. Our problem consists in identifying banks' business models within a high-dimensional space, provided by our proprietary data set, and therefore we employ methods suited in a Big Data context for dimensionality reduction. We derive from this approach a manageable number of variables which possess the highest discriminatory power as regards the clustering of our sample of banks. Existing attempts, both in the academic literature and in policy analysis, rely primarily on expert judgment and on clustering in a low-dimensional space. We focus on an objective, statistical (data-driven) procedure for classification.

The literature on classifying banks into sets according to their business models is scant. The concept of 'strategic groups', introduced by Hunt (1972) and the concomitant concept of 'mobility barriers' (Caves and Porter 1977), introduced to explain persistent performance differentials between firms within one industry, has been 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 in order to examine performance differences. 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.

Roengpitya et al. (2014) also 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). 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).

Even if the method followed by Roengpitya et al. (RTT, henceforth) provides results which are intuitive, it relies significantly on expert judgment; therefore there is the risk that the results are highly influenced by the priors of the researcher. In particular, the choice of input variables and clustering partitions presents a high degree of arbitrariness. In contrast, our aim is to develop a data-driven approach, minimising the subjective component, in order to extract the maximum possible information from our data and let the data determine the final classification of banks. The statistical challenge is to provide such a method in the Big Data context i.e. when a large number of input variables are used, by appropriately reducing data dimensionality in a data-driven way while also performing cluster analysis.

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 and their differences with respect to performance and risk indicators. Finally Section 5 concludes.

2. Review of the literature

The concept of the business model is widely used in management studies but less frequently in economics. The expanding management literature has identified the business model as a new "unit of analysis", like a country, a sector, a firm, while also stressing the holistic nature of the concept and the focus on their activities (see Zott and Amit (2011) for a comprehensive review of the management literature). Teece (2010) argues that mainstream, general equilibrium approaches in economics do not provide any theoretical grounding for the 'business model' concept, given the "ubiquity of theoretical constructs that have markets solving the problems that - in the real world - business models are created to solve" (Teece 2010, p. 175). An important distinction that emerges in this literature is between the business strategy and the business model. According to Magretta (2002) the business model refers to how "the pieces of a business fit together" (Magretta 2002, p. 89) while business strategy is defined with reference to the market competition that a firm faces and the concomitant need to differentiate from its peers.

The concept of 'strategic groups' (Hunt 1972) is akin to that of the 'business model' in the sense that refers to structures within an industry, which are determined by the strategic choices of the firms within a group. However, at the conceptual level, the strategic groups concept is used to analyse industries at the regional or national level and focuses to the varying degree of competition which characterises firms within and across groups (see e.g. Newman 1978). The business model concept on the other hand can be used to characterise firms operating in different jurisdictions which do not compete with each other. On the other hand, the empirical methodologies which can be used to analyse sectors using both concepts are similar, as will be elaborated below.

The evolutionary literature on the analysis of the firm goes beyond the neoclassical assumption of firms optimising uniformly across a sector (in both the perfect competition or oligopolistic versions) and provides insights which are relevant for thinking about the emergence of business models. The seminal contribution by Nelson and Winter (1982) provided the impetus for evolutionary analyses of corporate and industrial change which assume learning through practice and bounded rationality and are able to accommodate the coexistence of different business models. Nelson and

Winter link the persistence of “routines” within firms to the problem of assessing and storing knowledge. Moreover, they note that the production set can only be defined over the set of activities or techniques that a firm knows how to operate. This set is not identical across firms, and therefore their respective activities differ (see Nelson et al. 1976 and the literature originating from this work).

Consistently with the existing economics and management literature, we define the business model with respect to the *activities* that a firm undertakes. The dimensions which specify in further detail its mode of operation, like the degree of efficiency, pricing policy, effectiveness, revenues, costs, which refer to its business strategy and the outcomes of the business model and strategy do not form part of the business model determinants.

There is a limited number of empirical studies based on the business model concept. Burt et al. (2015) study the sector of international retailers and associate the business model with the underlying ‘activities, processes, behaviours and outputs’. They consider how firms adapt their business models into different institutional environments and illustrate this process with a case study. 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. Ayadi and de Groen (2014) also define banks’ business models based on their activities. They examine a set of 147 European banks and select six dimensions (loans, trading assets, liabilities to other banks, customer deposits, debt liabilities and derivative exposures) to perform an hierarchical clustering. Both RTT and Ayadi and de Groen, clearly distinguish between “activities”, which determine the business model, and “outcomes”, the latter measured by profitability and performance indicators.

The distinction between choice and outcome variables described above is also consistent with the strategic groups literature strand whereby a set of variables reflecting strategic choices are used to classify banks into strategic groups while the existence of differential performance is investigated ex post. The empirical strategies followed to classify banks into strategic groups also usually focus on balance sheet 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.[3] Different variants of clustering algorithms have been used in these studies, but in all cases the dimension of the input set was small and always much smaller than the number of banks examined which is the standard set-up for most clustering methodologies.

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. Other studies usually restrict themselves either to qualitatively describe the outcome of the classification procedure or compare results with widely accepted external classifications or simply investigate only performance differences.

Clustering methods represent the statistical technique used for grouping banks into business model types or strategic groups. Clustering methods belong to the class of unsupervised learning methods (Hastie et al. 2009, pp. 485-586). There is a variety of clustering methods which can be effectively implemented for different purposes, taking into account the nature of our data.

Agglomerative hierarchical methods like the Ward's clustering method (Ward 1963), which minimises the variance within clusters, usually gives spherical clusters. This method is employed by RTT but is not particularly suitable in a high dimensional context (the so called "fat data" setup, where the number of variables can be larger than the number of observations), since it is not so easy to characterize clusters on the base of a very high number of variables. Therefore, RTT select a priori subsets of eight variables representing bank assets and liabilities, excluding highly correlated variables.[4] A similar methodology is used in the study of Ayadi and de Groen (2014) while in the strategic groups literature, the input variables set is also low-dimensional. The Ward's

[3] Reger and Huff (1993) should be considered separately in this strand of the literature as it focuses on the cognitive dimension of the managers and utilises data originating from interviews with bankers.
[4] 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.

method is intuitive but it relies significantly on expert judgment and on the arbitrary selection of a small set of input variables based on the researchers' priors.

In the direction of classifying large dimensional objects, clustering methods which incorporate a dimension reduction process have also been proposed. For this purpose, PCA-based or factor-analysis based methods could be employed since they extract a small set of (uncorrelated) variables, named factors, from a large set of correlated ones. The factors enable to optimally compress the information content of the initial large set of variables. Another class of methods is the centroid-based e.g. as proposed by MacQueen (1967). In our dataset the objects (banks) possess a large number of dimensions, and our aim to exploit the richness of the information contained, without the imposition of our priors about the variables which should be used to determine the clusters. The nature of our data set necessitates a careful consideration of the clustering methodology, as will be explained in Section 3.1.

3. Methodology

3.1. Input set

We select the input set based on the conceptual framework for a business model as defined above i.e. comprising of "choice variables" (in the terminology used by RTT) meaning variables which reflecting strategic choices made by the banks' management with respect to the activities undertaken by the bank as reflected in the composition of the balance sheet.

We use for our study a set of proprietary supervisory data which are collected in the context of the SSM i.e. in the new architecture of banking supervision characterised by a step towards centralisation of supervision for a set of systemically significant banks. These supervisory data have been codified by the European Banking Authority (EBA) and employ harmonised definitions, thus representing an ideal set for a comparative analysis across countries.

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, on a prudential scope of consolidation. Our sample consists of 365 banks residing in the 19 euro area countries. All systemically significant banks, as defined by the SSM (based on their absolute and within-country, relative size), are included in this sample. Our data set is cross sectional with reference date the last quarter of 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. On the other hand, risk/solvency (i.e. Common Reporting - COREP) and performance variables (income variables in FINREP) could be viewed as "outcome" variables i.e. the results of the "choice" variables. We expect that these risk and performance metrics will exhibit identifiable difference across the resulting clusters of banks.

In particular, we use data from the FINREP templates which contain information on the banks' balance sheet composition under four breakdowns, specifically accounting portfolios (see below), instruments (loans, securities etc), counterparties (households, non-financial corporations etc) and products (mortgage loans, credit cards etc). We do not impose any priors as regards the type of breakdown which is more relevant in identifying business models.

We start from an initial set of 1039 variables for each bank and n=365 banks. Appendix A presents more information on the input data set.

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', using some measure of distance, 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.

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 possible to make distributional assumptions for the data in our setting. Consequently, density-based clustering methods, which are based on normal mixtures, are hard to be implemented. Alternative methods, like hierarchical and partitioning (i.e. centroid-based) methods, do not provide any dimension reduction by themselves.

In RTT, banks are clustered according to the Ward (1963) hierarchical clustering algorithm. Hierarchical clustering algorithms produce a tree of nested partition, whereby the base of the tree contains n clusters with 1 member and the top contains one cluster containing all the n elements of the object set. As explained in Section 2, this approach does not encompass a dimensionality reduction mechanism and the clustering is performed in a restricted space of variables (n>>p) which have been already selected based on the priors of the researcher.

