BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Learning from revisions: a tool for detecting potential errors in banks' balance sheet statistical reporting

by Francesco Cusano, Giuseppe Marinelli and Stefano Piermattei

March 2021

Number 611

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Learning from revisions: a tool for detecting potential errors in banks' balance sheet statistical reporting

by Francesco Cusano, Giuseppe Marinelli and Stefano Piermattei

Number 611 – March 2021

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

LEARNING FROM REVISIONS: A TOOL FOR DETECTING POTENTIAL ERRORS IN BANKS' BALANCE SHEET STATISTICAL REPORTING

by Francesco Cusano*, Giuseppe Marinelli** and Stefano Piermattei*

Abstract

Ensuring and disseminating high-quality data is crucial for central banks to adequately support monetary analysis and the related decision-making process. In this paper we develop a machine learning process for identifying errors in banks' supervisory reports on loans to the private sector employed in the Bank of Italy's statistical production of Monetary and Financial Institutions' (MFI) Balance Sheet Items (BSI). In particular, we model a "Revisions Adjusted – Quantile Regression Random Forest" (RA–QRRF) algorithm in which the predicted acceptance regions of the reported values are calibrated through an individual "imprecision rate" derived from the entire history of each bank's reporting errors and revisions collected by the Bank of Italy. The analysis shows that our RA-QRRF approach returns very satisfying results in terms of error detection, especially for the loans to the households sector, and outperforms well-established alternative outlier detection procedures based on probit and logit models.

JEL Classification: C63, C81, G21.
Keywords: banks, balance sheet items, outlier detection, machine learning.
DOI: 10.32057/0.QEF.2021.611

Contents

* Bank of Italy, Directorate General for Economics, Statistics and Research, Statistical Data Collection and Processing Directorate.

** Bank of Italy, Directorate General for Economics, Statistics and Research, Statistical Analysis Directorate.

1. Introduction and motivation: detecting errors in banks' balance sheet data*

In the context of European Regulations, National Central Banks regularly collect and disseminate monthly data on monetary and financial institutions' (MFIs) balance sheet items (BSI).[1] BSI data represent a crucial source for the production of both national and euro-area monetary aggregates (M1, M2 and M3) and their counterparts, which have a major role for the ECB assessment of the risks to price stability and for deriving the Eurosystem's minimum reserve requirements for credit institutions. As a result, the Bank of Italy monthly collects, elaborates and analyses a huge amount of individual balance sheet data reported by the entire population of Italian banks (487 reporting banks at the end of 2019) in order to compile aggregate statistics to be provided to the ECB and the public. Table 1 shows a simplified scheme of a bank's balance sheet.

Table 1

Bank's balance sheet simplified scheme

Assets	Liabilities
1) Cash	1) Deposits
2) Loans	2) Debt securities issued
3) Debt securities held	3) Capital and reserves
4) Equity	4) Remaining liabilities
5) Investment fund shares	
6) Non financial assets	
7) Remaining assets	

Among the items reported in Table 1, particular attention is devoted to the analysis of the developments of the loans granted by banks to the private sector, mainly non-financial corporations and households. In the last three years, in Italy these loans have accounted, on average, for two thirds of total consolidated loans and 35 per cent of total assets of the banking system. The development of credit to the private sector is one of the key factors monitored and analyzed in the context of the ECB monetary analysis as it explains part of the variation in the

* We thank Gianluca Cubadda, Silvia Fabiani, Alessio Farcomeni, Francesca Monacelli, Giorgio Nuzzo, Valeria Pellegrini, Riccardo Piermattei and Roberto Sabbatini for useful comments on earlier versions of the paper.

[1] See the ECB Regulation ECB/2013/33 on the balance sheet of the monetary financial institutions sector, and the ECB Guideline of 4 April 2014 on monetary and financial statistics (ECB/2014/15).

holdings of liquidity by the private sector, which may have different implications for price stability. It also embeds relevant information on economic growth, economic agents' confidence and financial stability. Credit growth analysis based on banks' balance sheet data can be particularly challenging, since it has to take into account and be combined with ancillary information on the dynamics of securitisations and other loan transfers, write offs/write-downs and statistical breaks due to, among the others, changes in the reporting population or in the characteristics of the counterparties.

Ensuring and disseminating high quality data is therefore crucial for adequately supporting the monetary authority's supervision and decision-making process for monetary policy and financial stability. In this respect, over the last few years the necessity of adequate and timely data revealed by the global financial crises and the improvements of more performing IT infrastructure led to an increasing collection of granular banking data that need to be checked and validated. In turn, identifying outliers that can signal potentially wrong data in the individual banks' supervisory reports has become more challenging due to the increasing amount and layers of data to elaborate.

Currently, the Bank of Italy's error detection approach in the BSI statistical production is based on automated procedures whose performance, although satisfactory, suffers from two main shortcomings. First, the absolute and percentage variations of the variables are usually assessed against predetermined thresholds that are often identical for all banks and variables; these procedures are therefore relatively inefficient, since the possible outliers and errors they signal have to be re-assessed – case by case – by the analysts in order to select plausible mistakes to be reported to banks. Second, as we will discuss later, these procedures do not process all the huge amount of available relevant information that instead some advanced *machine learning* techniques could exploit.

Within Central Banks the use of *big data* and *machine learning* (ML) techniques has become very popular over the last few years (Chakraborty and Joseph, 2017; Bank for International Settlements, 2019). The range of applications of ML methodologies' covers, among others, policy analysis evaluation, forecasting problems and statistical production. Big data and text mining techniques are used, for instance, to evaluate the impact of authorities' speech and reports on financial markets and public's sentiment (Correa et al., 2017) or to build up economic and uncertainty indicators by analysing the frequency of specific news or

keywords in prominent web sources (Baker et al., 2016; Ardizzi et al., 2019). Neural networks, random forest and gradient boosted tree algorithms have been exploited to forecast macroeconomic variables (Salzano, 1999; Moody, 2012) or corporate default (Moscatelli et al., 2019). Finally, ML techniques have been also used for enhancing the quality of statistical production (Cagala, 2017). Examples relate to the imputation of missing information (Carboni and Moro, 2018; Giudice et al., 2020) or data quality and errors detection procedures (Zambuto et al., 2020).

The goal of this paper is to develop and test an automated procedure able to identify in *real time* potential errors in the banks' supervisory reports used in the Bank of Italy's BSI production process. As anticipated, we focus on the outstanding amounts of loans to non-financial corporations and households reported by banks, given their relevance in monetary analysis and their weight in bank assets. Like in Zambuto et al. (2020), we make use of the Quantile Regression Random Forest (QRRF) algorithm in order to estimate prediction intervals (acceptance regions) associated to banks' reported data. With respect to the standard quantile regression, the QRRF algorithm offers relevant time-saving computation advantages deriving from the lack of parameter estimation, an important feature when processing millions of data. By employing a QRRF supervised learning algorithm, we are able to exploit several advantages related to the BSI production process. First, as the Bank of Italy has collected a huge amount of data since the very beginning of the BSI monthly production, we can test and train an algorithm on a big dataset. Second, the focus on an important balance sheet variable – bank loans granted to the private sector, i.e. a phenomenon widely studied in the economic literature – makes it relatively easy to specify a valid underlying model and a set of relevant independent variables.

Additionally, we have the possibility to test an algorithm and to make inference on past values reported by banks for which we already know whether they were wrong or correct data. In other words, for each past record initially reported by banks, we know the *true response* (the value finally confirmed or revised by banks) that we can use to evaluate the predictions of the QRRF procedure. Moreover, during the last years of BSI monthly production, for each bank in the sample the Bank of Italy has been collecting the entire history of *errors* and *revisions*, which provides a relevant information set indicating how "on average" the bank has been "imprecise" in reporting the data. The novelty of the approach followed in this paper is the elaboration of a two-stage - "Revisions Adjusted" - QRRF, where this huge "errors and

revisions information set" is used to estimate – for each bank and each month – an "imprecision rate" through which we *calibrate* the specific prediction interval resulting from the QRRF methodology.

A final advantage stems once again from the specific structure of the statistical production process. In building up a model specification that pins down the relevant independent variables able to explain banks' credit dynamics and to identify potential errors, we adopt the "supply-side" theory of credit (Adrian and Shin, 2010; Bonaccorsi di Patti and Sette, 2012; Jimenez et al., 2012; Bofondi, Carpinelli and Sette, 2013; De Bonis, Nuzzo and Stacchini, 2014; Cingano, Manaresi and Sette, 2016; Affinito, Albareto and Santioni, 2016). According to the latter (as opposed to the demand-side theory), loans dynamics depends mainly from bank balance sheet characteristic and not from demand variables. These independent "balance sheet" indicators are promptly available during the BSI production round, enabling the implementation of a *"real time"* outlier identification process.

The procedure we develop provides very satisfying results in terms of error detection, especially when we consider loans to the household sector: in the "optimized"[2] scenario, we identify up to 75 per cent of banks' errors and 92.8 per cent of the correctly reported values – that is, the procedure does not signal false positives. Concerning loans to non-financial corporations, the model identifies 93.3 per cent of the values correctly reported by banks and 40 per cent of the wrong values. As we will discuss later, the less satisfying results for these loans are partly due to specific features of the data that would make error-identification difficult for any procedure. Notwithstanding, our algorithm returns much better results than other alternative procedures, like, for instance, the well-established probit and logit models.

The rest of the paper is organized as follows: Section 2 discusses the process of BSI statistics' monthly production to show how its temporal structure can be exploited to test and train an automated error-detection algorithm. Section 3 illustrates the Revisions Adjusted - Quantile Regression Random Forest (RA-QRRF) algorithm that we use in the paper and discusses the variables employed in the specification. Section 4 shows the empirical results. Section 5 runs a robustness analysis by comparing the RA-QRRF results with the prediction of probit and logit models. Section 6 concludes.

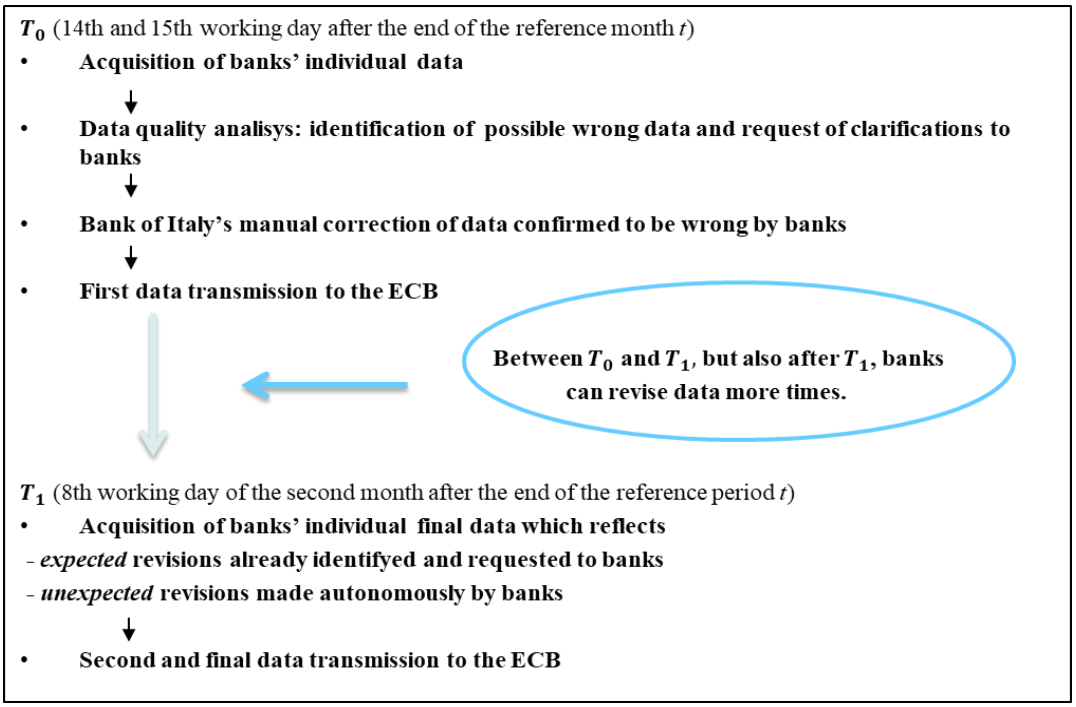
[2] We will see in Paragraph 4 what "optimized" means.

