Clustering Austrian Banks' Business Models and Peer Groups in the European Banking Sector

Robert Ferstl,
David Seres[1]

As the European banking sector is becoming increasingly intertwined, the degree of interdependence is also rising. Consequently, it is key to conduct comparisons for a timely identification of emerging patterns of this development. Furthermore, the product range of banks has expanded so that heterogeneity across the banking sector has also been growing rapidly. This rising heterogeneity makes it increasingly impractical to carry out comparisons on an aggregate level. A more efficient approach is identifying one or more "common denominators" of similar banks and establishing groups of banks which share this (these) common denominator(s). In this paper, we consider the business models of banks as one such common denominator, which can be described by a set of variables. These variables span a high-dimensional space where each bank represents a point, which can be measured by a statistical distance. Points close to each other may constitute a group, while points distant from these points will not belong to that group. Therefore, the objective of this study is, on the one hand, to define an efficient set of variables correctly reflecting the business models of banks and, on the other hand, to find subsets of high similarity. By applying statistical clustering techniques we aim to understand banks' business models, thereby gaining new insights into the design of the European banking sector and, in particular, identifying peer groups relevant to the top Austrian banks. Assessing the distribution of risk and identifying certain business patterns within those groups allows a meaningful ranking of Austrian banks in comparison to their European competitors.[2] The analysis in this paper is conducted on the basis of a purely quantitative methodology and the results should be interpreted accordingly.

JEL classification: C02, C44, C58
Keywords: Austrian banks, cluster analysis, data-driven decision support

The European banking sector is currently undergoing rapid changes due to recent financial market developments and the introduction of new legislation, such as the Capital Requirements Regulation/Capital Requirements Directive IV. This study aims to identify emerging patterns of these changes in the European banking sector via clustering analysis. There is a broad stream of literature analyzing the efficiency of banks. Tortosa-Ausina (2002), for instance, focuses on cost efficiency and calculates product mix clusters. Several authors analyze the influence of the recent financial crisis on banks' business models. Altunbas et al. (2011) focus on bank risk and use probit and linear regressions. Ayadi et al. (2011) perform a screening analysis of business models in European banking using clustering methods. Their sample consists of 26 European banks, and they use end-of-year data. Ayadi et al. (2012) extend their previous paper by using a larger sample and focusing on the impact of new regulatory measures.

This study builds on the existing literature and is based on a very large sample of European banks. We provide a detailed motivation for the selection of variables and validate the robustness of the statistical methodology used. The paper is structured as follows: Section 1 describes the statistical methodology used for identifying banks' busi-

Refereed by:
Kurt Hornik,
Vienna University of Economics and Business

[1] *Oesterreichische Nationalbank, Off-Site Banking Analysis and Strategy Division, robert.ferstl@oenb.at, david.seres@oenb.at.*

[2] *The data were acquired via Bankscope and are publicly available from banks' financial statements.*

ness models, section 2 discusses the database and the definition of variables, section 3 gives the results for the five identified business models and for a peer group of the top 3 Austrian banks, and section 4 provides a summary.

1 Methodology

This section deals with the mathematical and statistical background of our analysis and explains the clustering algorithm used to determine the different business models and the methodology to identify the peer groups of certain banks. Cluster analysis is a statistical tool for detecting groups that is especially helpful when the data exhibit no visible natural clusters. It is important to note that clustering is as much an art as it is a mathematical methodology. There exist numerous algorithms for clustering, ranging from classical hierarchical ones to highly sophisticated algorithms. There is no globally optimal strategy for choosing the correct methodology. Ideally, it should be in line with the underlying data and with the expectations of the results. For example, with data of small dimensions (i.e. with only a handful of observations and variables), hierarchical methods have provided reasonable solutions in some cases, whereas with data of large dimensions drawn from a high-dimensional space, the application of partitioning clustering algorithms is advisable. These considerations make the process of choosing a clustering method a trial-and-error procedure, in which the results of many algorithms have to be checked for consistency with the economic expectations from expert judgement. For a detailed summary of widely used clustering algorithms see Everitt et al. (2011).

In this paper we use a partitioning clustering algorithm, namely *k*-centroids clustering developed by Leisch (2006). It was chosen for several reasons:

- Simplicity: It employs the popular *k*-means algorithm for determining the clusters, which is a straightforward optimization technique.
- Robustness: It provides generally similar outputs for different random seeds.
- Speed: The algorithm is fast, which is particularly useful when bootstrapping is necessary.
- Results: It delivers the most plausible clustering results compared to other techniques.

1.1 *k*-Centroids Clustering

Given n column vectors of observations $X = \{x_1,...,x_n\}$ from a set of m random variables $\mathbf{X}$ spanning a space χ, the aim is to find a partition set $\mathbf{P}$ of high similarity within the set X. Dissimilarity usually is quantified by a function $d : X \times X \rightarrow \mathbb{R}^+$ fulfilling the following properties:

$$d(x, y) \geq 0,$$
$$d(x, y) = 0, \text{ iff } x = y,$$
$$d(x, y) = d(y, x)$$

The function d is then called a metric dissimilarity or distance. A partition set $\mathbf{P} = \{P_1,...,P_k\}$ of k disjoint clusters is formed by setting k centroids

$$C_k = \{c_1,...,c_k\}, c_k \in \chi$$

and assigning each point $x_i \in X$ to the cluster P_j with the closest centroid $c(x_i) = c_j$, i.e.

$$x_i \in P_j \text{ if } \min [d(x_i, C_k)] = d(x_i, c_j) \quad (1)$$

We aim to find a good set of centroids C_k which minimizes the entropy in each cluster of the partition induced by C_k. The optimization problem is then as follows:

$$\min_{C_k} d(X, C_k) = \sum_{x_i \in X} d\,(x_i, c(x_i))$$

To be able to compute the objective function, an appropriate distance measure *d* has to be defined. In this paper we focus on the Mahalanobis (1936) metric since it accounts for different variances as well as for the covariance structure within *X*. The metric is defined as

$$d_M\,(x_i, c(x_i)) = \sqrt{\big(x_i - c(x_i)\big)^T \sum_X^{-1} (x_i - c(x_i))}$$

with Σ_X as the covariance matrix of *X*. A rather desirable side effect of this metric is that due to the term Σ_x^{-1}, it becomes scale-invariant and hence there is no need to normalize the variables beforehand. It is easy to see if Σ_x^{-1} is a diagonal matrix, it divides each term in x_i by its corresponding variance. If $\Sigma_x^{-1} = I$ then the metric is the Euclidian distance. Hence the Mahalanobis distance is a generalized form of the Euclidian. Thus the objective function can be stated as

$$\min_{C_k} d_M\,(X, C_k) = \sum_{x_i \in X} d_M\big(x_i, c(x_i)\big) \qquad (2)$$

1.2 Implementation

There exists no closed-form solution for the optimization problem in equation (2) and thus an iterative estimation procedure must be used to find the optimum. A well-known algorithm is the *k*-centroids algorithm in its general form:

1. set *k* and initialize a random set of centroids;
2. apply equation (1);
3. update the centroids holding the clusters $c(x_n)$ fixed:

$$c_j = \operatorname*{argmin}_c \sum_{n:c(x_n)=c_j} d(x_n, c), \quad j = 1,\ldots,k$$

4. repeat steps 2 and 3 until convergence.