The most obvious way by which dimension reduction issues can be incorporated into a clustering methodology could be through applying a principal component analysis or a classical factor analysis before conducting the clustering. These techniques provide an r-ranked approximation of the sample covariance matrix (Hotelling 1933) in the initial, large input set. 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, pp. 212-219) and De Sarbo et al. (1990), this approach may not be the most efficient for the 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.

Research in the class of clustering techniques labelled as “k-means” techniques has produced algorithms which can incorporate dimension reduction within the classification schema. K-means clustering is a partitional, as opposed to the hierarchical, clustering method i.e. a method producing an unnested classification of the given objects into clusters. The name k-means refers to the prototype-based nature of this class of algorithms, which work by iteratively identifying the (k) centroids of each cluster as the means of the participating objects.

The standard k-means algorithm (MacQueen, 1967) does not provide any low-dimensional representation of the data, however this feature has been incorporated in subsequent versions of the algorithm. For example, De Soete and Carrol (1994) develop a method for clustering objects in a low-dimensional space while simultaneously identifying the subspace of the observed data having the largest clustering power. However, this method is not optimal when the clustering power is concentrated in directions which are orthogonal to the identified ones, because the latter may offer poor explanation of the overall variance.

Vichi and Kiers (2001) develop the factorial k-means algorithm, where a subspace is defined such that the projected data 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.

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

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

where $\boldsymbol{A}$ is a p $\times$ r column-wise orthonormal matrix (coefficients matrix), U is a n $\times$ 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 $\times$ r matrix $\overline{\boldsymbol{Y}}$ contains the centroids of the cluster 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-rank space. The low-dimensional spaced is the one spanned by the columns of the column-wise orthonormal matrix $\boldsymbol{A}$. Consequently, the model can be specified as follows

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

where E is a residual matrix. Equation (2) describes the partition in the original space.

The optimal partition therefore is sought by minimising

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

which can be equivalently expressed as

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

since $\bar{Y} = (U'U)^{-1}U'XA$. This minimisation is performed under the constraints that A$'$A=I_r and U is binary with only one non-zero element per row.

The above formulation of the clustering problem searches in the low rank space for a partition such that the objects are assigned to the group with the closest centroid. This is equivalent to seeking for the partition of objects having the lowest within-class deviance. 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.

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 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 to choose cannot be larger than $c-1$. The reason is that

$Rank((U'U)^{-1}U'XA)) = \min(c-1, r)$. 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.

The algorithm for solving the minimization problem (Vichi and Kiers, 2001), belongs to the class of Alternated Least Squares (ALS) algorithms.

In Step 1, we minimize $F(\boldsymbol{A}, \boldsymbol{U}, \overline{\boldsymbol{Y}})$ with respect to U given the values of $\boldsymbol{A}$ and $\overline{\boldsymbol{Y}}$. For each row i of U we set $u_{ij} = 1$, if $F(A, u_{ij}) = \min\{F(A, [u_{iv}]): v = 1, \dots, m\}$ and $u_{ij} = 0$ otherwise.

In Step 2, $F(\boldsymbol{A}, \boldsymbol{U}, \overline{\boldsymbol{Y}})$ is minimized keeping fixed U, to update *jointly* $\boldsymbol{A}$ and $\overline{\boldsymbol{Y}}$. Among all the linear combinations of $\boldsymbol{X}$, the ones closer to the centroids (in the transformed space) are derived by taking the first m eigenvectors of $\boldsymbol{X}'(\boldsymbol{U}\,(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}' - \boldsymbol{I_n})\boldsymbol{X}$ (see Ten Berge 1993). From the optimal $\boldsymbol{A}$, we then update $\overline{\boldsymbol{Y}}$ from the expression $\overline{\boldsymbol{Y}} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{X}\boldsymbol{A}$.

In Step 3, we compute $F(\boldsymbol{A}, \boldsymbol{U}, \overline{\boldsymbol{Y}})$ for the current values of $\boldsymbol{U}$, $\boldsymbol{A}$ and $\overline{\boldsymbol{Y}}$. If F has decreased, we go again with Step 1 and 2. Otherwise, the process has converged. In the latter case, we keep the values of $\boldsymbol{U}$, $\boldsymbol{A}$ and $\overline{\boldsymbol{Y}}$ from the previous iteration.

The algorithm as described above starts with some given values of $\boldsymbol{A}$ and $\overline{\boldsymbol{Y}}$. In order to avoid being trapped by a local optimum, we have to cover the parameter space. For this reason, a procedure for choosing a number of random initial points in a rational way is applied.

Matrix $\boldsymbol{A}$ is initialized in the following way. First, K=100 random permutation matrices P_k, $k= 1, \dots, K$ are generated. The Gram-Schmidt algorithm is applied on each of these matrices, in order to make them orthogonal. Our input matrix is the 382 × 382 unbiased sample covariance matrix $C = \frac{1}{n-1}X'X$. Its spectral decomposition is $V_r\ D_r V_r{}'$, where V_r is the matrix of eigenvectors and D_r is the matrix of eigenvalues. Thus, we post-multiply the random permutation matrix P_k by V_r. We run this initial procedure for K=100 times obtaining the same number of initial estimates $A_{0,k} = P_k V_r$, $k= 1, \dots, K$.

For each of the 100 random $A_{0,k}$, the group matrix U is initialised by computing for each bank the relevant quantity $t_i = n diff_i (C_F)^{-1} diff_i', i = 1, \dots, 365,$ where C_F is the computed covariance matrix on the columns of $\mathrm{F}_{0,k} = \mathrm{X}A_{0,k}$, and we set $diff_i = \mathrm{F}_{0,\mathrm{k,i}} - \overline{\mathrm{F}}_{0,\mathrm{k}}$ for k= 1, …,100 (denoting by $\mathrm{F}_{0,\mathrm{k,i}}$the i-th row of the matrix $\mathrm{F}_{0,k}$, and by $\overline{\mathrm{F}}_{0,\mathrm{k}}$ the mean across rows). Then, given the number of groups c, the 2 c quantiles of the distribution of t_i across banks are computed. The distance between each score in the low-dimensional space (F_i) and the first, third, fifth and seventh (i.e. the odd quantiles, $1,3, \dots 2c - 1$ in general) quantile is computed. So, each bank is assigned to the closer quantile, thus originating c initial groups for each A. The cluster corresponding to the seventh quantile is smaller than the other three, which are instead similar.[5] Finally, at each run r, the centroids $\overline{Y}$ are updated as $(U'U)^{-1}U'XA$.

3.3. *Robustified clustering with simultaneous multivariate 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 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. The term ‘outliers’ here refers to banks following unique business models. Detection of outlier banks represents a by-product of our classification methodology.

As well established in literature (see e.g. Rousseeuw and Leroy, 2003), the most used method for detecting multivariate outliers is via the Mahalanobis distance, $D = \sqrt{(x - \overline{x})\,'S^{-1}(x - \overline{x})}$, with D^2 being asmptotically 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 $t^2 = n(x -$

[5] The means of the percentages of cluster belongings across our 100 runs are, respectively 0.260, 0.254, 0.269, and 0.217. We can see that the last group tends to be penalised, while the third is the largest by a small margin. This is coherent with the empirical distribution of t_i across banks.

$\overline{x})'S^{-1}(x-\overline{x})$, 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 against rather than in a p-dimensional space.

For this reason, a method is developed here to overcome this issue, under the assumption that scores in the transformed space $\boldsymbol{XA}$ have different means across clusters but the same covariance matrix (which is equivalent to assuming that the clusters have similar within-variance). The rationale behind the method is the need to find the partition of the $100*(1-\alpha)\%$ most concentrated objects respect to the scores in the low dimensional space.

Let us suppose we have estimated the reduced space $\hat{\mathrm{F}}=X\hat{\mathrm{A}}$, which is a n x r matrix. We can exploit these low-dimensional representations of the objects for outlier detection purposes. If we call $\hat{\mathrm{F}}_i$ the i-th row of the matrix $\hat{\mathrm{F}}$, and $\overline{\mathrm{F}}_\mathrm{i}$ the mean of F across the cluster to which the i-th observation belongs, we can define $\mathrm{diff_i}=\mathrm{F_i}-\overline{\mathrm{F}}_\mathrm{i}$ for each $i=1,\dots,365$. We thus compute $\mathrm{C_F}$, the unbiased sample covariance matrix for these low-dimensional data.

The quantity $\mathrm{t_i}=\mathrm{n\ diff_i}(\mathrm{C_F})^{-1}\mathrm{diff_i'}$ is the relevant one to define the outliers. Under the normality assumption for X, we would have $\mathrm{t}\sim\mathrm{T_{p,n-1}}^2$ which is called Hotelling's T (1933). In this case, since this assumption is violated, we can compute these values for all observations, and then derive, for instance, as a threshold the $100*(1-\alpha)\%$ -th percentile of the empirical distribution of $\mathrm{t_i}$ across the sample. This value is our empirical quantile. The observations having a value for $\mathrm{t_i}$ exceeding it are flagged as outliers.

