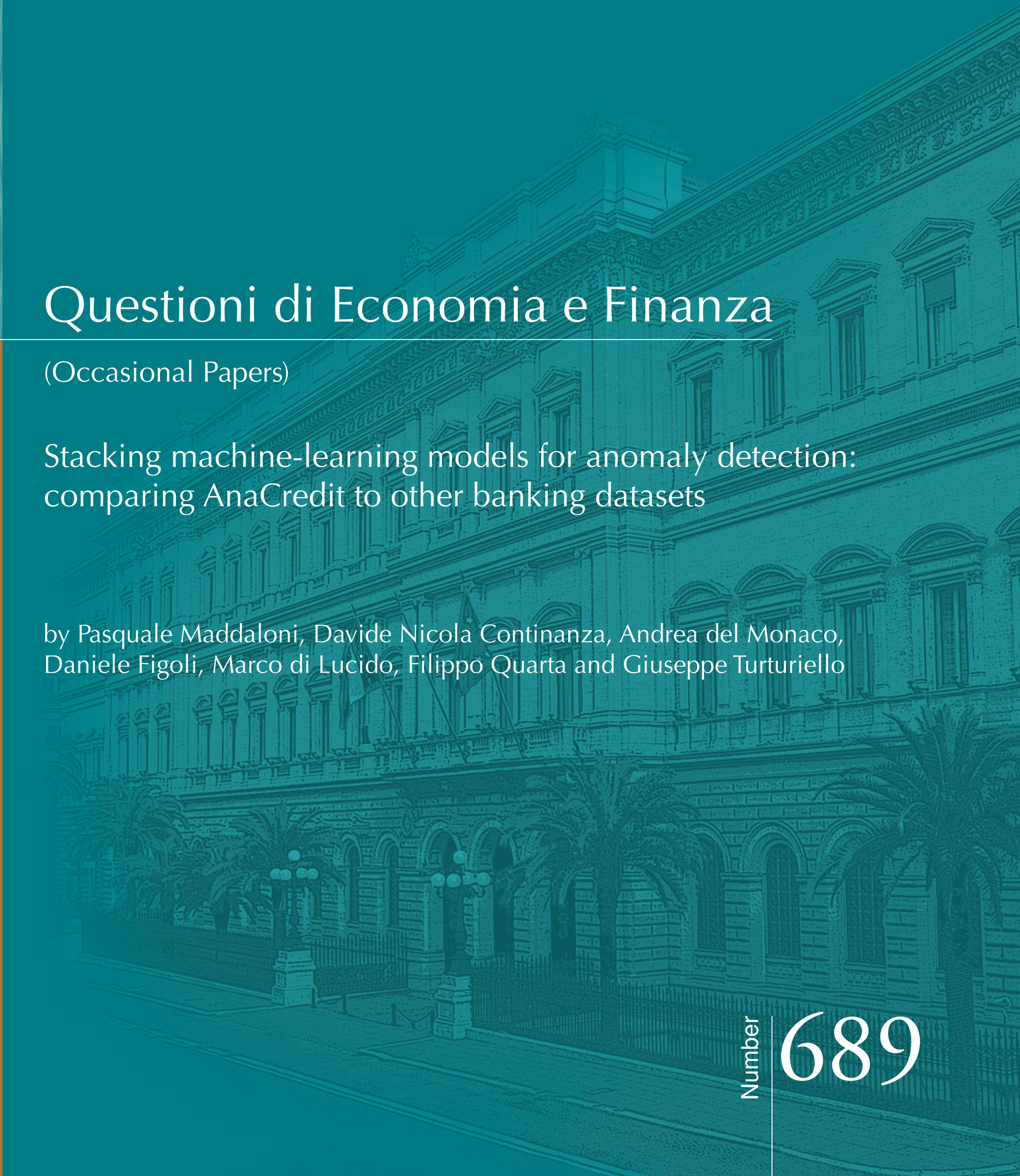
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking datasets

by Pasquale Maddaloni, Davide Nicola Continanza, Andrea del Monaco, Daniele Figoli, Marco di Lucido, Filippo Quarta and Giuseppe Turturiello

April 2022

Number 689

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking datasets

by Pasquale Maddaloni, Davide Nicola Continanza, Andrea del Monaco, Daniele Figoli, Marco di Lucido, Filippo Quarta and Giuseppe Turturiello

Number 689 – April 2022

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

STACKING MACHINE-LEARNING MODELS FOR ANOMALY DETECTION: COMPARING ANACREDIT TO OTHER BANKING DATASETS

by Pasquale Maddaloni[*], Davide Nicola Continanza[*], Andrea del Monaco[*], Daniele Figoli[*], Marco di Lucido[*], Filippo Quarta[**], Giuseppe Turturiello[**]

Abstract

This paper addresses the issue of assessing the quality of granular datasets reported by banks via machine learning models. In particular, it investigates how supervised and unsupervised learning algorithms can exploit patterns that can be recognized in other data sources dealing with similar phenomena (although these phenomena are available at a different level of aggregation), in order to detect potential outliers to be submitted to banks for their own checks. The above machine learning algorithms are finally *stacked* in a semi-supervised fashion in order to enhance their individual outlier detection ability.

The described methodology is applied to compare the granular AnaCredit dataset, firstly with the Balance Sheet Items statistics (BSI), and secondly with the harmonised supervisory statistics of the Financial Reporting (FinRep), which are compiled for the Eurosystem and the Single Supervisory Mechanism, respectively. In both cases, we show that the performance of the stacking technique, in terms of F1-score, is higher than in each algorithm alone.

JEL Classification: C18, C81, G21.

Keywords: banking data, data quality management, outlier and anomaly detection, machine learning, auto-encoder, robust regression, pseudo labelling.

DOI: 10.32057/0.QEF.2022.0689

Contents

[*] Bank of Italy, Statistical Data Collection and Processing Directorate.

[**] Bank of Italy, IT Development Directorate.

1 Introduction[1]

Big-data analytics is increasingly being adopted within the community of central banks, for several purposes (Cagala, 2017; Chakraborty *et al.*, 2017). An important area regards the application of machine learning techniques in order to improve the quality of data collected on the basis of regulatory reporting. Over the last few years, such surveys have become more granular and complex, in order to allow a better understanding of economic developments and, more in general, to improve the assessment of the actual and potential impact of policies on the economy[2]. As regards banking data, a key role is played by credit disbursement to the economy that, in Italy, represents more than two thirds of banks' total assets.

The main sources of credit data currently used at the Bank of Italy are the Eurosystem's collection of Balance Sheet Items (BSI), the EU harmonized Financial Reporting (FinRep), the Italian Central Credit Register data (CCR) and, for a couple of years now, the Eurosystem's granular collection AnaCredit.

This paper investigates the possibility of building statistically founded cross-checking between the highly granular AnaCredit survey and the aggregated BSI and FinRep statistics by exploiting the similarities shared by the three surveys with respect to the phenomena that are covered. Originally, the three surveys were designed for different purposes and so the actual data collections follow different reporting rules and definitions with regard to the types of loans that are collected, the reporting population, the data model and the transformation rules. More importantly for our purposes, BSI and FinRep are very well established and mature data collections, whereas AnaCredit is quite a recent one, so it might not have achieved the same high quality standards of the other two yet. This is why in this paper, in defining a new set of quality checks, we try to exploit the information available in BSI and FinRep to improve the quality of AnaCredit data through outlier detection techniques.

To set up a new set of data quality checks, the expertise of the analysts needs to be complemented with the use of advanced statistical tools that allow us to handle the complexity of a highly granular survey such as AnaCredit. In this respect, the basic idea of the paper is to resort to machine learning techniques to carry out systematic cross-checking between series on the same phenomena although pertaining to different data collections in order to identify potential outliers to be submitted to reporting banks for their own checks[3].

From a methodological point of view, we rely on machine learning methods (Bishop, 2011; Hastie *et al.*, 2001 and 2013) in order to overcome some of the limits recognized in the statistical literature on outlier detection as regards the identification of the boundary separating 'normal' observations from outliers. These limits are related both to the possibility that 'normal behaviour' might not be static but, rather, evolve over time and also to the lack of labelled data for training models (Chandola *et al.*, 2009). Within this research field (Cusano *et*

[1] The authors are grateful to Gianluca Cubadda and Alessio Farcomeni (University of Tor Vergata, Rome), Francesca Monacelli and Roberto Sabbatini (Bank of Italy) for their useful comments and fruitful discussions on a preliminary draft of the paper. The views expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy.

[2] For a recent discussion see Cœuré, 2017.

[3] Namely, records that are considered anomalous because they are significantly different from the other points of the dataset (Aggarwal, 2017).

al., 2021; Zambuto *et al*., 2020; Farnè *et al*., 2018; Goldstein *et al*., 2016), the novelty of this paper lies in the development of a general approach that makes a pairwise comparison between datasets containing information on similar phenomena.

We show that the proposed methodology, based on an ensemble learning technique, detects anomalies with a higher level of precision than the single methods used as baselines. Since anomalous observations are rare, the main metric considered to evaluate the performance of our developed models is the F1-score. With reference to this metric, the ensemble technique adopted also yields better results than the single baselines. In sum, we will show that the actual implementation of this methodology can contribute to improving the quality of AnaCredit data to the extent that the pairwise comparison with BSI and FinRep databases can lead to a more accurate list of potential outliers to be submitted to the cross-checking of reporting banks. It is worth remarking how the approach developed in this paper can be applied, more generally, to all those situations in which it is possible to exploit the information contained in aggregated datasets to detect potential outliers in a highly granular dataset.

The paper is organized as follows. Section 2 describes the three datasets under consideration and the deterministic pre-processing treatment carried out in order to make it possible to compare the aggregated series available for each of them. Section 3 explores the different strategies considered for detecting outliers and illustrates the developed ensemble machine learning techniques within a semi-supervised setting. Section 4 presents the results of the proposed approach. Section 5 summarizes the main conclusions, outlining the advantages of the proposed method and the possible directions for future research.

2 Data

Bank of Italy, in the context of the harmonized collections at the European level, collects aggregated credit information mainly within the scope of two 'surveys'[4]: the monthly Balance Sheet Items (BSI), which is used for the common monetary policy analysis, and the quarterly Financial Reporting (FinRep) used for SSM supervisory purposes. Both surveys capture credit phenomena at aggregated level; different contractual forms of loans (i.e. overdrafts, mortgages, repurchase agreements) are added together by the amount paid out and then they are broken down by the relevant characteristics of the borrowers (i.e. sector and the residence) and by the main contractual features (currency, maturity, etc.). The global financial crisis of 2007-08 and the European debt crisis of 2009-10 showed that such aggregated data had been not sufficient to fulfill users' need. This consideration led to the issue of Regulation (EU) 2016/867 on the collection of monthly granular credit and credit risk data (ECB/2016/13), the so-called AnaCredit Regulation, aimed at making available a new granular and multipurpose dataset containing loan-by-loan information on credit. Indeed, AnaCredit focuses

[4] For the purpose of this paper, a 'survey' is a collection of homogenous data for a given purpose and disciplined by a reporting framework.

on the single credit instrument issued by credit institutions, within a contract stipulated vis-à-vis a given borrower (Di Noia et al., 2020). The main innovation brought by AnaCredit, as compared to BSI and FinRep data, rely on a larger number of details, at the level of single loan granted to counterparty provided that it is above the reporting threshold of 25,000 euros. This new unprecedented granular credit data collection allows the European System of Central Banks (ESCB) to carry out its tasks having a view of the entire distribution of this financial phenomenon. Furthermore, this data is more suitable to shed light on lending dynamics to legal entities and on the accumulation of risky debts in the banking sector.

In the following sub-sections, we describe the pre-processing steps carried out to build the two datasets used for our analysis. In particular, we use for our comparison only AnaCredit data starting from December 2018, although the first reporting date was September 2018. We decided to skip first reporting dates that, as it is often the case, present a very high degree of instability in terms of the quality of data, which is typically connected to the effective implementation and settlement of the compilation rules by reporting banks.

2.1 AnaCredit vs. BSI

The first comparison we carry out is between AnaCredit and BSI data collections from Italian banks. The latter refers to monthly aggregated stocks on assets and liabilities of Italian banks' balance sheets and it is used to compile the national contribution to Eurosystem's monetary statistics. BSI loans aggregates are based on data provided by reporting banks, which are then aggregated by amount according to some relevant loan information: the characteristics of the underlying contracts (type of instrument, duration and currency) and some classification variables of the contract counterparty (sector and residence). As anticipated, loans are particularly relevant being the core business of banks as well as the largest fraction of their assets.

For the purpose of this work, we take into account the main BSI time series of loans broken down by original maturity of credit instruments, the residence and the institutional sector of the borrower. In particular, we consider the following breakdowns: 1) Domestic Monetary Financial Institutions (MFIs), excluding Central banks; 2) Domestic Central Banks; 3) Other Euro area MFIs; 4) Domestic General Government; 5) Other Euro area General Government; 6) Euro area Other Financial Institutions and non-Money Market investment funds; 7) Euro area Insurance Corporations and Pension Funds; 8) Domestic Non-Financial Corporations (NFCs), original maturity up to 1 year; 9) Domestic NFCs, original maturity over 1 to 5 years; 10) Domestic NFCs, original maturity over 5 years; 11) Other Euro area NFCs, original maturity up to 1 year; 12) Other Euro area NFCs, original maturity over 1 to 5 years; 13) Other Euro area NFCs, original maturity over 5 years.

The ECB and the National Central Banks (NCBs) have already developed cross-checks between BSI and AnaCredit based on a deterministic approach: outliers are identified when the figures of interest exceed a pre-specified threshold that, for each BSI time series, is the same across all banks and reference dates. Typically, such thresholds are expressed in terms of percentage changes in the values detected in the two surveys. The current quality-control system would largely benefit of a statistical approach aimed at identifying acceptance thresholds that are bank-specific and can change over time.

