Validation of Credit Rating Systems Using Multi-Rater Information

Kurt Hornik[a], Rainer Jankowitsch[b], Manuel Lingo[b], Stefan Pichler[b], Gerhard Winkler[bc]

December 2005

[a] Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, Augasse 2–6, A-1090 Vienna, Austria
[b] Department of Finance and Accounting, Vienna University of Economics and Business Administration, Nordbergstrasse 15, A-1090 Vienna, Austria
[c] Oesterreichische Nationalbank, Otto Wagner Platz 3, A-1090 Vienna, Austria

Any views expressed represent those of the authors only and not necessarily those of Oesterreichische Nationalbank

---

Abstract

We suggest a new framework for the use of multi-rater information in the validation of credit rating systems, applicable in any validation process where rating information from different sources is available. As our validation framework does not rely on historical default information it appears to be particularly useful in situations where such information is inaccessible. We focus on the degree of similarity or—more generally—proximity of rating outcomes stemming from different sources and show that it is important to analyze three major aspects of proximity: agreement, association, and rating bias. We suggest using a weighted version of Cohen's κ to measure the agreement between two rating systems and we introduce a new measure for rating bias. In contrast to the existing literature we suggest τ_x as a measure of association which is based on the *Kemeny-Snell metric* and, opposed to other measures, is consistent with a set of basic axioms and should therefore be used in the context of multi-rater information. We provide an illustrative empirical example using rating information stemming from the Austrian Credit Register on partially overlapping sets of customers of 27 banks. Using a multi-dimensional scaling technique in connection with a minimal spanning tree we show that it is possible to consistently detect "outliers", i.e., banks with a low degree of similarity to other banks. The results indicate that banks which are less diversified across the size of their loans are more likely to be outliers than others.

Keywords: Rating validation, benchmarking, rating agreement, rating association, rating bias, Kemeny-Snell metric, multi-dimensional scaling, minimal spanning tree.

JEL classification: G20, C49.

---

[1]Corresponding Author: Gerhard Winkler, Oesterreichische Nationalbank, Otto Wagner Platz 3, A-1090 Vienna, Austria. *Email adress:* gerhard.winkler@oenb.at

1 Introduction

Ensuring accuracy and reliability of credit ratings by means of validation is of critical importance to many different market participants motivated by their specific objectives. In light of the requirements under the new regulatory Basel II framework banks and their supervisors have to assess the soundness and appropriateness of internal credit risk measurement and management systems.

According to this framework which is also currently implemented in an EU directive banks have several incentives to make use of internal rating systems to estimate risk parameters which are the essential input to calculate their regulatory capital requirements (Bank for International Settlements, 2004b). The most important risk parameter is the probability of default (PD) which intends to measure the likelihood of the occurrence of a default event for a certain customer over a one year horizon (Bank for International Settlements, 2004b, § 452). While the estimation of the PD is the main task of all internal rating systems, more advanced banks are allowed to estimate additional risk parameters, e.g., loss-given-default (LGD) or exposure-at-default, as well. However, the accuracy of the PD estimation is the core property of the quality of the entire rating system. Another important motivation for many banks to estimate customer-specific PDs follows from the changes in lending policy. Banks are in the process of discarding their cut-off based approach in order to adopt risk-based lending (see, e.g., Jordão and Stein (2003) for an analysis of the two approaches). In a cut-off based approach loans with an equal margin set to cover the average expected loss rate are granted to all customers who exceed some minimum level of creditworthiness. In a risk-based lending framework margins are set to cover the individual expected loss rate and thus depend on the borrower's rating. Obviously, the accuracy of rating based PD estimates is of increasing importance for banks if they adopt a risk-based lending approach. This is true for banks developing proprietary rating systems based on their own default data as well as for banks seeking to purchase third-party systems from specialized rating tool providers.

Since exposure to credit risk continues to be one of the leading sources for problems in banks world-wide it will be one of the main challenges for supervisory authorities to validate banks' abilities to measure the creditworthiness of their borrowers accurately. Safeguarding that banks correctly assess the credit risk of their loan portfolios helps to prevent individual banks from running into serious financial distress and thereby fosters financial market stability on the whole.

Furthermore, there is a rising interest from central banks since the liquidity-providing operations of central banks are usually based on underlying assets provided by the counterparties as collateral. With the aim of protecting the central banks from incurring losses in their monetary policy operations, the

underlying assets must usually meet certain minimum credit standards. The future Eurosystem-wide Credit Assessment Framework (ECAF), which defines criteria for including bank loans in the set of eligible collateral, will rely on credit quality assessments provided by different sources, like bank internal rating systems or rating agencies, using different methods. The accuracy of ratings is therefore in general essential for central banks as well (Governing Council of the European Central Bank, 2005).

Finally, this subject is of increasing importance for the long-established rating industry consisting of the rating agencies and the rating tool providers. Living on precise and timely information on the creditworthiness of borrowers they have a natural interest in validating their systems and comparing their results to those of other credit quality assessment sources.

We conclude that the development of methodologies for validating external and internal rating systems apparently is an important issue. Validation methods may roughly be divided into qualitative and quantitative (or empirical) methods (e.g., Bank for International Settlements, 2005). Qualitative methods comprise, e.g., the assessment of the model design, its internal consistency, its relation to economic theory, and the appropriateness of the related processes. As rating systems are used to provide quantitative output, qualitative validation methods are useful to check minimum requirements but provide limited information about the accuracy or reliability of estimated parameters. As a consequence, quantitative validation methods will be dominant. A common approach to validate rating systems is *backtesting*, i.e., the comparison of estimated parameters with empirical realizations. Several backtesting methods have been proposed in the literature, most of them focusing on the discriminatory power of the PD estimates. Among the best-known methods of this type are the Accuracy Ratio, the Receiver Operations Characteristic, and the Brier Score (e.g., Bank for International Settlements, 2005). To make meaningful statistical inference the requirements on the available data, however, are rather high, in particular for rating classes with small default frequencies. More advanced methods aim at directly assessing the *calibration quality*, i.e., the accuracy and reliability of PD estimates (e.g., Stein, 2002). Usually, the data requirements for these methods are even higher than for methods measuring discriminatory power. There are many cases, however, where backtesting methods are not applicable or at least too costly.

First of all this could be caused by data restrictions. For portfolios with a small number of borrowers and/or very low PDs, available data might be insufficient. This is particularly likely to happen in the first years after the implementation of the new regulatory framework. In other cases, e.g. when banks change the statistical methodology of their rating system or apply an existing system in new market segments, historical default information will not be meaningful. Further, banks seeking to purchase third-party

rating systems usually also do not have access to suitable backtesting information which enables them to compare different rating systems. Finally, in the market for collateralized short-term money market transactions, central banks have to deal with rating information simultaneously provided by different credit quality assessment sources and therefore do not have access to historical default data.

Second, backtesting could appear costly, e.g. for supervisors because they will be obliged to check the validity of numerous banks' internal rating systems simultaneously but are constrained by their limited human and financial resources.

Third, backtesting in general—as the name suggests—conveys information about a rating system's ability to predict the number of defaults in the past. Every credit quality assessment source (e.g., rating agencies) as well as supervisors might, however, additionally be interested whether banks' current estimates of the future creditworthiness of certain borrowers are in accordance among one another and whether different models produce similar results at a certain point in time.