This calculation is included directly in the iterative clustering procedure. At each step, if an observation o_i is identified as an outlier, we set $u_{ij} = 0$ for j= $1, \dots, c$, such that the observations pervasively distant from the centroid of any cluster are identified. In this way, outliers are excluded from the computation of the coefficient matrix A and the centroid matrix $\overline{Y}$.

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. Results and discussion

4.1. Clustering results

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. 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. 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 when an initialisation is specified. However, given that we run the clustering algorithm with a number of initialisation in order to search in the space of solutions and that we use the covariance matrix of the input set for this purpose, the presence of linear dependencies among the input data is not welcome as the sample covariance matrix becomes ill-conditioned.

We follow a procedure to minimise the redundant information present in the input data set. To achieve this we perform a selection of variables according to their 'importance', measured for each variable as the sum of the absolute correlations respect to all the others. This is intended to avoid identities (i.e. perfectly collinear variables) and redundant variables, excluding the less important among the most correlated variables. Taking also into account also the fact that some of the initial data points were very sparsely populated, we narrowed down the initial set of 1039 data points into a final input set of 382 data points (see Appendix B).

Our sample consists of a cross section of n=365 SSM banks. All institutions labelled as ‘Significant Institutions’ in the context of the SSM are included in the sample. The reference date of the data is 2014 Q4.

The number of clusters and the number of factors have to be determined simultaneously in our approach. For this selection we used a step-wise approach, utilising a partition-based criterion to select the number of clusters combined with search in the number of factors space and then selecting the number of factors. It has to be noted that in the current setting which involves simultaneous clustering and dimensionality reduction, the approaches of deciding on the number of factors based on the initial data set and its covariance matrix are not suitable. In contrast, our adopted approach involves first selecting the number of clusters as recommended in Vichi and Kiers 2001. Consideration of the ensuing range of cluster numbers and their interpretability is also a consideration when using r.

A measure of the fitness of a specific clustering is given by the sum of within-cluster distances to centroids:

$$W_c(P, \overline{\boldsymbol{Y}}) = \sum_{j=1}^{c} \sum_{i \in P_j} d(i, \bar{y}_j)$$

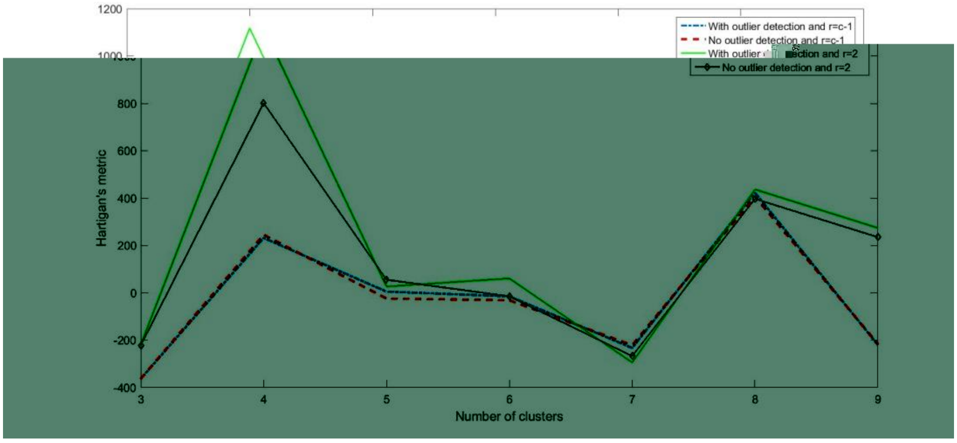
where $P=\{P_1, P_2, \ldots, {}_c\}$ is$P$a partition of c clusters and d is the Euclidean distance measure. Hartigan (Hartigan 1975) has proposed a heuristic rule to select the number of clusters which has been shown to be effective in subsequent simulation studies (see e.g. Chiang and Mirkin 2010). The idea of the method is that when the optimal c* is used a decrease of W with respect to c<c* will be observed because “coarse” clusters defined by c<c* will be split further while when c>c* the value of W will be relatively less volativel given that “optimal” clusters will be split in a random way. Hartigan suggested the calculation of the metric

$$H_c = (W_c / W_{c+1} - 1)(p - c - 1)$$

and selecting c when a large increase in this metric is observed when $c+1$ are used.[6]

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 that 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 c – 1 =3, however when examining the singular values of the n x r matrix $(U'U)^{-1}U'XA$) (see Vichi and Kiers 2001), it is clear that clusters nearly fall in a subspace with dimensionality lower than 3.[7] Therefore we select $r = 2$, 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. With the chosen value, it was found that the set of outlier banks indeed contains institutions with idiosyncratic

[6] Hartigan suggested specifically that when $H_c > 10$, the number of clusters should be selected to be equal to $c+1$. Chang and Mirkin (2010) find that this criterion works well also for different values than 10.

[7] 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 3,1.

features and that these institutions were 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 as usual in such cases. 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) therefore discriminates banks with respect to the imbalance of these standard items on their two sides of their balance sheet. Given that all variables in the input set are expressed as a percentage of assets, banks' size does not distort the factor composition.

We obtain a deeper understanding of the economic meaning of the factors by looking at the variables which are more highly correlated with them.[8] For the factor 1, we present in Table 1 the variables with correlation absolute values above 0.4. For the factor 2, we present in Table 2 the variables with correlation absolute values above 0.3.

After investigating the common characteristics of banks belonging into the four identified business models we label the clusters as follows:

Wholesale funded banks (Cluster 1) are generally large banks, their asset side consists mostly of loans (second to Cluster 3, see below), they rely much more than other types of banks on debt for their funding and less on household deposits. 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.

Securities holding banks (Cluster 2) have a relatively large securities portfolio and cash buffer, fund themselves with deposits and do not use derivatives much. They are usually small, but this cluster is the most heterogeneous cluster as regards their size. The number of banks which follow this model is 86.

[8] Alternatively we could analyse the economic content of the factors by looking at the respective weights of the variables. However this may be misleading because the variables are unscaled with respect to each other, therefore the correlation measure as used in the text is preferable.

Table 1: Variables with the maximum (absolute value above 0.4) correlation with the level factor

Variable	Correlation with the level factor	Data point in ITS templates
Loans and advances (Assets, carrying amount)	-0.465	{F01_01_r200_c010}
Derivatives – hedge accounting (Assets, carrying amount)	-0.419	{F01_01_r240_c010}
Derivatives – hedge accounting (Liabilities, carrying amount)	-0.467	{F08_01a_r010_c037}
Debt securities issued (Liabilities, amortised cost)	-0.418	{F08_01a_r360_c030}
Loan commitments given (Off-balance sheet exposures, nominal amount)	-0.496	{F09_01_r010_c010}
Loan commitments given - households (Off-balance sheet exposures, nominal amount)	-0.440	{F09_01_r080_c010}
Derivatives – hedge accounting: Interest rate OTC derivatives, except OTC options (Notional amount of total hedging)	-0.5296	{F11_01_r030_c030}
Derivatives – hedge accounting (Notional amount of total hedging)	-0.656	{F11_01_r500_c030}
Derivatives – hedge accounting: of which OTC with credit institutions (Notional amount of total hedging)	-0.630	{F11_01_r510_c030}

Table 2: Variables with the maximum (absolute value above 0.3) correlation with the contrast factor

Variable	Correlation with the contrast factor	Data point in ITS templates
Loans and advances - households (Assets, carrying amount)	0.364	{F05_00,r80,c60}
Deposits (Liabilities, contractual amount at maturity)	-0.883	{F08_01a,r50,c50}
Debt securities issued (Liabilities, contractual amount at maturity)	-0.322	{F08_01a,r360,c50}

Traditional commercial banks (Cluster 3) are medium-sized, have loans on their asset side more from all other banks, fund themselves mostly with deposits 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.