The comparison between Anacredit and BSI surveys requires the pre-processing of their differences in order to make data more comparable. To this end, we build a joint dataset and then we focus on the following differences. Firstly, we drop out from AnaCredit all those loans that are not recognized in the bank's individual balance sheet, since BSI includes only loans for which banks bear credit risk. Secondly, since AnaCredit contains accounting information only for end-of-quarter months, we impute the status of recognition of loans for the other two months of the quarter[5]. Thirdly, as new loans purchased on market are present in BSI at purchase price but in AnaCredit they are reported at nominal value, we discount the corresponding AnaCredit values by the difference between the nominal value and the price at the time of purchase. Fourthly, we derive in AnaCredit the classes of original maturity of loans present in BSI as the time between the settlement and the final legal maturity date of the contract expressed in years. Finally, reconciliation of data structures is performed in order to obtain comparable aggregates, by mapping the same subportfolio (e.g. interbank loans) of loans. Following the above preliminary adjustments, we can aggregate AnaCredit data by amount for the same bank and reference date and for the same characteristics of the BSI series, then obtaining the 'AnaCredit equivalent' specification of the 13 BSI series listed above. For both BSI and 'AnaCredit equivalent', the 13 considered series are further broken down by the sector of economic activity (NACE[6]) of the counterparty and the currency of the instrument[7]. The comparison between the two sets of indicators is carried out over the time span December 2018-March 2020 (monthly observations).

2.2 AnaCredit vs. FinRep

FinRep is the harmonized supervisory financial reporting that each credit institution must report on a quarterly basis according to the instructions of Regulation EU 680/2014 (Implementing Technical Standards - ITS) and International Financial Reporting Standards (IFRS). FinRep comprises accounting data on assets, liabilities, equity and statement of profit and loss. Within the assets of the balance sheet statements, reporting banks are also required to provide detailed information on loans, broken down by accounting portfolio, institutional sector and economic activity (NACE classification) of the counterparty, type of instrument, credit quality status and past due bands. It is relevant to underline that we exclude from the comparison all FinRep loans referred to households (institutional sectors S.14 and S.15 according to ESA 2010 classification) since not in all cases they are reported in AnaCredit. In order to derive the counterparty sector in FinRep, we resort to the detailed reporting rules defined by PUMA2 documentation[8].

[5] We assume that the last accounting evaluation is still valid for the non-end-of-quarter months and that all the new financial instruments are recognized. In general, it is quite rare that the recognition status of a financial instrument change from a month to another. Furthermore, the quota of new financial instrument is small and these new instruments are almost surely recognized.

[6] See: Statistical Classification of Economic Activities in the European Community, Rev. 2 (2008) (NACE Rev. 2).

[7] See Figures A1, A2 in Appendix A.

[8] The main goal of PUMA2 process is to generate financial information for the production of several different statistical and supervisory reports. PUMA2 documentation provides detailed transformation rules to generate the final statistical reports from a granular input layers. For more details, see https://www.cooperazionepuma.org/.

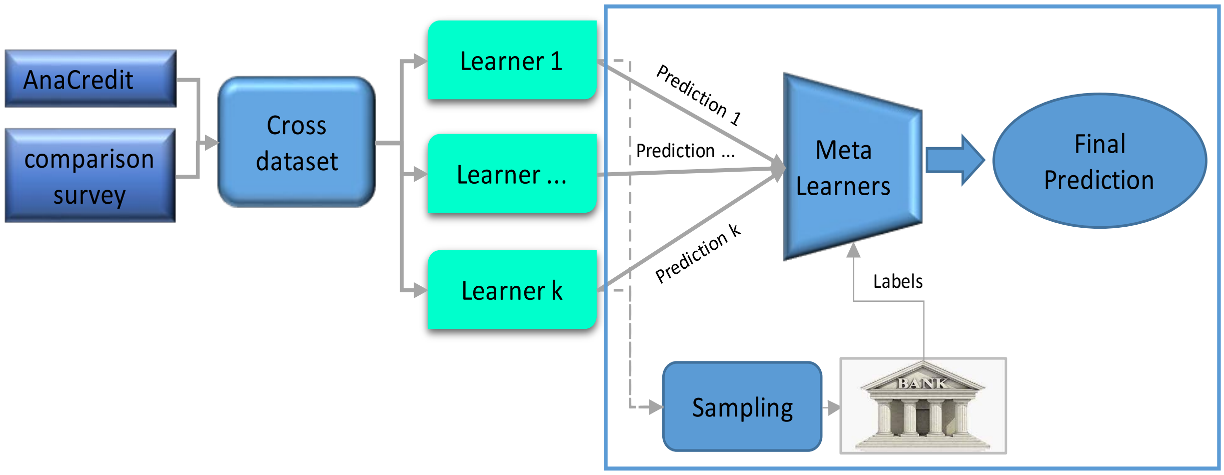
The three main measures of the accounting AnaCredit data, i.e. (1) net and gross carrying amount, (2) accumulated impairment amount and (3) accumulated changes in fair value due to credit risk, are compared to the equivalent measures of FinRep. We report some examples of disaggregated series elaborated for the comparison of the two dataset in Figure A3 of Appendix A.

As for the BSI comparison, a pre-processing of data is necessary to overcome a few differences between the two surveys. In particular, the main information that needs to be reconciled refers to: counterparty sector; type of instruments; the evaluation of past-due bands; the evaluation of gross carrying amount; the evaluation of accumulated negative changes on fair value due to credit risk on non-performing exposures. As in the case of BSI, the output is a joint dataset containing FinRep series and their equivalent (reconstructed) aggregation based on AnaCredit data. The comparison is carried out with reference to end-of-quarter dates, over the time span December 2018-March 2020.

3 Anomaly detection strategies: definitions and estimation procedures

Our cross-checking is based on two preliminary considerations. Firstly, BSI, FinRep and AnaCredit contain similar information on loans; therefore, we can assume that the patterns of the series referred to the same phenomena are similar. Secondly, the quality of BSI and FinRep datasets is very high, as improvements have been introduced over many years. So we assume that the potential outliers identified on the basis of 'divergences' between the compared series – BSI vs. AnaCredit and FinRep vs. AnaCredit, respectively – can be attributed to anomalies in AnaCredit data. The statistical approach that is followed combines supervised and unsupervised methods for the identification of regular patterns. In particular, for the supervised approach we develop a robust regression model, whereas for the unsupervised approach we resort to two autoencoder models. The three base models above ('learners') are then combined via a 'stacking algorithm' consisting of an additional classifier ('meta-classifier') trained on the base models' outputs (Figure 1).

*Figure 1: Workflow**



* In our work only 3 learners are used as base models.

The meta-classifier allows us to synthesize the complementary insights derived from the different base models and to outperform each one of them in making the final prediction[9]. In particular, the meta-classifier is trained in a semi-supervised setting by using a dataset enriched with the binary labels 'anomalous' or 'not-anomalous' that, with reference to sample cases[10], are attached to each observation on the basis of cross-checking with the intermediaries and pre-assessments based on the domain knowledge.

It is worth anticipating that the strength of the above approach lies in its versatility in terms of use cases to which it can be applied and 'learners' that can be considered. Actually, this method can be easily adapted to any comparison between data collections sharing similar information and each specific base models ('learners') could be swapped with others yielding better predictions and their number can also be changed.

It is worth noting that the prediction of the model, i.e. the list of potential outliers, refers to aggregates which are themselves a decomposition of BSI or FinRep aggregates. This detail allows the anomaly to be contextualized by elements that better explain it (such as the information that helps the reporting banks to identify the erroneous records in their own archives and the reasons behind the data verification request), allowing the intermediaries for a more effective and faster evaluation of the case.

3.1 Robust Regressions

As mentioned in previous paragraphs, loans to legal entities reported in AnaCredit, BSI and FinRep refer to the same information content, although at a different level of aggregation. The conceptual relationship between the phenomena, confirmed empirically by the high correlation between AnaCredit aggregates, on one side, and BSI and FinRep series, on the other[11], can be statistically exploited within a linear regression framework. In our model BSI (or FinRep) series represents the independent variable, given its ascertained high level of quality, whereas the equivalent AnaCredit aggregate is regarded as the dependent variable.

Despite the pre-processing steps described in Section 2 to build comparable credit statistics, there are inevitable and permanent structural differences between the two datasets under comparison (e.g. the reporting threshold effect and the exclusion of natural persons in AnaCredit). Therefore our linear regression introduces a specific explanatory variable to capture such structural differences. This variable is able to consider such differences as intrinsic and normal instead of as reporting mistakes.

We end up with the following equation, using a log transformation of the original variables[12]:

$$log(A_{i,j,t}) = \beta_0 + \beta_1 log(F_{i,j,t}) + \beta_2 log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t} \,, \tag{1}$$

where I denotes a bank, j a sub-portfolio of loans and t is a reference date. The AnaCredit aggregate for a particular sub-portfolio of loans at time t $(A_{i,j,t})$ is compared with the correspondent amount of BSI $(F_{i,j,t})$

[9] See Lessmann *et al.*, 2015.

[10] In future developments of this paper, data revisions could also be considered to further enrich the prior knowledge.

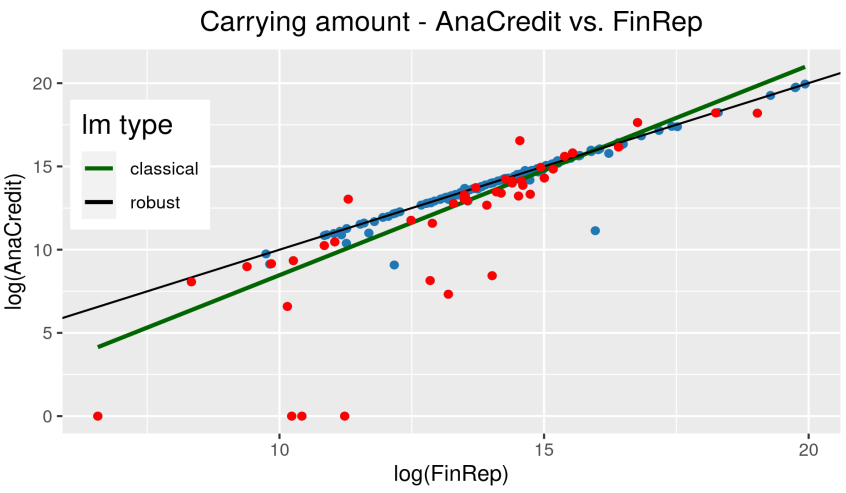
[11] See Figure 4 in Appendix A for the distributions of such index.

[12] For the derivation of the complete model see Appendix B.

(the same comparison holds for FinRep). The second explanatory variable is added to capture the definitional and structural differences between the datasets. The reporting mistakes remain isolated and they are contained only in the error component $\epsilon_{i,j,t}$. Unfortunately, we cannot identify the structural difference at time t, because of the presence of (potential) reporting errors in $A_{i,j,t}$. Instead, such differences are well identified at previous times, *t-1*, *t-2*, etc., on the basis of the findings of the data quality validation process in those periods. For the sake of simplicity, our model considers only the differences at time *t-1*. This way, the equation (1) could be read as an error correction model for the cross comparison of the two aggregates at time *t* (for more details, see Appendix B). We do not consider some seasonality form in equation (1), as we have short series available: only 5 dates for the FinRep/AnaCredit and 16 for BSI/AnaCredit comparison. Indeed in our stock data, we do not observe such element to a significant extent[13] (see Figures A2 and A3 in Appendix A).

In an 'ideal' context, i.e. without anomalous data, the relationship in equation (1) would be correctly estimated. The presence of anomalous data in the dependent variable spoils practically this relationship. However, the literature on statistical robust estimation helps us to handle this issue (Hampel 1985; Hampel *et al.*, 1986; Farcomeni, 2015; Gschwandtner, 2012; Maechler, 2021), as shown in Figure 2, where the logarithm of AnaCredit carrying amounts (y-axis) and the logarithm of FinRep carrying amounts (x-axis) for the time series of 'loans versus central governments, evaluated at amortized cost' are plotted according to classical and robust linear predictions.

Figure 2: Classical vs. Robust regression



The robust regression (black line) is not affected by anomalous data like high leverage data points - such as red dots that are lying on the x-axis- as it happens in the classical linear regression (green line). The presence of high leverage data points has a significant impact on the parameters of the regression.