2. Exploiting the temporal structure of BSI data production through a machine learning approach

The *temporal structure* of Bank of Italy's BSI monthly statistical production – the process of acquisition, analysis, validation and transmission of data – is crucial for the delineation of the algorithm testing strategy we use. The BSI statistical production consists of two main phases, the *first transmission* (done at time T_0) and the *second transmission* (at time T_1) of the aggregated data to the ECB (Figure 1).

Figure 1

Temporal structure of the BSI statistical production



At time T_0, on the 14th working day after the end of the reference month t, individual bank's data referred to month t are received and analyzed in order to identify possible outliers and wrongly reported data by banks. Currently, the errors detection is run through automated procedures based on absolute and percentage thresholds exogenously determined and equal for all the banks and the phenomena. In this relatively time-consuming procedure, a large part is left to the analyst's expertise, since detected outliers and errors are actually re-gone through by the analysts and plausible mistakes are selected. Request of clarifications are then sent to banks in order to have revisions (i.e. admissions of mistakes) or confirmations of the data

initially reported. During the 15th working day after the end of the reference month, manual corrections are applied on data confirmed to be wrong by banks and the corrected aggregated series are finally sent to the ECB in what is called the *first transmission*.

During the following days, banks are requested to correct and send revised correct data; revised data are finally elaborated and aggregated at time T_1 (the 8th working day of the second month after the end of the reference period) and re-transmitted to the ECB for what is called the *second (final) transmission* of BSI data for month t. It is important to underline that between T_0 and T_1, banks can revise autonomously *all* data and not only the ones previously identified as errors for which a request of revisions was forwarded. Moreover, it is worth remarking that banks can revise the same data *more times*. Banks can also send revisions of data not only between T_0 and T_1, but even after the final transmission of data to the ECB.

Such a production structure offers two clear advantages that a *supervised* learning process of outlier detection can exploit. First, in testing an error detection procedure by using past months data samples, we have the advantage to make inference on banks' reported data (i.e., the value of the outstanding amount of loans granted reported at time T_0 by bank i, $\widehat{L_{i,0}}$) for which we already know the true response (the final *correct* revised data $L_{i,1}$ sent for the second transmission at T_1). This actually implies that we have a series of monthly samples to work with in which: a) the "training set" consists of data on loans initially reported by the entire sample of Italian banks; b) we can immediately evaluate the predictions about the potential outliers $\widehat{L_{i,0}}$ by observing the final true value $L_{i,1}$.

Second, the structure of the production process enables us to collect, for each bank, its own history of errors and subsequent revisions. This history – encompassing the frequency and the magnitude of revisions – offers a very relevant information indicating how imprecise – "on average" – the bank used to be in reporting data. We make use of this information to build an "imprecision rate" for each bank i at each time t through which we calibrate their acceptance region resulting from the QRRF methodology. We name this procedure a Revisions Adjusted Quantile - Regression Random Forest (RA-QRRF).

Apart from the advantages stemming from the specific structure of the production process and the type of information available, the RA-QRRF error detection procedure also takes advantage of the huge dimension of the data to elaborate. Since the beginning of the BSI

production process, Bank of Italy has been collecting a huge amount of individual banks' balance sheet items, a large part of it related to loan items. Also, since the beginning of the collection process, Italian banks have transmitted millions of revisions, a big part of it related to loans items.[3]

3. The QRRF methodology and the specification of the model

Regression Trees have become very popular as long as they have proven to be powerful nonparametric tools for regression (estimation of conditional means) and classification analyses. *Random Forests*, introduced by Breiman (2001), represents the application of *bootstrap* to regression trees through the building of a large collection of uncorrelated trees and then averaging them. Bootstrap is a well-established technique for reducing the variance of the estimator function. The idea behind the bootstrap is that an average of Z i.i.d. random variables, each with variance σ^2, has overall variance σ^2/Z. If the variables are simply i.d., but not necessarily independent, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{Z}\sigma^2. \tag{1}$$

As Z grows and ρ reduces, the variance of the estimator function becomes smaller. This is exactly the goal of the Random Forest technique when building a large number of uncorrelated trees from the data (this is obtained by a particular variables selection in growing the tree) and then averaging them to obtain an estimation of the *conditional mean* of the response variable[4]

$$\hat{E}(y/X = x) = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}[\,Y_i/X = x]w_n(x;\,\partial_k) \tag{2}$$

where Y_i is the observable response variable in the unit $i = 1, \dots, N$; N is the total number of observations in the sample, K is the number of built trees, $X = x$ is a given realization of the independent variables and

[3] We make use of a (big) subset of this huge amount of data because, as we will see, we had to focus on a dataset starting from December 2017.

[4] For the details about the specific algorithm to grow trees and random forests, see Hastie et al. (2001) and James et al. (2013).

$$w_n(x;\partial_k) = \begin{cases} 1/(\text{number of observations in the leaf of tree } k \text{ where } X = x); \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

in which ∂_k represents the set of parameters that determines how tree k has grown. In this paper, like in Zambuto et al. (2020), we employ the Quantile Regression Random Forests (QRRF) proposed by Meinshausen (2006) in order to derive acceptance regions for outliers and errors detection of the response variable. In QRRF, trees are grown exactly as in the standard random forest algorithm. However, built trees are then averaged to obtain an estimation of the *conditional distribution* of the response variable for a given determination $X = x$ of the independent variables

$$\hat{F}(y/X = x) = \hat{P}(Y \leq y/X = x) = \hat{E}(1_{\{Y \leq y\}}/X = x) = \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N}[\,1_{\{Y \leq y\}}/X = x]w_n(x;\,\partial_k) \quad (4).$$

By applying this methodology to our sample, the estimation of the conditional distribution allows to compute the conditional quantiles

$$\widehat{Q_\alpha}(y/X = x) = sup\{\, y:\, \hat{F}(y/X = x \leq \alpha)\,\} \quad (5)$$

that constitute the limits of the *prediction interval PI* for any given variable $y_{i,t}$ reported by bank i for month *t*

$$PI_{i,t} = [\widehat{Q_\alpha}(y/X = x_{i,t}),\ \widehat{Q_{1-\alpha}}(y/X = x_{i,t})] \quad (6).$$

Bank's i reported values that fall outside the prediction interval in (6) are potential outliers, encompassing genuine anomalous observations and wrong reported data.

In our analysis the dependent variable is $\Delta L^0_{i,s,t}$, the absolute variation of the outstanding amount of loans reported in occasion of the *first data transmission* by bank i, for month *t,* with counterpart sector *s* (as said, we focus on the two more relevant segments of the private sectors, households and non-financial corporations). In order to perform the QRRF algorithm and obtain an estimation of the quantiles

$$\widehat{Q_\alpha}(\Delta L_{i,s,t}/X = x_{i,s,t}) = \widehat{Q_\alpha}(\Delta L_{i,s,t}, X^1_{i,s,t},\ X^2_{i,s,t},\ \ldots\ldots\ldots,\ X^j_{i,s,t}) \quad (7)$$

we need to select a set of variables $(X^1, X^2 \ldots\ldots\ldots, X^j)$ able to explain the monthly variation of bank's credit. To this aim, we can rely on a vast literature on the determinants of bank's

credit. In this literature, we can distinguish between two main theoretic points of view: while some authors stress the main relevance of demand-side variables in driving bank's credit, others argue that banks loans almost entirely depends on bank's own balance sheet characteristics. To our purposes, sharing the supply-side view and relying on balance sheets variables make possible to develop an automated procedure that could be run in *real-time* on the day of the statistical production, given that both response variables $\Delta L_{i,s,t}$ and the balance sheet control variables $(X^1, X^2 \dots \dots, X^j)$ would be contemporaneously available in the BSI supervisory reports provided by banks.

In the "supply-side determinants" view of banks' credit, loan growth is positively dependent on the banks' level of capital, which measures the ability to expand credit by maintaining the desired level of capital ratio (Bonaccorsi di Patti and Sette, 2012). It also depends on the level and the structure of banks' funding (Bonaccorsi di Patti and Sette, 2012; De Bonis, Nuzzo and Stacchini, 2014), i.e. the amount of interbank and retail deposits. Retail deposits are generally considered as an indicator of the type of specialization of the banks: the more a bank relies on retail funding, the more it tends to be specialized in traditional credit activities to the private sector rather than other forms of financial investments (Infante et al., 2020; Farné and Vouldis, 2017; Altunbas et al., 2011). Moreover, a business model where funding is mostly based on retail deposits instead of the recourse to the wholesale interbank market is most of the time associated with a more stable pattern of loans on the asset side due to the higher stability of the former with respect to the latter, especially during periods of crisis. Other types of liquid assets, such as the amount of public bonds held, may instead have a negative relationship with the magnitude of loan variation, as they represent an alternative form of profitable investment easy to substitute with loans if the bank wants to expand credit (Affinito, Albareto and Santioni, 2016; Bonaccorsi di Patti and Sette, 2012). Bad loans also have, in theory, a negative relationship with the growth of credit, given that they represent an inverse measure of the quality of bank's credit assets and, consequently, an inverse measure of the ability to expand it (Bonaccorsi di Patti and Sette, 2012). Bank's size variables, such as the amount of total assets, also play a role in determining the amount of credit granted (Cingano, Manaresi and Sette, 2016). Finally, the model should also consider some variables representing the specificity of the relation between the bank and clients belonging to different economic sectors such as the share of bank's credit to firms or households to the overall credit (Bofondi, Carpinelli and Sette, 2013). Hence, the model can be represented by the following

equation, where in brackets we report our *a priori* beliefs on the signs of the relationships between the explanatory variables and the outcome variable based on the cited studies:

$$\widehat{Q_\alpha}(\Delta L/X) = \widehat{Q_\alpha}(\Delta L/CAPITAL(+), INTERBANK\ FUNDING(+), RETAIL\ FUNDING(+), PUBLIC\ BONDS(-), BAD\ LOANS(-), ASSET(+), SIZE\ (+), SHARE\ (+)) \quad (8).$$

As mentioned in previous paragraphs, the novelty of our contribution lies in the idea of exploiting, for each bank, its own history of supervisory reports' errors and revisions drawn from the tables of the Bank of Italy Statistical Datawarehouse (SDW). In the SDW all data reported by banks are logged with a timestamp and in an incremental way with respect to the first data transmission on a specific item and its attributes. Such functionality, which is quite common in data base infrastructures, is usually employed to ensure the replicability of elaborations over time when revisions might come in between two different elaborations. The log tables of the SDW can provide relevant information for the estimation of the "likelihood" that the bank's last reported data could be – *ceteris paribus* – the result of errors or not. To the best of our knowledge, there is no other application in the literature exploiting such information in the context of outlier detection in statistical microdata. To this aim, for each bank i at month t we computed different measure of an "*imprecision rate"* (or "*score*"), $\theta_{i,t}$. We use these rates in a two-stage - "Revisions Adjusted" - Quantile Regression Random Forest, where the prediction intervals are function of the score

$$PI_{i,s,t} = [\widehat{Q_\alpha}(\Delta L_{i,s,t}, X_{i,s,t}, F(\theta_{i,t})), \widehat{Q_{1-\alpha}}(\Delta L_{i,s,t}, X_{i,s,t}, F(\theta_{i,t}))] \quad (9).$$

While in the first stage prediction intervals are obtained as standard output of the QRRF algorithm, in the second one the computed imprecision rates are used through specific "penalty functions" $F(\theta)$ to calibrate the intervals. Their role is to penalize banks that tended to make lot of errors in the past by restricting - *ceteris paribus* - the intervals and, at the same time, by benefiting banks with few errors by expanding the prediction intervals.