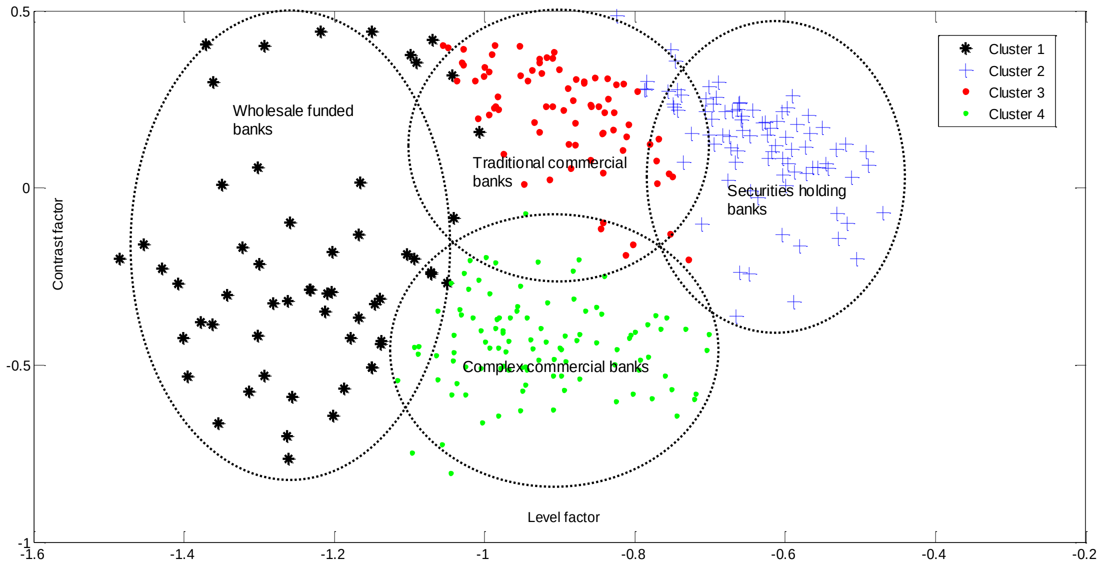
Complex commercial banks (Cluster 4) are medium sized, possess a significant percentage of loans on their asset side but lower compared to Cluster 3 banks because they also own securities to a larger extent, fund themselves mostly with deposits (but less than Cluster 3 banks) and use derivatives mostly for trading purposes. This is a hybrid category, between Clusters 3 and 1. It is the largest cluster and includes 108 banks.

Finally, 36 'outlier' banks were identified. 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, an institution in a run-down mode and central clearing counterparties. Therefore, this is a highly heterogeneous set of institutions which are idiosyncratic and would contaminate the other clusters had they remained within them.

Further insight into the composition of the outlier set can be gained by examining the initial classification of the banks which were re-classified by the outlier detection algorithm to the outliers set. 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.

Figure 2 presents a graphical illustration of the positions of all banks and clusters in the factor space. In addition, we report the centroids of the various clusters in the factor space below the Table. Moreover, Figures 4 and 5 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.

Figure 2. 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.



Centroids of clusters in the factor space

	Level factor	Contrast factor
Cluster 1	-1.07	-0.13
Cluster 2	-0.64	-0.04
Cluster 3	-0.82	-0.08
Cluster 4	-0.96	-0.16

Looking at the relative positions in the x-axis (level factor), it is noted that the wholesale funded banks are located leftwards compared to all other categories. Both types of commercial occupy approximately the same range across the x-axis while the securities holding banks are clearly located rightwards relative to all other types. This relative positioning conforms to the composition of the level factor as explained above. In particular, while commercial banks and wholesale funded, all contain similar amounts of loans, deposits and issued debt, they differ with respect to the use of derivatives (higher for wholesale funded banks)[9] 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.

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

[9] 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.

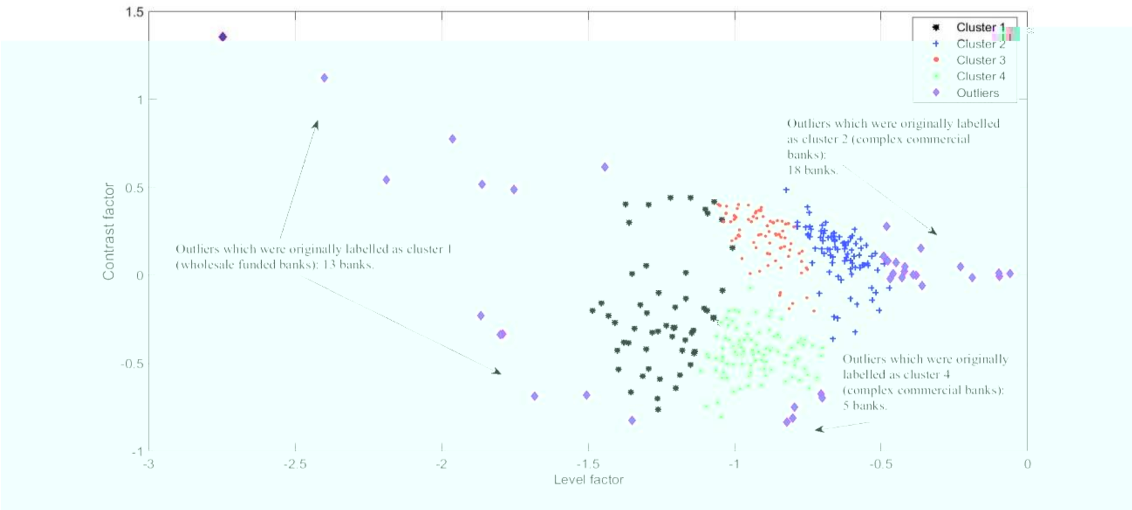
commercial ones, due to the more pronounced presence of loans on their asset side (loans enter with a positive sign in the contrast factor). On the other hand, wholesale funded banks occupy a wide range of positions with respect to the y-axis, reflecting varying contrasts between loans and standard liabilities. 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 of this item like e.g. amounts payable in respect of future settlements of transactions in securities or foreign exchange transactions.[10]

Regarding the "outlier" group, small investment banks and specialised lenders were identified. For example, some state-owned banks which provide refinancing for loans to public (and semi-public) entities are included in this group as is one clearing house. The ability of the algorithm to detect automatically institutions which follow idiosyncratic models based on quantiles of the identified clusters is only made possible by the richness of the input set and is reassuring as regards the reliability of the obtained results. Figure 3 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.

Figures 4 and 5 present an alternative form of description of the identified clusters. In particular the "median" composition of the two sides of the balance sheets for each cluster is shown, enabling the definition of 'benchmark' banks' balance sheets for each business model. In Figure 2 it is shown that traditional commercial and wholesale funded banks have more than 70% of their assets in the form of loans. The complex commercial banks possess a lower percentage of loans while the securities holding banks even less. The latter category holds the higher amount of cash, mainly to be able to carry out its trading activities.

[10] 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 3. Location of banks and clusters in the two-dimensional factor space, including outliers.



On the liability side, the traditional commercial banks hold the larger amount of deposits and the wholesale funded banks the larger amount of wholesale funding, as one would expect. The liability side of the securities holding banks looks pretty 'traditional' with a significant amount of deposits. 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.