Considering the possible drawbacks of rating validation using backtesting one could alternatively compare ratings of certain customers stemming from one specific credit quality assessment source with ratings on the same set of customers obtained from other sources. In this context the Basel Committee on Banking Supervision (BCBS) suggests a benchmarking method where banks compare their internal ratings with ratings from external rating agencies like Moody's or Standard & Poor's which are commonly available (Bank for International Settlements, 2005). Additionally, banks or supervisors may compare ratings or PD estimates with respect to ratings from a peer group of raters. Given such panels of ratings of the same objects or customers by different raters a variety of similarity or—more generally—proximity measures (Everitt and Rabe-Hesketh, 1997) can be derived, like e.g. the degree of agreement or the degree of association between two or more rating systems. Specifically, the BCBS suggests to making use of Kendall's τ and Somer's d as measures of association. As pointed out by the BCBS potential outliers may be derived, i.e., banks with rating systems with a very low degree of similarity with respect to a peer group. However, the use of multi-rater panels does not provide any information about the accuracy of the rating systems per se. Obviously, if the statistical properties (discriminatory power, calibration quality) of the peers are known additional conclusions could be drawn about the statistical properties of other members of the panel as well.

It is the main purpose of this paper to suggest a formal rating validation approach based on multi-rater information. We show how to use a multi-rater panel to derive suitable bivariate measures of proximity across different rating systems. As a measure of agreement we propose a version of Cohen's κ and

as a measure of association we suggest an extension of Kendall's τ obeying some important axiomatic requirements concerning the treatment of ties in a ranking which is particularly important for rankings based on ratings. Whereas these measures are based on a symmetric treatment of deviations between individual rating assignments, we additionally provide a method to measure the degree of "rating bias", i.e., whether a rating system on average produces too high or too low results compared to a benchmark.

For the purpose of aggregating bivariate measures of proximity we present the use of unweighted averages as one-dimensional representations and in a further step we show how to apply *Multi-Dimensional Scaling* (MDS) in combination with *Minimal Spanning Trees* as an example for a higher dimensional illustration of the results. Based on these methods one can identify rating sources with an apparently low degree of overall similarity to other rating sources, which could be seen as potential outliers, and investigate possible economic interpretations of the results.

In addition to a motivation and a formal description of our framework we present an extensive empirical example making use of a unique data set provided by the Austrian National Bank. These data are part of the Austrian Credit Register which contains information on all major loans including their ratings. Since many supervisory authorities have access to comparable data sets[2], our empirical example is also relevant for supervisors in other countries. We provide empirical estimates of the proposed bivariate measures of proximity and illustrate the aggregated results. In a final step we discuss possible economic interpretations of our results given the available bank specific information. We show evidence that a bank's diversification among the size of its customers could be an important factor explaining the quality of its rating system.

The paper is organized as follows. Section 2 motivates and formally describes the use of several measures of proximity between different rating systems. Section 3 contains a description of the data set and the empirical results. Section 4 concludes the paper.

---

[2]In fact, in many countries comparable credit registers exist. For an overview see, e.g., Jappelli and Pagano (2000).

2 Rating Validation Using Multi-Rater Information

In a rating process banks classify their customers with respect to their anticipated credit quality. Customers are grouped into rating classes of equal or indistinguishable creditworthiness. Rating classes are typically labeled by integers or capital letters like A, B, C or AAA, AA, A and so-called modifiers, '+' and '−'. In general, banks use different rating scales, i.e., different labels and/or a different number of rating classes. It is an important common property of all rating systems that rating classes are ordered with respect to some measure of creditworthiness of the customers, like the PD. Although most rating systems are used to produce PD estimates and the corresponding PD estimates might be used instead of the rating labels, it is in general not feasible to treat ratings as cardinally scaled for several reasons. Firstly and most importantly, many ratings are not aimed to primarily produce reliable PD estimates for a standardized horizon. In fact, some ratings reflect one-year PDs whereas others are deemed to correspond to longer termed PD estimates. Other ratings may reflect LGD related information such as collateral or seniority as well and therefore actually intend to measure expected loss. As an example, external rating agencies like Moody's and Standard & Poor's always point out that their rating methodology is not focused on the estimation of one-year PDs. Secondly, some banks may only provide a mapping of their rating scale into the scale of an external rating agency rather than a direct PD estimate. Thirdly, in some cases no PD information is provided at all. This applies in particular to data based on credit registers because many banks having not qualified for the Basel internal ratings based approach report their ratings to the supervisor. Thus, for the aims of our paper credit ratings are treated as ordinally scaled variables.

Following Cook et al. (1986) a *linear ordering* of a set of objects is defined as one where all objects are ranked and no ties are allowed. If unlimited ties are allowed the ranking is called a *weak ordering*. If the restriction that all objects are ranked by all raters is removed, the result is a *partial ordering*. It is straightforward to conclude that in general, the banks' internal ratings in a multi-rater panel are partial weak orderings. There are ties (usually many customers are assigned the same rating and therefore have the same rank) and not all the objects are rated by all banks. Therefore for practical applications it is essential to have the computational tools which handle such cases. Note that this case is similar to the case of preferential voting where each ranker selects only the top few candidates from a longer list (e.g., of cinema movies or books), see e.g. Fagin et al. (2003).

Consider a subset of ratings of four customers a, b, c, d provided by three banks X, Y, Z on a common rating scale where rating classes are labeled 1, 2, ..., 5. Bank X rates the customers <1 2 3 4>, bank Y rates <2 3 4 5>, and the rating provided by bank Z is <4 2 3 1>. The level of *agreement* between two

rating sources relates to the degree of similarity which for example could be expressed using the relative frequency of exactly equal assignments (i.e., proportion concordant). In the example the proportion concordant between X and Y is 0%, between X and Z it is 50% (2 equal assignments out of 4). Thus, the degree of agreement between X and Z is deemed higher than that between X and Y. Although agreement is a natural way to measure the proximity between two ratings the example shows that Y might be more similar to X than Z if other measures like *association* are used. Association relates to the number of equal pairwise comparisons among all possible pairwise comparisons (Emond and Mason, 2002). There is perfect association between X and Y (6 out of 6 pairwise comparisons are equal), whereas the degree of association between X and Z is much lower (only 1 out of 6). Finally, the (arithmetic) average rating class assigned to customers of X and Z is equal (of course "average" could be defined in other suitable ways). The ratings assigned by Y, however, are in general one rating class higher than those assigned by X. Obviously, there is a bias between the ratings X and Y.

In practical applications all three aspects of proximity are of relevance. If the relative ordering among customers is most important, one should use measures of association. This allows for the measurement of proximity even if rating systems use different rating scales (with a different number of rating classes). On the other hand, if the absolute meaning of the ratings is of key interest (e.g., for collateral valuation) measures of agreement should be used. Finally, the inspection of potential rating bias helps to determine whether deviations are caused by noise or by systematic factors.

In the simplest case when rating outcomes of only two sources are compared it is useful to select customers which both banks have in common and to aggregate and display the rating information in a contingency table, i.e., a matrix C whose elements c_{ij} count the number of cases where the first bank rates a specific customer into class i whereas the second bank rates the customer into class j.

The elements on the main diagonal count the cases where the two rating sources agree. The off-diagonal elements count cases where the two rating sources disagree by $|i-j|$ rating classes. An informal inspection of the contingency table allows for a preliminary judgment about the amount of proximity between the two rating sources. To gain further insight it is necessary, however, to obtain formal measures of agreement, association, and rating bias.