We classify each bank by using the following $m = 5$ variables:

- net interest income (as a percentage of operating income)
- trading income (as a percentage of operating income)
- income from fees and commissions (as a percentage of operating income)
- loan-to-deposit ratio
- loans (as a percentage of total assets)

These variables cover the aggregate income structure and the aggregate loan structure of banks. For a more detailed description, see section 2. According to this classification, each bank represents a point in a 5-dimensional Hilbert space. We can measure the distance – in other words, the similarity – between any two banks by employing any distance measure (in this analysis we employ the Mahalanobis distance) and determine which banks are close to each other and which banks are scattered further away from the others. The clustering algorithm aims to find and group those points which are relatively close to each other (and hence exhibit high similarity). It is crucial to choose an appropriate number of clusters as the results may vary strongly when *k* is low, which is usually the case when clustering. The impact can be high when setting $k = 4$ in contrast to $k = 3$. Thus it is necessary to validate any choice of *k* to allow a stable and meaningful solution. A robust validation can be achieved with many methods and measures but one of the simplest ones is bootstrapping, which we use in this paper (for details, see section 3.1). The resulting clusters of the algorithm should exhibit convexity as well as low intra-cluster and high inter-cluster entropy, meaning that the elements of one cluster should be as similar as possible, whereas the elements of two different clusters should be as dissimilar as possible. If this is the case then the result is assumed to be stable. Such clusters can be interpreted as groups that share distinctive business models.

1.2.1 Peer Groups Based on the Distance Measure

The peer group G_i of bank x_i is a group consisting of banks with a business structure similar to that of x_i. As already mentioned above, similarity can be derived from the distance of one bank to another bank, thus it is sufficient to set up an *m*-dimensional sphere with radius *r* (an ellipsoid in the Mahalanobis definition) around x_i and check which banks are located within that sphere. More formally, we write:

$$x_s \in G_i \text{ if } d_M(x_s,x_i) \leq r \qquad (3)$$

Depending on the position of a bank within a cluster it is possible that this bank might have peers in different clusters. Such would be the case if, e.g., bank *i* is on the boundary of a certain cluster and bank *s* is on the verge of a different cluster but the distance between x_s and x_i is small. Even if these two banks belonged to different clusters, they would be found in the same peer group.

1.2.2 Software

Most of the implementation tasks were carried out in the programming language *R*.[3] The clustering technique employed is a *k*-centroids cluster analysis provided in the *R*-package *flexclust* (see Leisch, 2006). We implement the Mahalanobis measure manually as it was not available in the package. Validation via bootstrapping is also provided in this package (see function *bootFlexclust*).

1.2.3 Validation

As mentioned earlier, the results may significantly depend on the choice of *k* (number of clusters). Thus it is necessary to validate the choice of *k* by both statistical tools and by qualitative means. Statistical methods, such as bootstrapping, are powerful tools but they lack the ability to express the meaningfulness of a solution from an economic point of view. The validation tool used in this analysis serves as a guidepost rather than as an irrefutable solution. The bootstrapping algorithm provided in the *flexclust* package estimates *k* based on the similarity of two subsets (with the possibility of non-empty intersection, i.e. elements of one subset can also occur in the other subset). A *k* with the highest Rand index indicates the best possible separation of clusters given the number of groups the algorithm has to divide the set into. The pseudo-code of the algorithm for *B* iterations (bootstrapping samples) is as follows:

1. sample (with replacement) two subsets S_1 and S_2 of the original set S^*;
2. apply the clustering algorithm for these two subsets, resulting in partitions p_{S_1} and p_{S_2};
3. calculate the Rand index for each pair of partitions as

$$I_R = \frac{\left(\sum_{i=1}^{c}\sum_{j=1}^{c} n_{ij}^2 - n\right) - \frac{1}{2}\left(\sum_{i=1}^{c} n_{i.}^2 - n\right) - \frac{1}{2}\left(\sum_{j=1}^{c} n_{.j}^2 - n\right) + \binom{n}{2}}{\binom{n}{2}}$$

where *c* is the length of both subsamples and $c \times c$ is the matrix $N = n_{ij}$, where n_{ij} is the number of objects in group *i* of partition p_{S_1} and group *j* of partition p_{S_2}.

4. Repeat steps 1 to 3 *B* times and return the Rand index for each trial.

[3] *R is an open-source statistical programming language, see R Development Core Team (2012).*

2 Data

In this section we explain the construction of the database,[4] how we defined each variable and the characteristics these variables exhibit through descriptive statistics.

2.1 Database and Data Preparation

The time frame of the data set used for our analysis ranges from year-end 2005 to year-end 2011 on a yearly basis. There is a total of 234 European banks covered, including the top 6 Austrian banks. Data points featuring more than two missing values in any of the variables are removed from the sample. The remaining missing values are then interpolated (trailing missing values extrapolated) across time. Based on the results in the existing literature (e.g. Ayadi et al., 2012) and on expert judgement, we construct the following five variables in table 1, bearing in mind that they should be representative of the expected cluster structure (i.e. the business model).

Table 1

Construction of Variables

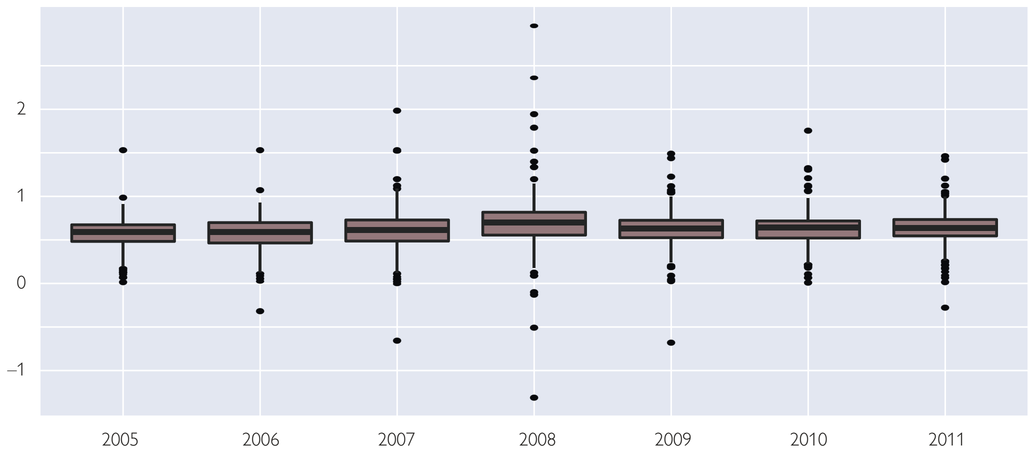
Parameter	Symbol
Net interest income	NII
Trading income	TI
Income from fees and commissions	CI
Operating income	OPINC
Customer deposits	CUSTDEP
Total loans	TOTLOANS
Total assets	SIZE

Variable	Symbol
NII/OPINC	NTR
TI/OPINC	TTR
CI/OPINC	CTR
TOTLOANS/CUSTDEP	LDR
TOTLOANS/SIZE	LAR

Finally, the three-dimensional data cube *(DATA)* comprises the banks in the first dimension, the variables in the second dimension and time in the third dimension. Data normalization was obsolete since the Mahalanobis distance measure used in this clustering analysis is scale-invariant.