In our work, the imprecision scores are estimated through a linear probability model (LPM)[5] with high-dimensional fixed effects *á la* Gaure (2013), where the dichotomic outcome

[5] Despite the weaknesses of the LPM, for instance its heteroskedasticity and the fact that it does not constraint the predicted probabilities to lie between 0 and 1 (Johnston and Di Nardo, 1996), we decided to use it rather than other dichotomous output models such as the probit or the logit because, in these non-linear models, the estimated derivative effects on the probability vary with the level of the dependent variables (Greene, 2002).

variable $y_{i,t}$ is equal to 1 if the reporting agent i has revised the observation in month t and is equal to 0 otherwise

$$Prob(y_{i,t} = 1) = \beta X_{i,t} + \sum_j \gamma_j \eta_{j,i,t} + \varepsilon_{i,t} \quad (10).$$

It is worth clarifying that in the estimation of (10), the variable $y_{i,t} = 1$ encompasses all possible reporting errors and revisions made by banks. In other words, not only revisions in the total amount of the reported loans (i.e. $L^1_{i,s,t} \neq L^0_{i,s,t}$) that we consider in our outlier detection algorithm, but also revisions in all the sub-details of the bank's reporting. Indeed, bank's reported information covers not only the amount and the counterpart sector, but also a set of sub-details such as the maturity of the loan, the purpose (consumer, credit, mortgage, etc.), the currency, the specific subsector (i.e. producer vs consumer households) that can be initially misreported and later revised. As we will see in the next paragraph, while errors and revisions in the total reported amount (more relevant to the goal of the BSI statistical production) are not so common, the same is not true for the reporting of the sub-details, which represent the vast majority of the errors.

In (10), the outcome variable is regressed on a set of explanatory variables: $X_{i,t}$ represent variables related to the *value* of the balance sheet item, in particular the final (log) amount of the balance-sheet item and the ratio of the revision over its final amount. $\eta_{j,i,t}$ is a vector of dummy and categorical variables representing specific characteristics of the reporting agent, the reference period and of the balance sheet items, such as the institutional subsector of the counterparty, the currency of the item and the nature of the instrument (securities, mortgages, etc…). The estimation is carried out for each reference period on a mobile window of 12 months, that is, when the reference period is T then the sample is limited to all the observations between $t = T - 12$ and $t = T - 1$. The imprecision rate - i.e. the likelihood associated to a revision by a reporting agent - is derived from the results of the LPM by extracting the marginal effect of a bank-level fixed effect (dummy) $\eta_{k,i,t}$ included among the independent variables, which represent an idiosyncratic bank-specific and time-invariant characteristic[6]

Therefore, they are not useful for the estimation of banks' imprecision rates. Moreover, heteroskedasticity issues can be dealt with through the estimation of robust standard errors.

[6] The idea is to disentangle from the probability to make a reporting error all the effects that can be due to characteristics other than an idiosyncratic feature of the bank, for example some characteristics of the reported

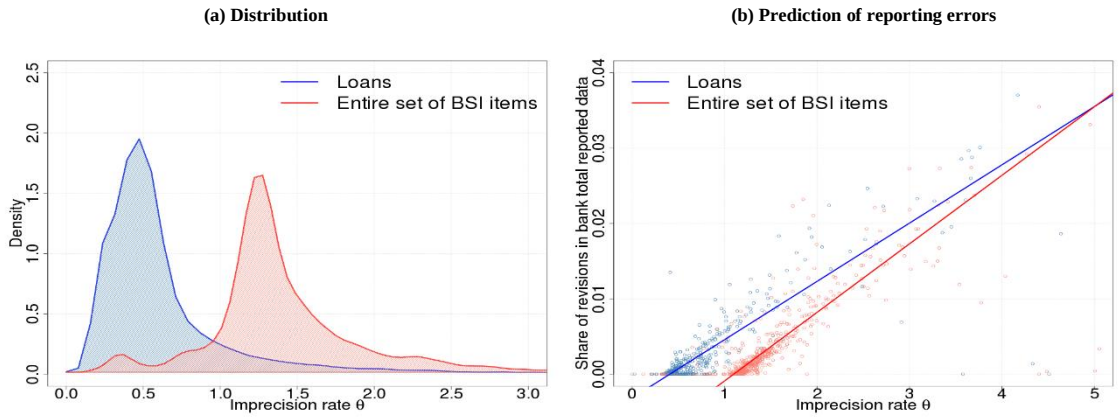
$$\theta_{i,t} = \widehat{\gamma_k}\eta_{k,i,t} \quad (11).$$

We computed two different versions of the imprecision rate: $\theta_{i,t}$, which is computed by considering errors and revisions in the entire set of balance sheet items reported by banks, and $\theta^{l}_{i,t}$ which considers only errors and revisions in the reporting of loans (the focus of the present paper). The left panel of Figure 2 depicts the distribution of the imprecision scores computed through the kernel density estimation. The distribution of the score related to the errors in all the balance sheet items has a mean higher than 1. Intuitively, it is higher than the mean of the imprecision rate estimated only considering data on loans, which is around 0.5. This is consistent with the intuition that the likelihood of making errors when considering *all* the BSI items is obviously higher than considering only a subset of them (e.g. data on loans). In Figure 5 of the Appendix we report some examples of the dynamics of the individual score θ_l over time and its relationship with banks' reporting errors.[7]

Figure 2

Estimated imprecision rates (scores)

(Banks'monthly rates from December 2017 to March 2020)



Note: the left panel of the figure depicts the distribution of the estimated imprecision rates on the entire set of BSI items and on the subset of items concerning loans. The distributions were derived on the basis of a kernel density estimation on the two variables. The right panel depicts the relationship between the estimated imprecision rates and banks' share of reporting errors in total reported data, i.e. the number of revised records divided by the total number of records at bank level. The lines represent the linear regression lines of the share of reporting errors on the imprecision rates.

instruments that can affect the probability to make a mistake (i.e. loans reported in foreign currency instead of euro, etc..).

[7] We do not report examples for the overall score θ because it is not straightforward to represent graphically its relationship with the revisions in the entire set of balance sheet items.

Once a reporting error has occurred, the score jumps and the penalty has a 12-month persistence while gradually falling to 0 in the absence of further errors. The estimated scores are positively correlated with the incidence of reporting errors at the bank level, as shown in the right panel of Figure 2. The imprecision rates significantly predict banks' reporting errors on the entire set of BSI items and on the subset of those concerning loans.

As we will see in the next paragraph, we tested different functional forms of the "penalty" (or "correction") functions. In the next paragraph, we illustrate the "results optimization" strategy to find the optimal model parameters and penalty function, and we illustrate the empirical findings.

4. Empirical results

In our analysis the QRRF algorithm is run on the following model specification

$$\widehat{Q_\alpha}(\Delta L^0_{i,s,t}/X = x_{i,s,t}) = \widehat{Q_\alpha}(CAP_{i,t}, INTB_{i,t}, RETAIL_{i,t}, PBONDS_{i,t}, BADL_{i,s,t}, TASS_{i,t}, SHARE_{i,s,t-1}, SIZE_i, (\Delta L/L)^1_{i,s,t-1}) \quad (12)$$

where $\Delta L^0_{i,s,t}$ is the variation in the stock of loans reported at the first data transmission (hence the apex "0") by bank i to sector s at month t. Given their relevance, we focus on two sectors, households and non-financial corporations. It is important to stress that changes in the stock of loans reported by banks in their balance sheets could be due to financial transactions, i.e. new loans being granted minus reimbursements, but also to write-downs/write-offs, securitizations and other loan transfers that are also reported by banks on a monthly basis. In order to avoid the algorithm signaling these reductions as possible outliers, loans variations are corrected by re-adding write-offs and net loans disposals. $CAP_{i,t}$ is the capital and reserves to assets ratio of bank i at the end of month t; $INTB_{i,t}$ are the interbank deposits over total asset net of capital. $RETAIL_{i,t}$ are the retail deposits (of households and non-financial corporations) over total liabilities net of capital and reserves. $PBONDS_{i,t}$ represents the holdings of public bonds over total assets. $TASS_{i,t}$ is the total amount of assets. $BADL_{i,s,t}$ is the amount of bad loans of sector s in the balance sheet of bank i at the end of the month t over total loans to that sector (at time $t-1$). $SHARE_{i,s,t-1}$ is the share of bank i loans over total loans granted to sector s in the previous month. $SIZE_i$ is a categorical variable used in

Bank of Italy's publications to classify banks in five assets categories.[8] $(\Delta L/L)^1_{i,s,t-1}$ is the percentage growth rate of bank i 's loans to sector s in the previous month.

Our dataset includes balance-sheet data on 28 months spanning from December 2017 to March 2020, a sample that consists of 257811 observations with an average of 485 reporting banks in each month.[9] Table 9 and Table 10 in the Appendix report the summary statistics and the correlation matrix of the variables.

The RA-QRRF algorithm aims at identifying potential outliers in banks' data reporting, i.e. anomalous data that need to be investigated and eventually corrected. In this respect, "wrong" data can be classified into two categories. On the one hand, they can be values misreported by banks, that is, errors in the reported stocks that have been later corrected autonomously by banks or because detected by Bank of Italy's analysts. The other type of data that the analyst must identify to apply a correction (and that therefore the algorithm must detect) are variations in the stock of loans that have to be handled with "statistical reclassifications". Statistical reclassifications come in every time there are variations of stocks not explained by financial transactions, price revaluations, or exchange rate fluctuations in case of foreign currency-denominated instruments. Examples are an increase in the stock of financial instruments due to a bank's acquisition by another bank (alternatively a reduction due to the selling of a bank), variations in the reported stocks due to a reclassification of a financial instrument (i.e. a repos reported as a bond issued up time t and then reported as loan debt from time $t+1$), etc. Statistical reclassifications of financial instruments are series transmitted by Bank of Italy to ECB in the BSI statistical production. However, reclassifications are not directly reported by banks but are identified by Bank of Italy's analysts when looking for anomalous stock variations that are not explained by other factors such as price revaluations. Statistical breaks not reflecting economic transactions that give rise to reclassifications are therefore outliers that the QRRF algorithm is requested and expected to detect as well.

