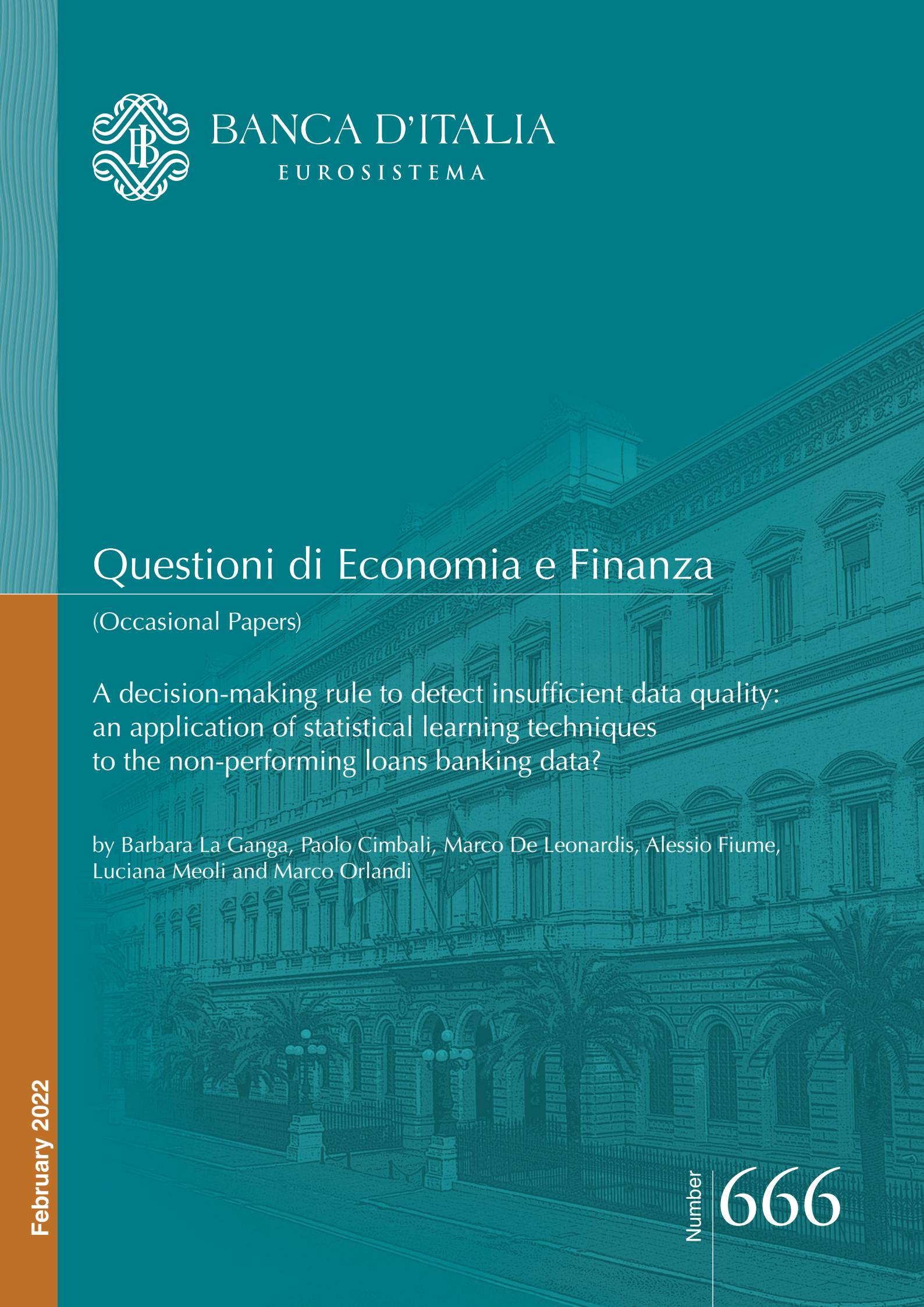
BANCA D'ITALIA
EUROSISTEMA



Questioni di Economia e Finanza

(Occasional Papers)

A decision-making rule to detect insufficient data quality: an application of statistical learning techniques to the non-performing loans banking data?

by Barbara La Ganga, Paolo Cimbali, Marco De Leonardis, Alessio Fiume, Luciana Meoli and Marco Orlandi

February 2022

Number 666

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

A decision-making rule to detect insufficient data quality: an application of statistical learning techniques to the non-performing loans banking data?

by Barbara La Ganga, Paolo Cimbali, Marco De Leonardis, Alessio Fiume, Luciana Meoli and Marco Orlandi

Number 666 – February 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

A DECISION-MAKING RULE TO DETECT INSUFFICIENT DATA QUALITY: AN APPLICATION OF STATISTICAL LEARNING TECHNIQUES TO THE NON-PERFORMING LOANS BANKING DATA

by Barbara La Ganga[*], Paolo Cimbali[*], Marco De Leonardis[‡], Alessio Fiume[*]
Luciana Meoli[§] and Marco Orlandi[*]

Abstract

The paper presents a decision-making rule, based on statistical learning techniques, to evaluate and monitor the overall quality of the granular dataset referring to the Non-Performing Loans data collection carried out by the Bank of Italy. The datasets submitted by the reporting agents must display a sufficiently high level of quality before their release to users. The study defines a decision-making rule to distinguish the cases where the corrections applied to the original dataset improve its overall quality from those where the revisions (unexpectedly) make it worse. The decision-making rule is based on a new synthetic data quality indicator, based on past evidence accumulated on data quality management activity, which makes possible the assessment and monitoring of the overall quality of the Non-Performing Loans dataset. The proposed indicator takes into account different metrics that influence the overall quality of the dataset, specifically the number of remarks (potential outliers) detected by the Bank of Italy's internal procedures, their degree of severity and the expected number of confirmations of underlying data, the latter based on the estimation provided by the logistic regression model.