Figure 4. "Median" asset composition per cluster

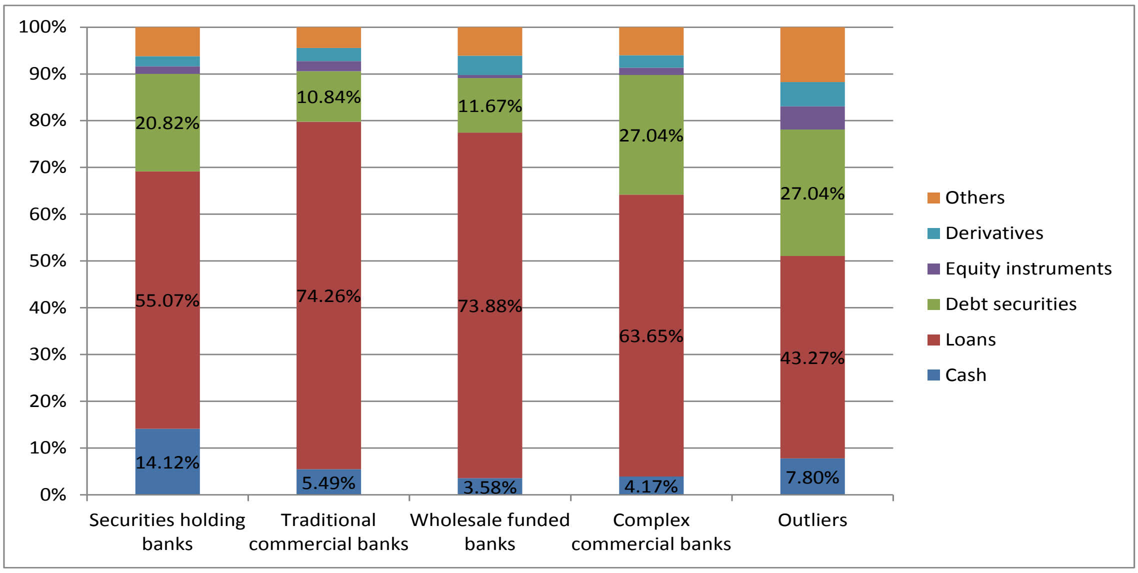


Figure 5. "Median" liabilities composition per cluster

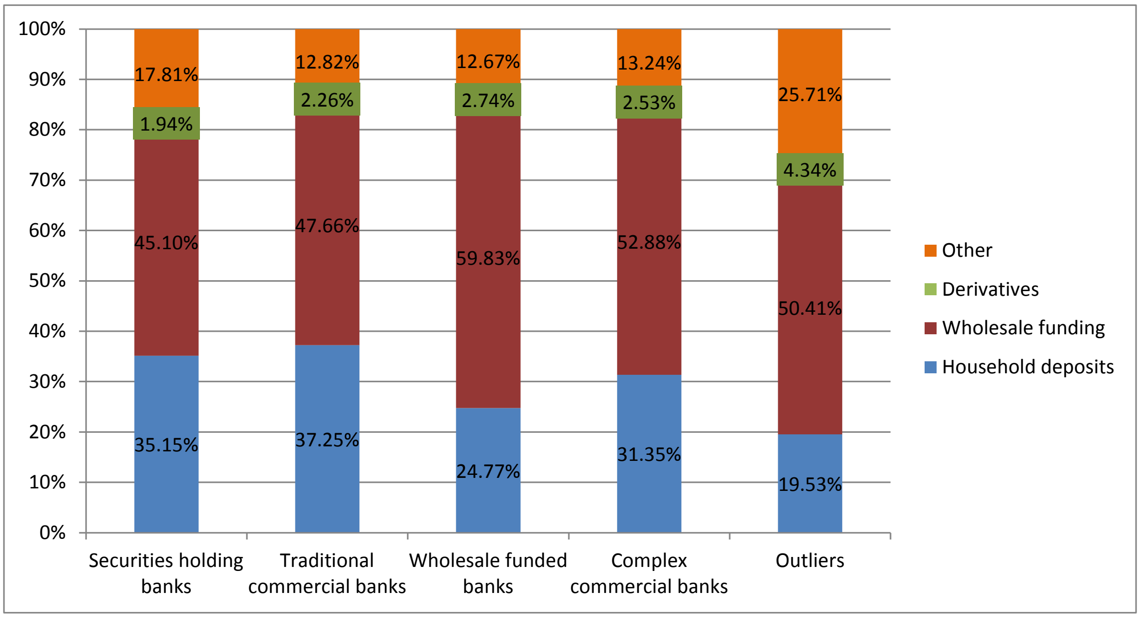
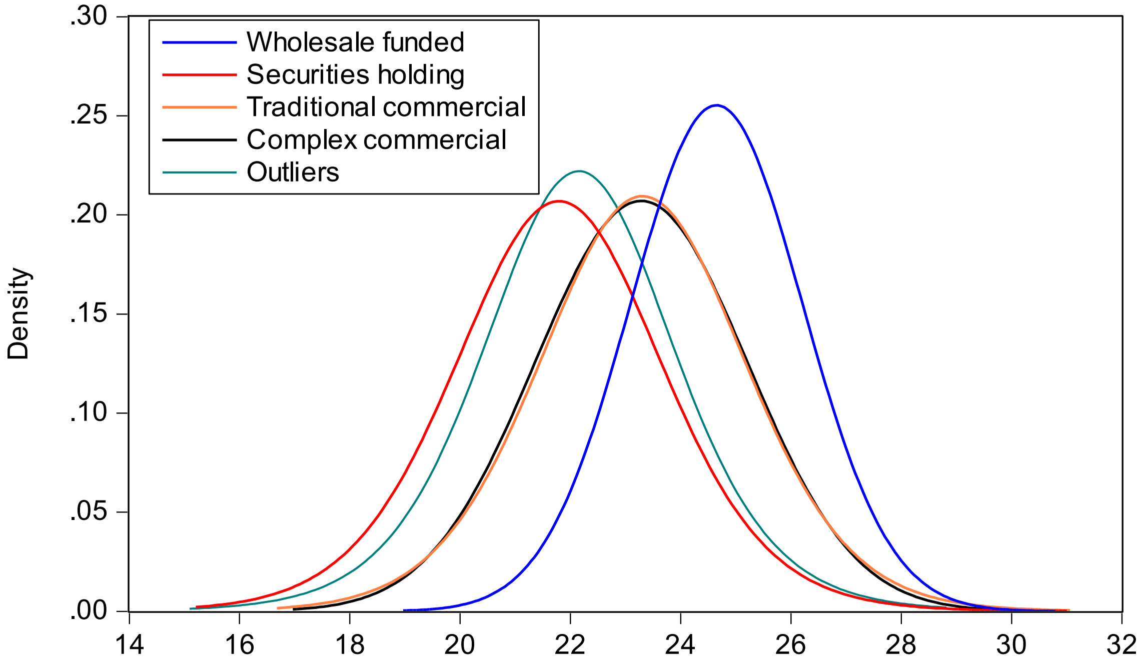


Figure 6 looks in more detail on the size distributions of each cluster. The fitted t-distributions of size for each cluster are depicted.[11] In general there is heterogeneity within each cluster regarding the sizes of the participating banks i.e. size is not a clear determinant of the business model. However, there are broad patterns, like for example the larger, on average, size of wholesale funded banks. The two types of commercial banks are in general smaller and very similar among themselves. Finally, the securities holding banks are on average the smallest ones from the four clusters. The outlier group consists also from generally small banks. Overall, there is no evidence that there is a straightforward relationship between the size and business model, but there seems to be a tendency for large banks to be mainly loan-granting institutions as opposed to having a sizeable securities portfolio.

[11] These results are robust to the distribution assumption. Specifically, fitting a t or a normal distribution to the population distributions leads to qualitatively similar results.

Figure 6. Fitted t-distributions of the log(size) per cluster



4.2. Comparative analysis of the performance of different business models

We investigate the effect of the business model adopted by banks on a number of 'outcome' variables. The outcome variables under examination fall into the following categories: leverage and solvency indicators, risk decomposition, credit risk, performance and efficiency and profit sources. Overall they capture the most important aspects of banks' behaviour in the dimensions of risk-taking, performance and efficiency.

We will comment below in detail on the aggregate features of each business model based on the selected indicators. These results have been obtained after removing the outliers, using the procedure described in Section 3.3. The group of outliers will be commented as a separate category, although due to the underlying methodology, it is not accurate to consider this group as a separate business model, rather it should be considered as an 'others' category. For each indicator, the mean, median, standard deviation and coefficient of variance is calculated. For each indicator, the mean, median, standard deviation and coefficient of variance is calculated. In addition, the nonparametric Wilcoxon rank sum test is performed to test for the pairwise equality of the population medians. The number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level is indicated – representing the 'uniqueness' of each cluster's distribution for each specific indicator.

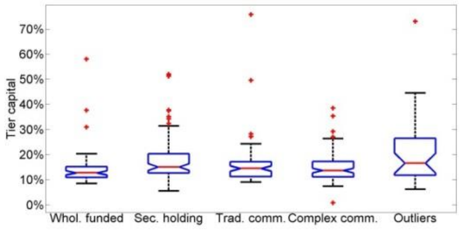
As expected, the securities holding banks hold, on average, the largest amount of capital, irrespectively of whether we measure capital by the risk-adjusted metrics (Tier 1 Capital ratio and the Total Capital Ratio) or the leverage ratio (see Table 3). This reflects a worse credit quality on average and a higher composition of operational risk and partly of market risk which are more 'lumpy' and volatile, respectively. This feature is also present when the median values are considered. Therefore, the pattern of higher capital levels of securities holding banks is a robust feature of our classification. The traditional commercial banks hold, on average, the second largest amount of capital. This result is also robust to the capital measure used and the use of mean or median values. Finally, the wholesale funded banks hold the lower amount of capital, on average, with a mean leverage ratio of less than 7%.[12] These results are consistent with the findings of RTT, where the 'trading' banks hold the higher amounts of capital and the 'wholesale-funded' banks the lower.

The decomposition of the underlying risks (using the risk weight densities), presented in Table 4, is consistent with previous results on the RWA components (e.g. Le Leslé and Avramova 2012). Specifically, credit risk represents the largest fraction of the total risk exposure with median values ranging from 87.3% for securities holding banks to 88.6% for traditional commercial banks i.e. the median fraction of credit risk is very close across the different business models but the ranking of the percentage across business models is as expected.

[12] Given the larger size, on average, of the wholesale funded banks, this observation raises naturally the question whether this pattern reflects too-big-to-fail incentives. The empirical investigation of this conjecture falls, however, outside the scope of this article.