[8] The classification in dimensional classes is based on the composition of the banking group. The categories are: "First five banking groups", "Big banks", "Foreign banks' branches", "Small banks" and "Minor banks" (Bank of Italy, Annual Report 2018).

[9] We actually collected balance sheet data starting from December 2014. However, given that data on securitizations and loans are available only from December 2017, a corrected response variable is available only starting from that month.

As mentioned earlier, an advantage of our analysis is the ex-ante knowledge of actual outliers in our dataset. In particular, from December 2017 to March 2020 banks reported 22 "wrong" values related to loans to households (14 errors in the reported values subsequently corrected and 8 reclassified values) and 43 wrong values related to loans to non-financial corporations (37 errors and 6 reclassifications) out of 13569 banks' reporting of loans to household (and 13569 to non-financial corporations).[10]

An optimal outlier detection procedure has to aim at the minimization of two types of error the output of the algorithm could provide. First, obviously it should minimize the number of non-identified real errors; we can label these errors as "non-identified true negatives" (or I-type errors). Second, it should also minimize the number of "identified false positives" (or II-type errors), that is observations being flagged as outliers by the procedure when they are actually not anomalous.

If we consider the simplified scheme reported in Table 2, in which we assume the total number of banks' reported data to analyze is N, with E misreported values among them ($E \subseteq N$), the goal of an optimal outlier detection procedure should be to minimize the sum of the elements on the secondary diagonal, i.e. the two types of errors $\omega + \Omega$ (or equivalently the maximization of the numbers in the principal diagonal). It is straightforward that, to our scope, the minimization of I-type errors Ω, that is the identification of all *real* errors of banks, has much more importance than the other type of error.

Table 2

Simplified scheme of I-type and II-type errors of the outlier detection procedure

	Correct reporting	**Errors in reporting**
Not signaled as outlier by the algorithm	$N - E - \omega$	Ω
Signaled as outlier by the algorithm	ω	$E - \Omega$

[10] In our analysis we considered as "reporting errors" revisions made between the first transmission at T_0 and the second transmission at T_1 higher than euro 20 millions.

One could argue that the only goal of the algorithm should be to reduce Ω to zero. This is actually possible, but given the specification of the model, and *ceteris paribus* the parameters of the model, it is clear the existence of a trade-off between ω and Ω; the reduction of Ω to zero could be accomplished with the cost of a procedure that signals "too many" II-type errors. This is not desirable, given that too many II-type errors imply too many (false) requests of clarification to be reported to banks, determining inefficiencies in a BSI production process that has to be completed only within two working days.

We can sum up by saying that an optimal outlier detection procedure should minimize some *loss function* of the type

$$\min LOSS(\omega\,, \lambda \cdot \Omega) \tag{13}$$

where λ is a parameter expressing the relative importance of the I-type errors with respect to the II-type ones. Once the specification of the variables relationship is fixed as in (12), the output of the RA-QRR algorithm – ω, Ω and hence the value of the loss function in (13) – depends on a set of model parameters. First, the confidence level α in (12), since the higher the confidence level is, the smaller will be – *ceteris paribus* – the prediction intervals for each observation and therefore more total outliers signaled. Second, the specific imprecision rate we decide to use (either θ, computed considering banks' errors in all the balance sheets reported items, or θ_l, computed only on errors in loans reporting). Third, the functional form $F(\theta)$ of the correction (penalty) function that involves θ. Fourth a possibly varying parameter n of this function.[11] This implies that we need to possibly find a solution to the minimization problem

$$\min_{\alpha,\theta,F,n} \{\ LOSS(\omega(\alpha;\theta;F;n)\ ,\ \lambda \cdot \Omega(\alpha;\theta;F;n))\ \} \tag{14}.$$

Our strategy in solving (14) could not be derived formally, but rather through computation. We figured out different alternatives for the functional form of the correction $F(\theta)$ and run a cycle of computations of the RA-QRRF algorithm over (12) by varying the type of imprecision

[11] To be precise, among the parameters we should also include those of the QRRF algorithm, that is the number of trees grown, the number of seeds, the number of variables pick up for splitting at each tree node (usually called the *mtry* parameter). These parameters are not included in (14) because not relevant for the rest of the discussion.

rate, the parameters n of $F(\theta)$ and the confidence level α.

We took in consideration different forms for the correction function $F(\theta)$. In the rest of the paper we focus only on the two that have proven to obtain better results in terms of (14). The first one is

$$PI_{i,t} = \left[\frac{\widehat{Q_\alpha}(\Delta L^0_{i,s,t}/X=x_{i,t})}{(\theta_{i,t})^n}, \frac{\widehat{Q_{1-\alpha}}(\Delta L^0_{i,s,t}/X=x_{i,t})}{(\theta_{i,t})^n}\right] \quad (15).$$

The correction in (15) has a different impact depending on whether the imprecision score θ is less or greater than 1. Banks that have made many errors up to time t and have values of the score greater than 1 are penalized by the correction, in the sense that the prediction interval is restricted and it increases the probability that its reported values is flagged as an outlier. On the contrary, banks with few errors and a score less than 1 "benefit" from the correction by the enlargement of the prediction interval. In order to avoid infinite values of the limit of the prediction intervals, in the case of banks with θ equal to zero in some months we set $\theta = 1$ implying no correction.[12] The disparity of treatment between banks with θ lower or higher than 1 is amplified as the exponential parameter increase. The case $n = 0$ is equivalent to the case of no correction, that is the standard output of the QRRF.

The second penalty function we consider is

$$PI_{i,t} = [\,\widehat{Q_\alpha}(\Delta L^0_{i,s,t}/X = x_{i,t}) * (1 - \frac{(\theta_{i,t})^n}{100}),\ \widehat{Q_{1-\alpha}}(\Delta L^0_{i,s,t}/X = x_{i,t}) * (1 - \frac{(\theta_{i,t})^n}{100})] \quad (16)$$

that also represents a penalty correction increasing in θ and n. Also for (16), the case $n = 0$ can be considered as an approximation of the standard QRRF output with no correction.[13]

Figure 3 illustrates the results of the iterated computations run on the banks reported loans to the household sector for the time span December 2017 – March 2020. The left side of the figure shows the results of the computations run by employing the imprecision rate θ_l (i.e. computed only on the reported loans); the right side illustrates the results of the algorithm that

[12] Usually banks have an imprecision score equal to zero at the very first months of their reporting history. The idea is that the probability to have errors after few months is low and that a premium is more deserved by banks with a lot of reporting but few errors (that is θ less than 1).

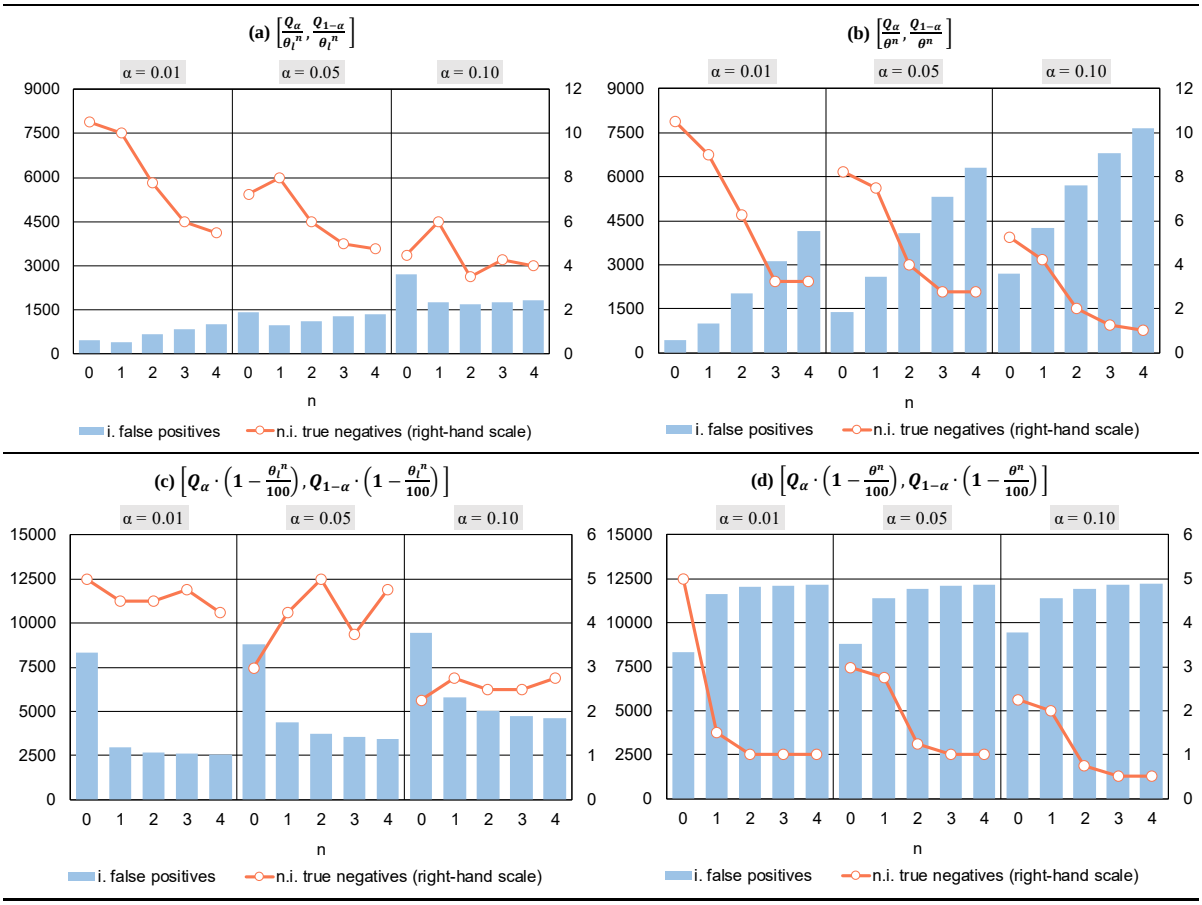
[13] Indeed, when $n = 0$ we have $[\widehat{Q_\alpha} \cdot \frac{99}{100}, \widehat{Q_{1-\alpha}} \cdot \frac{99}{100}\,] \approx [\widehat{Q_\alpha}\,, \widehat{Q_{1-\alpha}}\,]$.

employed the θ computed on the entire set of reported balance sheet items. The results obtained by employing the penalty function in (15) are shown in the upper part of the figure, while those employing the correction (16) are presented in the lower part. It is important to highlight that given the same set of parameters, two computations of the QRRF algorithm can produce different results that are nonetheless very robust.[14] For this reason, the numbers presented in the figure are obtained as average of four computations of the algorithm with the same fixed parameters (α, F, θ and n). It is evident the trade-off between the non-identified true negatives and the identified false positives; this negative relationship is much more evident as the confidence level α (measuring the width of the prediction intervals) increase.

Figure 3

Banks' loans to household: non-identified true negatives (I-type errors) and identified false positives (II-type errors) from the RA-QRRF

(average absolute values of 4 computations for each set of parameters; monthly observations from December 2017 to March 2020)



[14] This depends on the random choice of the subset of variables that the algorithm pick up for splitting at each tree node.