Chart 1

Net Interest Income over Operating Income



Source: Authors' calculations.

[4] *The data used are from Bankscope and stem from banks' publicly available financial statements. Bankscope provides information on over 27,600 banks around the world spanning up to 16 years, including detailed accounts (country specific "as reported" and standardized), ratios, ratings and rating reports, ownership, country risk and country finance reports.*

2.2 Descriptive Statistics of the Data Set

2.2.1 Net Interest Income over Operating Income (NTR)

Net interest income is the difference between the revenues from interest-bearing assets and the expenses on interest-burdened non-trading assets. It represents the part of a bank's operating income generated by the interest payment structure. Net interest income as a percentage of operating income is termed NTR, defined as

$$NTR = \frac{\text{gross interest income} + \text{dividend income} - \text{total interest expense}}{\text{operating income}}$$

A high NTR suggests that a bank generates a large part of its overall income through interest income and would therefore be classified as retail bank.

Table 2

NTR Descriptive Statistics

Year	μ	σ	E[X³]	E[X⁴] − 3
2005	0.563	0.192	−0.083	3.050
2006	0.560	0.206	−0.356	3.069
2007	0.603	0.241	−0.421	4.612
2008	0.693	0.357	1.051	14.626
2009	0.621	0.187	−0.256	1.144
2010	0.626	0.186	−0.332	1.290
2011	0.635	0.212	−0.124	3.356

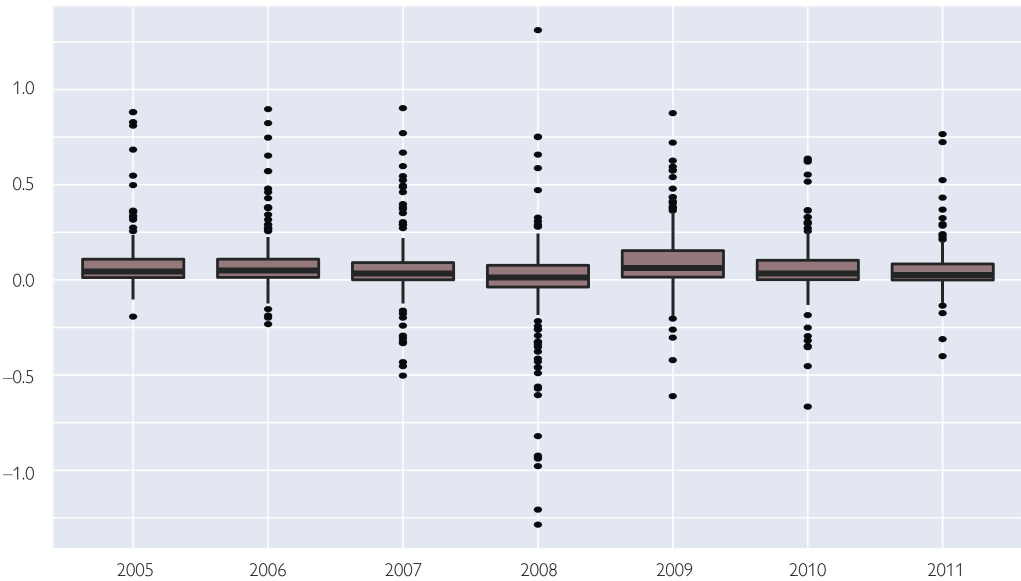
Source: Authors' calculations.

Chart 1 shows the NTR for the whole sample over six years.[5]

Table 2 summarizes the moments of the empirical distributions. One can observe on average a steady increase in the ratio except in the crisis year of 2008, when due to losses in other income structures the NTR increased by approximately 9% compared to 2007.

Chart 2

Trading Income over Operating Income



Source: Authors' calculations.

[5] *In the boxplots, the black horizontal bars represent the median, the red boxes represent 50% of the data around the median (lower part of the box: 25% quartile, upper part: 75% quartile). The upper and lower bounds of the vertical black lines (whiskers) span 95% of the data. The filled circles represent outliers.*

2.2.2 Trading Income over Operating Income (TTR)

Trading income is defined as the net gains (losses) on trading and derivatives. It includes marking to market of derivatives, currently related transactions, interest-rate instruments, equities and other trading assets, excluding insurance-related trading income. The TTR is the fraction of net trading income to operating income and is defined as

$$TTR = \frac{\text{trading income}}{\text{operating income}}$$

A high TTR suggests that a bank generates a large part of its overall income through trading and would therefore be classified as an investment bank. Chart 2 shows the TTR for the whole sample over six years.

Table 3 summarizes the moments of the empirical distributions. One can observe on average a constant ratio again except in the crisis year of 2008, during which the ratio dropped by approximately 7% compared to 2007 and became even negative on the aggregate level.

Table 3

TTR Descriptive Statistics

Year	μ	σ	$E[X^3]$	$E[X^4] - 3$
2005	0.084	0.139	3.416	15.104
2006	0.088	0.147	2.738	9.938
2007	0.057	0.157	1.827	8.137
2008	–0.015	0.266	–0.754	7.669
2009	0.094	0.144	1.614	5.863
2010	0.056	0.130	1.521	6.309
2011	0.049	0.109	1.530	9.776

Source: Authors' calculations.