Measuring Agreement. As a measure of agreement we suggest using a weighted version of Cohen's κ (hereafter, simply "κ") employing the Cohen-Fleiss weights (Fleiss and Cohen, 1973). This measure is obtained as follows. Let $N_c = \sum_{i,j} c_{ij}$ be the number of customers rated by both banks, and

$p_{ij} = c_{ij}/N_c$ be the joint proportion of customers rated into classes i and j, respectively. Also, write $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$ for the marginal proportions of customers rated into class i by the first bank and into class j by the second bank, respectively. Weighted κ (Spitzer et al., 1967) measures agreement as

$$\kappa_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} \tag{1}$$

and compares the observed agreement $P_{o(w)} = \sum_{j=1}^{R} \sum_{i=1}^{R} w_{ij} p_{ij}$ to that expected if the ratings were independent (and hence $p_{ij} = p_{i\cdot} p_{\cdot j}$), given by $P_{e(w)} = \sum_{j=1}^{R} \sum_{i=1}^{R} w_{ij} p_{i\cdot} p_{\cdot j}$, where in general the weights w_{ij} satisfy $0 \leq w_{ij} \leq 1$ with $w_{ii} = 1$ and $w_{ij} = w_{ji}$, and are used to describe closeness of agreement. R denotes the number of rating classes. We use the weights proposed by Fleiss and Cohen:

$$w_{ij} = 1 - \frac{(i-j)^2}{R-1} \tag{2}$$

which measure the strength of disagreement as quadratic in the difference in rating classes.

Measuring Association. For any reasonable measure of association a suitable distance metric has to be defined. In our case, such a metric measures the distance between any two elements of the set S of all partial weak orderings (of the same set of objects) with a finite, non-negative value such that the distance between any element of S and itself is zero. Since S is finite, distances are bounded by some constant $D_{\max}$. Given such a distance d, the transformation $d \mapsto 1 - 2d/D_{\max}$ converts d into an association measure with values in $[-1, 1]$ such that the association between identical rankings is 1. Desirable properties of distance metrics are inherited by the corresponding association measure, and vice versa (Emond and Mason, 2002).

Text books and statistical software packages most commonly suggest using Spearman's ρ or Kendall's τ as association measures. Spearman's ρ corresponds to a simple distance metric where for linear orderings all ranks are treated as natural numbers and for weak orderings ratios are used to break the ties (Kendall and Gibbons, 1990). Kendall's τ is based on a metric which is defined as the minimal number of interchanges or flips necessary to achieve a zero distance, i.e., to make one ranking equal to the other (e.g., Emond and Mason, 2002). It can be shown that both distance metrics do not fulfill very basic and widely accepted axioms first outlined in Kemeny and Snell (1962). One of these basic properties of a distance metric is the familiar triangle inequality which states that the distance from any element to any other cannot be greater than the distance calculated by going through a third element. Due to the treatment of ties both Spearman's ρ and Kendall's τ do not meet this requirement (Kemeny and Snell, 1962; Emond and

Mason, 2002).

Along the lines of Emond and Mason (2002), we make use of an extended version of Kendall's τ, called τ_x, which corresponds to the metric proposed by Kemeny and Snell. To calculate τ_x for two banks (A and B) we have to calculate a score matrix for each bank using only the rating information of the customers the two banks have in common. These matrices contain the following elements depending on the pairwise comparisons of the customers of bank A:

$$a_{uv} = \begin{cases} 1 & \text{if customer } u \text{ is ranked ahead of customer } v, \\ 0 & \text{otherwise,} \end{cases}$$

with b_{uv} defined similarly for bank B. Using these score matrices, τ_x can be calculated as

$$\tau_x(A,B) = \frac{\sum_{u=1}^{N_c}\sum_{v=1}^{N_c} a_{uv}b_{uv}}{N_c(N_c-1)/2} \tag{3}$$

Measuring Bias. Based on the logic behind the usage of Cohen's κ to measure the degree of agreement between two rating sources we define a near-at-hand measure of rating bias. For any *co-rating* (customer rated by both banks) the deviation between the two rating classes, $i-j$, is obtained. The average deviation among all co-ratings is then scaled to the interval $[-1,+1]$. This defines our measure of rating bias.

$$\theta(A,B) = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R}(i-j)c_{ij}}{N_c(R-1)} \tag{4}$$

Having introduced the above measures for the three different aspects of proximity, it is important to note that in many applications the computation of these bivariate measures might not be sufficient to draw the desired conclusions. If n denotes the number of different raters or banks in an analysis, a comparison of all raters yields $n \cdot (n-1)/2$ bivariate measures. As the number of resulting measures increases with the square of the number of raters for applications with more than a few raters there is a need for a suitable aggregation of the results. In a simple approach one could use arithmetic averages (weighted or unweighted) to represent the aggregated measure of proximity of a certain rater. Raters with a very low average τ_x or κ, or with an average θ large in absolute value, are deemed to be "outliers". Weights might be used to represent the economic importance of different raters dependent on the specific scope of the analysis. However, these averages are one-dimensional representations of the bivariate measures and more dimensions may be needed to capture the main features of the measures.

An alternative procedure to overcome this problem is to employ multi-dimensional scaling techniques. In a canonical application of MDS the raters are displayed on a two-dimensional map where the distances between the locations of the raters are related to the differences in the underlying proximities. If the explanatory power of a two-dimensional representation is too weak a minimal spanning tree might be used additionally to identify potential outliers. In the empirical example presented in the following section we make use of both techniques to illustrate their application in a real-world setup.

3 Empirical Analysis

3.1 Data Description

For our empirical analysis we were provided with a unique data set from the Austrian banking market by the Austrian National Bank. Compared to the whole European banking market the Austrian banking sector is rather small in size. In terms of total assets which amounted to EUR 635 bn at the end of 2004 on a consolidated basis the Austrian sector represents a fraction of 2.2% of the whole EU banking sector (European Central Bank, 2005b). The Austrian market, however, is highly developed and its banks play a prominent role in the process of financial intermediation: Relating the size of the Austrian banking sector to the economic output of Austria as measured by the GDP shows that intermediation depth (total assets over GDP) is fairly high with 268% at the end of 2004 (median intermediation depth for all EU countries: 227%, see European Central Bank (2005b)). With a share of the five largest credit institutions in total assets of 43.8%, concentration in the Austrian banking market is below the average of the EU member states' banking sectors (52.9% at the end of 2004, see European Central Bank (2005b)) whereas profitability as measured by the ROE of 14.6% at the end of 2004 is slightly above the EU average of 14.2% (European Central Bank, 2005a). Given these characteristics one may conclude that the market under consideration is representative for other developed banking markets.

The data set we use comes from the Credit Register of the Austrian National Bank. This database contains monthly information on all major loans (in excess of EUR 350,000) granted by credit and financial institutions as well as contract insurance companies in Austria. Like many other credit registers the Austrian Credit Register provides detailed information on the characteristics of each individual borrower and each loan including: the name, the address, the legal form, as well as the economic sector affiliation of each obligor, the type of instruments used for refinancing, and the outstanding as well as the undrawn amounts of credit commitments. With a total amount of EUR 373 bn at the end of 2004 approximately

67% of the total volume of all bank lending activities in Austria was covered by the credit register's recordings (Oesterreichische Nationalbank, 2004). In total 51% of all the major loans to Austrian borrowers was granted to corporates, 33% to financial institutions, 9% to the private sector, and the remaining 7% to public sector entities (Oesterreichische Nationalbank, 2004).