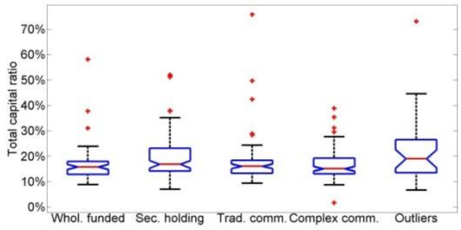
Table 3: Leverage and solvency indicators

(a) Tier 1 capital ratio



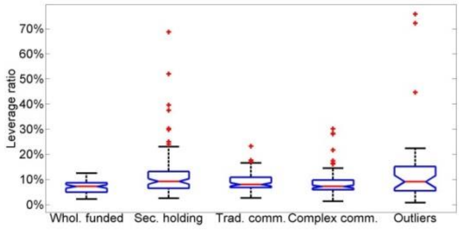
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	14.43%	18.39%	16.19%	15.10%	21.25%
median	12.74%****	15.07%***	14.48%**	13.66%***	16.62%***
std	7.59%	9.76%	9.09%	5.80%	13.74%
c.v.	0.53	0.53	0.56	0.38	0.65

(b) Total capital ratio



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	16.55%	19.95%	17.51%	16.42%	22.11%
median	15.76%***	16.77%***	16.01%*	15.04%***	18.95%***
std	7.38%	9.57%	9.20%	5.55%	13.23%
c.v.	0.45	0.48	0.53	0.34	0.60

(c) Leverage ratio



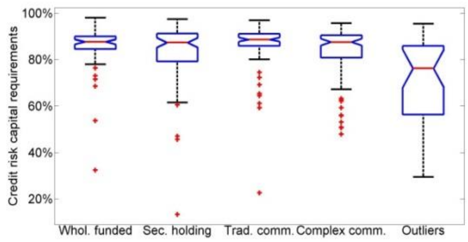
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	6.95%	12.40%	8.87%	8.51%	13.91%
median	7.16%****	9.2%***	7.98%**	7.15%**	9.09%**
std	2.53%	10.46%	3.56%	4.80%	16.79%
c.v.	0.36	0.84	0.40	0.56	1.21

Notes: The asterisks show the number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level.

Operational risk is the second largest risk component, and is higher for securities holding banks (median equals 10.5%) and lower for wholesale funded banks (median equals 8.6%). Finally, market risk has, on average, the lowest values (the median value is lower than 2% for all business models), and presents the largest variability, as measured by the standard deviation, for the securities-holding and complex commercial banks, meaning that there are banks in these groups which undertake a much larger percentage of market risk.

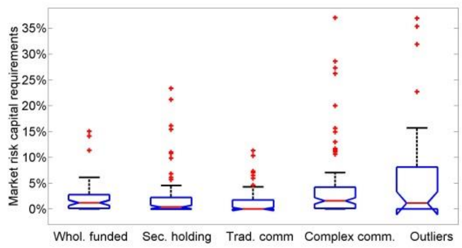
Table 4: Risk composition indicators

(a) Credit risk capital requirements (percentage)



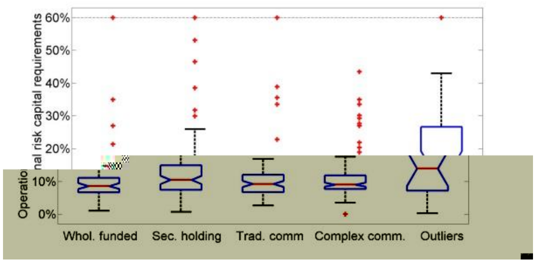
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	85.34%	83.32%	86.18%	83.38%	71.76%
median	87.64%**	87.32%**	88.62%***	87.55%***	76.17%*****
std	10.00%	12.86%	10.56%	10.89%	19.44%
c.v.	0.12	0.15	0.12	0.13	0.27

(b) Market risk capital requirements (percentage)



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	2.21%	2.37%	1.33%	3.79%	6.55%
median	1.22%**	0.4%**	0.01%****	1.6%***	1.16%**
std	3.16%	4.57%	2.37%	6.51%	10.74%
c.v.	1.43	1.93	1.78	1.72	1.64

(c) Operational risk capital requirements (percentage)



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	10.37%	13.75%	11.11%	11.07%	18.48%
median	8.59%***	10.46%****	9.24%***	9.02%**	13.94%*****
std	8.71%	12.09%	9.98%	6.84%	15.71%
c.v.	0.84	0.88	0.90	0.62	0.85

Notes: The asterisks show the number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level.

The statistics for credit risk seem to reflect specialization and risk-appetite effects (see Table 5). The banks which are on average mostly focused on loan-granting, namely traditional commercial and wholesale funded banks exhibit relatively high credit risk but they also possess relatively high allowance buffers against these potential losses. On the other end, securities holding banks, which are not so much focused on loans, hold low allowances relative to their realised credit risk compared to the other banks reflecting a higher risk appetite.

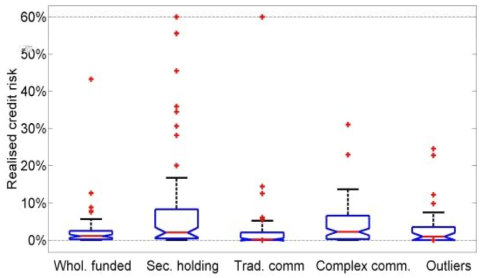
Specifically, looking at realised credit risk, measured by the percentage of impaired and past due loans to net loans (that is loans minus allowances), we find that securities holding banks and complex traditional banks are characterized by the highest amount of ex ante credit risk. This conclusion is reached both when the mean and median values are considered. The median percentage is above 2% for both business models, while wholesale funded and traditional commercial banks exhibit median rates which are lower than 1%. On the other hand, the traditional commercial banks present a high variance for ex ante credit risk which is comparable to that for securities holding banks, which means that for some banks of the former type the ex-ante credit risk is also relatively high. Wholesale funded banks seem to be the relatively better protected from credit risk. When we use the default rate to gauge ex ante credit risk[13], we find that securities-holding banks stand out on average, as being exposed to realised credit risk, however wholesale-funded banks emerge as also seriously exposed to credit risk. Finally, when we use a weighted PD metric, which is a forward looking expectation of credit risk, we find that wholesale-funded and securities-holding banks are more exposed to credit risk, while traditional commercial banks face the lowest exposure, when the median values are considered.

Regarding their performance, securities-holding and wholesale-funded banks seem to be the best performing business models in general compared to the commercial banks (see Table 6). The return on assets (RoA) of securities-holding banks has a median value of 0.4% and a mean of 0.7% which are the highest. Looking at the return on equity (RoE), the wholesale-funded banks are the best performing, on average, with a mean RoE of 4.8%. Wholesale funded banks show also the highest median RoE (5.02% compared to 4.78% for securities-holding banks). The commercial banks occupy the last two positions with respect to both the RoE and RoA metrics. These performance differentials are found to be statistically significant, a result which is consistent with the mobility barriers literature which posits that there could exist structurally differential profits within an industry (Caves and Ghemawat 1992).

[13] This measure does not take into account allowances, in contrast to the previously considered 'realised credit risk' measure, therefore represents credit risk which will 'eat up' banks' capital as it is expected to lead to future losses.

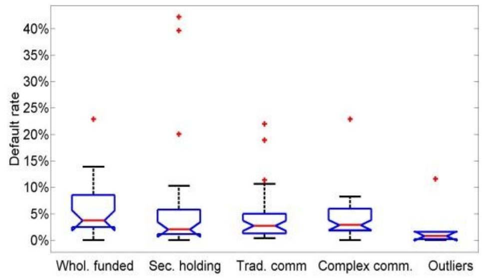
Table 5: Credit risk indicators

(a) Realised credit risk



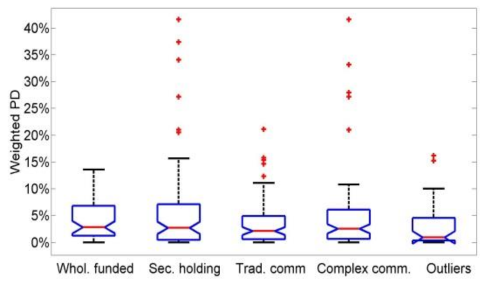
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
	2.45%	7.28%	3.07%	3.74%	2.61%
	1.09%****	2.04%****	0.19%****	2.22%***	0.96%**
	6.03%	14.16%	14.44%	5.31%	7.49%
	2.46	1.95	4.70	1.42	2.87

(b) Default rate



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	6.01%	6.50%	4.09%	3.85%	2.47%
median	3.7%***	2.01%*	2.71%**	2.85%*	0.77%**
std	5.23%	11.22%	4.85%	3.88%	4.50%
c.v.	0.87	1.73	1.18	1.01	1.82

(c) Weighted probability of default (PD)



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	4.12%	5.52%	3.64%	4.59%	3.12%
median	2.82%**	2.71%*	2.11%*	2.55%*	0.93%**
std	3.80%	8.30%	4.34%	6.93%	4.46%
c.v.	0.92	1.50	1.19	1.51	1.43

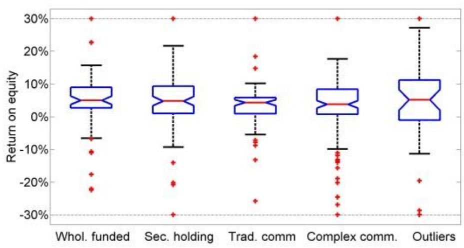
Notes: The asterisks show the number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level.

When it comes to cost efficiency, the securities holding banks also present the best (lowest) value, with respect to the median, with a cost-to-income ratio, lower than 60%. However, their mean value is the highest (68.4%). Overall, there is no clearly discernible pattern as regards cost efficiency, given the large variability of this metric across banks.