2.2.3 Income from Fees and Commissions over Operating Income (CTR)

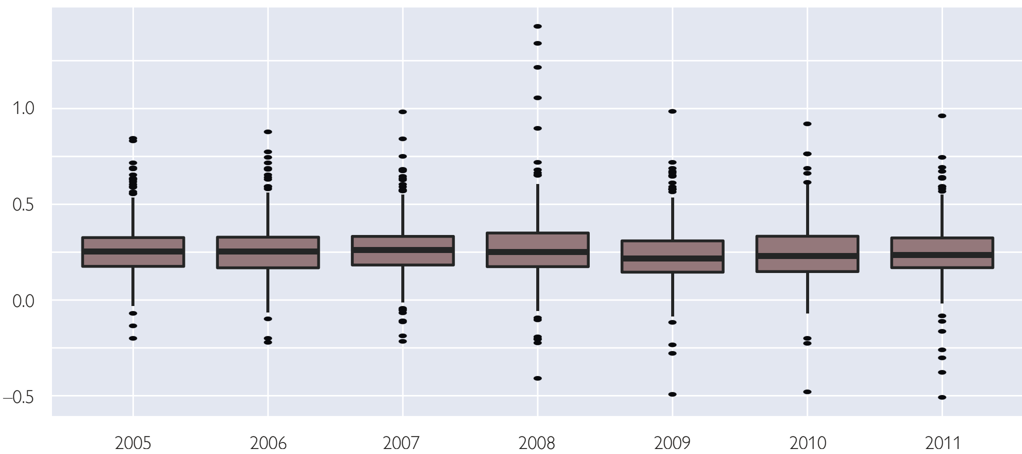
Income from fees and commissions is defined as the income from net fees and commissions which are not related to loans. The CTR is income from fees and commissions over operating income and therefore defined as

$$CTR = \frac{\text{income from fees and commissions}}{\text{operating income}}$$

A higher CTR indicates that a bank generates a larger portion of its income through commissions and fees. This holds true, e.g., for private banks.

Chart 3

Income from Fees and Commissions over Operating Income



Source: Authors' calculations.

Table 4

CTR Descriptive Statistics

Year	μ	σ	$E[X^3]$	$E[X^4] - 3$
2005	0.267	0.148	1.114	2.159
2006	0.264	0.149	1.034	2.274
2007	0.270	0.148	0.750	1.649
2008	0.275	0.212	1.909	8.685
2009	0.234	0.158	0.910	2.821
2010	0.248	0.156	0.762	2.379
2011	0.249	0.160	0.111	4.581

Source: Authors' calculations.

Chart 3 shows the CTR for the whole sample over six years.

Table 4 summarizes the moments of the empirical distributions. One can observe on average a constant ratio over the observed years.

2.2.4 Loan-to-Deposit Ratio (LDR)

The loan-to-deposit ratio (LDR, more commonly also LTD) is defined as

$$LDR = \frac{\text{gross loans}}{\text{total customer deposits}}$$

The numerator includes residential mortgage loans, other mortgage loans, other consumer or retail loans and corporate and commercial loans. The denominator is the sum of current, savings and term deposits. This ratio measures the liquidity of banks. A high ratio indicates low liquidity to cover unexpected funding requirements whereas a low ratio might suggest unrealized profitability in the income structure. Chart 4 shows the LDR for the whole sample in each year.

Table 5 summarizes the moments of the empirical distributions. The LDR increased steadily until end-2007 and then dropped by approximately 25% until end-2011.

Table 5

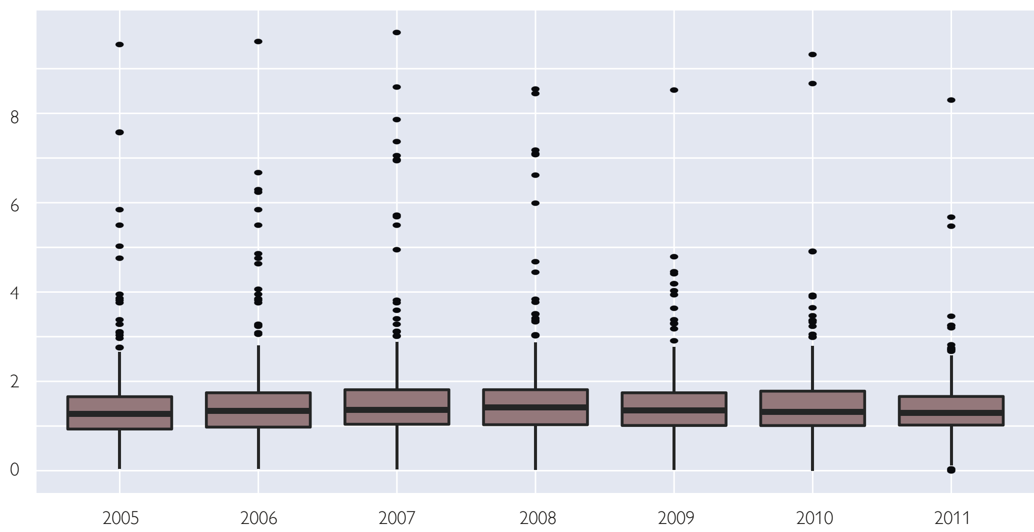
LDR Descriptive Statistics

Year	μ	σ	$E[X^3]$	$E[X^4] - 3$
2005	1.505	0.970	3.891	25.700
2006	1.584	0.826	1.896	6.079
2007	1.699	0.818	2.307	11.223
2008	1.680	0.983	3.123	15.923
2009	1.508	0.704	1.457	4.287
2010	1.511	0.706	1.682	5.666
2011	1.433	0.717	2.248	10.365

Source: Authors' calculations.

Chart 4

Loan-to-Deposit Ratio



Source: Authors' calculations.

2.2.5 Loan-to-Asset Ratio (LAR)

The loan-to-asset ratio (LAR) is defined as

$$LAR = \frac{\text{gross loans}}{\text{total assets}}$$

The numerator is defined identically to the numerator of the LDR. The denominator is the sum of on- and off-balance sheet items. A high LAR indicates that loans represent the bulk of balance sheet items. Chart 5 shows the LAR for the whole sample in each year. Table 6 summarizes the moments of the empirical distributions. Again, one can observe on average a constant ratio over the years.

Table 6

LAR Descriptive Statistics

Year	μ	σ	$E[X^3]$	$E[X^4] - 3$
2005	0.528	0.210	−0.396	−0.649
2006	0.540	0.212	−0.502	−0.554
2007	0.555	0.218	−0.590	−0.548
2008	0.570	0.219	−0.690	−0.417
2009	0.562	0.215	−0.702	−0.230
2010	0.566	0.214	−0.722	−0.198
2011	0.570	0.218	−0.693	−0.196

Source: Authors' calculations.