According to the Austrian Banking Act, all financial institutions have to supply monthly information to the credit register on any changes in the status of their outstanding credits as well as information on the new loans granted during the period. In general, the high level of confidentiality and a cost-free use by the participating institutions to obtain information on the total bank debt of any of their customers in exchange makes the credit register an appealing instrument for the participating institutions. The many different uses of credit registers and the positive contribution to the credit market by credit registers have been thoroughly discussed in the theoretical and empirical literature (for an overview see Bank for International Settlements, 2004a). Clearly, credit registers also have enormous potential to allow supervisory authorities to address many of the validation and benchmarking issues that Basel II entail.

The Austrian Credit Register is unique in this respect as information on credit risk has been recorded in a very detailed form since the beginning of 2003. Every participating institution has to report borrower specific information on provisions for loan losses, on the value of collateral it has at hand and—most important for the aims of this paper—on its own assessment of the individual borrower's creditworthiness expressed by a rating. Additionally, banks have to provide the supervisor with detailed information on the set-up of their rating scales concerning its granularity (i.e., the number of rating classes) and the level of risk associated with each class. Based on the exact descriptions of banks' rating scales the banks' ratings are consistently mapped on a common scale by the central bank to allow for better comparison of ratings issued by different institutions. This standardized harmonized rating scale, called the Austrian National Bank's master scale, is closely related to the major rating agencies' rating scales. It consists of eight buckets, four for investment grade entities (AAA–BBB) and four for non-investment-grade debtors (BB–D).

The sample provided by the central bank for the purposes of the empirical study stems from the reportings of 27 large Austrian banks to the credit register as of May 2005. Taken together the volume of major loans reported by these 27 banks to the credit register accounted for 67% of the total volume of all major loans at this point in time. In terms of total assets the 27 banks represent 69% of the total banking assets of the whole Austrian banking market for that year. Both ratios suggest that the sample banks and their loan portfolios can be considered to be highly representative for the market under consideration. For our

empirical example we were supplied with rating information on the borrowers of our sample banks. In order to be able to measure the proximity between the ratings based on the methodology outlined in the previous section we can only take rating information on those obligors into account which were assessed by at least two banks in the sample. Thus, we had to restrict the 45,746 ratings on 14,194 borrowers reported by our sample banks to 15,576 ratings on 5,911 borrowers. For each and every borrower out of these 5,911 at least two ratings are available. This data set forms the basis for further analysis.

Table 1 shows descriptive statistics for the composition of our data set.

— insert Table 1 —

Note that even the smallest bank has at least 157 debtors in common with one or more of the other institutions. Apart from looking at the number of co-ratings on a bank level, we also compute the number of co-ratings on an obligor level. The median number of these co-ratings is 3, suggesting that most customers have business relations to only a small number of banks.

In addition to the rating data on an individual obligor level, certain bank specific attributes provided by the central bank complete our data set. From both the size variables at hand, total assets and total volume of major loans granted, one could conclude that the sample banks are indeed amongst the larger ones on a national level, but would be considered as small- to medium-sized institutions on an EU level (see European Central Bank (2005a) for further details regarding the main characteristics of the EU banking sector). Finally, to capture the degree of diversification in the banks' loan portfolios, we furthermore obtained three different (standardized) Herfindahl indices from the central bank which measure relative concentration of each bank across different loan sizes, industries and regions. These indices are computed as

$$HI = \frac{\sum_{i=1}^{N} \left(\frac{X_i}{\sum_{j=1}^{N} X_j} \right)^2 - 1/N}{1 - 1/N} \tag{5}$$

where $X_1, \ldots, X_N$ denote the levels of the relevant variables, i.e., the size of each individual loan, and the aggregated volume of loans per industry and region, respectively. These indices take on values between 0 (representing perfect diversification) and 1 (indicating total concentration, i.e., no diversification). Overall it seems that the sample banks manage well-diversified portfolios, especially regarding the degree of relative concentration with respect to size. Although somewhat higher than the values for diversification across size, the Herfindahl indices measuring diversification across different industries and regions also take on rather low values indicating again considerable diversification. This comes as no surprise as all

banks operate at least on a nationwide basis, most of them supranationally, and service customers from all different industries and parts of the country.

3.2 Empirical Example

In this section we present the results for our empirical example. First of all we compute the contingency tables and the bivariate measures as described in Section 2 and show the results for two banks of our sample. In the next step these measures are aggregated to provide a suitable assessment of all banks with respect to agreement, association, and rating bias. For this purpose we present unweighted averages as one-dimensional representations and in a further step we apply multi-dimensional scaling in combination with minimal spanning trees as an example for a higher dimensional illustration of our results. Based on these methods we identify banks which could be seen as potential outliers, and provide some economic interpretation of the results. The presented methodology could be applied by banks to benchmark their rating systems or by supervisory authorities to identify banks with a potentially lower quality rating system.

As the first step in measuring proximity we present the contingency table for bank 1 and bank 27 as an illustrative example (see Table 2). The contingency table shows the number of cases where the first bank rates a specific object into class i whereas the second bank rates the object into class j.

— insert Table 2 —

In this contingency table the number of elements below the main diagonal is higher than the number of elements above, indicating that bank 27 on average assigns higher ratings than bank 1. Bank 1 considers none of the 848 jointly rated customers to be in the highest rating class whereas bank 27 assigns 335 customers to this class. Of these 335 customers bank 1 considers 304 customers to be in rating class 2 and 31 in rating class 3. Even more interesting is the default class, where bank 1 assigns 31 customers in contrast to the 26 customers assigned by bank 27. Out of these customers only 19 are classified as default by both banks. However, it is not possible to conclude whether bank 1 or bank 27 has the superior rating system using this information alone. Nevertheless, one can see that there are substantial differences between the results of different raters.

Based on the information of the contingency table we compute the bivariate measures between bank 1 and bank 27 presented in Section 2. We find a rather high positive association ($\tau_x = 0.768$) and agreement ($\kappa = 0.781$) but also a remarkable rating bias ($\theta = 0.099$) which on average corresponds to 0.693 (=

$0.099 \cdot 7$) rating classes on our eight-stage master scale. Note that it is still not possible to judge which bank has the superior rating system.

Therefore one could compute the bivariate measures for bank 1 and bank 27 with respect to all other banks (i.e., 26 pairwise combinations for each bank) and use these resulting measures to evaluate the overall level of proximity for these two banks with all other banks. Figures 1 and 2 show the resulting measures.