Moreover, it is clear that the parameter θ_l computed only on the reported loans (left-side of the figure) performs much better in terms of minimization of I and II-type errors with respect to the overall θ. Second, the preferable correction function seems to be the one in (15) despite the one in (16) proved to be better in identifying the real errors (less I-type errors). Third, it is very important to notice that in general, and especially when θ_l is employed, there is a clear gain when n increases from 0 to 1. This is a very important result, since it justifies the use of the imprecision rate θ and the elaboration of a Revisions Adjusted – Quantile Regression Random Forest.

The best results in terms of outlier identification of banks' reported loans to household are those depicted in the upper-left box of Figure 3. Reducing α does not bring a significant gain in terms of unidentified real errors but an improvement in terms of false positives reduction. At the same time, increasing the exponent n minimizes I-type errors. We can conclude that for banks' loans to the household (HH) sector the best model in terms of outlier identification is

$$PI_{i,t} = \left[\frac{\widehat{Q_{0.01}}(\Delta L^0_{i,HH,t}/X=x_{i,t})}{(\theta^l_{i,t})^4}, \frac{\widehat{Q_{0.99}}(\Delta L^0_{i,HH,t}/X=x_{i,t})}{(\theta^l_{i,t})^4}\right] \quad (17).$$

Table 3

Banks's loans to household: ratio of correct reporting and errors correctly identifyed by the algorithm
(average percentage values of 4 computations for each set of parameters; monthly observations from December 2017 to March 2020)

α	n	$\left[\frac{Q_\alpha}{\theta_l^n}, \frac{Q_{1-\alpha}}{\theta_l^n}\right]$		$\left[\frac{Q_\alpha}{\theta^n}, \frac{Q_{1-\alpha}}{\theta^n}\right]$		$\left[Q_\alpha \cdot \left(1-\frac{\theta_l^n}{100}\right), Q_{1-\alpha} \cdot \left(1-\frac{\theta_l^n}{100}\right)\right]$		$\left[Q_\alpha \cdot \left(1-\frac{\theta^n}{100}\right), Q_{1-\alpha} \cdot \left(1-\frac{\theta^n}{100}\right)\right]$	
		Id. correct reportings	**Identified errors**	**Id. correct reportings**	**Identified errors**	**Id. correct reportings**	**Identified errors**	**Id. correct reportings**	**Identified errors**
0.01	0	96.75	52.27	96.75	52.27	38.52	77.27	38.52	77.27
0.01	1	97.14	54.55	92.54	59.09	78.06	79.55	14.08	93.18
0.01	2	95.28	64.77	84.97	71.59	80.25	79.55	11.26	95.45
0.01	3	93.80	72.73	76.94	85.23	80.73	78.41	10.44	95.45
0.01	4	92.77	75.00	69.35	85.23	81.04	80.68	10.15	95.45
0.05	0	89.74	67.05	89.76	62.50	34.84	86.36	34.83	86.36
0.05	1	92.93	63.64	80.83	65.91	67.69	80.68	15.96	87.50
0.05	2	91.99	72.73	70.00	81.82	72.51	77.27	12.05	94.32
0.05	3	90.76	77.27	60.73	87.50	73.93	82.95	10.77	95.45
0.05	4	90.14	78.41	53.50	87.50	74.57	78.41	10.05	95.45
0.10	0	80.10	79.55	79.99	76.14	30.17	89.77	30.27	89.77
0.10	1	87.20	72.73	68.49	80.68	57.06	87.50	15.71	90.91
0.10	2	87.63	84.09	57.79	90.91	62.81	88.64	11.88	96.59
0.10	3	87.06	80.68	49.80	94.32	64.99	88.64	10.29	97.73
0.10	4	86.57	81.82	43.52	95.45	66.00	87.50	9.60	97.73

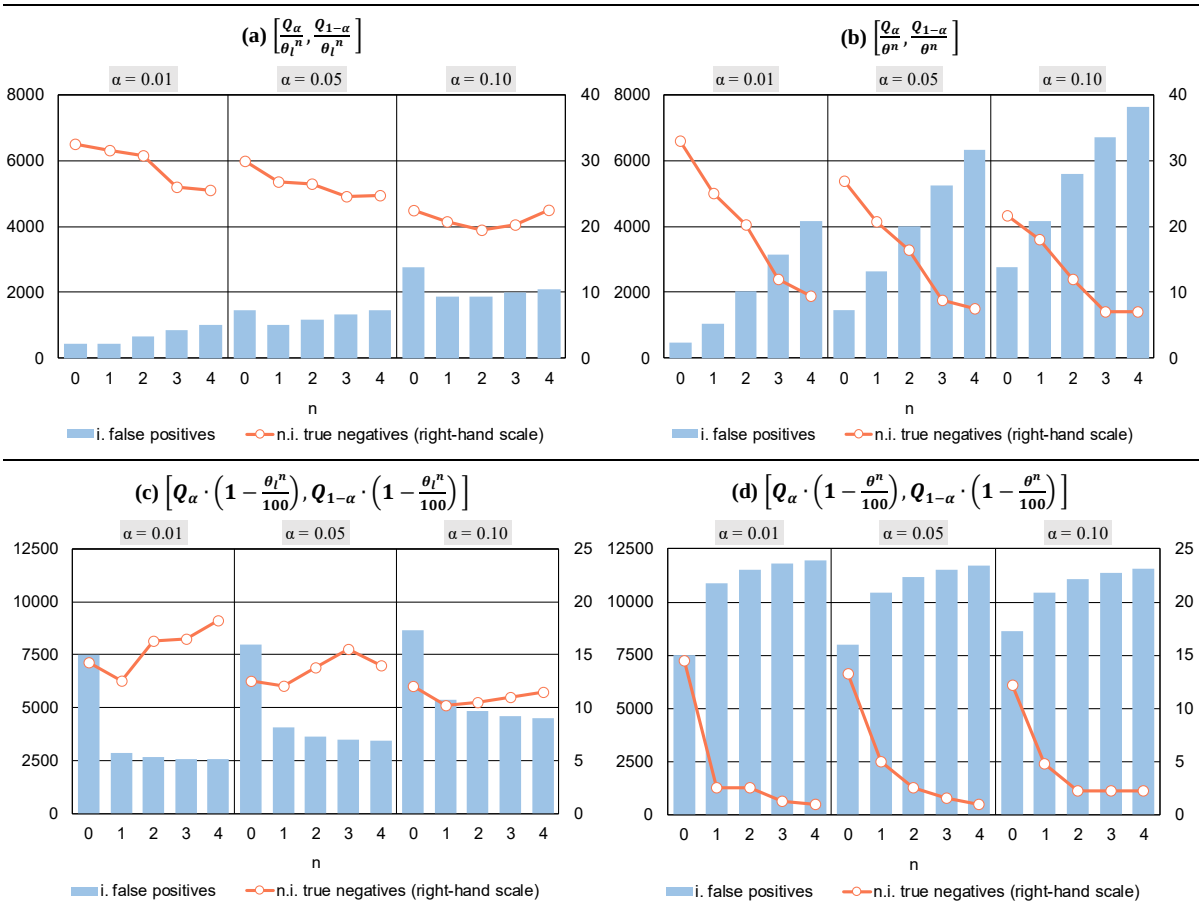
Overall, the RA-QRRF returns satisfying results in terms of identifying errors. Table 3 shows the percentage of correct and incorrect reported loans to the household sector correctly identified by the RA-QRRF procedure. In the optimized scenario, represented by (17), we correctly detect the 75 per cent of banks' errors and 92.8 per cent of the correctly reported values. Considering the parameters selection corresponding to (17), the model leads – on average - to 35 requests of clarifications to be transmitted to reporting agents every month.

Figure 4 shows the results of the same computations run on the banks reported loans to non-financial corporations. The trade-off between the two types of errors is again evident, as well as the gain of employing a Revisions Adjusted - Quantile Regression Random Forest (moving from $n = 0$ to $n = 1$). It is also evident that the automated procedure does not perform as well as in the case of the loans to the household sector.

Figure 4

Banks' loans to non-financial corporations: non-identified true negatives (I-type errors) and identified false positives (II-type errors) from the RA-QRRF

(average absolute values of 4 computations for each set of parameters; monthly observations from December 2017 to March 2020)



It can be argued that this is due to three main reasons, two of which are related to specific features of the data sample and one is presumably related to the model specification. First, in analyzing our data, we observed that among the 43 errors in loans to non-financial corporations collected since December 2017, a group of them consists of stocks repeated almost identically in the first transmission of the subsequent month (that is, banks initially reported $L^0_{i,s,t} \cong L^1_{i,s,t-1}$, implying zero growth) and then strongly revised in occasion of the second transmission. Identifying repeated values with no (or small) variations between two subsequent months as "outlier" is a tough goal for every outlier detection technique. Second, loans to non-financial corporations show a much higher variance with respect to loans to households. Bigger banks may grant larger amount loans to firms and we often observe huge amounts both in absolute values and in terms of variations. In this case, if a wrongly reported value happen to be classified in a leaf with some of these very big observations, and therefore in its same estimated distribution, it is difficult for the wrong observation to "fall" outside the α-th quantiles (i.e. the limit of the prediction intervals) of this distribution.

Finally, the last reason stems from the way we decided to specify the model in (8). Loans to non-financial corporations are probably driven by demand factors (GDP growth, firms' confidence, etc..) more than the loans to households. Fully adopting the supply theory, if on the one hand it is necessary to develop a procedure that can be run in real-time during the day of production, on the other hand it obviously penalizes the prediction ability of the algorithm. This is due to the fact that very few economic indicators referred to a given reference date are published earlier than the time of BSI production which, on average, takes place on the 20th day of the following month.

As for loans to households, the best results for loans to non-financial corporations (NFC) are those presented in the upper-left part of Figure 4. As reported in Table 4, by selecting the model

$$PI_{i,t} = \left[\frac{\widehat{Q_{0.01}}\left(\Delta L^0_{i,NFC,t}/X=x_{i,t}\right)}{(\theta^l_{i,t})^4}, \frac{\widehat{Q_{0.99}}(\Delta L^0_{i,NFC,t}/X=x_{i,t})}{(\theta^l_{i,t})^4}\right] \tag{18}$$

we are able to identify the 93.3 per cent of the correctly reported values and the 40 per cent of the wrong reported data, implying an average of 35 requests of clarifications to be transmitted to banks in each month.