3 Results

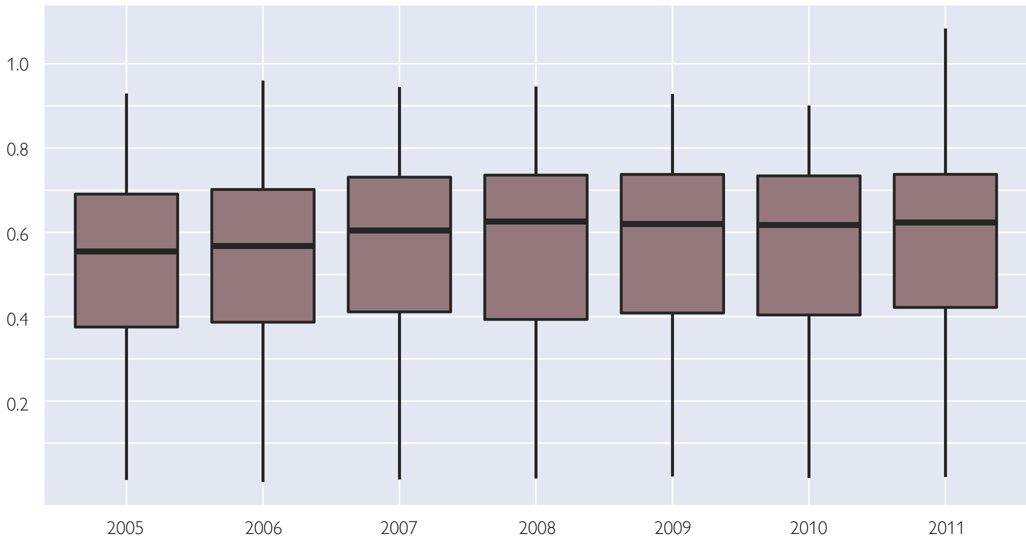
This section first explains how we fixed the optimal number of clusters and then gives an interpretation of each business model derived from the clustering algorithm. Furthermore we identify the peer groups of Austrian banks.

3.1 Determining the Optimal Number of Clusters

As underlined in section 1.2.3 we need to determine k in order for the clustering algorithm to produce plausible results. To this end, we apply the bootstrapping algorithm described in that chapter. For every year spanning the data set we calculate the Rand index, see equation (4), for $k^* \in \{3, 4,..., 8\}$ number of clusters in $B = 300$ iterations. The k^* with the highest Rand index is chosen conditional to plausibility and robustness. Chart 6 shows the results of the bootstrapping algorithm. In most years k between 4 and 6 appears optimal so we choose $k = 5$ for each year to ensure comparability between the clus-

Chart 5

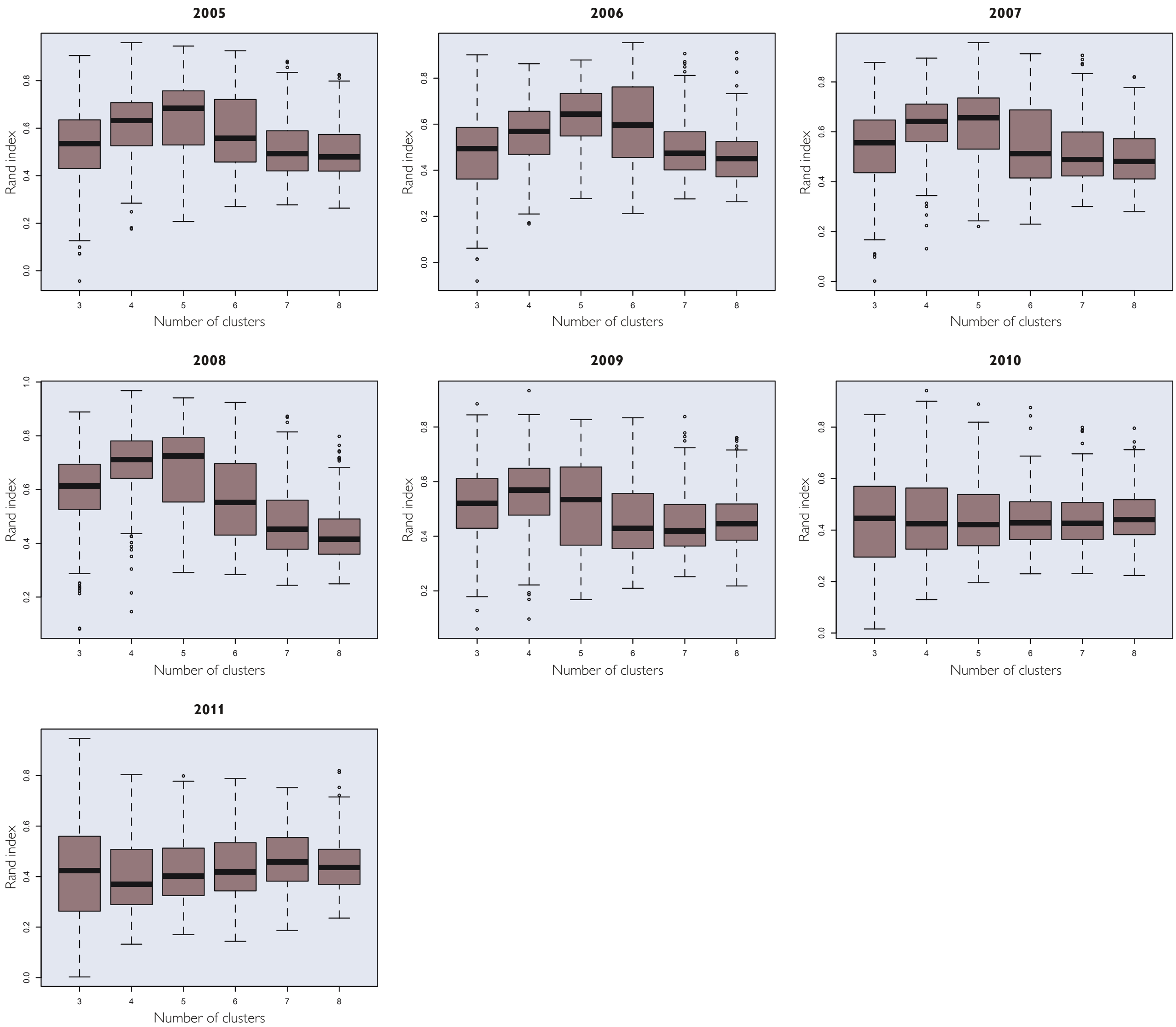
Loan-to-Asset Ratio



Source: Authors' calculations.

Chart 6

Bootstrapping Results with B = 300 Iterations



Source: Authors' calculations.

tering results of each year. The result for 2011, when the optimum was *k = 3,* is an exception. This is partly due to what appears to be a merging of two distinct business models in 2011 and will be explained in detail later in this chapter (section 3.2.4).

3.2 Business Models

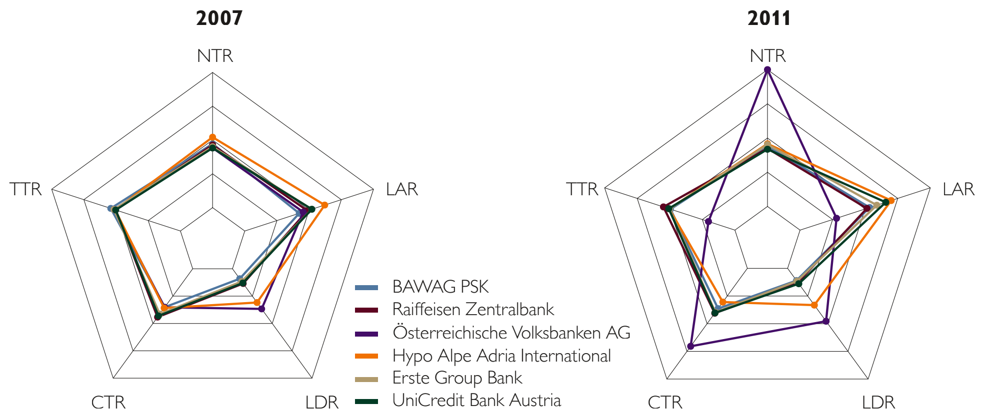
Based on the bootstrapping results we set *k = 5* and apply the clustering algorithm on the data for each year. The clusters generated in this manner should be easily separable from each other and exhibit different features. From each cluster we try to deduce unique business models for further analysis.