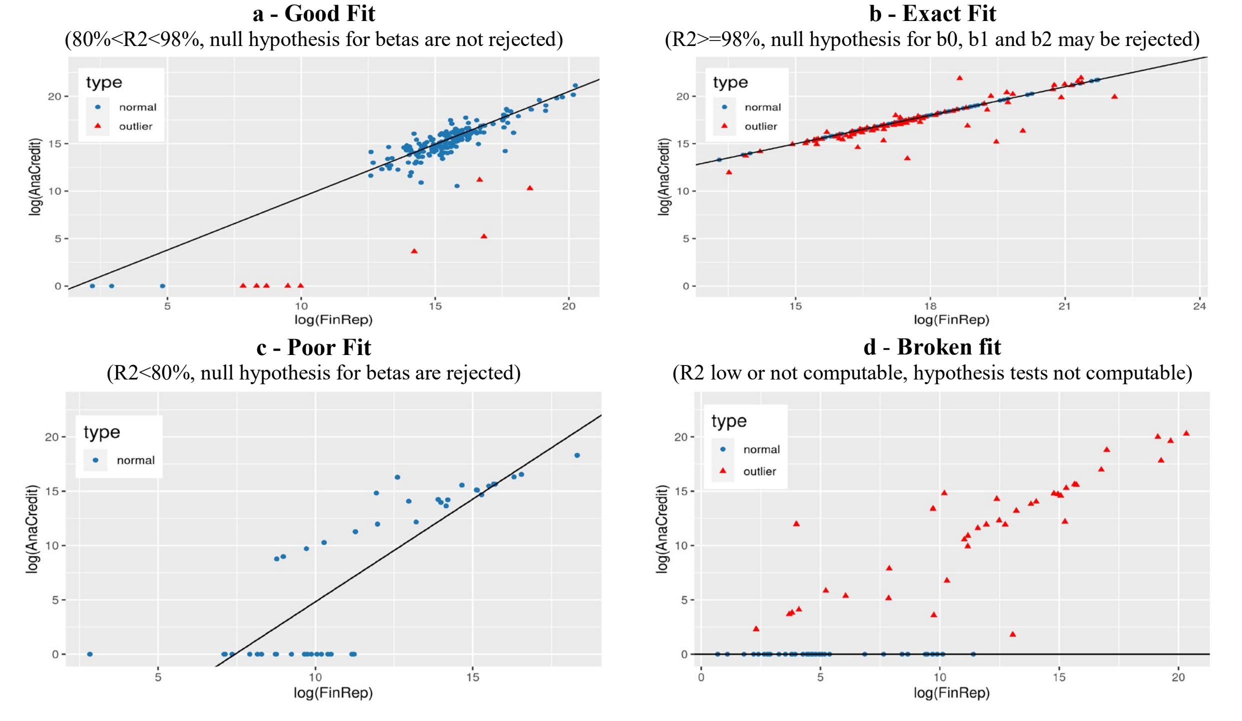
[13] When more data is available, as future work we will be able to model also the seasonality of the series.

In particular, we consider the SMDM estimation proposed by Koller e Stahel (2011) that shows both high asymptotic efficiency and high breakdown point (BP; see Hampel *et al*. 1985). Furthermore, as derived in Appendix B we expect that the coefficients β_0 and β_1 should be equal, respectively, to zero and one, while the β_2 term should be less or equal to zero. Using the robust covariance matrix it is possible to test the null hypothesis that these conditions are met ($\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 \leq 0$). When these conditions are not met, we cannot consider the corresponding regression as reliable, and then we do not analyze the correspondent aggregates that are compared.

Four types of outcomes are obtained from our robust regressions (Figure 3):

- ‘Good fit’ (top left): the regression estimates are coherent with the prior knowledge on the betas and the R-squared is high;
- ‘Exact fit’ (top right): the SMDM algorithm tends to classify as ‘outlier’ those observations close to the regression line;
- ‘Poor fit’ (bottom left): the robust regression is affected by a high number of leverage points (BP is almost at 50%);
- ‘Broken fit’: there are more outliers than good observations (BP greater than 50%).

Figure 3: Types of robust regression fit



a - Good Fit
(80%<R2<98%, null hypothesis for betas are not rejected)

b - Exact Fit
(R2>=98%, null hypothesis for b0, b1 and b2 may be rejected)

c - Poor Fit
(R2<80%, null hypothesis for betas are rejected)

d - Broken fit
(R2 low or not computable, hypothesis tests not computable)

In our analysis, we face mainly the problem known as ‘exact fit’ (see, for example, Maronna *et al*., 2006). Because of the high correlation between the AnaCredit aggregates and the corresponding BSI (and FinRep) series, our regressions often show a very high *R-squared* value (close to one), as the observed points are concentrated around a very small radius of the regression line. As a consequence, even in presence of a small distance between an observed point and the regression line, that observation is marked as an outlier although it does not show an anomalous behavior. This is due to the fact that the distance is greater than the one recognized as 'normal' by the model in such situation. A second consequence has to do with the impossibility to perform reliable tests of hypotheses on the regression coefficients, as the robust estimates of their standard errors are close to zero, leading to the wrong rejection of the theoretically expected relationship (i.e. the model under the null hypothesis) when instead it should be regarded as valid.

To cope with the cases of exact fit, together with the cases of 'poor fit' and 'broken fit' when outliers are clustered (see Figures 4 and 5), we investigate three possible solutions:

(1) resorting to the 'Bonferroni correction' for the *chi-squared* test of residuals (see Cerioli and Farcomeni, 2011);
(2) adding Gaussian noise to the dependent variable (a procedure known as *jittering*);
(3) de-noising through the removal of observations (a procedure known as *thinning*; see Cerioli and Perrotta, 2013).

Approach (1) relies on the assumption that the squared regression residuals follow approximately a *chi-squared* distribution with one degree of freedom. As discussed in Cerioli and Farcomeni (2011), in order to control for the probability of making one or more false rejections, a simple but effective method is to adopt the 'Bonferroni correction' for the confidence level over the sample size α/n. The observation is labelled as outlier if the statistics $T_{ijt} = \hat{e}_{ijt}^2 > \chi^2_{(1-\alpha/n,1)}$.

Strategy (2) consists in adding a Gaussian random noise ε, $\varepsilon \sim N(0, \sigma_\varepsilon)$, to the dependent variable and, then, perform a robust regression on the new transformed variable. For the calibration of σ_ε, first we run a robust regression without adding noise and compute the Mean Absolute Deviation (MAD) of residuals, then we multiply it by a constant factor k, considering a floor positive value δ: $\sigma_\varepsilon = \min\left(\delta, k * MAD\left(\hat{e}_{ijt}\right)\right)$. The empirical values of k and δ depend on the nature of the datasets under inspection: in this application, according to the abovementioned literature, we use $\delta \geq 0.1$ and $k \, \epsilon \, [1.48;\ 5]$.

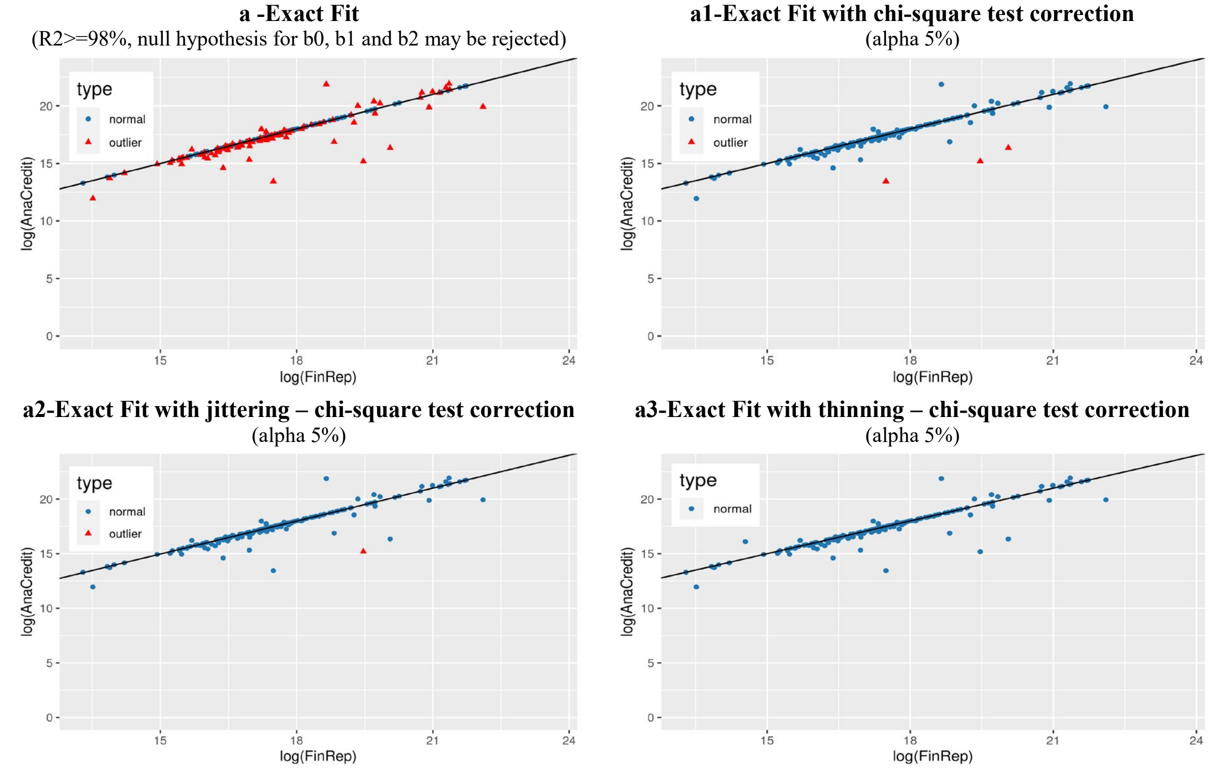
Finally, the procedure (3) consists in removing observations in order to down-weight the influence of high-density regions. To this end, it is necessary to define and evaluate a retain probability for each observation $\{y_i, x_i\}$, i.e. the logarithm of A_{ijt} and F_{ijt}. Following Cerioli and Perrotta (2013), we consider the retain probability $p(y_i, x_i) \propto 1 - \lambda_d(y_i, x_i)$, so that points will be deleted mainly in high-density regions; for the estimation of $\lambda_d(.)$ we use the same isotropic Gaussian kernel[14] introduced by the authors.

To detect outliers from procedure (2) and (3) we apply the *chi-square* test as defined in (1).

The 'exact fit' problem is well addressed by all the three solutions, as shown in Figure 4, where panel *a* illustrates the standard robust regression and panels *a1* to *a3* indicate the corresponding, robust regression when applying the three abovementioned strategies, respectively.

[14] For its implementation, see function *density.ppp* of R package 'spatstat'.

Figure 4: Solving 'exact fit' issue



a -Exact Fit
(R2>=98%, null hypothesis for b0, b1 and b2 may be rejected)

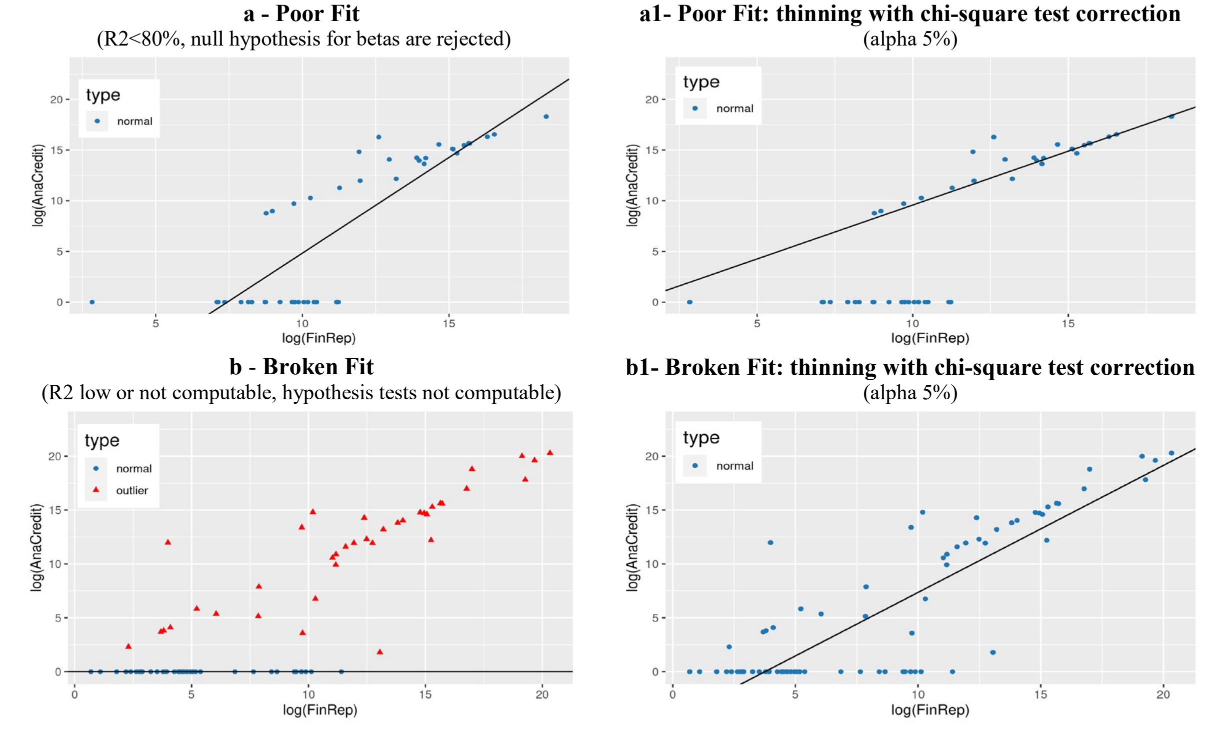
a1-Exact Fit with chi-square test correction
(alpha 5%)

a2-Exact Fit with jittering – chi-square test correction
(alpha 5%)

a3-Exact Fit with thinning – chi-square test correction
(alpha 5%)

The thinning procedure also allows to handle 'broken' and 'poor' fits (Figure 5): when extreme observations are clustered, the expected regression lines are correctly estimated both for 'broken fits' and for 'poor fits', allowing to correctly evaluate the hypothesis of presence of outliers.