Table 4

Banks's loans to non-financial corporations: ratio of correct reporting and errors correctly identifyed by the algorithm

(average percentage values of 4 computations for each set of parameters; monthly observations from December 2017 to March 2020)

α	n	$\left[\frac{Q_\alpha}{\theta_l^n}, \frac{Q_{1-\alpha}}{\theta_l^n}\right]$		$\left[\frac{Q_\alpha}{\theta^n}, \frac{Q_{1-\alpha}}{\theta^n}\right]$		$\left[Q_\alpha \cdot \left(1-\frac{\theta_l^n}{100}\right), Q_{1-\alpha} \cdot \left(1-\frac{\theta_l^n}{100}\right)\right]$		$\left[Q_\alpha \cdot \left(1-\frac{\theta^n}{100}\right), Q_{1-\alpha} \cdot \left(1-\frac{\theta^n}{100}\right)\right]$	
		Id. correct reportings	Identified errors	Id. correct reportings	Identified errors	Id. correct reportings	Identified errors	Id. correct reportings	Identified errors
0.01	0	96.66	24.42	96.60	23.26	44.75	66.86	44.70	66.28
0.01	1	96.84	26.74	92.19	41.86	78.70	70.93	19.38	94.19
0.01	2	95.21	28.49	84.91	52.91	80.32	62.21	14.74	94.19
0.01	3	93.73	39.53	76.75	72.09	80.80	61.63	12.66	97.09
0.01	4	92.64	40.70	69.28	77.91	80.95	57.56	11.58	97.67
0.05	0	89.28	30.23	89.27	37.21	41.06	70.93	40.97	69.19
0.05	1	92.42	37.79	80.61	51.74	69.79	72.09	22.67	88.37
0.05	2	91.46	38.37	70.33	61.63	73.13	68.02	17.55	94.19
0.05	3	90.23	43.02	61.19	79.65	74.22	63.95	14.86	96.51
0.05	4	89.30	42.44	53.31	82.56	74.64	67.44	13.24	97.67
0.10	0	79.58	47.67	79.60	49.42	36.12	72.09	36.09	71.51
0.10	1	86.25	51.74	69.12	58.14	60.15	76.16	22.75	88.95
0.10	2	86.14	54.65	58.68	72.09	64.35	75.58	18.27	94.77
0.10	3	85.18	52.91	50.42	83.72	66.04	74.42	15.89	94.77
0.10	4	84.49	47.67	43.70	83.72	66.61	73.26	14.47	94.77

In the next paragraph we carry out a robustness analysis by estimating a probit and a logit model in order to compare the efficacy and the predictive power of our procedure with respect to more standard (and more computational time-consuming) techniques.

5. Robustness analysis: probit and logit models

In this paragraph, we estimate alternative models that aim at predicting the probability that - conditional on the available information – the value reported by a bank could be wrong or not, i.e. probit and logit models. The specific goals of this robustness analysis are, on the one hand, to evaluate the predictive ability of our RA-QRRF with respect to alternative models and, on the other hand, to confirm the goodness of our selected predictors. Hence, we estimate

$$Prob(y_{i,s,t} = 1/x_{i,s,t}) = \Phi(\beta_1 CAP_{i,t} + \beta_2 INTB_{i,t} + \beta_3 RETAIL_{i,t} + \beta_4 PBONDS_{i,t} +$$

$$\beta_5 BADL_{i,s,t} + \beta_6 TASS_{i,t} + \beta_7 SHARE_{i,s,t-1} + \beta_8 SIZE_i + \beta_9 (\Delta L/L)^1_{i,s,t-1} + \beta_{10}\theta_{i,t}) \quad (19)$$

where $y_{i,s,t} = 1$ indicates that the value of loans to sector s reported by bank i at the end of month t is wrong at the first data transmission and Φ is the cumulative probability distribution of a standard normal in the case of the probit model and of a logistic distribution in the case of the logit one. It is worth noting that we also include the imprecision score θ among the independent variables of the model for which we want to evaluate its predictive significance and make a comparison between its different versions (i.e., the score computed on the errors on the overall balance sheet items or on the errors when focusing on loans items).

We estimate both a *pooled* probit/logit model where all the sample periods and observations are put together as if a cross-section specification applies and a panel probit/logit model. In addition, we do not perform any out-of-sample prediction, but include all available observation when fitting the model in order to estimate predicted probabilities of a mistake. As said, the goal of this analysis is indeed to have confirmation on the goodness of our predictors and on the predictive ability of our RA-QRRF with respect to alternative models: in doing this, we are not interested in "penalizing" the alternative models but let them exploit all the available information they can use.

Table 5 shows the estimation results of the probit model in which we consider the "failures" in banks' reporting of loans to the household sector. The correspondent estimations of the logit models, whose results are very similar to those of the probit, are reported in Table 11 in the Appendix. Results shown in Table 5 are robust in the various specifications. In column (1) and (2) we report the estimation of, respectively, the pooled and the panel probit where we included the imprecision rate θ_l estimated only on loan-related observations as a regressor. In column (3) and (4) we show the same models but considering the overall score θ, estimated on the full sample of balance sheet items, as independent variables.

Estimation results confirm the goodness of the selected balance sheet variables that are all significant except the lagged growth of loans. The signs of the coefficients seem to indicate a negative relationship between bank's probability to misreport data (or, in general, to be in presence of an outlier) and its size and its degree of activity (different from that of granting loans). The bigger the bank, as measured by capital, total asset and the categorical variable of size, the lower the probability to report errors.

Table 5

Probit estimation on banks reporting of loans to the household sector

(monthly observations from December 2017 to March 2020)

	(1) Pooled Probit (θ_l)	(2) Panel Probit (θ_l)	(3) Pooled Probit (θ)	(4) Panel Probit (θ)
Capital	-0.0114***	-0.0234***	-0.00955***	-0.0229***
	(0.000)	(0.006)	(0.000)	(0.008)
Interbank funding	-0.0261***	-0.0400***	-0.0244***	-0.0394***
	(0.000)	(0.000)	(0.000)	(0.000)
Retail funding	-0.0276***	-0.0575***	-0.0270***	-0.0573***
	(0.000)	(0.000)	(0.000)	(0.000)
Public bonds	-0.0237**	-0.0453**	-0.0257***	-0.0449*
	(0.011)	(0.048)	(0.006)	(0.051)
HH bad loans	0.123**	0.153	0.118**	0.156
	(0.010)	(0.131)	(0.012)	(0.124)
Total assets	-0.0000135***	-0.0000222**	-0.0000133***	-0.0000222**
	(0.000)	(0.038)	(0.000)	(0.038)
HH share of loans	0.483***	0.886**	0.472***	0.887**
	(0.000)	(0.014)	(0.000)	(0.013)
Bank's size	-0.391***	-0.241***	-0.365***	-0.232***
	(0.000)	(0.002)	(0.000)	(0.003)
HH growth of loans$_{(t-1)}$	-0.0000171	-0.00245	-0.0000154	-0.00242
	(0.964)	(0.873)	(0.962)	(0.872)
θ_l	0.0570***	-0.00128		
	(0.004)	(0.968)		
θ			-0.00393	-0.0394
			(0.921)	(0.531)
Observations	13569	13569	13569	13569

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Similarly, the higher the degree of activity in the interbank funding, retail funding and public bonds market, the lower the probability of wrong data. A higher probability is associated to higher levels of the variables related to the granting loans activity, that is, the level of bad loans in the balance sheet (not significant in the panel specification) and the market share of the total loans to that sector. Consistently, in the pooled specification, higher levels of the imprecision score θ_l are associated to a higher probability of errors. It is the only relevant result, since in columns (2), (3) and (4) the imprecision rates turned out to be not significant.[15]

[15] Given that the computed θ_l and θ are estimated random variables, the significance levels could suffer from incorrect estimated standard errors. For this reason, we also estimated the model in (1) by employing bootstrap resampling on the standard errors. Results are basically identical to those reported in Table 5.

In Table 6 we present the summary statistics and some selected predicted probabilities of the estimated probit model for the reported values of banks' loans to the household sector (see Table 12 in the Appendix for the statistics on the predicted probabilities of the logit specification). All the four estimations return, on average, very small values of the predicted probabilities to make errors in reporting, with the means of the predicted probabilities varying between 0.004 and 0.009 among the four models. We then choose to report in Table 6 the probabilities predicted by the models higher than 0.3. With respect to the entire set of predicted values, these values signal some significant probability that the data can be an error. As it is evident, despite the estimation is run entirely in-the-sample, the estimated models are able to correctly identify only 1 actual errors out of the 22 outliers and reclassifications we observed in the sample period (4.5 per cent of the errors). In this respect, our RA– Quantile Regression Random Forest algorithm has proven to perform much better than these alternative models.

Table 6

Banks' loans to household: predicted probabilities of the probit model

(predicted probabilities lower than 0.3 are omitted)

	(1) Pooled Probit (θ_l)	(2) Panel Probit (θ_l)	(3) Pooled Probit (θ)	(4) Panel Probit (θ)
Mean	0.004	0.009	0.004	0.009
St. Dev.	0.021	0.039	0.021	0.039
Min	0.000	0.000	0.000	0.000
Max	0.678	0.817	0.663	0.810
Errors		Predicted probabilities		
0	0.447	0.356	0.367	0.344
0	0.436	0.342	0.356	0.331
1	**0.678**	**0.817**	**0.663**	**0.810**
0	0.468	0.457	0.468	0.447
0	0.467	0.456	0.467	0.447
0	-	0.401	-	0.386
0	-	0.535	-	0.511
0	-	0.536	-	0.512
0	0.402	-	-	-
0	0.398	0.394	0.402	0.384
0	0.403	0.399	0.407	0.389
0	0.519	0.517	0.512	0.509
0	0.503	0.497	0.496	0.488
0	0.503	0.497	0.496	0.489
0	0.440	0.403	0.427	0.393
0	0.447	0.356	0.367	0.344
0	0.436	0.342	0.356	0.331

Table 7 shows the results of the estimation for the reported loans to non-financial corporations.[16] Results are consistent with those of the estimation for the reported loans to households. Size variables (capital, total asset and bank's size) have a negative relationship with the probability of reporting an error, as well as the variables measuring the degree of activity like the outstanding amount of interbank funding, retail funding and public bonds. Differently from the loans to household sector, in Table 7 the lagged growth of loans to the non-financial corporations sector shows a negative relationship with the predicted probabilities. The share of the market has a positive correlation in the pooled specification, while bad loans have a non-significant effect.

Table 7

Probit estimation on banks reporting of loans to the non-financial corporations sector

(monthly observations from December 2017 to March 2020)

	(1) Pooled Probit (θ_l)	(2) Panel Probit (θ_l)	(3) Pooled Probit (θ)	(4) Panel Probit (θ)
Capital	-0.00885***	-0.0179**	-0.00744***	-0.0168**
	(0.000)	(0.013)	(0.001)	(0.021)
Interbank funding	-0.0225***	-0.0434***	-0.0202***	-0.0414***
	(0.000)	(0.000)	(0.000)	(0.000)
Retail funding	-0.0256***	-0.0578***	-0.0238***	-0.0572***
	(0.000)	(0.000)	(0.000)	(0.000)
Public bonds	-0.0170***	-0.0389**	-0.0176***	-0.0375**
	(0.002)	(0.020)	(0.002)	(0.024)
NFC bad loans	0.00990	0.0256	0.00556	0.0280
	(0.774)	(0.747)	(0.877)	(0.724)
Total assets	-0.00000459***	0.000000197	-0.00000469***	0.000000103
	(0.005)	(0.963)	(0.004)	(0.981)
NFC share of loans	0.208***	0.168	0.215***	0.172
	(0.000)	(0.182)	(0.000)	(0.169)
Bank's size	-0.266***	-0.0482	-0.228***	-0.0389
	(0.000)	(0.488)	(0.000)	(0.575)
NFC growth of loans$_{(t-1)}$	-0.560**	-1.062**	-0.653**	-1.135**
	(0.033)	(0.042)	(0.015)	(0.028)
θ_l	0.0177	0.00725		
	(0.370)	(0.833)		
θ			-0.114**	-0.0696
			(0.018)	(0.250)
Observations	13569	13569	13569	13569

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[16] The corresponding logit estimations and predicted probabilities are illustrated in Table 13 and Table 14 in the Appendix.