3.2.1 Top 6 Austrian Banks

Chart 7 shows a table and radar charts depicting the business model variables

Chart 7

Business Model Variables of Austria's Top 6 Banks



Bank	NTR		TTR		CTR		LDR		LAR	
	2007	2011	2007	2011	2007	2011	2007	2011	2007	2011
BAWAG PSK	71.00	75.57	11.99	0.03	17.13	19.61	89.24	108.18	46.31	58.27
Raiffeisen Zentralbank	67.46	67.92	2.17	13.95	34.75	27.62	131.97	125.30	53.18	56.03
Österreichische Volksbanken AG	58.20	336.66	5.20	–75.86	17.40	85.57	359.86	468.80	49.65	30.92
Hypo Alpe Adria International	91.99	83.63	3.21	2.27	18.62	8.12	302.71	325.84	67.61	76.06
Erste Group Bank	61.92	81.16	5.54	1.78	29.33	26.08	113.82	113.35	56.83	64.16
UniCredit Bank Austria	57.89	65.10	2.16	3.43	32.56	27.09	127.58	136.19	56.85	71.59

Source: Authors' calculations.

of Austria's top 6 banks. The axes in the chart are scaled by the minimum and the maximum of the total data set. It should be noted that Österreichische Volksbanken AG's data for 2011 are classified as an outlier in our data set and were therefore not used in the cluster analysis. The business models of Erste Group Bank, Raiffeisen Zentralbank, UniCredit Bank Austria and BAWAG PSK are very similar in their composition compared to the rest of the data set. Österreichische Volksbanken AG and Hypo Alpe Adria International exhibit higher loan-to-deposit ratios compared to the other Austrian banks.

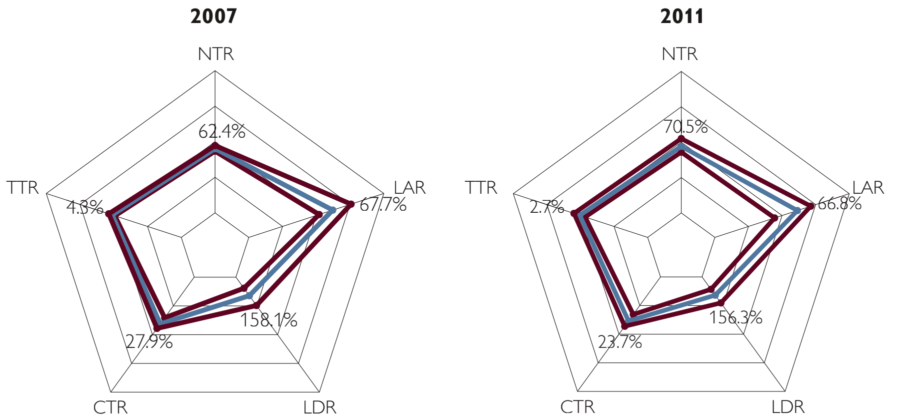
3.2.2 Business Model A

Business model A is characterized by a relatively high NTR and a high LAR, indicating that these banks generate a large portion of their income through interest income. Another notable feature is a medium CTR, pointing out business activities in wealth management. The LDR of approximately 150% is low compared to the whole sample and can be regarded as a balanced ratio.

The business model A group is the largest of the five clusters; it consisted of 125 banks in 2007 but shrank to 81 banks in 2011. Business model A shows the least changes in the five variables over the years. Chart 8 shows that between the pre-crisis year 2007 and 2011 the NTR increased by roughly 8% and the CTR decreased by 4% while the other variables exhibited a constant value. Chart 8 also shows a list of banks representative of the cluster (i.e. in terms of business model). Austria's top 3 banks (Erste Group Bank, UniCredit Bank Austria and Raiffeisen Zentralbank) are members of this group.

Chart 8

Business Model A

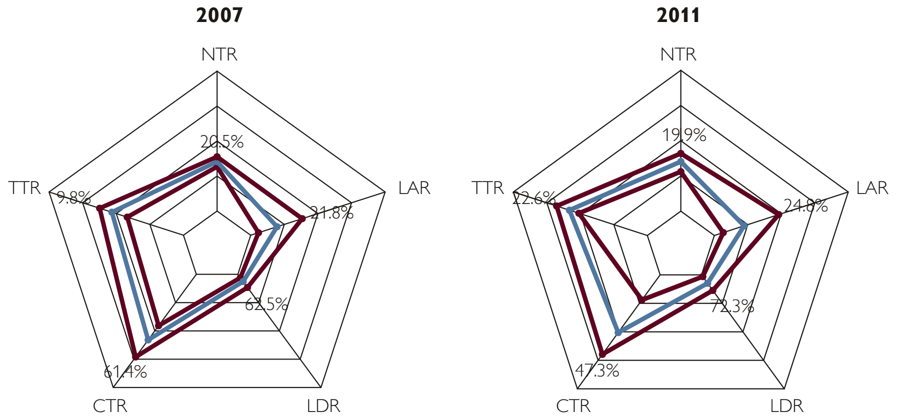


Bank name	Country	Total assets 2011 in EUR billion
Banco Santander	ES	1,251
Uni Credit SpA	IT	926
Nordea Bank	SE	716
Intesa Sanpaolo	IT	639
Danske Bank	DK	460
Svenska Handelsbanken	SE	275
DnB	NO	274
Sberbank of Russia	RU	260
Erste Group Bank	**AT**	**210**
UniCredit Bank Austria	**AT**	**199**
Raiffeisen Zentralbank	**AT**	**150**
Banco Popolare	IT	134
Banco Popular Espanol	ES	130
Le Crédit Lyonnais	FR	122
PKO Bank Polski	PL	43
Jyske Bank Group	DK	36

Source: Authors' calculations, Bankscope.

Chart 9

Business Model B



Bank name	Country	Total assets 2011 in EUR billion
Deutsche Bank AG	DE	2,164
Barclays Bank Plc	GB	1,868
UBS AG	CH	1,166
Credit Suisse Group AG	CH	862
DekaBank Deutsche Girozentrale	DE	134
Investec Plc	GB	26
Banque Internationale à Luxembourg SA	LU	24
EFG International	CH	17
Schroders Plc	GB	17
Vontobel Holding AG-Vontobel Group	CH	15
Union Bancaire Privée – UBP	CH	15
Banque Privée Edmond de Rothschild S.A., Geneve	CH	12

Source: Authors' calculations, Bankscope.