Figure 5: Solving 'poor fit' and 'broken fit' cases



a - Poor Fit
(R2<80%, null hypothesis for betas are rejected)

a1- Poor Fit: thinning with chi-square test correction
(alpha 5%)

b - Broken Fit
(R2 low or not computable, hypothesis tests not computable)

b1- Broken Fit: thinning with chi-square test correction
(alpha 5%)

Since the series analyzed are different in terms of sample sizes, variance of residuals and level of contamination, and since each of the three methodologies has its own pros and cons, in order to take most out of them, we apply them according to the following hierarchical order:

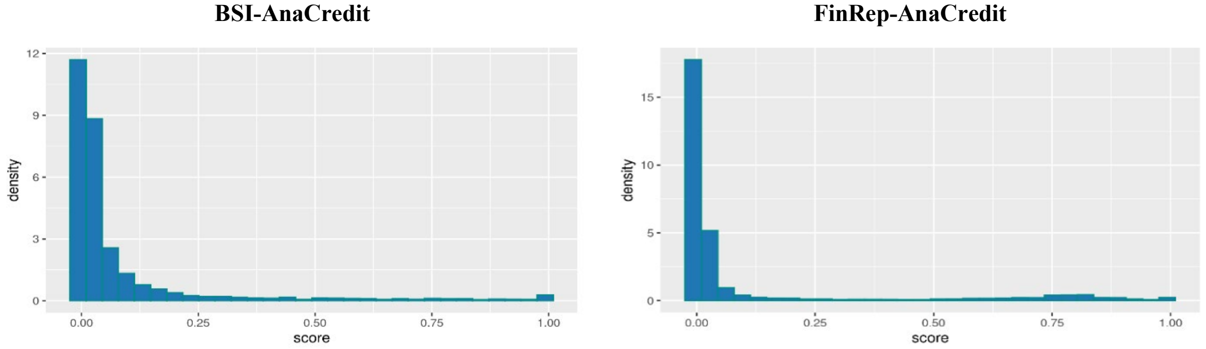
i) we run the jittering procedure, check that the null hypothesis for $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ is not rejected and the R-squared is above 80%. If such conditions are met, we classify the observation as an outlier or not on the basis of the chi-square test on residuals; if not, we move to (ii);

ii) we run the standard robust regression, check that the null hypothesis for $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ is not rejected and the R-squared is above 80%. If such conditions are met, we classify the observation as an outlier or not on the basis of the chi-square test on residuals; if not, we move to (iii);

iii) we run the thinning procedure, check that null hypothesis for $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ is not rejected and the R-squared is above 80%. If such conditions are met, then we classify the observation as outlier or not on the basis of the chi-square test on residuals; if not, we do not evaluate the observation.

Finally, in order to put the results of the robust regression in stacking with those of other models, we introduce a measure of anomaly for each data point inspected (score). Such scores are obtained by applying a min-max scaler to the absolute value of residuals of each ‘acceptable’ regression (i.e. that satisfies the conditions above). In Figure 6 the final scores of all the series for the BSI-AnaCredit and FinRep-AnaCredit comparison are reported.

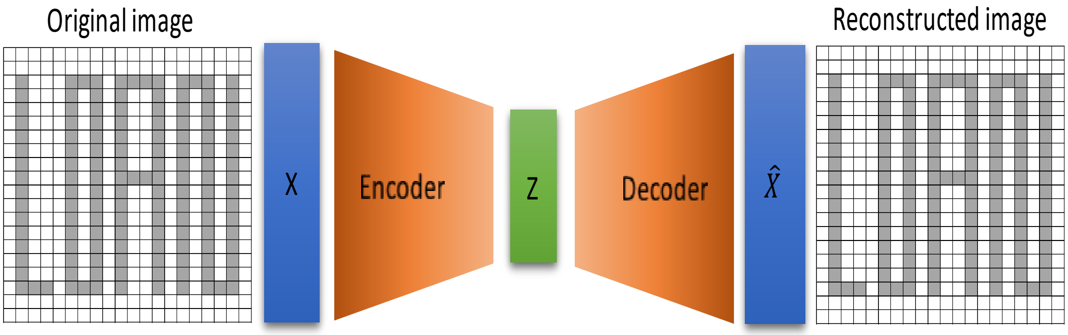
Figure 6: Scores distributions



3.2 Autoencoders

An autoencoder (AE) is a special type of multi-layer neural network performing hierarchical and nonlinear dimensionality reduction of data. The goal of an autoencoder is to replicate a given input as an output. Therefore, the output is the input itself, reconstructed. Typically, the model architecture is layered and symmetric, with the same number of nodes in the output and in the input layer, while nodes in middle layers are fewer in number. Therefore, the only way to reconstruct the input is by learning weights so that the intermediate outputs of the nodes in the middle layers consist in reduced representations. Figure 7 illustrates a fully connected autoencoder architecture.

Figure 7: Autoencoder architecture



Autoencoders are unsupervised models, as they do not need labels since the target variable is the input itself. Note that the outputs of the bottleneck layer represent the reduced representation, also known as ‘compressed representation’.

Because of its reduced representation of data, autoencoders represent a useful approach to detect outliers (Russo *et al*. 2019). The basic idea is that, in such dimensionality reduction, it is much harder to reproduce outliers than inliers (normal points), so the error of outliers’ reconstruction will be larger and, therefore, better identifiable.

Formally, autoencoders attempt to reconstruct an input image $x \in R^{k\times h\times w}$ through a bottleneck, effectively projecting the input (image) into a lower-dimensional space, called ‘latent space’. The projection (dimensionality reduction) occurs through an encoder function $E: R^{k\times h\times w} \rightarrow R^{d}$ and the reconstruction through an inverse decoder function D: $R^{d} \rightarrow R^{k\times h\times w}$, where *d* denotes the dimensionality of the latent space and *k*, *h* and *w* denote, respectively, the number of channels (equal to 3 in the case of Red, Green and Blue - RGB – images, to 1 for grayscale images), the height and the width of the input image. Choosing $d \ll k \times h \times w$ prevents the architecture from simply copying its input and forces the encoder to extract meaningful features from the input patches that facilitate accurate reconstruction by the decoder. The overall process can be summarized as

$$\hat{x} = D\big(E(x)\big) = D(z), \tag{2}$$

where *z* is the latent vector and $\hat{x}$ the reconstruction of the input *x*. In our project, two models parameterize the functions E and D: the convolutional autoencoder (AE-CNN) and the dense autoencoder (AE-DNN). In the AE-CNN, ‘strided’ convolutions are used to downsample the input feature maps in the encoder and to upsample them in the decoder, while in the AE-DNN, dense layers are used for the same tasks.

We propose to measure the reconstruction accuracy with the Structural Similarity Index Metric (SSIM), as in Wang *et al*. (2004). This measure is designed to capture perceptual similarity; it captures inter-dependencies between local pixel regions that are disregarded by the current state-of-the-art unsupervised defect segmentation methods based on autoencoders with per-pixel losses. The measure is not very sensitive to edge alignment and attaches importance to salient differences between input and its re-construction. The SSIM works in the spatial domain and, given two image patches $x = \{x_i | i = 1, \dots, P\}$ and $y = \{y_i | i = 1, \dots, P\}$, it is defined as

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{3}$$

where μ, σ are, respectively, the sample mean and sample standard deviation, σ_{xy} is the sample covariance between x and y and (c_1, c_2) are two positive stabilizing constants. The resultant SSIM index is a real value that could be normalized within the range $[0, 1]$, where 1 indicates perfect structural similarity that can be achieved only in case of two identical sets of data, while 0 denotes the absence of any degree of structural similarity. Following this approach, a loss function is derived as a mean of structural dissimilarity indexes (DSSIM), i.e. as the mean of the complement to 1 of the SSIM[15] over all images

$$\mathcal{L}^{SSIM}(I) = \frac{1}{N}\sum_{i \epsilon I}\big(1 - SSIM(i, \hat{\imath})\big), \tag{4}$$

where I is the set of all images. This loss function is used for AE-CCN models, while in the case of AE-DNNs the mean square error (MSE) is used.

The key idea of our model is to train an autoencoder on BSI (and FinRep) data in order to learn their 'normal' structure and to use it to identify abnormal structures (i.e. anomalous data) in AnaCredit. This method provides an *overall assessment* of all data reported by each bank for a given reference date. The identification of the reporting components that are anomalous occurs based on a score function, as described below.

The input data for the two neural networks are, respectively, the complete report of all series of a bank at a given reference date for the AE-DNN and their transformation in image for the AE-CNN. To create the image we use the collected scaled data to derive the auto cross-product: the result of this operation is a matrix whose elements are in the [0, 1] range. This matrix can be regarded as an image in grey scale where each pixel is originated from each value of the matrix; each value of the matrix gives the grey level to color the pixel. The entries of the matrix have also a statistical meaning: each element gives the contribution of the interaction of the single component of the data collected with respect to the others. The final selected architecture is different for AE-DNN and AE-CNN. In the case of AE-DNN, the encoder and decoder networks consist of 2 fully-connected layers with respectively 150 and 28 hidden units, with LeakyReLU as activation function. The first layer also contains a dropout (Srivastava *et al.*, 2014) of 5%. The bottleneck layer is set as one fully connected layer with 20 hidden units, resulting in a 20-dimensional latent space.

In the case of AE-CNN, the encoder and decoder networks are comprised of convolutional layers with batch normalization and a max-pooling window of 2x2 after each convolution. The sigmoid as activation function is

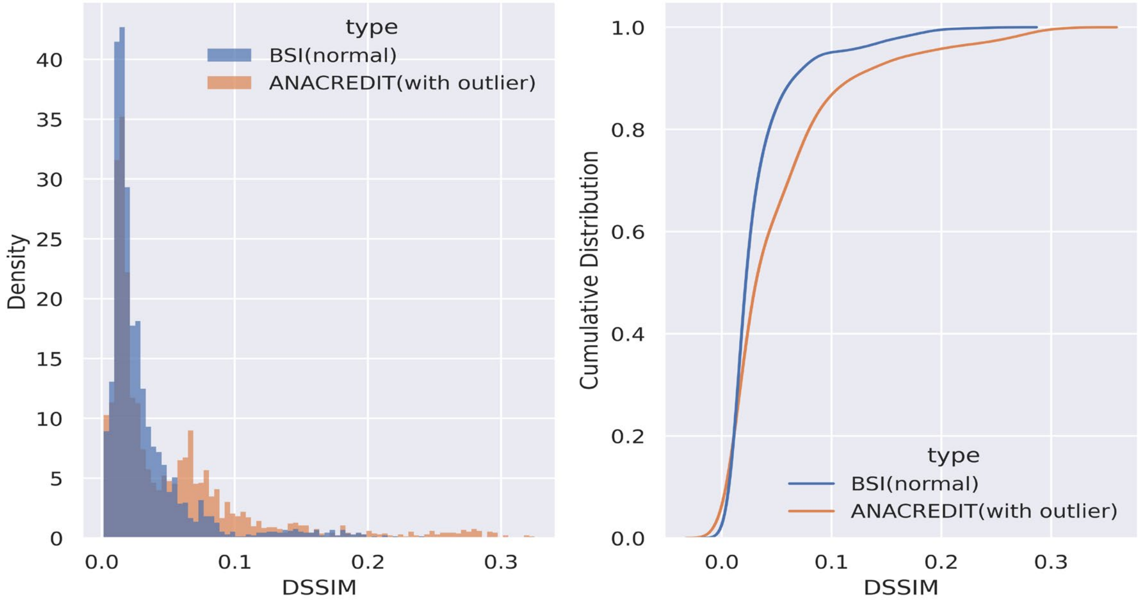
[15] For more details, see Brunet *et al.* (2012).

used and the padding to reproduce the same dimensions. The encoder network is structured in a stack of three hidden layers with convolutional filters of respectively 16, 32, 16 units and kernel sizes of 3x3. The bottleneck consists of a convolutional layer with 4 convolutional filters of 23x23 size. Regarding the decoder network, its structure mirrors the encoding part having an up-sampling step in substitution of the max-pooling.

Training these two networks on BSI (and FinRep) dataset, which we assume are of a better quality, generates a distribution of the losses (MSE or DSSIM) associated with the entire dataset reported by each bank. This distribution is compared with the one derived from the application of the model to the AnaCredit dataset.

Figure 8 illustrates, as an example, the loss distributions (and relative cumulative distribution functions) obtained by training the AE-CNN model on BSI and then applying it to AnaCredit. Similar distributions occur for the comparison between AnaCredit and FinRep and in the cases of AE-DNN networks. For low values of the dissimilarity index the two distributions overlap, as expected, while for high values we observe the difference that need to be investigated.