Finally, different from the households estimation, the imprecision score θ_l computed on the errors in reported loans, turns out to be non-significant, whereas – inconsistently – higher levels of the overall imprecision rate θ show a negative relation with the probability of reporting an outlier.

Concerning the predicted probabilities of the estimated probit for the non-financial corporations sector illustrated in Table 8, they also return, on average, small values. The mean of the predicted probabilities to report an outlier varies between 0.007 and 0.0116 among the four models. Also in this case, the predicted probabilities of the four estimates models higher than 0.3 detect only 1 *true* outlier out of 43, that is the 2.2 per cent of the actual errors.

Table 8

Banks' loans to non-financial corporations: predicted probabilities of the probit model

(predicted probabilities lower than 0.4 are omitted)

	(1) Pooled Probit (θ_l)	(2) Panel Probit (θ_l)	(3) Pooled Probit (θ)	(4) Panel Probit (θ)
Mean	0.007	0.016	0.007	0.015
St. Dev.	0.028	0.056	0.027	0.055
Min	0.000	0.000	0.000	0.000
Max	0.625	0.698	0.569	0.680
Errors		Predicted probabilities		
0	-	0.528	-	0.527
0	-	0.513	-	0.512
0	-	0.568	-	0.573
1	**0.625**	**0.698**	**0.569**	**0.680**
0	0.475	0.467	0.422	0.450
0	0.471	0.463	0.420	0.447
0	-	0.531	-	0.537
0	-	0.551	-	0.556
0	-	0.579	0.316	0.584
0	-	0.530	-	0.525
0	-	0.569	-	0.557
0	-	0.565	-	0.552
0	-	0.523	-	0.512
0	-	0.553	-	0.544
0	0.514	0.515	0.474	0.503
0	0.488	0.486	0.441	0.471
0	0.499	0.497	0.454	0.482

The robustness analysis run in this paragraph brings us three main results. First, the estimated probit and logit models confirm the goodness of the selected independent variables

which show, in general, significant coefficients associated to the probability that a given bank's reported values on loans be an error. Second, the analysis confirms the better predictive performance of the imprecision rate θ_l estimated on the errors in loan-related items reporting with respect to the imprecision rate θ derived on the entire set of reported items. Finally, most importantly, the robustness analysis highlights the better performance of the Revisions Adjusted - Quantile Regression Random Forest we developed in detecting outliers in BSI statistics with respect to alternative models.

6. Conclusion

Ensuring and disseminating high quality data is crucial in order to adequately support the monetary authority's supervision and decision-making process for monetary policy and financial stability. Over the recent years, the increasing collection of granular banking data made possible by more performing IT infrastructures has determined an increasingly challenging activity of error- and outlier-detection in bank's supervisory reports.

In this paper we develop and test an automated *machine learning* procedure able to identify potential errors in bank's supervisory reports on loans to the private sector employed in the Bank of Italy's Balance Sheet Information (BSI) production process. In particular, we develop a Revisions Adjusted – Quantile Regression Random Forest algorithm in which the predicted acceptance regions of each monthly reported value are calibrated through an individual *imprecision score.* This monthly score provides a measure of each bank's likelihood of making errors and is estimated by employing the entire history of its errors and revisions on BSI items collected by the Bank of Italy.

The algorithm we develop has two main advantages. First, it processes and uses the whole huge amount of relevant information at our disposal, that consists of millions of data (also including millions of revisions made by banks in the past). Second, by exploiting exclusively banks' balance sheet variables as explanatory variables, our procedure is able to identify outliers in real time, that is, during the day of statistical production. This real-time automated algorithm also improves the outlier detection process currently employed by the Bank of Italy, which is based on quite simple techniques that leave a large role to the analyst's expertise, hence implying a relatively time-consuming process that has to be carried out in a half working day of production.

Focusing on the monthly BSI statistics from December 2017 to March 2020, our results are very satisfying as far as loans to households are concerned: by computationally optimizing the selection of the algorithm parameters, we are able to identify up to 75 per cent of banks' errors and 93 per cent of correctly reported values (i.e. the procedure does not signal them as false positives). Concerning loans to non-financial corporations, our results are not as good, since we are able to identify up to 93.3 per cent of banks' correctly reported values in the period, but only 40 per cent of the errors. As we argue, this is partly due to the specific characteristics of the data that would probably make it difficult to identify such errors for any outlier-detection procedure.

Finally, as a robustness analysis, we estimate alternative models, in particular a probit and a logit model. The analysis highlights a worse performance of these models with respect to the RA-QRRF approach in the outlier detection process, thus corroborating the latter.

References

Adrian, T. and Shin H. S. (2010), "Liquidity and Leverage", *Journal of Financial Intermediation*, 19(3), pp. 418-437.

Affinito M., Albareto G. and Santioni R. (2016), "Purchases of sovereign debt securities by Italian banks during the crisis: the role of balance-sheet conditions", Bank of Italy, Occasional Papers, No. 330.

Altunbas Y., Manganelli S. and Marqués-Ibáñez D. (2011), "Bank risk during the financial crisis: do business models matter?" European Central Bank, Working Paper Series, No. 1394.

Ardizzi G., Emiliozzi S., Marcucci J. and Monteforte L. (2019), "News and consumer card payments", Bank of Italy, Working Papers, No. 1233.

Baker S. R., Bloom N. and Davis S.J. (2016), "Measuring economic policy uncertainty", *The Quarterly Journal of Economics*, 131(4), pp. 1593-1636.

Bank for International Settlements (2019), "The use of big data analytics and artificial intelligence in central banking", IFC Bulletin, No. 50.

Bank of Italy (2019), "Annual Report 2018".

Bofondi M., Carpinelli L. and Sette E. (2013), "Credit supply during a sovereign debt crisis", Bank of Italy, Working Papers, No. 909.

Bonaccorsi di Patti E. and Sette E. (2012), "Bank balance sheets and the transmission of financial shocks to borrowers: evidence from the 2007-2008 crisis", Bank of Italy, Working Papers, No. 848.

Breiman L. (2001), "Random forests", *Machine Learning*, 45, pp.5–32.

Cagala T. (2017), "Improving Data Quality and Closing Data Gaps with Machine Learning", IFC Bulletin, No. 46.

Carboni A. and Moro. A. (2018), "Imputation techniques for the nationality of foreign shareholders in Italian firms", IFC Bulletins chapter in Bank for International Settlements, External sector statistics: current issues and new challenges, vol. 48.

Chakraborty C. and Joseph A. (2017), "Machine learning at central banks", Bank of England Staff Working Paper, No. 674.

De Bonis R., Nuzzo G. and Stacchini M. (2014), "Andamenti e determinanti del credito nell'area dell'euro" in "Le banche e il credito alle imprese durante la crisi" by A. Zazzaro, Il Mulino.

Cingano F., Manaresi F. and Sette E. (2016), "Does Credit Crunch Investment Down? New Evidence on the Real Effects of the Bank-Lending Channel", *The Review of Financial Studies,* vol. 29, n. 10, pp. 2737-2773.

Correa R., Garud K., Londonoy J.M. and Mislang N. (2017), "Sentiment in central banks' financial stability reports", International Finance Discussion Papers, Board of Governors of the Federal Reserve System, No. 1203.

Farné M. and Vouldis A. (2017), "Business models of the banks in the euro area", European Central Bank, Working Paper Series, No. 2070.

Gaure S. (2013), "lfe: Linear Group Fixed Effects", *The R Journal,* vol. 5(2), pp. 104-116.

Giudice O., Massaro P. and Vannini I. (2020), "Institutional sector classifier, a machine learning approach", Banca d'Italia, Occasional Papers, No. 548.

Greene W. H. (2002), "Econometric Analysis", Mac Millan.

Hastie T., Tibshirani R. and Friedman J. (2001), "The Elements of Statistical Learning". Springer Series in Statistics.

Infante L., Piermattei S., Santioni R. and Sorvillo B. (2020), "Diversifying away risks through derivatives: an analysis of the Italian banking system", *Economia Politica: Journal of Analytical and Institutional Economics,* vol. 37(2), pp. 621-657.

James G., Witten D., Hastie T. and Tibshirani R. (2013), "An Introduction to Statistical Learning", Springer Series in Statistics.

Jimenez G., Ongena S., Peydrò J. L. and Saurina J. (2012), "Credit Supply and Monetary Policy: Identifying the Bank Balance-Sheet Channel with Loan Applications", *American Economic Review,* n. 102, pp. 2301-2326.

Johnston J. and Di Nardo J. (1996), "Econometric Methods", McGraw – Hill.

Meinshausen N. (2006), “Quantile Regression Forest”, *Journal of Machine Learning Research*, Volume 7, pp. 983-999.

Moody J. (2012), “Forecasting the Economy with Neural Nets: A Survey of Challenges and Solutions”, in Montavon G., Orr G.B. and Müller K.R., Neural Networks: Tricks of the Trade, *Lecture Notes in Computer Science*, vol. 7700, pp. 343-367.

Moscatelli M., Narizzano S., Parlapiano F. and Viggiano G. (2019), “Corporate default forecasting with machine learning”, Bank of Italy, Working Papers, No. 1256.

Salzano M. (1999), “Neural Networks for Economic Forecasting”, in Marinaro M. and Tagliaferri R., Neural Nets WIRN, *Perspectives in Neural Computing*, pp 391-407.

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.

Appendix

Table 9

Summary statistics of the variables used in the analysis

	Mean	St. Dev.	Min	Max
HH ΔLoans	6.07	376.74	-34642.99	16230.56
NFC ΔLoans	5.75	501.49	-36635.00	34132.13
Capital/Asset	12.19	11.63	-180.01	106.83
Interbank funding/(Asset-Capital)	23.05	23.92	0.00	101.34
Retail funding/(Asset-Capital)	58.00	27.07	1.24	99.22
Public bonds/Asset	21.79	14.34	2.68	71.69
HH Bad loans/HH Loans $_{(t-1)}$	0.57	0.91	0.00	10.00
NFC Bad loans/NFC Loans $_{(t-1)}$	1.02	1.18	0.00	10.00
Total asset	7723.94	40554.81	0.00	618707.00
Bank's HH share of loans $_{(t-1)}$	0.21	1.04	0.00	18.76
Bank's NFC share of loans $_{(t-1)}$	0.21	1.04	0.00	19.73
Bank's size	3.44	3.65	0.00	99.00
HH growth of loans $_{(t-1)}$	8.09	938.35	-1.00	109303.80
NFC growth of loans $_{(t-1)}$	0.59	63.42	-1.00	7384.34
θ	1.64	1.25	0.00	27.00
θ_l	0.84	1.40	0.01	28.93

Table 10

Correlation matrix of the variables used in the analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) HH ΔLoans	1													
(2) NFC ΔLoans	0.76	1												
(3) Capital	0.01	0.00	1											
(4) Interbank funding	0.02	0.01	-0.15	1										
(5) Retail funding	0.00	0.00	0.11	-0.79	1									
(6) Public bonds	-0.01	-0.01	0.11	-0.36	0.49	1								
(7) HH Bad loans	0.00	0.00	0.22	0.09	-0.10	-0.03	1							
(8) NFC Bad loans	0.01	0.00	0.04	-0.07	0.07	0.04	0.53	1						
(9) Total asset	0.14	0.16	-0.03	0.04	-0.12	-0.14	0.00	-0.02	1					
(10) HH share of loans	0.13	0.14	-0.01	0.03	-0.10	-0.16	-0.01	0.00	0.84	1				
(11) NFC share of loans	0.11	0.12	0.00	0.05	-0.12	-0.16	0.02	0.01	0.86	0.93	1			
(12) Bank's size	-0.01	-0.01	0.06	-0.09	0.08	0.06	0.08	0.11	-0.09	-0.08	-0.09	1		
(13) HH growth of loans	0.00	0.00	-0.01	-0.01	0.01	-0.01	-0.01	-0.01	0.00	0.00	0.00	-0.01	1	
(14) NFC growth of loans	0.00	0.00	-0.02	-0.01	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	1