3.2.3 Business Model B

Business model B as depicted in chart 9 is characterized by a high CTR, a low LDR and a low LAR. The list of banks representative of this business model mainly consists of private banks or wealth management companies with smaller total assets. But the cluster is also defined by some large universal banks with significant investment banking operations.

3.2.4 Business Model C

Business model C is mainly characterized by a high TTR. The banks representative of this group are large international banks with investment banking activities. This cluster is similar to business model B. As mentioned before, this similarity was also visible in the bootstrapping results, which showed that only three clusters would have been optimal for the 2011 data.

3.2.5 Business Model D

Business model D in chart 11 is characterized by a high NTR, a low LAR and a medium LDR. Universal banks with strong retail operations are typical representatives of the business model D group. Of the Austrian banks, BAWAG PSK is a member of this group.

3.2.6 Business Model E

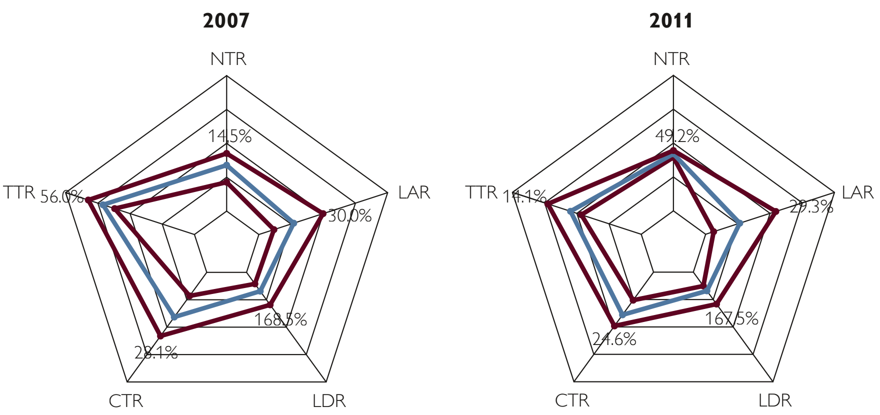
Business model E is characterized by a very high LDR and a high NTR, see chart 12. Some of these banks received state guarantees, which could explain their high loan-to-deposit ratios.

3.3 Clustering with Pooled Data

For the results presented above, we ran the clustering algorithm for each year. The shortcoming of this method was that we had to identify the order of the business models for each year because the group labels (i.e. the cluster numbering) could switch over time. A simple approach to avoid this is to pool the observations over time, i.e., our data cube becomes two-dimensional (banks, variables). The majority of Austrian banks belongs to the largest cluster, which represents – as before – business

Chart 10

Business Model C

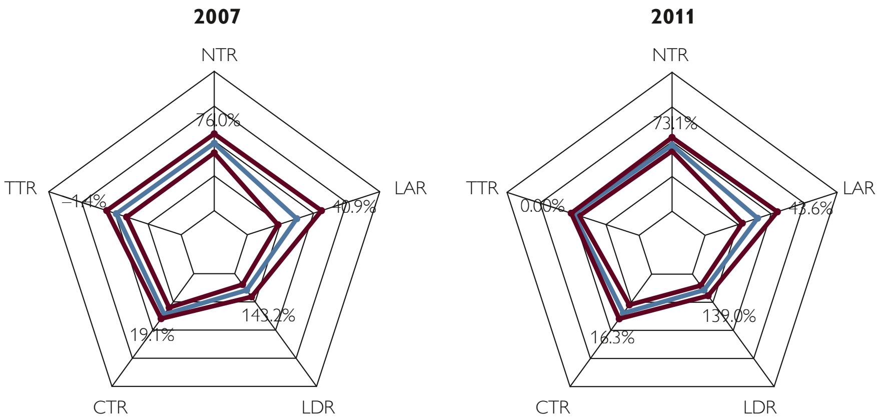


Bank name	Country	Total assets 2011 in EUR billion
Société Générale	FR	1,181
Lloyds Banking Group Plc	GB	1,160
Credit Agricole Corporate and Investment Bank-Credit Agricole CIB	FR	826
Natixis	FR	508
Fédération du Crédit Mutuel Centre Est Europe	FR	468
Merrill Lynch International Bank Limited	IE	459
HSBC France	FR	221
Standard Bank Plc	GB	21

Source: Authors' calculations, Bankscope.

Chart 11

Business Model D



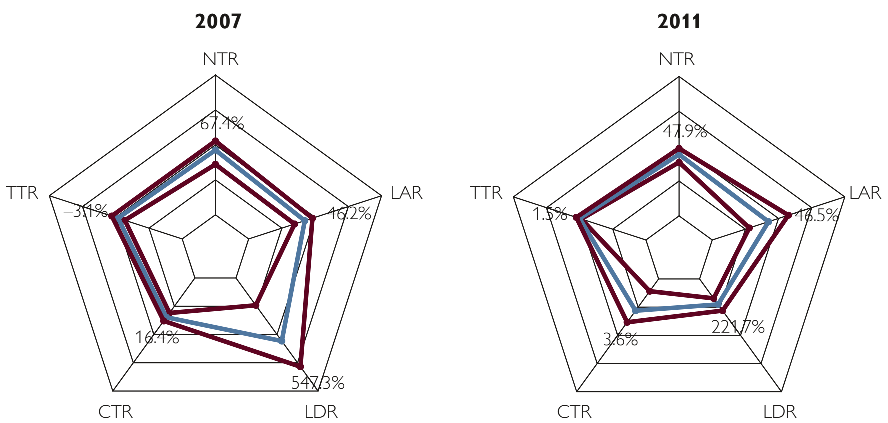
Bank name	Country	Total assets 2011 in EUR billion
BNP Paribas	FR	1,965
ING Bank NV	NL	961
DZ Bank AG-Deutsche Zentral-Genossen-schaftsbank	DE	406
UniCredit Bank AG	AT	386
Landesbank Baden-Württemberg	DE	373
Bayerische Landesbank	DE	309
Landesbank Hessen-Thüringen Girozentrale – HELABA	DE	164
WGZ-Bank AG Westdeutsche Genossenschafts-Zentralbank	DE	94
AXA Bank Europe SA/NV	BE	42
BAWAG PSK Group	**AT**	**41**

Source: Authors' calculations, Bankscope.

model A. BAWAG PSK Group is in the business model D group until 2009, but is assigned to business model A group for the years afterwards. This seems plausible because until 2009 the bank clustered more with retail-oriented banks, but in 2010 and 2011 BAWAG PSK Group showed a higher degree of similarity with the majority of banks in the business model A group. Chart 13 displays the bootstrapping results for determining the optimal number of clusters. For the pooled data setting $k = 3$ would be optimal, which is the