— insert Figures 1 and 2 —

These results can be used to analyze bank 1 and bank 27 with respect to agreement, association, and rating bias by comparing their respective bivariate measures. Starting with association represented by τ_x it seems that bank 1 and bank 27 perform nearly equally well. To compare the twenty-six bivariate τ_x of each bank one could compute the arithmetic average (see Table 3) resulting in an average τ_x of 0.602 for bank 1 and 0.617 for bank 27. Comparing agreement measured by κ one can see that on average bank 1 yields considerably higher values than bank 27 (0.722 for bank 1 versus 0.631 for bank 27, see Table 3). This difference in agreement can mainly be explained by rating bias. Whereas the rating bias for bank 1 oscillates around zero we observe an average rating bias of -0.1 for bank 27 (see Table 3), indicating too favorable ratings. This result can be used to eventually interpret the bivariate contingency table. Therefore one could conclude that the rating system of bank 27 is able to properly rate a customer relative to its other customers in a way that the bank's ratings allow for an accurate ordering which is in line with orderings produced by the rating systems of the other banks. However, even if bank 27 performs well when it comes to association, its ratings clearly show a upward bias, which also has an obvious effect on its agreement measure and leads to the conclusion that its rating system may not be well calibrated. Based on these results one would have more confidence in the rating system of bank 1.

The comparisons of the bivariate measures of bank 1 and bank 27 enable us to gain valuable information about the quality of the rating systems of these two banks. Additionally, one can use these measures simultaneously to judge the rating systems of the other banks as well. As an illustrative example, a simple visual inspection of Figures 1 and 2 shows that the rating system of bank 24 deviates considerably from those of banks 1 and 27, for all three aspects of proximity. This indicates that the rating system of bank 24 is potentially of lower quality.

In the next step these measures have to be calculated for all possible pairwise combinations of banks. We therefore obtain a unique τ_x, κ, and θ for each combination. To formally compare the rating systems for

different banks with respect to a certain measure one needs to aggregate the bivariate measures to come to a conclusion which enables one to identify banks with low degree of similarity. We start with a heuristic approach where we calculate the unweighted arithmetic average of the bivariate measures for each bank (as we have done in the previous comparison of bank 1 and bank 27) which yields a one-dimensional aggregated measure for τ_x, κ and θ for each bank. We also employed the median and a weighted average, using the number of co-ratings as weights, to aggregate the bivariate measures. Since the results are virtually identical in particular for banks with a low degree of similarity we only present the results for the unweighted average. Table 3 shows the average τ_x, κ, and θ for all banks.

— insert Table 3 —

Using the average value of the bivariate measures provides a very useful tool to detect banks which are *outliers* with respect to a certain measure, which could be interpreted as an indicator for a weak rating system and could be chosen by the supervisory authorities for further analysis. However, there is no defined minimum level or cut-off level for the average measures to identify outliers. Therefore a prespecified number of banks with the lowest average τ_x or κ, or the largest average θ in absolute value, can be taken as potential outliers. The particular number can be chosen in accordance with the relevant objective, e.g., the supervisory authorities could choose a number dependent on their on-site audit resources. For the purpose of this paper the number of outliers is set to five which represents approximately 20% of the banks in this sample. In Table 3 the five banks with the lowest degree of similarity are highlighted. Bank 24 has the lowest average τ_x confirming the bivariate results presented for banks 1 and 27. Comparing the five banks with the lowest τ_x (banks 3, 17, 19, 21, and 24) to the results for κ we find that four out of the five banks with the lowest τ_x are also in the group of the five banks with the lowest κ (banks 3, 19, 21, and 24), showing an overall low degree of similarity. Nevertheless the two groups are not identical, emphasizing the importance of measuring both agreement and association. Bank 22 for example, has a high τ_x and a low κ whereas the reverse is true for bank 7. The average θ of the five banks with the highest rating bias (banks 2, 3, 20, 21, and 27) corresponds to half a rating class which indicates a severe miscalibration of their rating systems. We find that three of the five banks with the highest bias are not in the group of the five banks with the lowest agreement and association. Therefore the results for the rating bias show that association and agreement cannot detect all potential shortcomings of a rating system. There are significant biases in some rating systems which do not result in low agreement or association (bank 27 is the best example for this case). Again this indicates that it is important to measure all three aspects of proximity. Emphasizing this importance, Figure 3 shows the

three measures for each bank, where banks are ordered with respect to τ_x, illustrating that certain banks with rather high τ_x have low κ and/or θ.

— insert Figure 3 —

Depending on the objective of the benchmarking process the importance of the results for agreement, association, and rating bias can be defined. For some objectives agreement will be crucial while for others association or rating bias might be of more interest.

In addition to the marginal analysis based on average proximities, one can attempt modeling the *joint* collection of proximities in order to clarify, display and possibly "structure" the underlying numerical proximity values. A very popular family of such models is provided by MDS (e.g., Everitt and Rabe-Hesketh, 1997) which can be characterized as the search for a low-dimensional space in which each point in the space represents one original entity of interest (here, a bank) such that the distances between the points in the space match as well as possible, in some sense, the original proximities (distances or similarities). One can then e.g. plot the first two coordinates of these points to obtain a visual representation of the (proximity structure of the) banks.

— insert Figure 4 —

Here, we use classical (metric) MDS. Figure 4 shows the MDS plot using $1-\tau_x$ as a distance measure. As described in the previous section, banks with similar rating systems should be close to each other whereas banks that use completely different rating systems are expected to be rather distant on the MDS map. Interpreting these results one can see that most banks are concentrated in the center. Situated on the edge of the plot the banks 10, 19, 20, 21 and 24 are clearly different and could be considered as potential outliers. This seems to be the case for the banks 19, 21 and 24 which had the lowest average τ_x. Banks 10 and 20, however, showed a rather high average τ_x, being among the best banks using the aggregated τ_x to measure the overall level of association.

This indicates that one can not use a two-dimensional MDS plot alone for identifying outliers. In fact, the first two components only account for about 28.8% of the total variation in the proximities, and considerably more components are necessary to capture "most" of the variation. One thus needs to indicate the cases in which the visual map poorly represents the underlying proximity structure. One such approach is based on the so-called Minimal Spanning Tree (MST). A spanning tree connects pairs of points (here, banks) in a way that all points are connected and no closed loops occur. A minimal

spanning tree is one where the sum of the "lengths" of the connecting edges (here, the sum of the distances between the banks connected) is minimal. If the edges of the MST are added to the two-dimensional scaling representation, distortions are indicated when nearby points on the plot are not joined by an edge of the tree. See e.g. Chapter 4 in Everitt and Rabe-Hesketh (1997) for more information.

Looking at the MST in Figure 4 we can clearly see that banks 10 and 20 have multiple connections with other banks whereas the banks 19, 21 and 24 only show one connection on the MST. This illustration provides additional support for the arithmetic average results and shows that the combination of aggregated τ_x figures and the MDS/MST representation yield a powerful set of tools to identify outliers.

— insert Figure 5 —

Comparing these results with the MDS plot in Figure 5 where $1 - \kappa$ was used as a distance measure, we can see that the set of outliers is exactly the same. The banks 10, 19, 20, 21 and 24 are on the edge of the plot of which three banks were among the five banks with the lowest average κ (banks 19, 21 and 24). The MST shows a slightly different picture, but banks 19, 21 and 24 still have far less connections to other banks than banks 10 and 20 which validates that the latter have a higher degree of overall similarity. However, the plot is not fully compatible with aggregated κ figures since banks 3 and 6 cannot be identified as outliers using only the MDS/MST plot. Nevertheless the MDS/MST plots show that the average measures though being a heuristic approach are able to produce consistent results.