JEL Classification: C18, C81, G21.
Keywords: potential outliers, non-performing loans, data quality, supervised machine learning, logistic regression.
DOI: 10.32057/0.QEF.2022.0666

Contents

1. Introduction 5
2. From data collection to the release of the information 7
3. The dataset used in this study 10
4. Descriptive data analysis 12
5. Supervised statistical techniques in a nutshell 13
6. Model Selection 15
7. Main results for the selected model 17
Conclusions 21
References 23
Appendix A - Additional descriptive analysis and results 24
Appendix B - Application of the decision-making rule in cases A, C and D 26

[*] Bank of Italy, Statistical Data Collection and Processing Directorate.
[‡] Bank of Italy, IT Development Directorate.
[§] Bank of Italy, Campobasso.

1. Introduction[1]

The Bank of Italy collects a large array of statistical and supervisory data from banks and other financial institutions on a regular basis (Reporting Agents – RAs) to support its analyses and policy decisions. The data are organized in datasets, also called reports, each obeying different reporting regulations. Assessing the Data Quality Level (DQL) is key to enabling users to carry out thorough and robust analyses. RAs are required to ensure high-quality data; however, the reliability of information, assessed upon arrival by the Bank of Italy using a set of automatic Data Quality Checks (DQCs), may be impaired by errors.

As soon as a DQC detects a potential outlier, a remark is generated and transmitted to the RA for its assessment and possibly for action to be taken. Indeed, anomalies may be due to actual errors or they may just be a sign of an unusual, yet correct, content of the data. In the first case, the RA delivers a new report with the corrected values; otherwise, for each remark it sends a confirmation of the original data together with a text explaining the phenomenon causing the anomaly. It is worth pointing out that, depending on the nature of the DQCs applied to the data, the generated remarks can determine either a confirmation ('confirmable remarks') or they must be closed with a correction of the original data ('non-confirmable remarks'). In addition, the remarks are classified as 'serious' or 'non-serious' depending on their predefined degree of severity.[2] An example of a serious and non-confirmable remark is related to the DQC, which verifies the co-presence of both the collateral and the corresponding credit line: if the credit line is missing, then a serious remark is generated and it can only be closed with a correction.

The data quality cycle aims at reducing the number of pending remarks to zero and it is structured as follows: first, for each data submission, the software re-applies the whole set of automated checks; second, the data manager evaluates the (possible) explanations that justify the correctness of the previously reported data. On the basis of the degree of severity of the pending remarks as well as of the analysis of the explanations, the data manager may conclude that the overall quality of the dataset is still not adequate and therefore reopen the dialogue with the RA in order to collect further information. This process continues until the overall quality reaches a level high enough for the publication of the data. In practice, the data manager faces a trade-off between on the one hand, the need to make the information promptly available to users, which implies that the above data quality cycle must be kept short, and on the other, the need for the data to be 'fit for use', i.e. they are as free as possible from significant errors.

[1] The authors are grateful to Gianluca Cubadda and Alessio Farcomeni (University of Tor Vergata, Rome), Laura Mellone, 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] The severity level is usually defined during the implementation of a DQC and is based on an a priori evaluation by data quality experts of the impact of errors on the overall quality of the underlying data.

Data quality is typically assessed by monitoring some stand-alone dimensions, such as timeliness, accuracy and consistency, all evaluated through metrics and indicators (Damia and Aguilar, 2006). In the international statistical context, some Institutions have developed specific strategies to analyse this topic. In particular, in 2012, the IMF Data Quality Assessment Framework (DQAF) proposed a flexible framework with the aim of identifying and assessing quality-related features of the governance of statistical systems, processes and products. The DQAF is organized with a set of prerequisites and with five dimensions of data quality – assurances of integrity, methodological soundness, accuracy and reliability, serviceability and accessibility. For each dimension, several relevant indicators can be identified (Carson, 2001). The European Statistical System (ESS) also identifies possible methods, tools and good practices in order to produce quality statistics based on a sound methodology, namely the Quality Assurance Framework (ESS QAF European Commission and Eurostat, 2019). A different approach to Data Quality Management (DQM) is envisaged by the application of Benford's law, also known as the 'first digit law'. This approach exploits an empirical regularity that can be found in sets of data describing naturally-occurring phenomena (Gonzalez-Garcia and Pastor, 2009). This law, under a set of hypotheses, could make it possible to identify data errors.

In this paper, we focus on the definition of a decision-making rule to evaluate the overall DQL of a dataset sent by an RA (Pipino *et al*., 2002) that could support the data manager in assessing whether its overall quality has improved from the first reporting to the next one. In other words, the algorithm should be suitable to assess whether a revision by an RA of previously transmitted data, on which remarks had been identified by the Bank of Italy and then communicated to the RA, improves or worsens the overall quality of the dataset. In particular, the study evaluates the application of statistical learning techniques that, starting from the actual evidence of previously reported datasets, can provide, as a first step, a prediction on whether a remark will be confirmed or will trigger a revision by the RA. Such a prediction is subsequently used to estimate the probability that confirmable remarks are indeed confirmed. This allows us to identify two groups of observations in our dataset: the first is composed of those reports whose DQL is greater than the previous one sent by the RA, the second contains those reports whose DQL is lower and are then likely to be revised. The implementation of this statistical technique leads to a synthetic data quality indicator that, together with the decision-making rule, provides guidance on the overall fitness for use of the information contained in the dataset.

With regard to the actual data used in the present study, we refer to the Non-Performing Loans (NPL)[3] supervisory report because of the large number and diversity of the automated checks applied to assess the quality of the incoming datasets. Although the methodology proposed in this paper concerns the DQM of NPL reporting, it is worth remarking that the method can be applied, with appropriate adaptations, to any other type of quantitative data collection.

[3] Non-performing loans are exposures to debtors who are no longer able to meet all or part of their contractual obligations because their economic and financial circumstances have deteriorated.

The paper is organized as follows. Section 2 describes the logical path that leads from the DQM process to the definition of the decision-making rule. Section 3 describes the data used in the empirical part of the study. Section 4 summarizes the main results of a descriptive analysis useful for appreciating the main characteristics of the dataset used for the model estimation. Section 5 provides a brief anthology of the main supervised statistical techniques explored for this study, listing the essential elements of each methods. Section 6 illustrates the main arguments that guided the choice of the final model by comparing different alternatives. Section 7 shows the empirical results, with some further details in the Appendix.