RTT find that commercial banks ('retail-funded banks' in the terminology of their paper) present the higher values for RoE and RoA, when a sample of banks from 34 countries is considered. In our study we find that this result does not hold, given our more homogeneous sample, consisting only of European banks. The securities holding banks (‘trading banks’ in the terminology of RTT) are the second best performers in RTT while they turn out to be the best performers within our sample when the RoA metric is considered. The complex commercial banks show the worst performance in our sample, rather than the wholesale funded banks as in RTT. The latter result is supportive of our decision to split the commercial banks into two clusters.

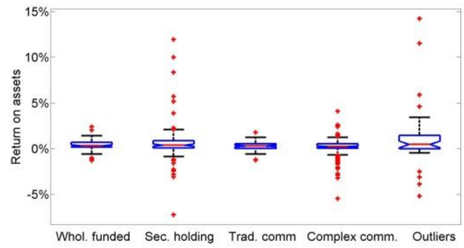
Table 6: Performance and efficiency indicators

(a) Return on equity



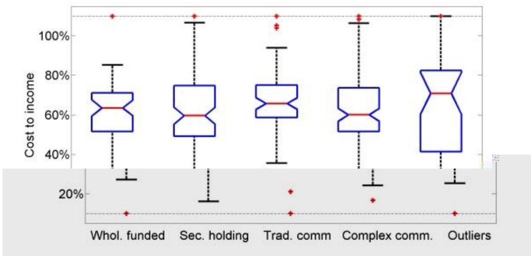
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	4.85%	3.67%	2.64%	0.75%	1.49%
median	5.02%**	4.78%*	4.28%**	3.79%*	5.15%*
std	9.61%	17.73%	6.03%	3.79%	28.01%
c.v.	1.98	4.83	2.29	1.19	18.58

(b) Return on assets



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	0.37%	0.71%	0.28%	0.07%	1.13%
median	0.33%*	0.4%**	0.31%*	0.24%**	0.48%*
std	0.66%	2.39%	0.48%	1.22%	3.53%
c.v.	1.79	3.38	1.69	16.94	3.13

(c) Cost to income ratio



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	25.38%	68.45%	65.43%	68.30%	56.69%
median	63.52%*	59.58%*	65.74%**	60.06%**	70.9%*
std	285.09%	49.29%	23.08%	53.97%	220.00%
c.v.	11.23	0.72	0.35	0.79	3.88

Notes: The asterisks show the number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level.

As regards the revenue sources, there is a large variability across business models, reflecting also national idiosyncratic features (see Table 7); therefore we will concentrate primarily on median values. The observed patterns reflect clearly the ‘median’ compositions of the balance sheets as presented in Section 4.1. Wholesale funded and traditional commercial banks, which hold the relatively higher amount of loans, earn the largest percentage of their income from interest rates (slightly above 60% for both types), while the respective figure for the securities holding and complex commercial banks, which conduct more trading business, is lower (slightly above 50%). On the other hand, fee and commission income is the lowest for wholesale funded banks. Interestingly, the median value is the highest for traditional commercial banks, reflecting the possibility of charging fees to the customers of these more ‘relationship-based’ institutions. In the case of traditional commercial banks, interest income and fees and commission income seem to be complements, mutually reinforcing components of income, rather than substitutes (see also the discussion in Louzis and Vouldis 2017).

On the other hand, securities holding banks clearly earn on average the highest (approximately 36%), as a percentage of total income. This feature is consistent with RTT who also find that ‘trading banks’ earn a substantially higher percentage of their income through fees and revenues (more than 40% on average, in their sample). However, these results are not robust when median values are considered, reflecting the heterogeneity across securities holding banks in earning fees and commissions from their trading business.

Finally, the trading income is the most variable component of income and while its median percentage is higher for the securities-holding and complex commercial banks (more than 2%), as it is expected given the noticeable presence of trading assets in these business models, the much higher variance of the percentage for wholesale funded and traditional commercial banks means that the latter types of banks follow much more risky strategies when it comes to trading.

The ‘outlier’ group holds on average large amounts of capital (both when measured as risk-adjusted capital and when the simple leverage ratio is used). In addition, these banks are in general much more exposed to market and operational risk, compared to the other banks and less to credit risk. These banks earn a distinctively higher proportion of their income from fees and commissions compared to the four other clusters of banks. Moreover, the median ‘outlier’ bank performs better than all the four other clusters, when performance is measured by the RoE and RoA. However, the variability of the performance distributions for these banks is for almost all indicators much higher compared to the four clusters, reflecting the heterogeneity of this group.

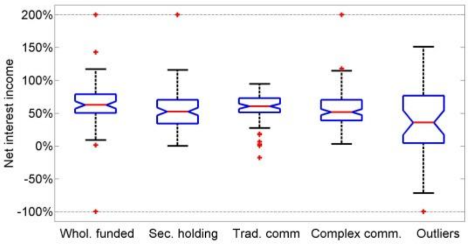
Broadly, we find that our decision to define a separate cluster for traditional and complex commercial banks was justified, in the sense that the two sets of banks exhibit distinctive characteristics with respect to their outcome variables. Specifically, traditional commercial banks hold more capital and undertake more credit risk compared to their more complex counterparts (which undertake more market risk). Also traditional commercial banks perform better overall, both with respect to their RoE and RoA, and their revenues originate to a larger extent on interest income.

Based on the descriptive statistics presented above one can attempt to place the different clusters of banks within a risk-performance space. In the context of this paper this mapping is undertaken based on expert judgment rather than a formal procedure. Specifically with respect to risk, we take into account the capital buffers, credit risk indicators (given that credit risk appears to be the most significant source of risk) and additionally, the percentage of credit risk as expressed in the decomposition of risk-weighted exposures per type of risk. For performance assessment, RoE and RoA are used. Based on these considerations, Table 8 summarises the characterisation of the different clusters with respect to risk and returns.

Specifically, we can characterise the securities holding banks as presenting competitive returns while also holding relatively high capital buffers (therefore, they excel on RoA outcome) and relatively risky assets. Wholesale funded banks also hold a risky portfolio, on average, while exhibiting high returns (especially with respect to RoE, given their relatively low capital). On the other hand, the traditional commercial banks hold, on average, the safest assets and they outperform the complex commercial banks although their returns are lower compared to both wholesale funded and securities holding banks. The complex commercial banks seem to present a non-optimal risk-performance combination.

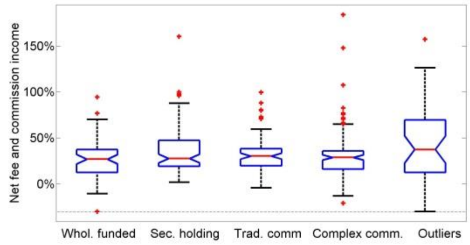
Table 7: Profit sources indicators

(a) Net interest income to total operating income



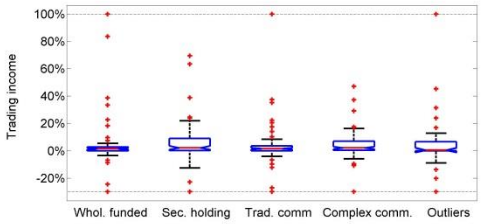
	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	-24.48%	54.71%	57.06%	55.10%	12.31%
median	62.42%****	52.5%**	60.56%**	51.57%***	35.58%****
std	693.41%	38.33%	22.95%	31.92%	183.94%
c.v.	28.32	0.70	0.40	0.58	14.95

(b) Net fee and commission income to total operating income



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	20.22%	36.42%	31.38%	32.67%	42.68%
median	26.93%*	27.71%*	30.36%*	28.81%*	37.43%*
std	58.51%	27.41%	20.18%	27.50%	40.24%
c.v.	2.89	0.75	0.64	0.84	0.94

(c) Trading income to total operating income



	Wholesale funded	Securities holding	Traditional commercial	Complex commercial	Outliers
mean	-45.98%	6.48%	6.39%	3.97%	-2.41%
median	1.46%**	2.19%**	1.57%*	2.14%*	0.44%*
std	399.66%	13.65%	37.22%	9.60%	51.63%
c.v.	8.69	2.11	5.82	2.42	21.41

Notes: The asterisks show the number of the other clusters for which the pairwise equality of medians is not accepted at the 5% significance level.

Table 8: Risk-performance levels per business model

	Risk	Performance
Wholesale funded	High	High
Securities holding	High	High
Traditional commercial	Low	Medium
Complex commercial	Medium	Low

5. Conclusion and future work

We present the first study which makes use of a very granular data set on banks in the euro area in order to infer the types of business models that they follow. We have defined the concept of the business model following insights from the strands of evolutionary economics and organisational studies. We have adopted a statistical, data driven clustering approach which combines 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. The approach minimises the impact of the researcher’s priors on the results.