Having identified outliers, supervisory authorities might be interested to find out why certain banks provide ratings of their customers which differ markedly from those of other banks for at least one aspect of proximity. A natural point to start with is to test whether the group of outliers differs from the remainder of the sample in terms of bank specific attributes. Given the bank-specific information available in our data set we formulate two hypotheses; one with respect to size as a potential explanatory variable, the other with respect to a bank's degree of diversification.

Hypothesis 1: *Outliers are smaller in size than the remainder of the banks.*

One possible explanation for this hypothesis is that smaller banks lack the resources to install a suitable rating system. This could either be due to the high cost of purchasing a third party rating system, which are often too expensive for small banks, and/or due to the lack of sufficient default data in a small bank's portfolio necessary for developing a proprietary rating system. To check hypothesis 1 we employ Wilcoxon's rank sum test (Wilcoxon, 1945), which is a nonparametric alternative to the two-sample t-test. Table 4 shows the p-values for the Wilcoxon tests as well as the medians of the two groups for the

bank specific size variables described in Section 3.1. Size1 measures the size of a bank's loan portfolio whereas Size2 measures a bank's total assets.

— insert Table 4 —

For the outliers we observe a median Size1 of EUR 2.76 bn for τ_x, 2.47 for κ, and 2.79 for θ compared to EUR 3.07, 3.26, and 2.64 bn for non-outlier banks. Considering Size2 we observe for the outliers a median Size2 of EUR 5.24 bn for τ_x and κ, and 11.18 for θ compared to EUR 9.69, 9.51, and 8.48 bn for non-outlier banks. Although all τ_x and κ outliers are rather small in size one is not able to find significant differences between the group of outliers and the regular banks using the Wilcoxon test. The reason for this is that a considerable number of small banks performed rather well, indicating that size is not the driving factor for being an outlier. Therefore one could consider other possible attributes that might be different for the group of outliers.

Hypothesis 2: *Outliers are less diversified than the remainder of the banks.*

This hypothesis can be motivated by the circumstance that less diversified banks have to rate relevant groups of customers that are not part of their core business. This situation can arise due to several factors. First of all, low diversification across loan size can lead to poor ratings, e.g., banks with mainly small and medium sized enterprise customers with small loan sizes might have problems in rating corporate customers with large loan sizes. The same argument holds for diversification across industries and across regions, e.g., banks focused on specific industries or regions might lack well calibrated rating systems for customers of other industries/regions.

— insert Table 5 —

Table 5 shows the results for the diversification variables of which diversification across loan size exhibits significant differences between the two groups based upon τ_x and κ (medians of the outliers: 0.0185 (τ_x), 0.0139 (κ), 0.0070 (θ) versus non-outliers: 0.0069 (τ_x), 0.0069 (κ), 0.0073 (θ)). Diversification across industries showed no significant differences for the outliers (median of outliers: 0.1902 (τ_x), 0.1902 (κ), 0.0699 (θ) versus non-outliers: 0.0889 (τ_x), 0.0889 (κ), 0.1094 (θ)). The degree of diversification of customers across regions, however is not significantly different for either of the two groups of outliers (median of outliers: 0.3092 (τ_x), 0.2864 (κ), 0.2472 (θ) versus non-outliers: 0.2875 (τ_x), 0.3306 (κ), 0.3403 (θ)). The reason for that could be that most banks operate nationwide and differences in diversification across regions might very well be quite unincisive. Interestingly there are no significant differences when

the groups are divided according to rating bias θ, which indicates that rating bias occurs independently from size and diversification.

Using bank-specific attributes to explain lower figures for association, agreement, and bias shows no dependence on the size of the portfolio or the total assets and some dependence on the diversification of the portfolio with respect to loan size. Thus it seems that banks which are focused on a certain market segment but still have relevant exposures in other segments have problems to produce sound ratings for all their customers. To provide deeper insights customer-specific information, which is not available in our data set, would be necessary to identify the customer groups for which a certain bank has potentially less consistent ratings.

4 Conclusion

In this paper we suggested a new framework to use multi-rater information for the validation of credit rating systems. This framework is useful for any validation process where rating information from different sources is available. This applies to validation of bank internal ratings in the Basel II supervisory review process, to the benchmarking of ratings in the context of the future Eurosystem-wide Credit Assessment Framework, and to the internal validation processes adopted by banks.

We showed that it is important to analyze three major aspects of proximity, i.e., agreement, association, and rating bias. We suggested using a weighted version of Cohen's κ to measure agreement between two rating systems and we developed a new measure θ for the rating bias, which represents the average difference in rating classes between the two rating systems. Obviously, measures of agreement and rating bias can only be applied if the rating information is measured on the same scale. We suggest τ_x as a measure of association which is based on the Kemeny-Snell metric. In contrast to the existing literature where other measures of association are promoted like Kendall's τ_b, Somer's d, Kruskal's γ, and Spearman's ρ, τ_x is consistent with a set of basic axioms and should therefore be used in the context of multi-rater information.

In an empirical example based on data of the Austrian Credit Register we compared rating information provided by 27 banks on partially overlapping sets of customers. We obtain 351 bivariate measures for every aspect of proximity. To visualize the results and to draw conclusions we used unweighted averages across banks and a multi-dimensional scaling technique in connection with a minimal spanning tree. We showed that it is possible to consistently detect "outliers", i.e., banks whose bivariate agreement or

association measures are much lower compared to other banks. The results indicate that banks with a low degree of diversification across the size of their loans are more likely to be outliers than others.

References

Bank for International Settlements. A review of credit registers and their use for Basel II. Working paper, Financial Stability Institute, Bank for International Settlements, CH-4002 Basel, Switzerland, 2004a. URL http://www.bis.org/fsi/awp2004.pdf.

Bank for International Settlements. International convergence of capital measurement and capital standards: A revised framework, 2004b. URL http://www.bis.org/publ/bcbs107.htm.

Bank for International Settlements. Studies on the validation of internal rating systems (revised), 2005. URL http://www.bis.org/publ/bcbs_wp14.pdf.

W. D. Cook, M. Kress, and L. M. Seiford. Information and preference in partial orders: A bimatrix representation. *Psychometrika*, 51:197–207, 1986.

E. J. Emond and D. W. Mason. A new rank correlation coefficient with applications to the consensus ranking problem. *Journal of Multicriteria Decision Analysis*, 11:17–28, 2002.

European Central Bank. EU banking sector stability. Press Release, October 2005a. URL http://www.ecb.int/pub/pdf/other/eubankingsectorstability2005en.pdf.

European Central Bank. EU banking structures. Press Release, October 2005b. URL http://www.ecb.int/pub/pdf/other/eubankingstructure102005en.pdf.

B. Everitt and S. Rabe-Hesketh. *The Analysis of Proximity Data*. Edward Arnold Publishers Ltd, 1997.

R. Fagin, R. Kumar, and D. Sivukumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

Governing Council of the European Central Bank. Eurosystem collateral framework: Inclusion of non-marketable assets in the single list. Press Release, July 22 2005. URL http://www.ecb.int/press/pr/date/2005/html/pr050722.en.html.

T. Jappelli and M. Pagano. Information sharing in credit markets: The European experience. CSEF Working Paper 35, Centre for Studies in Economics and Finance (CSEF), University of Salerno, Italy, 2000. URL http://ideas.repec.org/p/sef/csefwp/35.html.

F. Jordão and R. Stein. What is a more powerful model worth. Technical Report #030124, Moody's KMV, 2003. URL http://www.moodyskmv.com/research/whitepaper/MorePowerfulModel_TR030124.pdf.