2. From data collection to the release of the information

The Bank of Italy collects a large array of data from the banking and financial system; then, it guarantees that the quality of the data is high and constantly monitors it over time. The process ranging from the data collection phase to the release of the information to internal and external users is structured in various steps. As soon as a dataset is submitted by an RA, it undergoes the application of a large set of automatic DQCs to spot anomalies in the report; then, RAs are required to assess potential outliers detected by the Bank of Italy staff (data manager) and to correct or confirm them. In the latter case, RA has to provide suitable explanations for the underlying anomalous pattern that are assessed by the data manager. At each iteration the data manager faces the problem of establishing whether the overall quality of the dataset has improved compared to the previous transmission or not. The DQCs are classified in advance according to a scale increasing from 0 to 10 depending on the degree of severity. Those data reports that give rise to 'serious remarks', i.e. with a severity equal to 9 or 10, are not considered fit for use and data are not-released to users; in case of 'non-serious remarks', i.e. with a severity level below 9, data are released to the users.

This assessment represents the core issue tackled in the present study. It can be illustrated in a more formal way as follows. AN RA submits the k^{th} dataset related to a specific reference date *t*. After the first round of DQCs, the RA receives a set of remarks by the Bank of Italy data manager. Then, the RA submits the $(k+1)^{th}$ report containing the revisions to those data that had been recognized as actual errors. The critical point is that the revisions can either enhance or, in some extreme but still plausible cases, worsen the overall quality associated to the entire dataset. Then, the data manager faces the situation illustrated in Figure 1.

Figure 1: Overall quality from the k^{th} to the $(k+1)^{th}$ submission

		$(k+1)^{th}$ *submission*	
		Not-released	Released
k^{th} *submission*	Not-released	**D**	**C**
	Released	**A**	**B**

CASE A: the k^{th} data submission is considered fit for its use since the automatic DQCs have generated only non-serious remarks and information can be disseminated to the users. With the $(k+1)^{th}$ data submission of the same dataset the RA aims at closing the pending remarks; however, in this situation the new data do not improve the DQL, since the DQCs detect the presence of one or more serious remarks, which prevent data to be made available to users. In this scenario, the data manager cannot release immediately the data but has to send the remarks to the RA in order to request proper corrections. While waiting for the $(k+2)^{th}$ data submission in order to close the pending remarks, the k^{th} one remains the only data available to the users.

CASE B: the k^{th} data submission is considered fit for its use because the automatic DQCs have generated only non-serious remarks; this allows the release of the dataset to the users while corrections are nonetheless awaited to close the pending remarks. Differently from the previous case, here the $(k+1)^{th}$ data submission remains suitable for its use and the new report replaces the previous one.

CASE C: the k^{th} data submission is not-released to the users due to the presence of one or more serious remarks, which make the overall report not fit for its use. The data manager sends the pending remarks to RA requesting corrections. The $(k+1)^{th}$ data submission is, instead, suitable for its use because the automatic DQCs have only generated non-serious remarks.

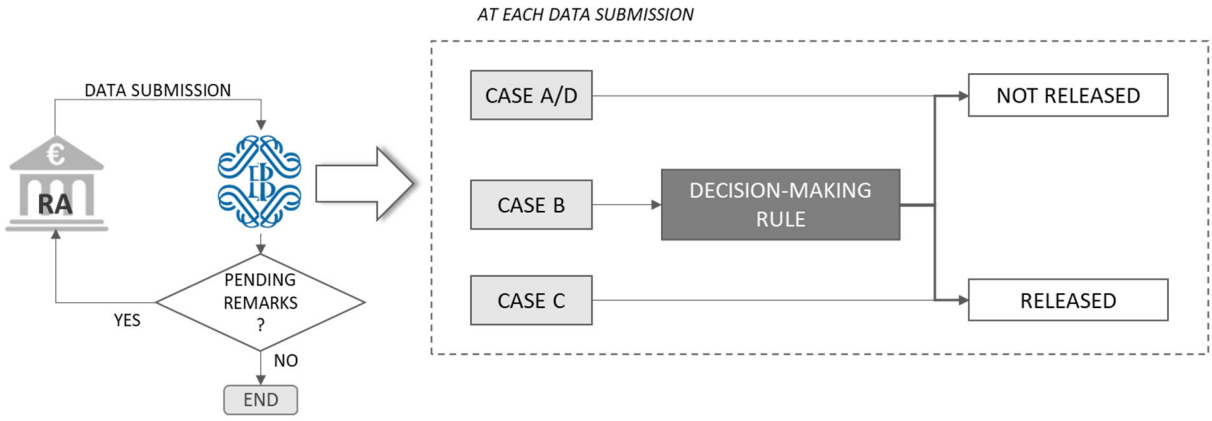
CASE D: the k^{th} data submission is not-released to the users because of the presence of one or more serious remarks that make it not fit for use. The data manager sends the pending remarks to the RA requesting corrections. In the $(k+1)^{th}$ data submission, DQCs spot again one or more serious error; therefore, also the new report is not suitable for its release to the users. The data manager and the RA carry on in their interaction, with the RA analysing the remarks and sending, at each stage, a new report bearing all corrections.

In cases A, C and D, the data manager's decision is not ambiguous; in case B the $(k+1)^{th}$ data submission could contain anomalies such that, due to their number, severity and type, reduce the overall quality of the report in comparison with the k^{th} submission. Hence, in this situation it is key that the data manager has an instrument to assess the variation in the DQL from one report to the next. The purpose of this study is to define a decision-making rule that supports the data manager in this evaluation.

The same rule can also be useful in situations that belong to case D. Here, a measure of the change in the overall quality of a report provides the data manager and the RA with useful indications on the impact of the correction activity.

Figure 2 depicts the DQM process in presence of a decision rule based on the change of the overall quality of a report.