Figure 8: Structural dissimilarity index



The distribution of DSSIM or normalized MSE[16] helps to label as anomalous or not anomalous the whole set of AnaCredit data reported by each bank at a reference date, but it does not provide information on what components (i.e. which of the BSI and the FinRep series) have contributed the most to such result. To this end, we consider a score function mixing the loss function (DSSIM or normalized RMSE - *l*) value with the absolute relative difference (f) between pairs of BSI (and FinRep) and equivalent AnaCredit series:

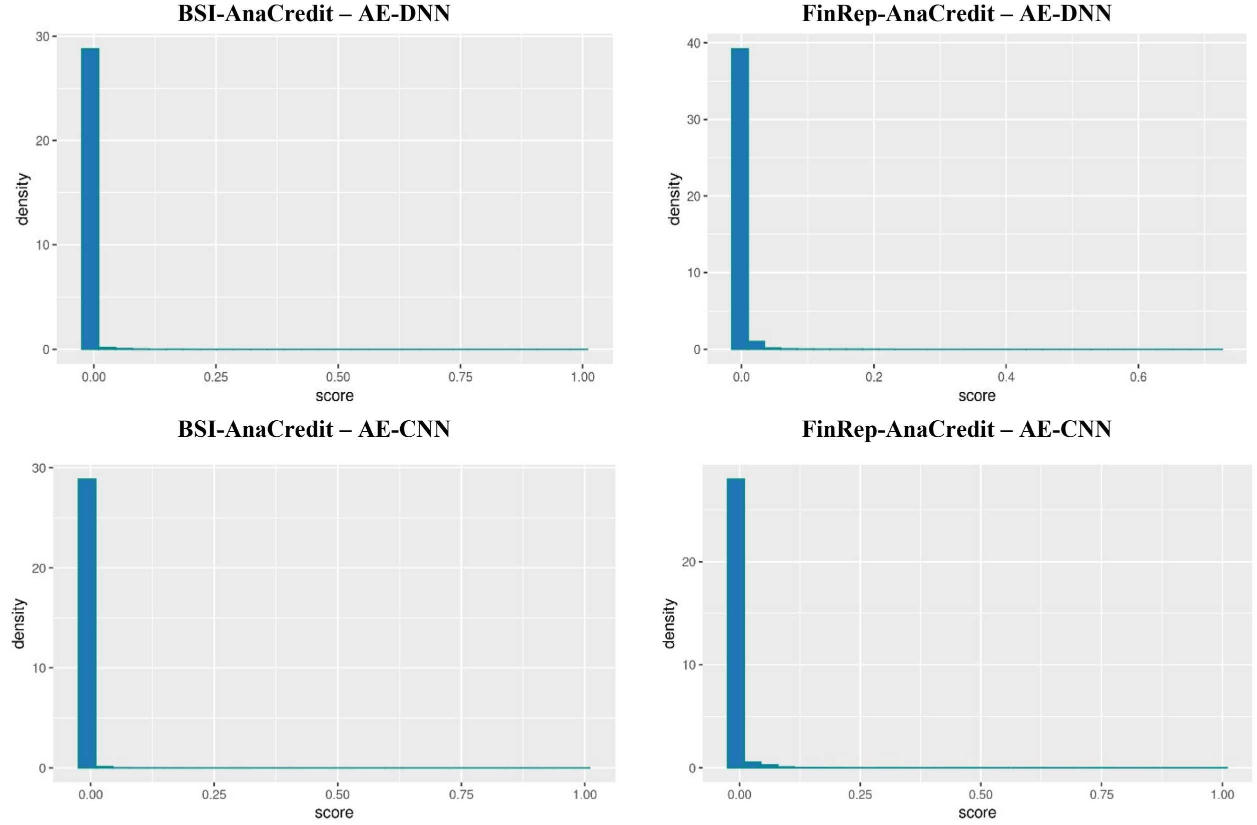
$$s(l,f) = l \cdot f^2 \tag{5}$$

[16] We normalize the RMSE with the mean to obtain the coefficient of variation.

where l is the generic loss function for the model considered and f is the absolute relative difference between a generic BSI (and FinRep) aggregate and the equivalent AnaCredit, considered squared to emphasize big differences. Such a defined score function allows weighing the global evaluation (loss function) over all the pairwise comparisons. Besides, we rescale the score values in the interval [0, 1] to have a normalized score, which can be compared to those produced by the other models developed.

Such normalized scores (see Figure 9) allow discriminating between good (low score values) and anomalous data (high score values).

Figure 9: Scores'distributions



BSI-AnaCredit – AE-DNN

FinRep-AnaCredit – AE-DNN

BSI-AnaCredit – AE-CNN

FinRep-AnaCredit – AE-CNN

In the following steps, the scores obtained above are used as input to the stacking models.

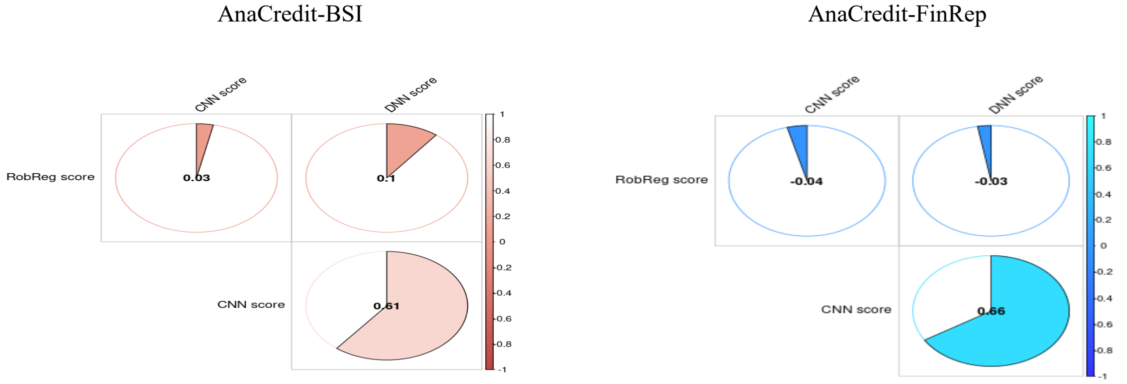
3.3 Stacking predictions in a semi-supervised learning setting

After training the basic models, for each observation (combination of reference date, reporting entity and compared aggregate) we have three anomaly scores, two produced by the autoencoders and one by the robust regression: these are the final predictions that are input to the meta-learners.

The robust regression scores exhibit low correlation with the autoencoders' scores, while the two autoencoders' scores do not have remarkable correlation among them. This is a common result in the two joint datasets (Figure 10). As the scores within the two comparisons map different information, they are combined together to produce a final forecast.

The combination takes place with the stacking technique (Wolpert, 1992; Dzeroski and Zenko, 2004), i.e. through a meta-learner using the abovementioned scores (predictions) as input for its forecast (Figure 1).

Figure 10: Correlation between learners' scores


AnaCredit-BSI
AnaCredit-FinRep

Stacking models typically yields a better performance than each of the input learners. Stacking models require the existence of labels, i.e. a response variable that is attempted to replicate. To get a response variable, some cases have been sampled and verified with banks, while others have been pre-labelled on the basis of the domain knowledge of the analysts. Such information is bound to learner scores, obtaining a final partially labelled dataset on which we can train meta-learners in a semi-supervised learning paradigm (Chapelle *et al*., 2006; González *et al*., 2019).

With regard to the pre-labeling, based on domain knowledge, we mark as 'correct' (label 0) observations for which the difference between the amounts in BSI (FinRep) and the corresponding amounts in AnaCredit is deemed very small. In addition, we have marked as 'anomalous' (label 1) those for which the same distance is far negative, i.e. less than a chosen percentile of the empirical distribution.

The sampled cases are selected following a stratified sampling approach using the Neyman's optimal criterion (Neyman, 1934). The stratification is obtained by considering the joint distribution of the scores appropriately discretized in binary values[17], so that each observation is classified to a specific stratum. Neyman's optimal criterion is used with reference to the different variability of the average scores and to the sampling cost of each stratum, the latter being inversely proportional to the median distance of the difference between the amount of BSI (and FinRep) and the corresponding amount of AnaCredit. At this step, the sample size remains determined[18] and the units are selected within each stratum with simple random sampling without replacement. Each sampled unit is analyzed with the bank that has reported it and this leads to the attribution of a response (R) that confirms or not the correctness of the AnaCredit data.

[17] The scores are converted into binary coding, choosing the value of 0.2 as threshold.
[18] For more details, see Table A1 in Appendix A.

Moving to the meta-learners, two different semi-supervised approaches are considered. Following the first semi-supervised approach, we develop and compare models in a context where the sample distribution of anomalous and not-anomalous cases within strata based on responses received have been previously reported to the universe of observations. In the second semi-supervised approach, the reporting of the sample responses to the universe and the estimation of the parameters of the models occurs in the training phase.

In the first setting, the completion of the labeling is performed following a Monte Carlo simulation approach. Pseudo labels (i.e. simulated responses) to the not-sampled and not-pre-labelled observations are assigned randomly by replicating the sample distribution of the responses received by banks within each stratum. By repeating this pseudo labeling over a sufficiently high number of times (we choose N = 1000), we obtain a dataset with this set of pseudo response variables, where each is composed by sampled/pre-evaluated labels (R) and simulated labels (Z).

On this enriched dataset, we train four different meta-learners: a logistic regression (*LOGIT*), a *k*-nearest neighbor (*KNN*), a random forest (*RF*) and a support vector machine (*SVM*). Each of these models has been optimized by cross-validation and the training is performed over the *N* simulated response variables. The final prediction is obtained as the central value of the *N* simulated predictions.

In this way, for a given reference date *t* and with respect to each bank *i* and sub-portfolio of loans *j*, we train four different models on the full vector of responses (*R* and *Z*) able to combine the predictions of the underlying scores exploiting different combination paths. Afterwards, we compare the model predictions and choose the best one with regard to some appropriate reference measures[19]. F1-score is our preferred reference measure to address the issue of the unbalanced classes (of 0 and 1) in the variables *R* and *Y*, but we provide the results also obtained for other measures, e.g. precision rate, recall, etc.

As regards the LOGIT, denoting by $p_{i,j,t} = \Pr(\hat{Y}_{i,j,t} = 1)$ the probability of an observation to be an outlier, where i, j, t represent, respectively, a bank, a sub-portfolio of loans and a given reference date, the model is described by the following equation:

$$l_{i,j,t} = \frac{p_{i,j,t}}{1-p_{i,j,t}} = \beta_0 + \beta_1 ScoreRobReg_{i,j,t} + \beta_2 ScoreCNN_{i,j,t} + \beta_3 ScoreDNN_{i,j,t} + \Theta\gamma + \varepsilon_{i,j,t}, \quad (6)$$

where Θ is a vector of control variables, i.e. dummy variables for the identification of re-aggregations of the analyzed aggregated series, and $\varepsilon_{i,j,t}$ is the noise term. In order to train the meta-learner, we use the weighted accuracy - with optimal weights[20] chosen for non-anomalous and anomalous data respectively - to cope with the imbalance in the responses. The optimal weights are obtained at the maximum value of the average F1-score calculated on the training set by employing the five-fold cross validation.

As regards the other three models, the generic function describing each of them is:

[19] See Figure A5 in Appendix A.

[20] See more details in Figure A6 in Appendix A.

$$\hat{Y}_{i,j,t} = f\left(ScoreRobReg_{i,j,t}, ScoreCNN_{i,j,t}, ScoreDNN_{i,j,t}, \Theta\right) + \varepsilon_{i,j,t}, \tag{7}$$

where $f(\cdot)$ is a different function based on the model type and $\varepsilon_{i,j,t}$ is the noise term. Also for these models, the meta-parameters are calibrated on the training set by employing the five-fold cross validation, maximizing the average F1-score.

In the second setting, semi-supervised models for the equation 7 are trained by using only the sampled and pre-labelled variable (R) instead of $\hat{Y}$, as the labeling is carried out within the model estimation process. In this class of models, the following algorithms have been used: Self-training, SETRED, Tri-training, Co-Bagging and Democratic-Co.

The Self-training model (Yarowsky, 1995) is probably the earliest idea about using unlabelled data in classification. This wrapper-algorithm, starting from only the labelled data, iteratively uses a supervised learning method trained only on the part of the dataset labelled until the current iteration. At each step, it labels a part of the unlabelled points according to the current decision function until the whole dataset is labelled.

Similarly, the SETRED algorithm (Li and Zhou, 2005) first learns from labelled examples, and then iteratively chooses to label a few unlabelled cases on which the learner is most confident in prediction and adds them to its labelled set for further training at next step. However, at each iteration SETRED does not completely accept all the pre self-labelled examples, but it actively identifies the possibly mislabelled examples by testing a predefined null hypothesis with the local cut edge weight statistic associated with each self-labelled example. If the result of the test falls in a left rejection, the example is regarded as a good one; otherwise, it is a possible mislabelled example and it should not be included to the learner's training set.

The Tri-training algorithm (Zhou and Li, 2005) generates three classifiers from the original labelled set and labels an unlabelled observation if the other two classifiers agree on the labeling, under certain conditions. This procedure is repeated until convergence (generally with the complete labeling of the dataset).

The Co-Bagging method (Blum and Mitchell, 1998) assumes that the feature space can be split into two different conditionally independent views and that each view is able to predict the classes on its own. It trains one classifier in each specific view, and then the classifiers learn from each other the most confidently predicted examples from the unlabelled pool. The process continues until a predefined number of iterations is reached.

The Democratic-Co algorithm (Zhou and Goldman, 2004) uses multiple algorithms, instead of multiple views, to enable learners to label data from each other. This technique leverages off the fact that different learning algorithms have different inductive biases and that better predictions can be made by the majority vote.

In the next Section, we present the empirical results of the abovementioned models.

4 Results and discussion

The empirical results of the models introduced in the previous section are presented in the following two sub-sections, the first devoted to the comparison between AnaCredit and BSI, the second to that between AnaCredit and FinRep.

4.1 BSI vs. AnaCredit: empirical evaluation

The data dimension used for training and testing the models are reported in Table 1.

Table 1: Observations in training and test set

	1° semi-supervised setting	2° semi-supervised setting
Training	5440	6199
Test	2331	1572
Total	7771	7771

The different size of the training and the test sets in the two approaches derives from the fact that in the first setting labels are available for all observations, being simulated via Monte Carlo. In the second setting, we need to consider only labelled observations in the test set; therefore, we assign all unlabelled observations to the training set and then split the pre-labelled observation between training and test set according to a 50% share.

Before evaluating the meta-learners, we measure the performance of the three learners (basic models) with standard performance metrics together with the balanced accuracy (in order to take into account the imbalance of the target variable *Y* in our dataset). Table 2 presents the performance metric results that can also be considered as benchmarks for the meta-learners results, presented in Section 3.

Table 2: Performance of the base models (learners)

	Sample cases		
	RobReg	DNN	CNN
Precision	0.09507	0.82353	1.00000
Recall	0.04500	0.02333	0.01333
Specificity	0.79965	0.81260	0.81122
Accuracy	0.73601	0.81266	0.81170
F1 score	0.06109	0.04538	0.02632
Balanced accuracy	0.42233	0.41797	0.41228

All the metrics considered in Table 2 are evaluated on a test set representing 30% of the available data. The Table clearly shows that the three basic models RobReg, DNN and CNN perform quite well in terms of accuracy, with values equal to 0.736, 0.813 and 0.811, respectively.. Unfortunately, since our input dataset is biased towards non-anomalous cases (0.79 and 0.21) in *Y* variable, such models perform poorly when metrics taking into account the imbalance are considered: for instance, the balanced accuracy drops to 0.422, 0.418, and 0.412, respectively.