Chart 12

Business Model E

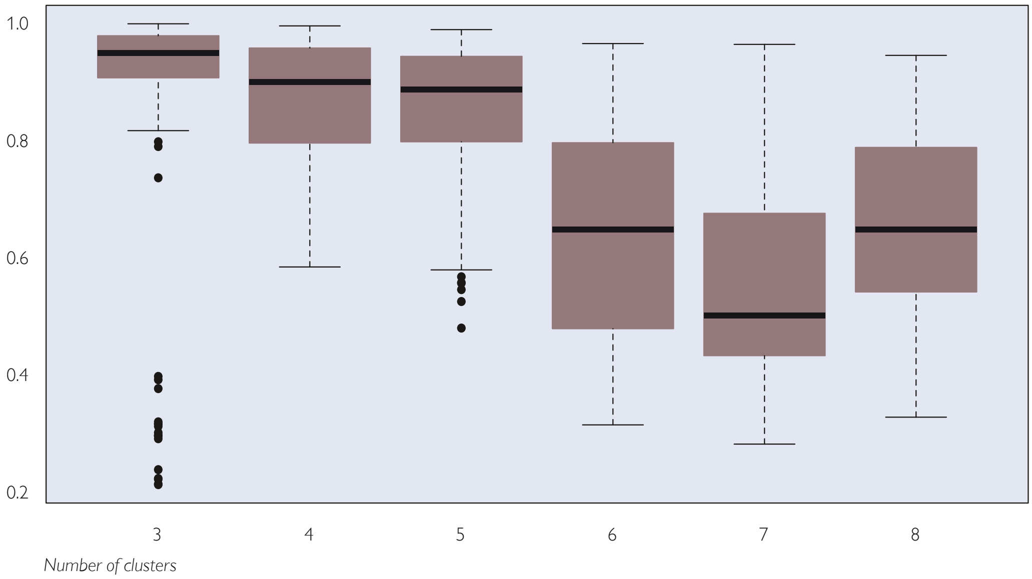


Bank name	Country	Total assets 2011 in EUR billion
Hypo Real Estate Holding AG	DE	237
Landesbank Hessen-Thüringen Girozentrale – HELABA	DE	164
Mediobanca SpA	IT	75
Pohjola Bank plc-Pohjola Pankki Oyj	FI	41
IKB Deutsche Industriebank AG	DE	31

Source: Authors' calculations, Bankscope.

Chart 13

Bootstrapping Results with Pooled Data (B = 300 Iterations)



Source: Authors' calculations.

result of combining the data for all years. In chart 6 we see that up to 2009 five business models were optimal. In the last three years the optimal number decreased to three, which is also reflected in the pooled data.

3.4 Peer Group of Austrian Banks

A peer of a certain bank is a bank whose ratios as defined in section 2 are similar to the other bank's ratios. Hence it is sufficient to consider those banks as peers that exhibit a small distance to this specific bank according to formula 3 in chapter 1.2.1.

Radius r is chosen according to the mean distances of the members of the business model group the peer group should belong to. If r is large (small) then the peer group contains many (few) elements. Erste Bank Group, Raiffeisen Zentralbank and UniCredit Bank Austria are close to each other, therefore we can define an aggregate peer group for the top 3 Austrian banks. As the banks' position in χ varies over the years, some banks belong to the peer group of the aforementioned banks in some years whereas in other years they do not. The top 3 Austrian banks are assigned to the business model A group (status of 2011). The members of this group have a mean pairwise distance of *1.4,* therefore setting the radius $r = 1.4$ seems to be most plausible. The peer group is shown in table 7 (omitting those banks with total assets below EUR 50 billion): Most of the banks in this peer group are linked with business model A and C. This suggests that the top three Austrian banks are mostly found within the business model A group, though they are not located exactly at the centroid but at some distance, deviating toward business model C.

Table 7

Peer Group of Austrian Banks

Name	Country	Business model	Total assets at end-2011
			EUR billion
KBC Bank	BE	A	241
Zurich Cantonal Bank	CH	A	101
Komerčni banka	CZ	A	28
Deutsche Postbank	DE	A	215
Jyske Bank	DK	D	33
Banco Santander	ES	A	1,218
Banco Bilbao Vizcaya Argentaria	ES	A	553
Caja de Ahorros y Pensiones de Barcelona – LA CAIXA	ES	C	286
Banco Popular Español	ES	A	130
HSBC Holdings	GB	D	1,837
Standard Chartered Bank	GB	B	386
OTP Bank	HU	A	33
UniCredit	IT	C	929
Intesa Sanpaolo	IT	C	659
Banca Monte dei Paschi di Siena	IT	C	244
Banco Popolare	IT	C	135
ING Bank	NL	A	933
DnB	NO	A	237
Powszechna Kasa Oszczędności Bank Polski	PL	A	43
Banco Espirito Santo SA	PT	C	83
Banco Comercial Português, SA-Millennium	PT	C	100
Sberbank of Russia	RU	A	212
Nordea Bank	SE	B	581
NLB dd-Nova Ljubljanska banka	SI	A	18
Türkiye Garanti Bankasi	TR	A	66

Source: Authors' calculations, Bankscope.

4 Summary

This paper shows the first results of a statistical methodology to cluster the business models of a large sample of European banks and to identify peer groups for selected banks. The analysis is based on publicly available data from banks' financial statements. We define five variables to describe the business model of a bank. A *k*-centroids clustering method based on the Mahalanobis distance is used for assigning the banks to groups that represent specific business models. We find that European banks can be grouped by five distinct business models. We provide a list of reference banks for each business model. Furthermore, we derive a peer group for Austria's top three banks based on our statistical methodology. The impact of the financial crisis on banks was clearly visible in our results, showing that banks have adapted their business models in the wake of the crisis.

References

Altunbas, Y., S. Manganelli and D. Marques-Ibanez. 2011. Bank risk during the financial crisis: do business models matter? ECB Working Paper 1394.

Ayadi, R., E. Arbak E. and W. P. De Groen. 2011. Business Models in European Banking: A pre-and post-crisis screening. In: Financial Markets, CEPS Paperbacks. Centre for European Policy Studies.

Ayadi, R., E. Arbak and W. P. De Groen. 2012. Regulation of European Banks and Business Models: Towards a new paradigm? In: Financial Markets, CEPS Paperbacks. Centre for European Policy Studies.

Everitt, B. S., S. Landau, M., Leese and D. Stahl. 2011. Cluster Analysis. 5th edition. Wiley Series in Probability and Statistics. Wiley.

Leisch, F. 2006. A Toolbox for K-Centroids Cluster Analysis. In: Computational Statistics and Data Analysis 51(2). 526–544.

Mahalanobis, P. 1936. On the generalized distance in statistics. In: Proceedings of the National Institute of Sciences of India 2(1). 49–55.

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org. Version 2.15.0, March 30, 2012.

Tortosa-Ausina, E. 2002. Cost efficiency and product mix clusters across the Spanish banking industry. In: Review of Industrial Organization 20(2). Springer. 163–181.