Figure 2: DQM -flow chart



To sum up, the problem is to establish, for a given RA and reference date, the quality variation between two subsequent reporting submissions, the k^{th} and the $(k+1)^{th}$, of the same dataset ($\Delta DQL_{k+1} = DQL_{k+1} - DQL_k$) when both submissions do not bear serious remarks and they are in principle fit for use.

As already illustrated, in case B a new report (i.e. the $(k+1)^{th}$ data submission) replaces the original dataset (k^{th}) and the automatic validation process returns a list of remarks that have to be addressed again to the RAs. The decision-making algorithm must mirror the reasoning made by the data manager when comparing the overall quality of data reported in two consecutive reports. Hence, the rule takes as inputs the number of remarks generated at each round of DQCs, the severity level for each of them and the number of confirmed remarks.

Let us define the dummy variable *R* (remark) taking the value 1 if the DQC *c*, applied at the reference date *t* for the k^{th} data submission by the RA *p*, is violated and 0 otherwise

$$R_{t,p,c,k} = \begin{cases} 1, & if\ c\ is\ violated \\ 0, & if\ c\ is\ satisfied \end{cases}$$

and the dummy variable *Conf* taking the value 1 if the remark *R* is confirmed by the RA and 0 otherwise

$$Conf_{t,p,c,k} = \begin{cases} 1, & if\ the\ R_{t,p,c,k}\ is\ confirmed \\ 0, & otherwise \end{cases}$$

Our decision rule (1) is such that the release of the $(k+1)^{th}$ data submission takes place when

$$\sum_{c=1}^{C_1} \tau_c \left(R_{t,p,c,k+1} - Conf_{t,p,c,k+1}\right) + \sum_{c=1}^{C_2} \tau_c R_{t,p,c,k+1} \leq \sum_{c=1}^{C_1} \tau_c \left(R_{t,p,c,k} - Conf_{t,p,c,k}\right) + \sum_{c=1}^{C_2} \tau_c R_{t,p,c,k} \qquad (1)$$

where: C_1 and C_2 denote the number of remarks generated by DQCs that are confirmable and non-confirmable, respectively; τ_c represents the severity level of the DQC *c;* $k \leq K_{t,p}$ with $K_{t,p}$ equals to the number of

submissions transmitted by RA *p* at the reference date *t*. In other words, the decision rule is such that the overall level of quality associated to the situation in which a few remarks are still pending cannot decrease from the k^{th} to the $(k+1)^{th}$ data submission.

The decision rule (1) assumes that the evidence of a confirmed remark is known in advance; however, in practice this cannot be the case since the submission of an explanation from an RA is always subsequent to the generation of the related remark. Hence, when the decision rule is applied to the $(k+1)^{th}$ dataset, we do not know if a remark generated by DQC *c* is confirmed or not, but this status can be estimated on the basis of the expected probability that a remark is confirmed by a given RA, on a specific reference date. Then, on the basis of the expected probability *p(Conf)*, we can define the estimated confirmation $\widehat{Conf}$ as follows

$$\widehat{Conf}_{t,p,c,k+1} = \begin{cases} 1 & if\ p\left(Conf_{t,p,c,k+1}\right) > cut\text{-}off \\ 0 & otherwise \end{cases}$$

where *cut-off* is a threshold lying within (0, 1). In order to obtain the estimated probability *p(Conf)* we will apply machine learning techniques to the available dataset including confirmable remarks actually observed in the previous reference dates; a cross-validation method allows to assess the *cut-off* level.

3. The dataset used in this study

3.1. The contents and the structure of dataset

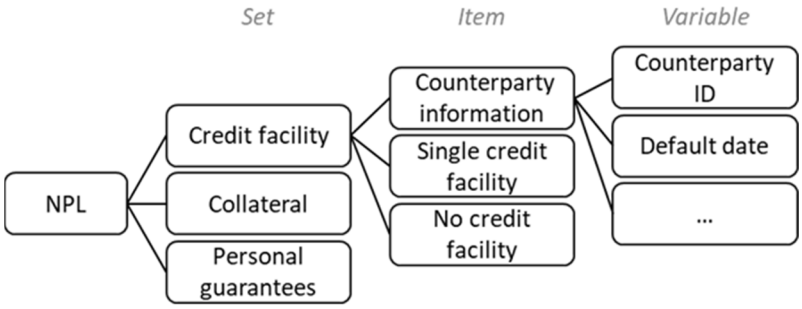
The NPL dataset [4] comprises detailed information on non-performing exposures and on the state of their credit recovery procedures. The data collection regards detailed data on both real estate collaterals and other types of guarantees that mitigate the credit risk. The NPL data collection focuses on gross non-performing loans attributable to single Italian resident counterparties and for a total amount exceeding 100,000 euros at the reference date. The submitted data shall refer only to cash exposures. The RAs are the parent company[5] of a banking groups that report on a consolidated basis and the individual banks on a stand-alone basis. Reporting follows a half-yearly periodicity with reference dates 31th December and 30th June.

Figure 3 reports the structure of the dataset and how the collected data are recorded and organized in order to be processed, aggregated and published. The reported variables are organized into *items* that refer to homogeneous granular information by credit facilities and financial instruments. In turn, the items are grouped into the following three information sets: credit facilities; real estate collaterals; personal guarantees.

[4] The reporting of non-performing exposures was introduced by the Bank of Italy on 29th March 2016; see the 'Instructions for the editing of reports on non-performing exposures' on the Bank of Italy website.

[5] The definition of parent company of a banking group is contained in Circular no. 285 of 17th December 2013.

Figure 3: The structure of the NPL dataset



3.2. The automatic DQCs on the collected data

As described in Section 2, the data sent by the RAs are subjected to automatic DQCs that can be classified into four groups:

i. formal and data structure DQCs;
ii. DQCs aimed at verifying the consistency of different variables belonging to the same item (coherence checks);
iii. cross-item DQCs that, within a specific set of information, compare variables belonging to different items[6];
iv. plausibility and reliability DQCs that compare, for any specific variable, the value reported by a given RA at the reference date to that reported in a previous date. In case the difference is higher than a predefined threshold, a remark is generated.

Only remarks generated by the first three groups of DQCs are regarded as 'non-confirmable remarks'.