J. G. Kemeny and J. L. Snell. *Mathematical Models in the Social Sciences*, chapter Preference Rankings: An Axiomatic Approach. MIT Press, Cambridge, 1962.

M. G. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Oxford University Press, 1990.

Oesterreichische Nationalbank. Statistisches Monatsheft 12/2004. Wien, 2004.

R. Spitzer, J. Cohen, J. L. Fleiss, and J. Endicott. Quantification of agreement in psychiatry diagnosis: A new approach. *Archives of General Psychiatry,*, 17:83–87, 1967.

R. Stein. Benchmarking default prediction models: Pitfalls and remedies in model validation. Technical Report #020305, Moody's KMV, 2002. URL http://riskcalc.moodysrms.com/us/research/crm/Validation_Tech_Report_020305.pdf.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

	Minimum	Median	Maximum
Characteristics of rating information provided by sample banks			
Number of obligors per bank for which at least one co-rating exists	157	395	3,006
Number of ratings per obligor	2	3	27
Characteristics of sample banks			
Size of banks measured by their total assets in EUR billions	2.834	8.660	142.238
Size of banks' loan portfolios according to credit register entries in EUR billions	0.816	2.794	40.011
Diversification of loan portfolios across sizes (Herfindahl Index)	0.002	0.007	0.030
Diversification of loan portfolios across industries (Herfindahl Index)	0.044	0.095	0.311
Diversification of loan portfolios across regions (Herfindahl Index)	0.058	0.290	0.581

Table 1: Descriptive statistics for the composition of our sample

	Bank 27							
Bank 1	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	304	12	3	0	0	0	0	0
3	31	33	73	28	1	0	0	0
4	0	9	45	74	16	1	2	0
5	0	5	25	66	40	2	4	6
6	0	0	2	11	14	4	2	2
7	0	0	0	2	0	0	0	0
8	0	0	1	5	2	1	3	19

Table 2: Contingency table for the ratings of bank 1 and bank 27 across rating classes. The elements on the main diagonal count the cases where the two rating sources agree. The off-diagonal elements count cases where the two rating sources disagree by $|i - j|$ rating classes.

	Measure of Proximity		
Bank	τ_x	κ	θ
1	0.602	0.723	0.001
2	0.550	0.586	0.040
3	0.539	0.495	0.082
4	0.666	0.680	0.011
5	0.621	0.721	−0.007
6	0.573	0.467	0.010
7	0.588	0.713	0.032
8	0.611	0.689	−0.027
9	0.647	0.674	0.010
10	0.696	0.696	0.010
11	0.637	0.682	0.025
12	0.610	0.710	0.004
13	0.611	0.732	−0.005
14	0.672	0.741	−0.020
15	0.681	0.663	0.026
16	0.638	0.709	−0.025
17	0.544	0.628	0.012
18	0.563	0.657	−0.026
19	0.500	0.517	0.025
20	0.647	0.625	−0.063
21	0.469	0.485	−0.033
22	0.625	0.623	0.006
23	0.669	0.727	0.019
24	0.449	0.478	0.014
25	0.610	0.635	−0.014
26	0.646	0.686	−0.006
27	0.617	0.631	−0.100

Table 3: Aggregated bivariate measures for all banks. The five banks with the lowest degree of similarity are highlighted.

	Bank Attributes	
Measure of Proximity	Size1	Size2
τ_x p-value	0.5239	0.1857
Median Regulars	3.0716	9.6855
Median Outliers	2.7562	5.2390
κ p-value	0.2572	0.2078
Median Regulars	3.2587	9.5050
Median Outliers	2.4671	5.2390
θ p-value	0.4472	0.4112
Median Regulars	2.6417	8.4765
Median Outliers	2.7942	11.178

Table 4: Wilcoxon p-values and medians of the two groups for bank specific size variables.

	Bank Attributes		
Measure of Proximity	div. loan size	div. industries	regional div.
τ_x p-value	0.0101	0.1653	0.4472
Median Regulars	0.0069	0.0889	0.2875
Median Outliers	0.0185	0.1902	0.3092
κ p-value	0.0468	0.1857	0.4848
Median Regulars	0.0069	0.0889	0.3306
Median Outliers	0.0139	0.1902	0.2864
θ p-value	0.6065	0.1857	0.5645
Median Regulars	0.0073	0.1094	0.3403
Median Outliers	0.0070	0.0699	0.2472

Table 5: Wilcoxon p-values and medians of the two groups for bank specific diversification variables.

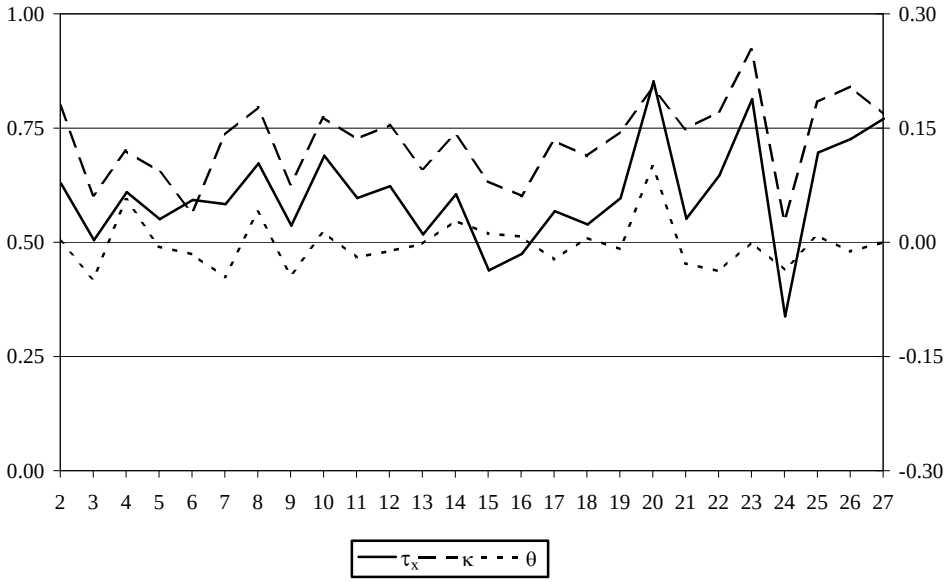

Figure 1: Bivarate comparison of bank 1 with all other banks using τ_x, κ, and θ. The left axis scales τ_x and κ whereas θ is scaled by the right axis. The x-axis shows the banks 2 to 27.