To strengthen the power of prediction of our models, a stacking step is further considered. In the first place, a baseline logistic model combining the predictions of the three basic models only on the sampled and pre-labelled data is developed. In particular, this baseline model is trained with a weighted accuracy in order to take into account the imbalance between non-anomalous and anomalous cases. Such model presents the following main results (Table 3): the F1 score is 0.997, the balanced accuracy 0.998, the precision 0.981 and the recall 0.988.

The performance over all the dataset is presented in Table 3 that shows the central values of the different meta-learners trained on the 1000 simulated pseudo labels.

Table 3: Performance of the models within the first semi-supervised setting

	Baseline*	LOGIT1	KNN	RF	SVM
Precision	0.981±0.019	0.901±0.016	0.848±0.014	0.864±0.018	0.830±0.021
Recall	0.988±0.011	0.433±0.161	0.459±0.016	0.463±0.014	0.514±0.017
Specificity	0.994±0.006	0.867±0.003	0.871±0.003	0.872±0.003	0.882±0.004
Accuracy	0.997±0.003	0.870±0.004	0.869±0.004	0.872±0.004	0.876±0.005
F1 score	0.997±0.002	0.585±0.017	0.596±0.015	0.603±0.015	0.635±0.017
Balanced accuracy	0.998±0.001	0.650±0.009	0.665±0.009	0.668±0.009	0.698±0.010

* Only on sampled cases.

All the metrics are evaluated on five test sets, each representing 30% of the data and obtained setting a different random seed: the central value of the different metrics is reported together with the deviation of this value from the minimum and maximum, in order to assess metrics' variability. The linear SVM yields somewhat better results in terms of F1 score and balanced accuracy when trying to replicate the baseline.

As regards the second setting of semi-supervised algorithms, the results highlighting their performances are shown in the following table:

Table 4: Performance of the models within the second semi-supervised setting

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Precision	0.995±0.005	0.993±0.007	0.993±0.007	0.989±0.004	0.993±0.007
Recall	0.995±0.005	0.988±0.012	0.987±0.011	0.995±0.005	0.989±0.008
Specificity	0.999±0.001	0.997±0.003	0.997±0.002	0.999±0.001	0.997±0.002
Accuracy	0.998±0.001	0.997±0.000	0.997±0.000	0.997±0.000	0.997±0.000
F1 score	0.995±0.003	0.991±0.003	0.991±0.002	0.992±0.002	0.992±0.002
Balanced accuracy	0.997±0.003	0.993±0.007	0.992±0.007	0.997±0.003	0.993±0.005

All the models use the support vector machine as learner (Democratic-Co use also a KNN and a C5.0) and all the metrics are evaluated on the test presented in Table 1. Additional four runs using different random seeds are used to generate new training and test sets, in order to assess metrics' variability. Based on the F1 score, the best performer is the self-training algorithm, showing a central value of 99.5%. If we compare it to the baseline model, we find a good replication in the F1-score and better results in terms of precision and recall.

Pairwise comparison of the various models is carried out by using the McNemar's Test for binary classification. Table 5 contains the p-values for the null hypothesis that two models do not have significant differences in their label predictions. A small p-value denotes that there is a statistical significant difference in the power of prediction between the two models.

Table 5: Predictions comparison (*)

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Self-training		0.009	0.138	0.003	0.013
SETRED			0.269	0.000	0.251
Tri-training				0.002	0.025
Co-Bagging					0.000
Democratic-Co					

(*) P-value mean over the five test set. A p-value lower than 0.05 indicates a significant disagreement between the model predictions.

Table 5 shows that the self-training predictions are statistically equivalent to the tri-training ones. Since the self-training model gets the higher F1-score, and all the differences with SETRED, Co-Bagging and Democratic-Co are statistically different from zero, we can conclude that the self-training model is the best performer.

4.2 FinRep vs. AnaCredit: empirical evaluation

The data dimension used for the training and test set are reported in Table 6.

Table 6: Observations in training and test set

	1° semi-supervised setting	2° semi-supervised setting
Training	26035	29680
Test	11201	7656
Total	37336	37336

As in the BSI comparison, the different size of the training and the test sets in the two approaches derive from the constraint of assigning, in the second semi-supervised setting, a share of 50% of pre-labelled observations to the test set and the rest of observations to the training set, while in the first setting it is not present.

In this case, we have a greater number compared to the BSI-AnaCredit case due to more disaggregated series considered in FinRep. As for BSI-AnaCredit, we first measure the performance of the three underlying models with standard performance metrics together with the balanced accuracy (imbalance of target variable Y). The results, reported in Table 7, are useful in terms of benchmark to the meta-learners presented in Section 3.

Table 7: Performance of the base models (learners)

	RobReg	DNN	CNN
Precision	0.05330	0.83333	0.75000
Recall	0.05263	0.00283	0.00170
Specificity	0.82913	0.84726	0.84711
Accuracy	0.71184	0.84725	0.84708
F1 score	0.05296	0.00564	0.00339
Balanced accuracy	0.44088	0.42505	0.42441

All the metrics are evaluated on the test set representing 30% of the available data. The Table clearly shows that the three models RobReg, DNN and CNN perform quite well in terms of accuracy, with values equal to 0.712, 0.847 and 0.847, respectively. Therefore, their predictive capacity, considering the two classes of anomalous and non-anomalous data at the same time, is quite high. However, the dataset we are considering is unbalanced towards anomalous cases (only 16%). When we move to measures that take into account such imbalance, their performance decreases; for instance, the balanced accuracy falls to 0.441, 0.425, and 0.424, respectively.

Moving to the stacking step, the results of the performance measures for a weighted logistic baseline (only on sampled and pre-labelled data) and for LOGIT, RF, KNN and linear SVM models (over all the dataset) are reported in Table 8.

Table 8: Performance of the models within the first semi-supervised setting

	Baseline*	LOGIT	KNN	RF	SVM
Precision	0.998±0.002	0.833±0.008	0.501±0.005	0.747±0.010	0.503±0.006
Recall	0.999±0.001	0.367±0.009	0.204±0.004	0.434±0.008	0.380±0.013
Specificity	0.996±0.004	0.890±0.002	0.860±0.002	0.899±0.001	0.884±0.002
Accuracy	0.998±0.003	0.886±0.002	0.836±0.002	0.885±0.002	0.837±0.002
F1 score	0.997±0.003	0.508±0.009	0.290±0.005	0.548±0.009	0.432±0.010
Balanced accuracy	0.997±0.003	0.628±0.005	0.532±0.002	0.667±0.005	0.631±0.006

* Only on sampled cases.

All the metrics are evaluated on five test sets, each representing 30% of the data, obtained by setting a different random seed; the central value over the five test sets is reported for the different metrics and the deviation of this value from the minimum and maximum value so as to assess metrics' variability. The various models attempt to replicate the baseline; although far from it, the model with the best results is the random forest, with an F1-score equal to 0.548 and a high precision of 0.747.

As regards the second approach of semi-supervised algorithms, the results highlighting their performances are shown in Table 9.

Table 9: Performance of the models within the second semi-supervised setting

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Precision	0.723±0.167	0.929±0.072	0.825±0.075	0.825±0.075	0.833±0.001
Recall	0.513±0.058	0.487±0.058	0.504±0.042	0.523±0.023	0.179±0.179
Specificity	0.605±0.079	0.656±0.106	0.653±0.097	0.653±0.097	0.500±0.107
Accuracy	0.697±0.054	0.733±0.054	0.715±0.036	0.715±0.036	0.518±0.125
F1 score	0.598±0.098	0.634±0.034	0.634±0.034	0.650±0.018	0.500±0.001
Balanced accuracy	0.563±0.065	0.582±0.072	0.595±0.053	0.595±0.053	0.335±0.139

All the models use C5.0 as learner (Democratic-Co use also a KNN and a SVM) and all the metrics are evaluated on the test presented in Table 6. Additional four runs using different random seeds are used to generate new training and test sets in order to assess metrics' variability. According to the F1-score, Co-Bagging is the best performers, followed by the SETRED and Tri-training algorithms.

Pairwise comparisons of various models is carried out by using the nonparametric McNemar's Test for binary classification. Table 10 contains the p-values for the null hypothesis that each pair of models does not show significant differences in their label predictions. A small p-value denotes a significant difference (improvement) in the prediction between the two models. Results reported in Table 10 show that the predictions of Co-Bagging is equivalent to the others. Therefore, there is not a clear winner from this competition and the Co-Bagging, SETRED and Tri-training algorithm seems to be equally efficient. Only the Democratic-co presents a rather poor performance.

Table 10: Predictions comparison (*)

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Self-training		0.473	0.787	0.833	0.080
SETRED			0.488	0.573	0.184
Tri-training				0.500	0.089
Co-Bagging					0.087
Democratic-Co					

(*) P-value mean over the five test set. A p-value lower than 0.05 indicates a significant disagreement between the model predictions.

5 Summary and conclusions

AnaCredit is a relatively recent ESCB dataset; it contains granular information (at contract and instrument level) on loans that banks grant to legal entities. Within the ESCB, two other 'historical', high quality datasets providing similar information on loans are the BSI and the FinRep, which are typically used in monetary policy and supervisory analyses respectively. Both of them can be exploited for a pairwise comparison with AnaCredit data to enhance the outlier detection process in the AnaCredit granular survey.

More specifically, we explore the use of machine learning techniques to carry out the above cross-checking with specific reference to loan portfolios, in a framework that takes the time dimension into account and is bank-specific. We resort to three models – a robust regression one and two autoencoder models – that grasp the existing relationships between each benchmark dataset – BSI and FinRep – and AnaCredit in order to identify potential outliers in the latter one.

Each model assigns an 'anomaly score' to each observation considered (uniquely identified by reporting date, entity, and loan aggregate). These anomaly scores are combined in order to yield a better forecast using a stacking approach under a semi-supervised learning context. Indeed, our approach lies in a semi-supervised environment having true anomalous or non-anomalous data labels only for a subset of the datasets. The true labels are based on both the domain knowledge of the analysts and the responses directly received from reporting entities on a number of observations, which are sampled by using a selection schema that is able to reproduce the distribution of scores assigned by the three models.

In this semi-supervised context, we consider two settings. In the first one, where the true labels have been reported to the universe of observations under a Monte Carlo simulation, we train logistic, random forest, KNN and SVM models and compare the results obtained from each of them. In the BSI-AnaCredit comparison, the SVM model has the highest F1-score, whereas in the FinRep-AnaCredit comparison the random forest is the better learner in terms of the same statistic. In the second setting, where the expansion of true labels takes place within the learning of the models themselves, we train Self-training, SETRED, Tri-training, Co-bagging and Democratic-co models and compare their results to the baseline. We find that for the BSI-AnaCredit comparison, the self-training gives the better F1-score, while for FinRep-AnaCredit comparison, the Co-Bagging model turns out to have the best performance. Considering all the models developed, we find that the algorithms of the second settings of semi-supervised models outperform those of the first settings.

Possible refinements of the paper, which are left to future developments, might consist in developing the current base learners: for the robust regressions we could move to a panel approach and for the autoencoders to the variational autoencoders. Further improvements are related to the optimization of the parameters underlying the second setting semi-supervised models and the use of other disaggregated BSI and FinRep series.

The framework developed in this paper is quite flexible and general and can be applied to carry out pairwise comparisons between datasets on similar phenomena but with different levels of granularity. As shown in our empirical exercise, this approach exhibits important advantages not only in terms of a more accurate detection of potential outliers in a highly granular database, but also from the point of view of reporting banks that will have to cross-check such anomalies and decide whether to confirm or revise the data.

References

Aggarwal C. (2017). “Outlier Analysis”, Springer.

Bishop, C.M. (2011). “Pattern Recognition and Machine Learning”, Springer.

Blum A. and Mitchell T. (1998). “Combining labeled and unlabeled data with co-training”, in Eleventh Annual Conference on Computational Learning Theory, COLT’ 98, pages 92–100, New York, NY, USA.

Brunet D. and Vrscay E. R. (2012). “On the Mathematical Properties of the Structural Similarity Index”, IEEE Transactions on Image Processing, vol. 21, no. 4.

Cagala, T. (2017). “Improving Data Quality and Closing Data Gaps with Machine Learning”, IFC Bulletin, 46.

Cerioli A. and Farcomeni A. (2011). “Error rates for multivariate outlier detection”, Computational Statistics and Data Analysis 55, pp. 544–553.

Cerioli A. and Perrotta D. (2013). “Robust clustering around regression lines with high-density regions”, Springer, Advances in Data Analysis and Classification volume 8, pp. 5–26.

Chakraborty C. and Joseph A. (2017). “Machine Learning at Central Banks”, Bank of England Staff Working Paper No. 674, https://doi.org/10.2139/ssrn.3031796

Chapelle O., Scholkopf B. and Zien A. (2006). “Semi-supervised learning”, MIT Press

Chandola V., Banerjee A. and Kumar V. (2009). “Anomaly detection: a survey”, ACM Computing Surveys, Vol. 41, No. 3, http://doi.acm.org/10.1145/1541880.1541882

Cœuré B. (2017). “Setting standards for granular data”, Opening remarks by Benoît Cœuré, Member of the Executive Board of the ECB, at the Third OFR-ECB-Bank of England workshop on “Setting Global Standards for Granular Data: Sharing the Challenge”, Frankfurt am Main, 28 March 2017, https://www.ecb.europa.eu/press/key/date/2017/html/sp170328.en.htm

Cusano F., Marinelli G. and Piermattei S. (2021*)*. “Learning from revisions: a tool for detecting potential errors in banks’ balance sheet statistical reporting”, Bank of Italy, Working Papers, No. 611.