Table 11

Logit estimation on banks reporting of loans to the household sector

(monthly observations from December 2017 to March 2020)

	(1) Pooled Logit (θ_l)	(2) Panel Logit (θ_l)	(3) Pooled Logit (θ)	(4) Panel Logit (θ)
Capital	-0.0252***	-0.0459***	-0.0214***	-0.0444***
	(0.000)	(0.003)	(0.001)	(0.005)
Interbank funding	-0.0565***	-0.0861***	-0.0537***	-0.0849***
	(0.000)	(0.000)	(0.000)	(0.000)
Retail funding	-0.0596***	-0.114***	-0.0576***	-0.117***
	(0.000)	(0.000)	(0.000)	(0.000)
Public bonds	-0.0640**	-0.0978**	-0.0690**	-0.0982*
	(0.032)	(0.050)	(0.022)	(0.052)
HH bad loans	0.265**	0.316	0.298**	0.321
	(0.034)	(0.133)	(0.019)	(0.130)
Total assets	-0.0000328***	-0.0000460**	-0.0000323***	-0.0000466**
	(0.000)	(0.034)	(0.000)	(0.034)
HH share of loans	1.206***	1.835***	1.183***	1.865***
	(0.000)	(0.010)	(0.000)	(0.010)
Bank's size	-0.824***	-0.475***	-0.771***	-0.449***
	(0.000)	(0.004)	(0.000)	(0.006)
HH growth of loans$_{(t-1)}$	-0.0000483	-0.00418	-0.0000458	-0.00430
	(0.970)	(0.891)	(0.970)	(0.884)
θ_l	0.117**	0.0444		
	(0.012)	(0.529)		
θ			-0.0331	-0.0415
			(0.789)	(0.752)
Observations	13569	13569	13569	13569

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12

Banks' loans to household: predicted probabilities of the logit model

(predicted probabilities lower than 0.3 are omitted)

	(1) Pooled Logit (θ_l)	(2) Panel Logit (θ_l)	(3) Pooled Logit (θ)	(4) Panel Logit (θ)
Mean	0.003	0.009	0.003	0.009
St. Dev.	0.018	0.037	0.017	0.038
Min	0.000	0.000	0.000	0.000
Max	0.804	0.838	0.789	0.834
Errors		Predicted probabilities		
0	0.415	0.367	0.307	0.342
0	0.400	0.352	-	0.328
0	-	0.341	-	0.353
0	-	0.367	-	0.376
1	**0.804**	**0.838**	**0.789**	**0.834**
0	0.458	0.459	0.450	0.455
0	0.456	0.458	0.449	0.454
0	-	0.453	-	0.424
0	-	0.454	-	0.425
0	0.363	0.390	0.357	0.387
0	0.369	0.394	0.365	0.392
0	-	0.349	-	0.354
0	-	0.518	-	0.513
0	0.504	0.498	0.489	0.492
0	0.504	0.498	0.489	0.493
0	0.415	0.408	0.390	0.401
0	0.415	0.367	0.307	0.342

Table 13

Logit estimation on banks reporting of loans to the non-financial corporations sector

(monthly observations from December 2017 to March 2020)

	(1) Pooled Logit (θ_l)	(2) Panel Logit (θ_l)	(3) Pooled Logit (θ)	(4) Panel Logit (θ)
Capital	-0.0177***	-0.0365***	-0.0144***	-0.0341**
	(0.000)	(0.010)	(0.007)	(0.017)
Interbank funding	-0.0438***	-0.0845***	-0.0390***	-0.0808***
	(0.000)	(0.000)	(0.000)	(0.000)
Retail funding	-0.0530***	-0.113***	-0.0489***	-0.113***
	(0.000)	(0.000)	(0.000)	(0.000)
Public bonds	-0.0478***	-0.0765**	-0.0487***	-0.0746**
	(0.004)	(0.025)	(0.005)	(0.028)
NFC bad loans	0.0389	0.0373	0.0390	0.0440
	(0.658)	(0.819)	(0.669)	(0.790)
Total assets	-0.00000930**	0.000000342	-0.00000922**	0.000000243
	(0.015)	(0.968)	(0.014)	(0.977)
NFC share of loans	0.457***	0.346	0.461***	0.352
	(0.000)	(0.168)	(0.000)	(0.160)
Bank's size	-0.634***	-0.134	-0.541***	-0.111
	(0.000)	(0.359)	(0.000)	(0.449)
NFC growth of loans$_{(t-1)}$	-1.133*	-2.050**	-1.257**	-2.232**
	(0.059)	(0.043)	(0.041)	(0.026)
θ_l	0.0566	0.0289		
	(0.220)	(0.676)		
θ			-0.255*	-0.118
			(0.078)	(0.353)
Observations	13569	13569	13569	13569

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 14

Banks' loans to non-financial corporations: predicted probabilities of the logit model

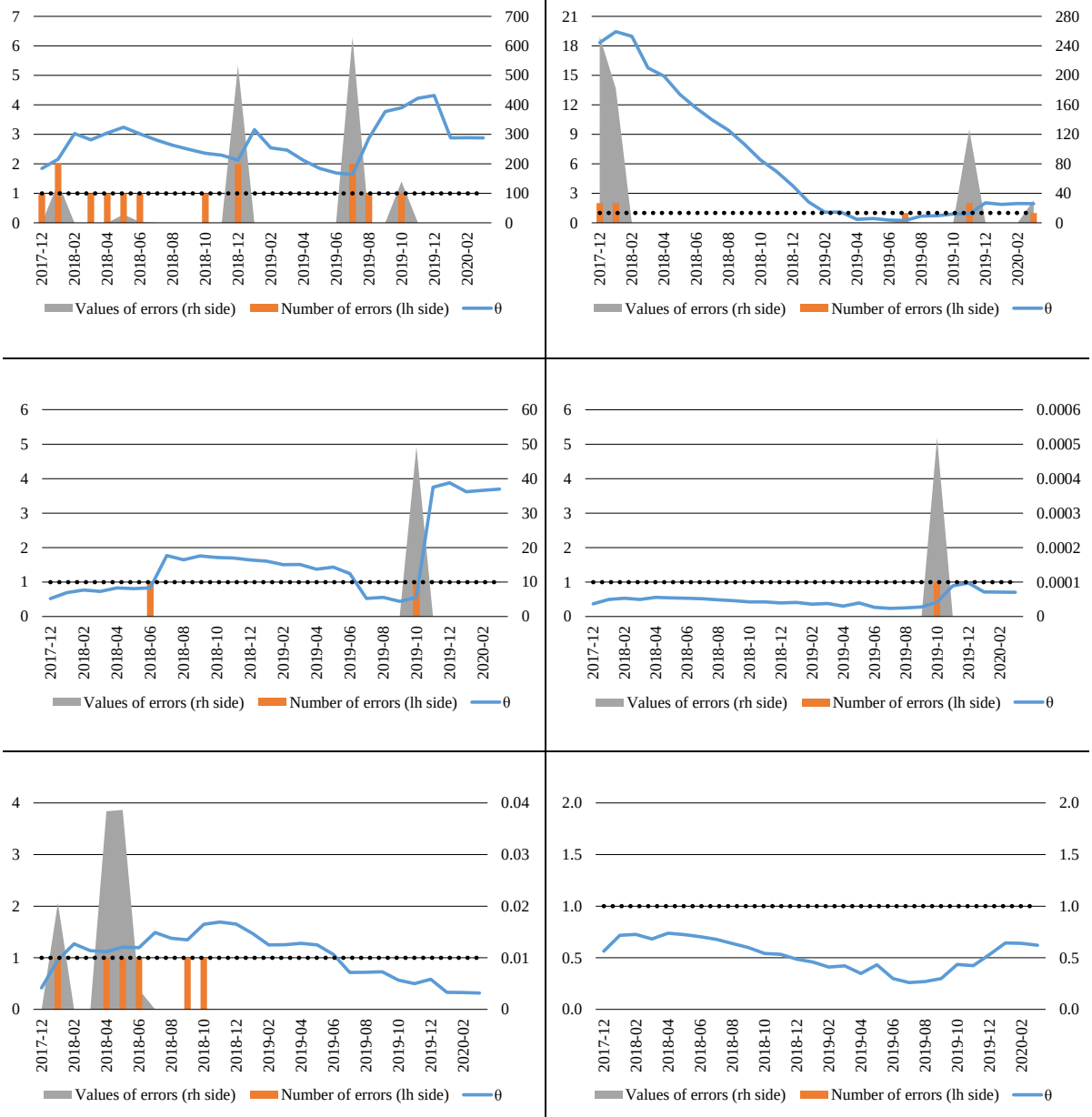
(predicted probabilities lower than 0.4 are omitted)

	(1) Pooled Logit (θ_l)	(2) Panel Logit (θ_l)	(3) Pooled Logit (θ)	(4) Panel Logit (θ)
Mean	0.007	0.016	0.007	0.015
St. Dev.	0.028	0.056	0.027	0.055
Min	0.000	0.000	0.000	0.000
Max	0.625	0.698	0.569	0.680
Errors		Predicted probabilities		
0	-	0.528	-	0.519
0	-	0.507	-	0.505
0	-	0.554	-	0.567
1	**0.685**	**0.711**	**0.607**	**0.695**
0	-	0.515	-	0.526
0	-	0.534	-	0.547
0	-	0.562	-	0.575
0	-	0.513	-	0.512
0	-	0.551	-	0.545
0	-	0.548	-	0.542
0	-	0.508	-	0.502
0	-	0.536	-	0.534
0	0.520	0.516	0.462	0.504
0	-	0.528	-	0.519
0	-	0.507	-	0.505
0	-	0.554	-	0.567
1	0.685	0.711	0.607	0.695

Figure 5

Some examples of the temporal behavior of the imprecision score θ_l and its relationship with bank's errors in reporting

(monthly scores, number and absolute value of errors from December 2017 to March 2020)



Note: the graphs illustrate six examples of the temporal relationship of the monthly score θ_l with banks' errors in reporting. We decided to illustrate only the score on loans θ_l because it was easier to represent graphically. The blue line represent the temporal behavior of the score (l.h. side) for six chosen banks. The orange bars represent the number of errors in the reporting (l.h. side; the value is 2 if the bank misreported both the loans to household and to non-financial corporations and 1 if the banks misreported one of the two). The grey area represents the absolute value of the errors in unit of euro (r.h side). We insert the horizontal line in correspondence of the unit because, as explained in the paper, in the case of the actual selected penalty function banks with a score higher than one are penalized and vice versa. Despite the temporal path of the score reflects all revisions made by banks in every sub-items, and not only errors in the total amount of loans (as explained in Section 3), it is nonetheless graphically evident that the score is quite able to catch the reporting behavior of banks in a satisfying way.