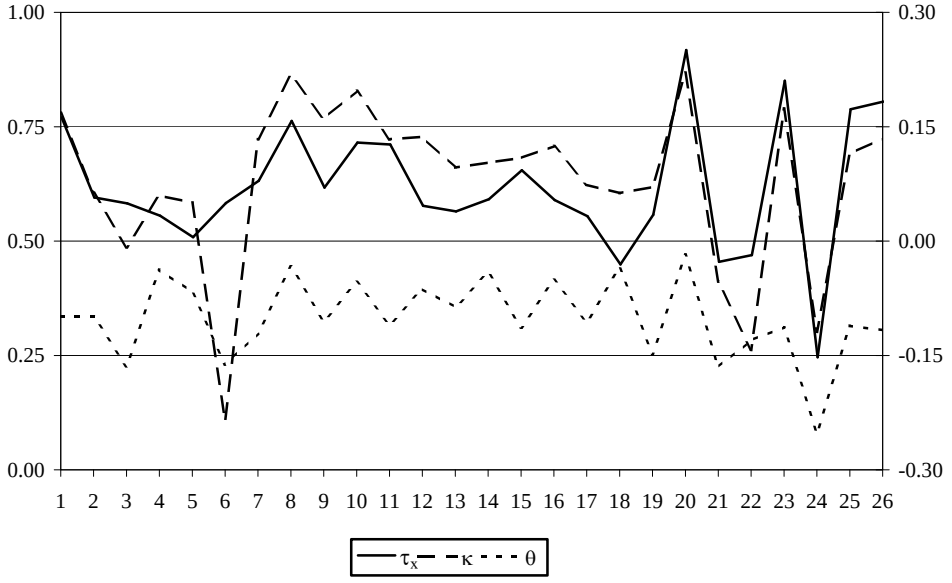

Figure 2: Bivariate comparison of bank 27 with all other banks using τ_x, κ, and θ. The left axis scales τ_x and κ whereas θ is scaled by the right axis. The x-axis shows the banks 1 to 26.

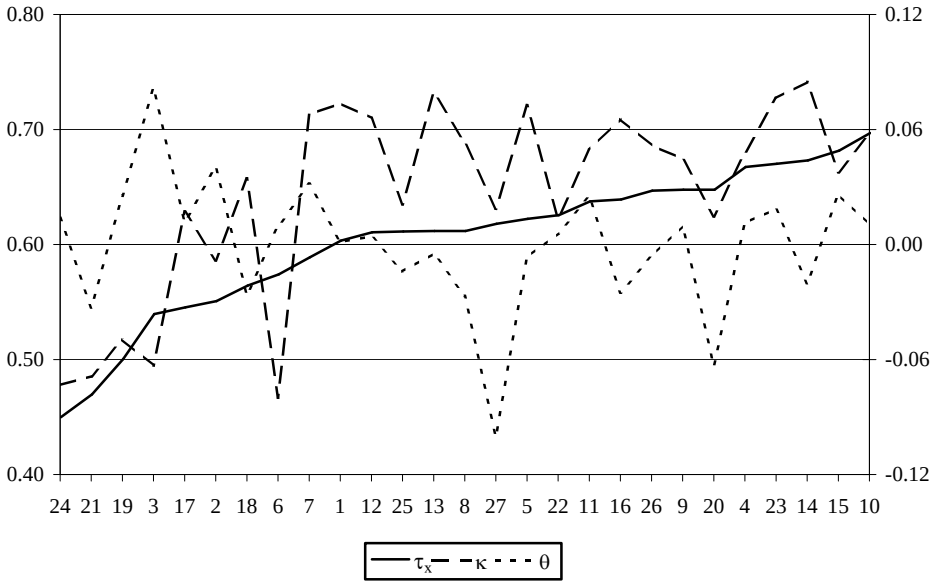

Figure 3: Aggregated τ_x, κ and θ ordered with respect to τ_x. The left axis scales τ_x and κ whereas θ is scaled by the right axis. The x-axis shows the banks ordered with respect to τ_x.

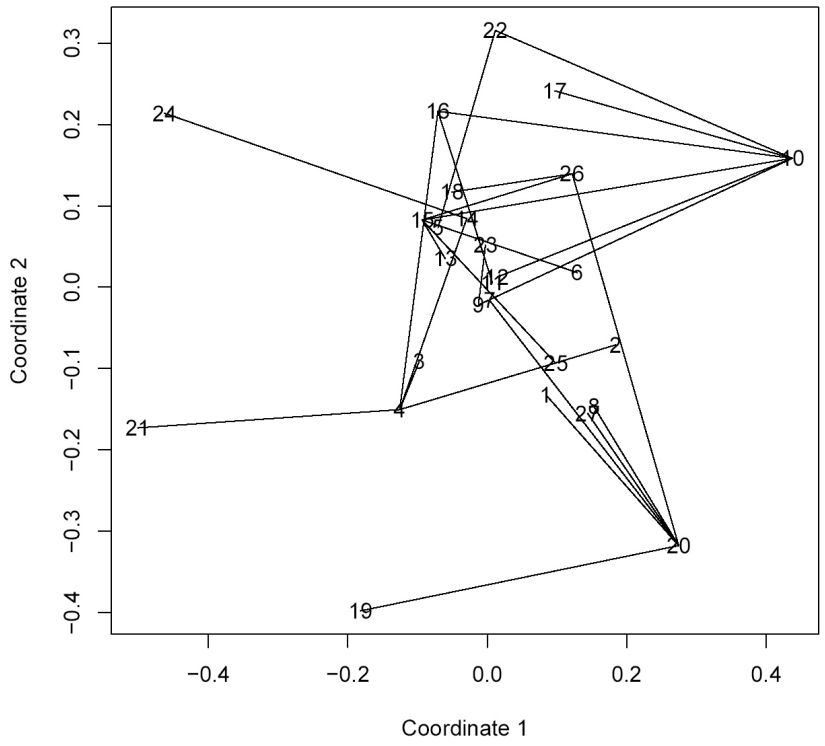

Figure 4: MDS plot for τ_x, including the Minimal Spanning Tree.

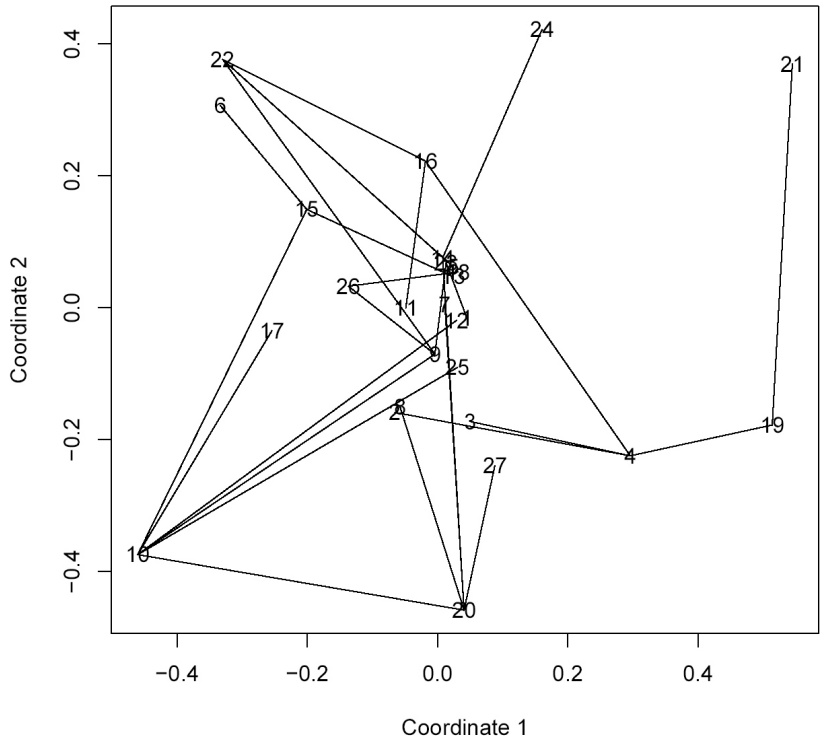

Figure 5: MDS plot for κ, including the Minimal Spanning Tree.