Dzeroski, S. and Zenko, B. (2004). “Is combining classifiers with stacking better than selecting the best one?”, Machine Learning, 255–273.

Di Noia M. and Moretti D. (2020). “Le informazioni statistiche della Banca d’Italia sul rischio di credito e la nuova rilevazione AnaCredit”, Banca d’Italia, Occasional Papers, No. 544.

Farcomeni A. and Greco L. (2015). “Robust methods for data reduction”, CRC Press.

Farnè M. and Vouldis A.T. (2018). “A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks’ supervisory data”, ECB Working Paper N. 2171.

Goldstein M. and Uchida S. (2016). “A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data”, Computer Science, Medicine, PLoS ONE.

González M., Rosado O., Rodríguez J. D., Bergmeir C., Triguero I. and Benítez J. M. (2019). “ssc: An R Package for Semi-Supervised Classification”, R package version 2.1-0.

Granger C.W.J. (1981). “Some Properties of Time Series Data and Their Use in Econometric Model Specification”, Journal of Econometrics, 28, 121-130.

Gschwandtner M. and Filzmoser P. (2012). “Computing Robust Regression Estimators: Developments since Dutter 1977”, Austrian Journal of Statistics, Volume 41, Number 1, 45–58.

Hampel, F. R. (1985). “The Breakdown Point of the Mean Combined With Some Rejection rules”, Technometrics, 27, 95-107.

Hampel, F., Ronchetti E., Rousseeuw P. and Stahel W. (1986). “Robust Statistics: The Approach Based on Influence Functions”, N.Y.: Wiley

Hastie T., Tibshirani R. and Friedman J. (2001). “The Elements of Statistical Learning”, Springer.

Hastie T., James G., Tibshirani R. and Witten D. (2013). “An Introduction to Statistical Learning”, Springer.

Koller, M. and Stahel W. A. (2011). “Sharpening wald-type inference in robust regression for small samples”, Computational Statistics & Data Analysis 55(8), 2504–2515

Lessmann S., Baesens B., Seow H.V. and Thomas L.C. (2015). “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research”. *European Journal of Operational Research*, 247 (1), 124-136.

Li M. and Zhou Z. (2005). “Setred: Self-training with editing. In Advances in Knowledge Discovery and Data Mining”, volume 3518 of Lecture Notes in Computer Science, pages 611–621. Springer Berlin Heidelberg.

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T, Koller M., Conceicao E.L. and Anna di Palma M. (2021). "robustbase: Basic Robust Statistics", R package version 0.93-7, http://robustbase.r-forge.r-project.org/

Maronna*, R.A.*, Martin, D.R. and Yohai, V.J. (2006). "Robust Statistics: Theory and Methods", Wiley, New York.

Neyman, *J. (1934).* "On the two different aspects of the representative methods. The method stratified sampling and the method of purposive selection", Journal of Royal Statistical Society, 97, 558-606.

Russo S., Disch A., Blumensaat F. and Villez K. (2019). "Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data". Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2019), Copenhagen, Denmark, September 1-4.

Srivastava N, Hinton G., Krizhevsky A., Sutskever I. and Salakhutdinov R. (2014). "Dropout: a simple way to prevent neural network from overfitting", Journal of Machine Learning Research, 15.

Tukey, J. W. (1977). "Exploratory Data Analysis", Addison- Wesley, Reading, MA.

Zambuto F., Buzzi M. R., Costanzo G., Di Lucido M., La Ganga B., Maddaloni P., Papale F. and Svezia E. (2020). "Quality checks on granular banking data: an experimental approach based on machine learning", Banca d'Italia, Occasional Papers, No. 547.

Zambuto F., Arcuti S., Sabatini R. and Zambuto D. (2020). "Application of classification algorithms for the assessment of confirmation to quality remarks", Banca d'Italia, Occasional Papers, No. 631.

Zhou and Goldman S. (2004). "Democratic co-learning", in IEEE 16th International Conference on Tools with Artificial Intelligence (ICTAI), pages 594–602.

Zhou Z. and Li M. (2005). "Tri-training: exploiting unlabeled data using three classifiers", IEEE Transactions on Knowledge and Data Engineering, 17(11):1529–1541.

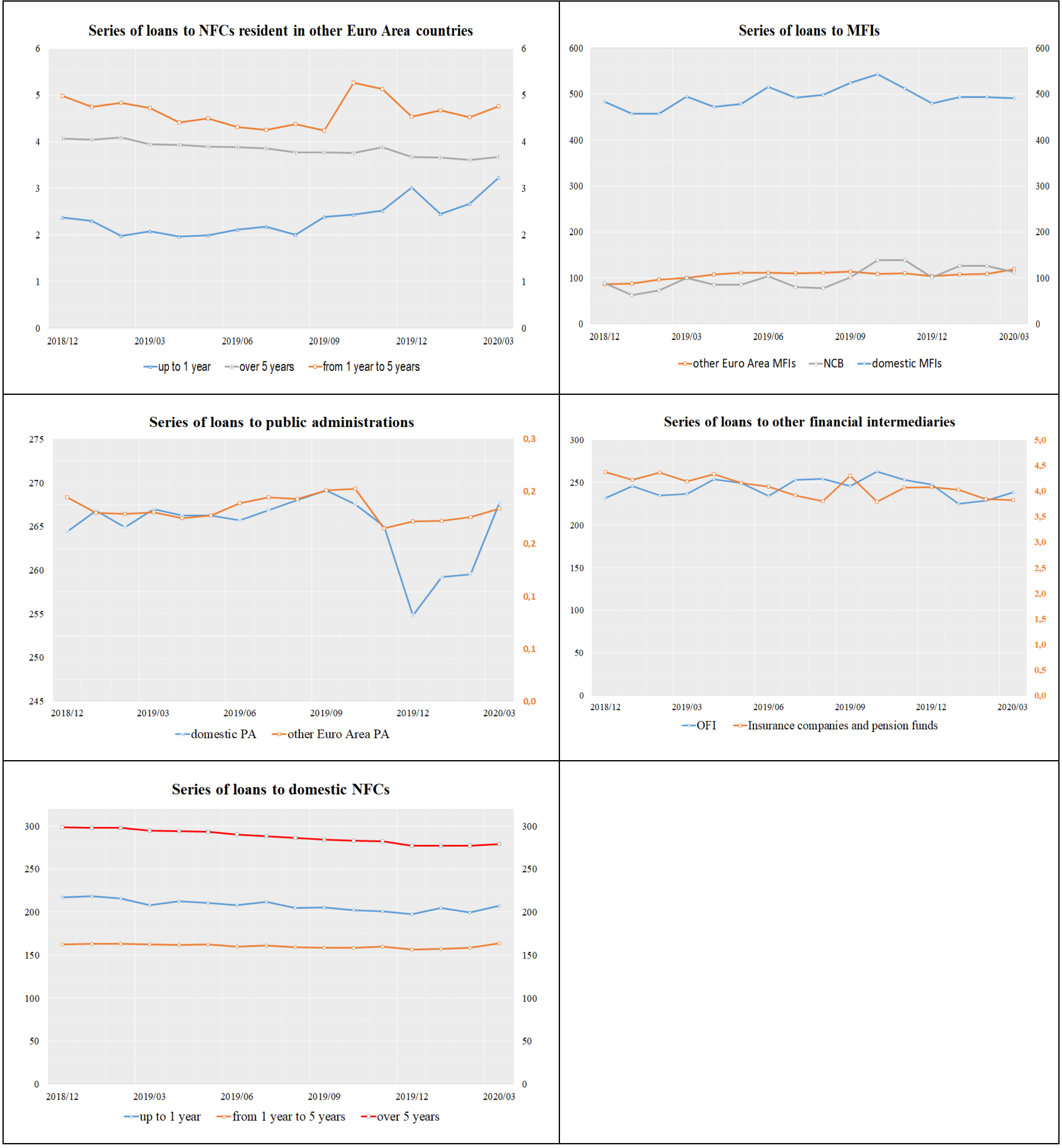
Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P. (2004). "Image quality assessment: from error visibility to structural similarity", IEEE transactions on image processing, 13(4):600–612.

Wolpert, D. (1992). "Stacked generalization", Neural Networks, 5, 241-260, https://doi.org/10.1016/S0893-6080(05)80023-1.

Yarowsky D. (1995). "Unsupervised word sense disambiguation rivaling supervised methods". In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196, Association for Computational Linguistics.

Appendix A - Tables and Charts

Figure A1: BSI aggregates



Series of loans to NFCs resident in other Euro Area countries

Series of loans to MFIs

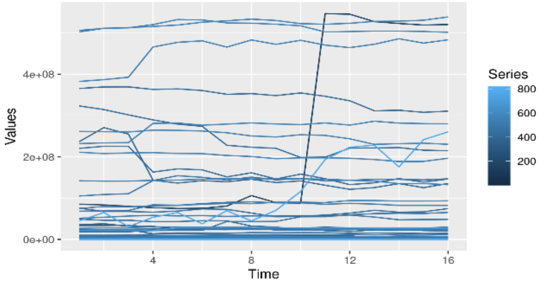
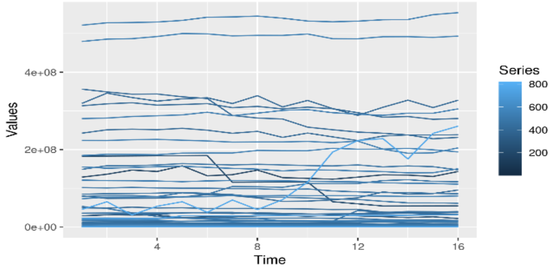
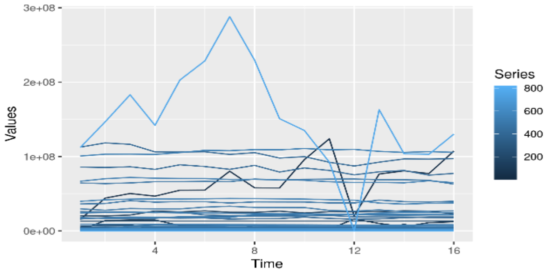
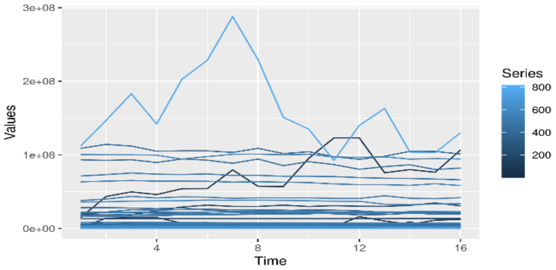
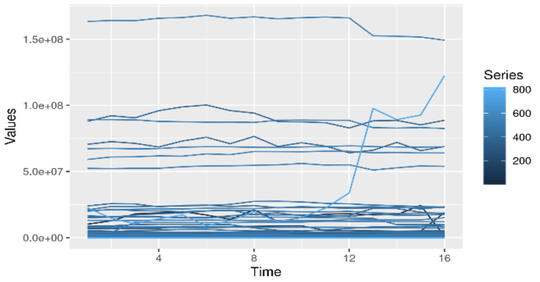
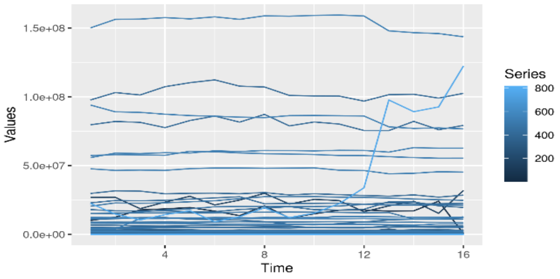
Series of loans to public administrations

Series of loans to other financial intermediaries

Series of loans to domestic NFCs

Amounts in billions of euros.

Figure A2: AnaCredit and BSI series (examples)

Bank	Reconstructed BSI series from AnaCredit	BSI series
1		
3		
5		

Amounts in billions of euros.