3.3. Dataset creation and variable selection

Five reference dates are considered for this study, from 30th June 2017 to 30th June 2019. Overall, in this time interval 445 RAs have reported the data, over 17 million of records were collected and 37 automatic confirmable DQCs were applied to each data submission.

The selection of the reported variables to be included in the model was the result of the dialogue with the data manager in charge of the NPL data management. The variables were analysed in order to minimize and, possibly, eliminate redundancies (Pereira *et al*., 2016). Then, the set of observed variables thus obtained was integrated with other calculated variables deemed suitable to explain some reporting behaviours, identified by the data managers and available in other datasets. The variables considered in the analysis are listed in Table 1.

[6] For example, if a subject is reported in the 'single credit facility' item - which belongs to the 'credit facilities' information set -, his personal data must also be present. Similarly, if a property is reported in the 'single collateral' item - which belongs to the 'real estate collaterals' information set -, the relative register must be present.

Table 1: List of variables included in the analysis

Name	*Type*	*Description*
DQC	Qualitative variable	Identification of the DQC
RA	Qualitative variable	Identification of the RA
Ref_date	Date	Reference date of the reporting data
Conf	Qualitative variable	The remark is confirmed, Yes=1/No=0
Severity level (a)	Quantitative variable	Predefined degree indicating the seriousness of the remark
Imbalance (b)	Quantitative variable	The difference among the aggregates of the remark
Log_Imbalance (b)	Quantitative variable	The logarithm of the absolute amount of Imbalance
Log_Imbalance2 (b)	Quantitative variable	The Log_Imbalance squared
Records	Quantitative variable	The average number of records sent by an RA for a specific Reference date
Log_Records	Quantitative variable	The logarithm of Records

(a) The severity level is usually defined during the implementation of a DQC and is based on a-priori evaluation by data quality experts about the impact of errors on the overall quality of the underlying data.
(b) The variable is available only for DQCs which compare two or more amounts.

4. Descriptive data analysis

A descriptive analysis was carried out in order to explore the main characteristics of the dataset used for the model estimation, in particular with reference to the number of remarks (Table 2). The total number of observations for which a remark was detected amounts to 65,705, 5,083 out of which are 'confirmable remarks'; in turn, 4,643 of the confirmable remarks are used for the model estimation ('training set') and 440, all included in the last reference date, for its validation ('validation set').

Table 2: Number of remarks

Dataset	*Reference date*	*Number of remarks (total)*	*Number of confirmable remarks*	*Number of non-confirmable remarks*
Training set	2017-06-30	31,306	758	30,548
	2017-12-31	8,679	857	7,822
	2018-06-30	18,706	1,398	17,308
	2018-12-31	5,576	1,630	3,946
Validation set	2019-06-30	1,438	440	998
Total		65,705	5,083	60,622

Tables 3 and 4 report the main statistics and the correlation matrix for the continuous explanatory variables in the training set (the detailed results on the validation set are available in Appendix A). Since the variables *Imbalance* and *Records* showed high skewness, we applied the logarithmic transformation. In detail, *Records* is an asymmetric variable since a few RAs presented high number of non-performing loans (Table 3). About the correlation, the *Imbalance* and the related transformations are weakly correlated with those related to the *Records* (Table 4).

Table 3: Training set – summary statistics

Variable	*Min*	*Max*	*Median*	*Mean*	*SD*
Imbalance	-194.85 bn	199.04 bn	-0.18 K	-0.10 bn	5.14 bn
Log_Imbalance	0.00	26.02	15.14	12.92	5.97
Log_Imbalance2	0.00	676.87	229.08	202.51	129.63
Records	0.001 K	1750.62 K	0.79 K	19.89 K	128.61 K
Log_Records	0.00	14.38	6.68	6.58	2.20

Table 4: Training set - correlation matrix

Variable	**Imbalance**	**Log_Imbalance**	**Log_Imbalance2**	**Records**	**Log_Records**
Imbalance	1.00	-0.04	-0.05	-0.02	-0.03
Log_Imbalance	-0.04	1.00	0.97	0.11	0.23
Log_Imbalance2	-0.05	0.97	1.00	0.16	0.31
Records	-0.02	0.11	0.16	1.00	0.45
Log_Records	-0.03	0.23	0.31	0.45	1.00

5. Supervised statistical techniques in a nutshell

In Section 2, we have proposed a decision-making algorithm to determine whether data revisions improve the DQL based on the number of remarks and on their severity level. In order to estimate the difference between two submissions in terms of data quality, it is necessary to predict the confirmation probability of a remark. The response variable is the dummy *Conf* ('the remark is confirmed, Yes=1/No=0') as defined in Section 2. Generally speaking, in supervised learning a model is developed either in order to predict an outcome response using known predictors or to better understand the relationship between the response variable (Y) and the predictors (X), both included in the dataset used for the model estimation. The supervised approach fits a model that relates the response to the predictors and maximizes the accuracy of the prediction of future observations.

Predicting a qualitative response for an observation can be regarded as classifying that observation, since it implies the assignment of the observation to a category or class (James *et al*., 2013; Hastie *et al*., 2009). While the logistic regression estimates the probability that Y is equal to the category of interest, a linear discriminant analysis models the distribution of the predictors X separately in each of the response classes (i.e. given Y) and then use Bayes' theorem to flip these around into estimates for $Pr(Y = k|X = x)$. When these distributions are assumed to be normal, it turns out that the model is very similar in form to a logistic regression.

When the classes are well-separated, Linear Discriminant Analysis (LDA) overcomes the instability of the estimates for the logistic regression model. Furthermore, LDA, which is popular in case of more than two response classes, is more stable than the logistic regression model if the number of observations is small and the distribution of the predictors X is approximately normal in each class. Conversely, logistic regression can outperform LDA if this assumption of Gaussian distributions is not met.

Quadratic Discriminant Analysis (QDA) offers an alternative discriminant analysis approach. Like LDA, the QDA classifier assumes that from each class the observations are drawn from a Gaussian distribution and obtains estimates of the parameters through the Bayes' theorem. Unlike LDA, QDA can lead to model a wider range of problems since it assumes that each class has its own covariance matrix and a quadratic decision boundary. Conversely, the LDA and logistic regression approaches tend to outperform when the true decision boundaries are linear.