The results provide an anatomy of the banking sector of the euro area and indicate the co-existence of four distinct business models: traditional commercial, complex commercial, wholesale funded and securities holding banks are present, alongside with specialised institutions such as state owned entities aimed at refinancing loans to semi-public and public entities. These specialised entities have been identified as outliers by the clustering algorithm. Wholesale funded banks are, on average larger while the securities holding banks the smallest and the two types of commercial banks lie in the middle, on average, with respect to their size.

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” (which include deposits and issued debt).

Moreover, our investigation of the outcome variables, regarding performance, efficiency and risk characteristics, provides empirical evidence of significant differences across business models. Overall, we find that the two model types which depart more clearly commercial banks perform better than the complex ones, given the underlying risk of their portfolio.

Our results are based on a comprehensive data set for a snapshot in time (2014Q4). Therefore, they reflect the history of macroeconomic developments and policy decisions which have taken place during the crisis period. An extension of this analysis in time, when past macroeconomic and policy shocks have been absorbed, complemented by a migration analysis across business models, would be valuable to understand the dynamics of the banking sector.

References

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

Arabie, P., Hubert, L., 1994. Cluster analysis in marketing research. In: Bagozzi, R.P. (ed.), *Handbook of marketing research*. Blackwell, Oxford.

Ayadi, R and W de Groen (2014): Banking business models monitor 2014. *Centre for European Policy Studies and International Observatory on Financial Services Cooperatives*

Bai, J., Ng, S. (2002) "Determining the number of factors in approximate factor models", *Econometrica*, vol. 70, no. 1, 191-221

Burt, S., Johansson, U., and Dawson, J. (2015) "International retailing as embedded business models", doi:10.1093/jeg/lbv008

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

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*, vol. 91, no. 2, 241-262

Caves, R. E., Ghemawat, P., (1992) "Identifying mobility barriers", *Strategic Management Journal*, vol. 13, no. 1, 1-12

Chiang, M.M., Mirkin, B. (2010). “Intelligent choice of the number of clusters in *K*-Means clustering: An experimental study with different cluster spreads”, *Journal of Classification*, vol. 27, 3-40

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

DeSarbo, W. S., Grewal, R. 2008. Hybrid Strategic Groups. *Strategic Management Journal*, vol. 29, No. 3, 293-317

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

ECB 2015, SSM SREP Methodology Booklet,

(Link:https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm_srep_methodology_booklet.en.pdf?6bdbd6a1833a3ff833921460ac1bd391)

Farnè, M. (2016) *Large covariance matrix estimation by composite minimisation*. PhD Thesis. University of Bologna.

Halaj, G., Zochowski, D., (2009). Strategic groups and bank’s performance. *Financial Theory and Practice*, vol. 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*, vol. 16, 581-600

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*, vol. 24, 417-441 and 498-520.

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

Le Leslé, V., and Avramova, S., (2012), Revisiting Risk-Weighted Assets: Why do RWAs differ across countries and what can be done about it?, *IMF Working Paper*

Louzis, D. P., Vouldis, A. T. (2017), "Profit strategy of Greek banks: cross-subsidization and diversification versus complementarity", *Applied Economics*, http://dx.doi.org/10.1080/00036846.2017.1284989

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

Magretta, J., (2002) “Why business models matter”, *Harvard Business Review*, vol. 5 86-92

Nelson, R. R., Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, Mass.: Harvard University Press.

Nelson, R. R., Winter, S. G. and Schuette, H. L. (1976). “Technical change in an Evolutionary Model”, *Quarterly Journal of Economics*, vol. 90, no. 1, 90-118

Newman, H. H. (1978). “Strategic groups and the structure-performance relationship”, *Review of Economics and Statistics*, vol. 60, no. 3, 417-427

Reger, R. K., Huff, A. S. (1993). Strategic groups: A cognitive perspective, *Strategic Management Journal*, vol. 14, no. 2, 103-123

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

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

Teece, D. (2010) “Business models, business strategy and innovation”, *Long Range Planning*, vol. 43 172-194.

Ten Berge, J.M.F. (1993) *Least squares approximation*. DSWO Press, Leiden.

Tywoniak, S., Galvin, P., Davies, J. (2007) “New Institutional Economics’ Contribution to Strategic Groups Analysis”, *Managerial and Decision Economics*, vol. 28, No. 3, 213-228

Vichi, Kiers, (2001). “Factorial k-means analysis for two-way data”, *Computational Statistics and Data analysis* vol. 37, 49-64

Ward, J.H. (1963), “Hierarchical Grouping to optimize an objective function”, *Journal of the American Statistical Association*, vol. 58, no. 301, 236-244

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

Zott, C., Amit, R., (2011) The business model: Recent developments, and future research. *Journal of Management*, vol. 37 no. 4 1019-1042

Appendix A

Table A-1 presents the templates which are used as the input set along with the number of variables from each template. The EBA templates along with the definitions of the contained data can be found at the EBA website: https://www.eba.europa.eu/regulation-and-policy/supervisory-reporting.

Table A-1: Contents of the templates defining the input set

Template code (as defined by the EBA)	Contents
F 01.01	Assets
F 04.01	Assets held for trading
F 04.02	Assets designated at fair value through profit or loss
F 04.03	Available-for-sale assets (carrying amount)
F 04.04	Loans and receivables & held-to-maturity assets
F 05.00	Loans and advances by product (on demand, credit card, leases, loans etc)
F 08.01.a	Liabilities
F 09.01	Off-balance items (loan commitments and guarantees)
F 10.00	Derivatives – trading
F 11.01	Derivatives – hedge accounting

F 01.01 and all F 04 templates provide the breakdown of assets across accounting portfolios, with additional breakdowns on instruments and counterparties. F 05 provides the breakdown of loans by product (credit card loans, collateralized loans, project finance etc). The liability side is covered by the F 08 template, which breaks down liabilities by accounting portfolio (the largest percentage of banks' liabilities are valued at amortised cost), instrument and counterparty. Off-balance sheet items, primarily loan commitments and guarantees are contained in template F 09.01. Finally, templates F 10 and F 11.01 provide detailed information on derivatives, distinguishing between trading and hedge accounting derivatives. The information is further broken down by type of derivatives (interest rate, equity, foreign exchange, credit and commodity) and by the type of market in which the derivatives are traded (OTC or organised markets).

Appendix B

The initial set of variables contains a number of highly correlated variables. We would like to automatically select the ones which are more "fundamental" in the sense of being more related overall to the remaining set of variables; for example, when there is a variable like "notional amount of total derivatives" and one of its subcategories like "notional amount of OTC derivatives", we would prefer to keep the broader category on the condition that it is more related to the set of the remaining variables. This selection of variables is also subject to the condition that we would like to exclude pairs of variables with the absolute level of correlations above a threshold, which was set to 0.95, in order to avoid bias in the results.

The threshold 0.95 is chosen in order to obtain a non-redundant sample covariance matrix. The thresholding step is intended simply to remove almost perfect identities for the computation of the sample covariance matrix rather than to create of set of variables which present low levels of correlation, since our algorithm is well suited to deal with sets of correlated variables because of the incorporation of factor analysis takes place simultaneously with clustering. This is in contrast to other clustering methods (most of those reviewed above), which require that the input set be 'cleaned' from correlated variables.

Specifically, the results are not much affected by the threshold choice because of the following mechanism: if one variable is particularly correlated with all the others, the multivariate data-driven procedure adjusts downwards their weights in the resulting factors, because there is less clustering power on that direction. This is one of the reasons why our adopted method is well suited for a large-dimensional context. In this respect, we remark that none of the top 15 variables by importance among the surviving ones has a correlation which exceeds 0.3 in absolute value with both of the two factors.

Consequently, we define a measure of the "importance" of each variable within the data set in order to operationalise the above selection criteria. The "importance" $\mathrm{I}(j)$ of each variable j is defined as the linear combination of the correlation absolute values with the other variables of the input set:

$$\mathrm{I}(j) = \sum_{\substack{k=1,..,P, \\ k \# j}} |Corr(j,k)|$$

Consequently we order the variables in a non-increasing order based on their $\mathrm{I}(j)$. Whenever the predefined level of correlation C^* is exceeded, then only one of the two correlated variables is retained, specifically the one with a higher level of $\mathrm{I}(j)$.

Acknowledgements

The authors are grateful to Doyne Farmer, Michael Fedesin, Fabio Franch, Patrick Hogan, Heinrich Kick, Christoffer Kok, Giancarlo Pellizzari, Aurel Schubert, Maik Zimmermann, Dawid Zochowski and a number of seminar participants within the ECB for insightful comments and discussions.

Matteo Farnè
University of Bologna, Department of Statistics; matteo.farne2@unibo.it

Angelos Vouldis
Directorate General Statistics, Supervisory Statistics Division, European Central Bank, Frankfurt, Germany; Angelos.Vouldis@ecb.int

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website.

ISSN	1725-2806 (pdf)	DOI	10.2866/386225 (pdf)
ISBN	978-92-899-2792-5 (pdf)	EU catalogue No	QB-AR-17-082-EN-N (pdf)