Figure A3: AnaCredit and FinRep series (examples)

Bank	Reconstructed FinRep series from AnaCredit	FinRep series

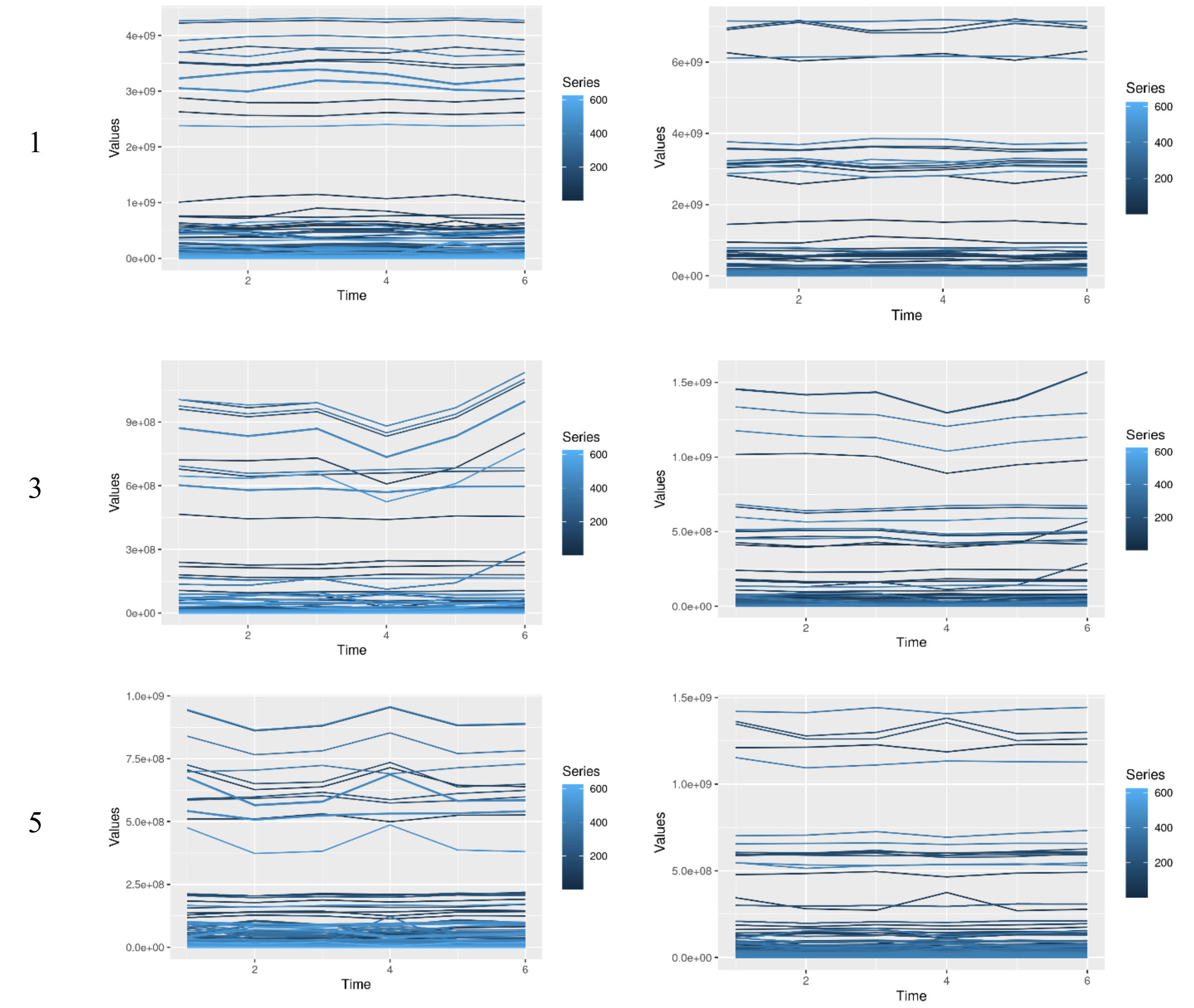
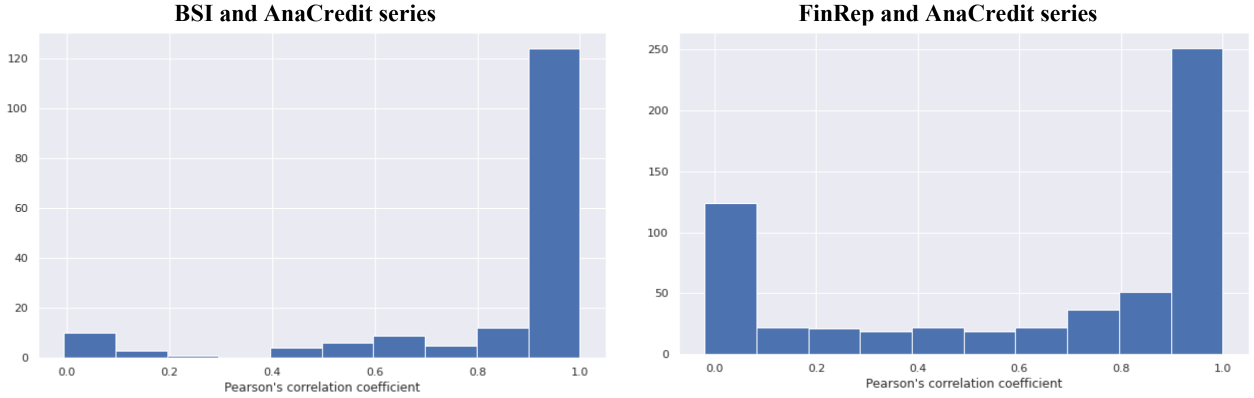


Figure A4: Distributions of the correlation between the compared aggregates



BSI and AnaCredit series

FinRep and AnaCredit series

Table A1: stratified sampling dimension

	AnaCredit-BSI		AnaCredit-FinRep	
Strata	**not sampled**	**sampled**	**not sampled**	**sampled**
0-0-0	4192	11	22210	9
0-0-1	2	2	35	2
0-1-0			13	2
0-1-1			4	2
1-0-0	425	11	3529	8
1-0-1	8	3	3	2
1-1-1	0	1		
0-cases		2529		9761
1-cases		587		1756
Total	**4627**	**3144**	**25794**	**11542**

x-y-z stratum is identified respectively by Robust Regression (x), CNN (y), DNN (z) by binary prediction (0 not anomaly, 1 anomaly). 0-cases previously classified not anomalies and 1-cases previously classified anomalies

Figure A5: performance metrics

		Predicted	
		Positive (1)	**Negative (0)**
Actual	Positive (1)	True positive (tp)	False negative (fn)
	Negative (0)	False positive (fp)	True negative (tn)

0=Not outlier, 1=Outlier

Metrics:

Precision $$\frac{\mathrm{tp}}{(\mathrm{tp}+\mathrm{fp})}$$

Recall $$\frac{\mathrm{tp}}{(\mathrm{tp}+\mathrm{fn})}=\frac{\mathrm{tp}}{\mathrm{p}} \quad tpr\ sensitivity$$

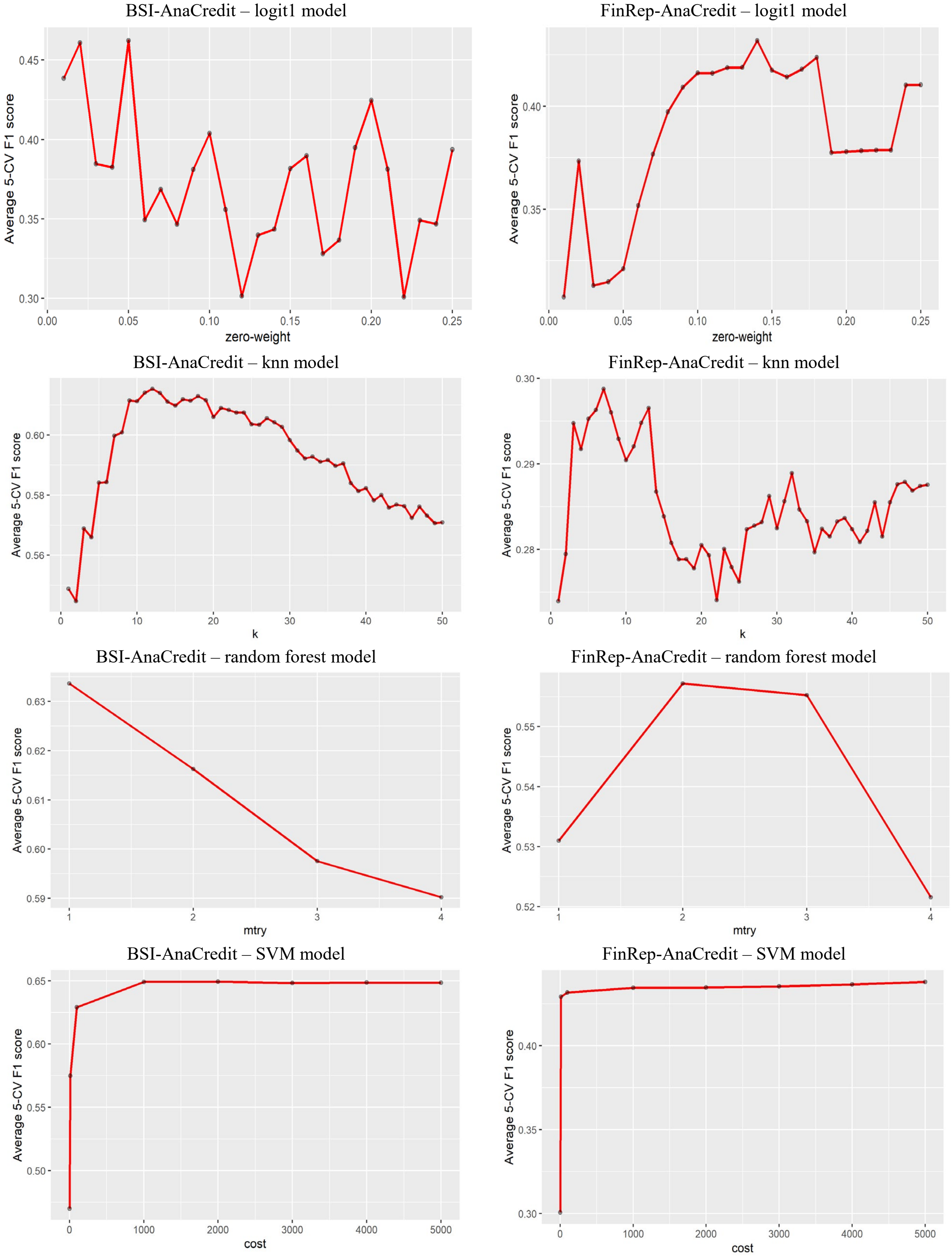
$$\frac{\mathrm{tn}}{(\mathrm{tn}+\mathrm{fn})}=\frac{\mathrm{tn}}{\mathrm{n}} \quad tnr\ specifity$$

Accurancy $$\frac{(\mathrm{tp}+\mathrm{tn})}{(\mathrm{tp}+\mathrm{fp}+\mathrm{fn}+\mathrm{tn})}$$

F1 score $$2*\frac{(\mathrm{precision}*\mathrm{recall})}{(\mathrm{precision}+\mathrm{recall})}$$

Balanced accurancy $$\frac{(\mathrm{tpr}+\mathrm{tnr})}{2}$$

Figure A6: optimizing meta-parameters


BSI-AnaCredit – logit1 model

FinRep-AnaCredit – logit1 model

BSI-AnaCredit – knn model

FinRep-AnaCredit – knn model

BSI-AnaCredit – random forest model

FinRep-AnaCredit – random forest model

BSI-AnaCredit – SVM model

FinRep-AnaCredit – SVM model

The parameters have been optimized with respect to the mean value of the simulated response variables.

Appendix B - Robust regression equation

Let $P_{i,j,t,s}$ be the amount of loan s of the subportfolio j granted from bank *i* at reference date *t*. Then the aggregated amount for bank *i* and the *j-th* subportfolio at date *t* is given by:

(B.1) $$F_{i,j,t} = \sum_{s=1}^{J_{i,j,t}} P_{i,j,t,s}\,,$$

where $J_{i,j,t}$ is the number of loans in the the j-th subportfolio.

In the AnaCredit framework, there are different criteria triggering a reporting obligation of loans (i.e. that regulatory threshold of 25,000 euros, the counterparty classified as legal persons, etc.). Therefore, the sum is only restricted to such eligible loans:

(B.2) $$A_{i,j,t} = \sum_{s=1}^{J_{ijt}} P_{i,j,t,s} \cdot I\,(s\!: s\ \epsilon\ eligible\ loans),$$

where $I(x)$ is an indicator function which is equal to 1 in case of eligible loan. $A_{i,j,t}$ is by definition less or equal to $F_{i,j,t}$. Since we have errors in AnaCredit data ($\xi_{i,j,t}$), we can express the amount observed: $A_{i,j,t}$ as the product of the 'true' value $\mathrm{A}^{*}_{\mathrm{i,j,t}}$ and an error $\xi_{\mathrm{i,j,t}}$: $A_{i,j,t} = A^{*}_{i,j,t} * \xi_{i,j,t}$. Therefore, the difference between the logarithm two aggregates of the two compared datasets can be expressed as follows:

(B.3) $$\log(A_{i,j,t}) - \log(F_{i,j,t}) = \log(A^{*}_{i,j,t}) + \log(\xi_{i,j,t}) - \log(F_{i,j,t})$$

For an unbiased estimator $T_{i,j,t}$ of the 'true' difference, in the form $T_{i,j,t} = \log(A^{*}_{i,j,t}) - \log(F_{i,j,t}) - u_{i,j,t}$, so the equation (B.3) can be written,

(B.4) $$\log(A_{i,j,t}) = \log(F_{i,j,t}) + T_{i,j,t} + \log(\xi_{i,j,t}) + u_{i,j,t}$$

The previous equation represents the theoretical model, in which the reporting error adds to a white noise $u_{i,j,t}$. An easy empirical specification to capture the relation between the two variables is by means of an Autoregressive Distributed Lag model (Granger, 1981) of order (1,1):

(B.5) $$log(A_{i,j,t}) = \alpha_0 + \alpha_1 log(A_{i,j,t-1}) + \alpha_2 log(F_{i,j,t}) + \alpha_3 log(F_{i,j,t-1}) + \epsilon_{i,j,t}$$

As we are interested to capture the differences in the aggregated amounts as independent variable, in order to specify T as function of past differences, we impose the restriction $\alpha_3 = -\alpha_1$. This restriction allows to obtain the following restricted model:

(B.6) $$\log(A_{i,j,t}) = \alpha_0 + \alpha_2 \log(F_{i,j,t}) + \alpha_3 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}$$

When $E(T_{i,j,t}) = \alpha_3 log(F_{i,j,t-1}/A_{i,j,t-1})$ and $\alpha_0 = 0$, $\alpha_2 = 1$, this last equation is equivalent to our theoretical model in B.4, where $\epsilon_{i,j,t} = log(\xi_{i,j,t}) + u_{i,j,t}$. Since T is non-positive and the difference referred to t-1 is non-negative, then the coefficient α_3 must be non-positive. It's worth to noting that the error reporting term *(*$\xi_{i,j,t}$) is only included in the error component term of equation B.6.

In our work, we have applied a log transformation[21] of the original data values.

[21] We adopt the *log1p* function of x, that is the natural logarithm of *x+1*.