The k-Nearest Neighbor classifier (KNN) relies on a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. Therefore, this approach is expected to dominate LDA and logistic regression when the decision boundary is highly non-linear; however, it is worth remarking how KNN does not tell anything on what predictors are more important than others.

Classification trees have the advantage of simplicity in representation and interpretability, due to the possibility of handling qualitative predictors without the need to create dummy variables, although in general they do not have the same level of predictive accuracy as some of the other regression and classification approaches. Moreover, they can be non-robust because a small change in the data can cause a large change in the final estimated tree. However, by aggregating many decision trees through methods such as bagging, random forests and boosting, the predictive performance of trees can generally be substantially improved.

Finally, ridge estimators are used in the logistic regression model to obtain more realistic estimates for the parameters and to improve the predictive value of the model (Le Cessie and Van Houwelingen, 1992). In particular, this aim is reached by a penalty parameter λ used in the estimation that allows to obtain a model with a lower mean squared error. The Ridge Logistic classifier (Hoerl and Kennard, 1970; Le Cessie and Van Houwelingen, 1992; Schaefer *et al*., 1984; Cule and De Iorio, 2012; Van Wieringen, 2020) improves the logistic regression model estimation in case of unstable parameters estimation, when the number of covariates is relatively large or when the covariates are highly correlated. The logistic regression can be considered as a Ridge's specific case when λ is set equal to 0.

6. Model Selection

The scope of this section is to solve the binary classification problem in order to predict whether a remark will be confirmed or not before the RAs provide such information.

The data used to estimate the probability of confirmation include, as observations, all the individual confirmable remarks that have been generated for each RA and reference date; more specifically, a remark is considered only once even if it is pending in more than one data submission[7].

The dataset contains 431 dummy variables, of which 15 for DQCs, 415 for RAs and 1 for the response variable *Conf*; it is also made up of 4 quantitative variables related to the imbalance (*Log_Imbalance* and *Log_Imbalance2*), the number of records sent (*Log_Records*) and the reference date (*Ref_date*). The resulting imbalance is different from zero for the DQCs which verify the plausibility of monetary amounts. Overall, we have 4,643 observations and 436 variables, 435 of which are predictors in the model estimation.

In order to estimate the probability that a remark will be confirmed, several methods have been considered and applied to the training set. The latter includes all those remarks that have been sent to RAs except those related to the last reference date, that has been used as validation set.

The techniques illustrated in Section 5 were considered and tested: a Ridge Logistic classifier; a linear and quadratic discriminant analysis; a decision tree classifier; a k-neighbors classifier and a random forest. The results are reported in Table 5.

The Ridge Logistic classifier outperforms the others methods in terms of accuracy, recall and precision. The Linear and Quadratic Discriminant Analysis (LDA and QDA) and the random forest do not predict properly the probability of confirmation of a remark in the validation set, as (about) all the remarks are expected to be confirmed. QDA should exhibit a better performance in case of non-linear decision boundary, and in our empirical analysis its performance turned out to be only slightly higher than the LDA considering the validation set. A decision tree classifier and a k-neighbors classifier provide similar results than the previous models, but they still underperform the Ridge Logistic classifier. A few studies suggested that the latter models may improve their performance when filtering or clustering of features is applied (Ala'raj *et al*., 2020; Rajaguru and Chakravarthy, 2019). However, this may be associated with enhanced skewness of the results, increasing their sensitivity and reducing their specificity.

For the model estimation, the logistic regression classifier has been selected as a special case of the Ridge Logistic classifier. Following Le Cessie and Van Houwelingen (1992) and considering that in our case the binary response *Conf* and a set of predictors X of U dimensions are measured on remarks; then, the

[7] The statistical platform used for DQM in the Bank of Italy foresees that when the RA submits the data several times without closing all the pending remarks, these remarks are sent again to the RA until they are solved with a data correction or confirmation.

estimated probability that a remark is confirmed is given by *p(Conf)*, where the probability function *p* follows the logistic regression model:

$$p(Conf) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

The Ridge logistic classifier is obtained by maximizing the likelihood function $l(\beta)$ with a penalized parameter λ applied to all the coefficients β except the intercept (β_0); in such case the estimator will be:

$$\beta_{RL} = argmax_{\beta}\left\{l(\beta) - \lambda \sum_{u=1}^{U} \beta_u^2\right\}$$

The logistic ridge regression estimator depends on the choice of a tuning parameter $\lambda \geq 0$ to be determined separately (Cule and De Iorio, 2013; Le Cessie and Van Houwelingen, 1992).

In order to maximize the model's performance, λ and the cut-off parameter were optimized by cross-validated *grid-search.* The best results are reached when the cut-off is 0.41 and λ is set to zero, i.e. when the standard logistic regression is applied.

This outcome may be conditioned by the fact that the dataset includes a large set of uncorrelated variables. This is a feature of our dataset that is strictly connected to the production of remarks as a result of the data quality: a remark is generated by a single DQC and this implies a low correlation among most covariates. This result is not new in the literature. Cornell-Farrow and Garrard (2020) apply several machine learning models to predict the classification of students' performance using a large dataset and the logistic regression classifier outperformed the other methods. Instead, Bradley (1996) evaluates the model performance of some machine learning algorithms using the area under the ROC curve and shows that the logistic regression outperforms methods like decision tree, random forest and neural net.

Table 5: Model comparison in the training and in the validation sets

	Model	**Logistic regression**	Ridge logistic classifier (λ=1)	Linear Discriminant Analysis	Decision Tree Classifier	Quadratic Discriminant Analysis	K Neighbors Classifier	Random forest
	Optimal cut-off	**0.41**	0.69	0.50	0.52	0.49	0.53	0.71
Training set	**Accuracy**	**0.8283**	0.8277	0.7265	0.7252	0.7256	0.7271	0.5044
	Recall	**0.9516**	0.9168	0.9902	0.9355	0.9938	0.9807	0.3944
	Precision	**0.8347**	0.8558	0.7293	0.7483	0.7274	0.7330	0.8346
	Negative predictive value	**0.7980**	0.7303	0.5541	0.5023	0.5435	0.5390	0.3325
Validation set	**Accuracy**	**0.8068**	0.7841	0.7636	0.7523	0.7795	0.7545	0.7795
	Recall	**0.9417**	0.8455	0.9738	0.9038	0.9942	0.9359	1.0000
	Precision	**0.8325**	0.8735	0.7786	0.8031	0.7821	0.7887	0.7795
	Negative predictive value	**0.6154**	0.5093	0.1818	0.3889	0.5000	0.3333	NA

7. Main results for the selected model

The rule described in (1) takes into account the main features that could affect the overall data quality of a report sent by an RA for a specific reference date: the number of occurring remarks; the expected number of these remarks which refer to correct data; the severity of the errors. In order to estimate the probability of future confirmation of the remarks, a logistic regression model is proposed and its results are showed in Section 7.1. The results related to the application of the proposed decision-making algorithm are presented Section 7.2.

7.1. Logistic regression: estimation of 'remark confirmed, Yes/No'

The logistic regression model used to estimate the probability that a remark is confirmed is defined as follows:

$$Logit\big(p(Conf)\big) = \beta_0 + \beta_1 \cdot Log\_Imbalance + \beta_2 \cdot Log\_Imbalance2 + \beta_3 \cdot Log\_Records + \\ + \beta_4 \cdot Ref\_date + \sum_{c=1}^{15} \gamma_c \cdot DQC_c + \sum_{p=1}^{415} \delta_p \cdot RA_p + \varepsilon$$

As mentioned before, the model is estimated by using the training set made up of confirmable remarks that occurred in the years 2017-18 (4,643 observations); the evaluation is carried out in the validation set consisting of remarks referred to the reference date of June 2019 (440 observations)[8].

Table 6 reports the confusion matrices computed in the training set and in the validation set; the main measures for assessing the goodness of the classification are reported in the right column of the two tables ('value'). The confusion matrices use the cut-off equal to 0.41, optimal since it maximizes the accuracy, in order to predict *Conf* based on the probability $p(Conf)$ estimated by the logistic model.

In particular, the estimated confirmation $\widehat{Conf}$ of a remark, generated by a DQC *c* for the $(k+1)^{th}$ data submission sent by the RA *p* for the reference date *t*, is given by the following:

$$\widehat{Conf}_{t,p,c,k+1} = \begin{cases} 1 & if\ p\big(Conf_{t,p,c,k+1}\big) > 0.41 \\ 0 & otherwise \end{cases}$$

In general, all the measures computed in the training and validation sets suggest a satisfying performance of the chosen classification model. The measures selected for assessing the goodness of the classification show high and similar values in the two sets, except for the negative predictive value (NPV) that is 0.7980 and 0.6154, respectively, in the training and in the validation set (Table 6).

[8] The remarks generated by non-confirmable DQCs that always detect reporting errors have been removed from both the training and the validation sets.

Since the purpose of the paper is to set a decision rule in order to determine whether the DQL has improved by means of a subsequent data transmission, the scenario where remarks of the new report are predicted as confirmed when they should not (false positive) is more dangerous than classifying a remark as not-confirmed when it should be (false negative). On the one hand, the presence of false positives could overestimate the DQL and it could lead to accepting the new data although they have a lower level of quality than the previous report. On the other, the presence of false negatives could lead to reject the new report and request further analysis of remarks to the RA. The latter case represents a precautionary approach, whereas the former is clearly not prudent.

In sum, the classification model presented in this section offers a reasonable approach to classify the remarks as 'confirmed' or 'not-confirmed' for the decision-making algorithm since it achieves a high level of precision (0.8347 in the training set; 0.8325 in the validation set) and, at the same time, an acceptable number of false negative (163 FN in the training set; 20 FN in the validation set).

Table 6: Confusion matrix computed in the training set and the validation set

Training set		***Conf***			**Measures**	**Value**
		No	Yes	Total	Accuracy	0.8283
$\widehat{Conf}$	No	644 (TN)	163 (FN)	807	Recall	0.9516
	Yes	634 (FP)	3,202 (TP)	3,836	Precision	0.8347
	Total	1,278	3,365	4,643	Negative Predictive Value (NPV)	0.7980

Validation set		***Conf***			**Measures**	**Value**
		No	Yes	Total	Accuracy	0.8068
$\widehat{Conf}$	No	32 (TN)	20 (FN)	52	Recall	0.9417
	Yes	65 (FP)	323 (TP)	388	Precision	0.8325
	Total	97	343	440	Negative Predictive Value (NPV)	0.6154

Note: TN true negative; FN false negative; TP true positive; FP false positive; Conf and $\widehat{Conf}$ are the actual and predicted values of the variable Conf, respectively.

7.2. Application of the decision-making rule

The decision-making algorithm (1) determines whether a new report *k+1* can be immediately disseminated to the users in substitution of the previous report *k,* sent by an RA *p* for the same reference date *t.*

In order to identify the reports on which the decision rule can be applied, all the eligible submissions (*k+1)* with $k \geq 1$ are considered (the first submissions are obviously excluded since they cannot be compared with a previous report). In particular, the eligible submissions taken into account are: 1,002 submissions with reference dates from 2017 to 2018; 53 reports sent for the reference date of June 2019 (Table 7).

Table 7: Distribution of reports sent by RAs

Type of reports	*Reference dates between 2017 and 2018*	*Reference date of June 2019*
First submission	378	74
Further submissions	1,002	53

The results of the current approach described in Section 2 are shown in Table 8. This method could be improved in the so-called 'case B' when both the previous and the further submissions are classified as not-released reports (275 cases observed between 2017 and 2018; 14 in June 2019). In particular, in this case the decision-making algorithm proposed is assessed in order to identify which further reports enhance the DQL compared to the previous submission. Therefore, the proposed decision rule is applied to 275 submissions with reference date between 2017 and 2018 and to 14 submissions sent with the reference date June 2019.

Table 8: Classification of eligible submissions in the current approach

Reference dates between years 2017 and 2018		***(k+1)th submission***		
		Not-released	Released	Total
kth submission	Not-released	269	407	696
	Released	51	275	326
	Total	320	682	1,002

Reference date June 2019		***(k+1)th submission***		
		Not-released	Released	Total
kth submission	Not-released	15	23	38
	Released	1	14	15
	Total	16	37	53

According to the proposed decision-making algorithm applied in the 'case B' (Table 9), for reference dates from 2017 to 2018, the overall quality of 246 further submissions is higher than that of the corresponding previous reports. Consequently, the release of the last submission is recommended. In 29 cases the data quality decreases with the subsequent submission; then, the new report should not be published and the data manager should contact the RA in order to urgently fix the data quality issues that compromise the quality of the transmitted data. Regarding the reference date of June 2019, in only one over 30 cases there is a reduction in the DQL. The results of the application of the decision rule are consistent with the data corrections that take place in practice. Indeed, the incidence of cases where a data quality reduction is identified after a new submission is moderate (11% between 2017 and 2018; 7% in June 2019) since the new submission is transmitted by the RA with the purpose of correcting previous reporting mistakes.

Table 9: Application of the decision-making algorithm to 'case B'
(number and percentage in brackets)

Results of the decision rule	***Reference dates between 2017 and 2018***	***Reference date of June 2019***
Released submission	246 (89%)	13 (93%)
Additional Not-released submission	29 (11%)	1 (7%)
Total	275 (100%)	14 (100%)

In order to assess the appropriateness of the classification of the submissions as released and additional not-released according to the decision rule, a comparison with a benchmark is carried out. In particular, the benchmark considered is obtained from the application of the decision rule when the observed variable *Conf* is considered instead of the estimated value in according to the logistic model of Section 7.1.

$$\sum_{c=1}^{C_1} \tau_c \left(R_{t,p,c,k+1} - \widehat{Conf}_{t,p,c,k+1}\right) + \sum_{c=1}^{C_2} \tau_c R_{t,p,c,k+1} \quad \leq \sum_{c=1}^{C_1} \tau_c \left(R_{t,p,c,k} - Conf_{t,p,c,k}\right) + \sum_{c=1}^{C_2} \tau_c R_{t,p,c,k} \qquad (2)$$

The application of rule (1) is compared with (2) and the results are shown in Table A.3 of the Appendix A for the time range from 2017 to 2018 and June 2019, respectively. The results confirm the performance of the decision rule that uses the estimated variable $\widehat{Conf}$. Indeed, for the reference dates from 2017 to 2018, in 97% of cases the decision to consider a submission as released or not-released for replacing the previous one, using the predicted $\widehat{Conf}$, is the same as the one we would take the actual *Conf* were known. Considering the reference date of June 2019, the decision is definitely the same in both cases.

Conclusions

In the applied statistical literature, data quality is normally approached by looking at its different dimensions, e.g. timeliness, accuracy and completeness of data. This paper contributes to this literature by exploring the possibility of assessing the overall DQL of a dataset reported by an RA to the Authorities and implementing a decision-making algorithm to support the data quality manager in the releasing of data to internal and external users even if the data quality process is not finalized yet.

Although the methodology is applied to a specific (granular) dataset – the Non-Performing Loans dataset of the Bank of Italy – the method presented here is rather general and can easily be adapted to the assessment of the quality of any dataset. Using machine learning techniques based on the results of the automatic validation process and data managers' past evaluations of explanations received by RAs, we compare the DQL of a report containing corrections with that of the original one sent by the same RA and for the same reference date. In the current practice, the comparison between two consecutive reports submitted by the RA is left to the judgment and expertise of the data managers and the interaction with the RA. The opportunity for an automatic algorithm based on the proposed decision-making rule leads to a less time-consuming and more harmonized approach.

From a methodological point of view, in order to estimate the difference in quality between two versions of the same dataset (the second carrying the revisions of the first), a logistic regression model has been used to predict the confirmation probability of a single remark. The final model has been selected by comparing different available alternatives and optimizing the goodness of fit (i.e. accuracy, recall and precision). In our case, the logistic regression model outperforms several models that are commonly used in machine learning studies.

In the second phase of the study, a decision-making rule has been developed by taking into account the estimated confirmation probability of a single remark, which emerged from the previous step, together with the total number and the severity of the remarks. The results show that a decision-making algorithm could actually support the data manager in deciding whether the new data have a sufficient DQL and can then be released to users. The rule may help the data manager to determine whether data revisions do improve the data quality of previously reported data and to distinguish the reports that most likely will need to be corrected from those that will instead be confirmed by the RA. The decision-making rule also provides guidance to data managers for prioritizing data quality activities for the identification of insufficient data quality reports that require a new submission from the RAs. At the same time, it provides a tool to identify reports that may not be immediately suitable for use.

In order to verify the appropriateness of the decision rule for classifying further submissions as released and not-released, a comparison with a benchmark is carried out. In particular, the benchmark is obtained by applying the decision rule when the observed variable, 'the remark is confirmed, Yes/No', is considered instead

of its estimation according to the logistic model. The results are remarkable, since in 97 per cent of cases, the decision would be the same as that in the situation in which the actual status was known.

This approach is particularly useful when two consecutive submissions are considered as released by the current method; hence, based on the results of this application, this rule may be integrated into the dashboards used by the data manager to monitor the DQL of incoming reports.

References

Ala'raj, M., Majdalawieh, M., Abbod, M. F. (2020), 'Improving binary classification using filtering based on k-NN proximity graphs'. J Big Data 7, 15 (https://doi.org/10.1186/s40537-020-00297-7).

Banca d'Italia (2016), 'Instructions for the editing of reports on non-performing exposures' (https://www.bancaditalia.it/statistiche/raccolta-dati/segnalazioni/normativa-segnalazioni/ISTR_COMP_SOFF.pdf).

Bradley, A. P. (1996), 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', The University of Queensland.

Carson C. S. (2001), 'Toward a Framework for Assessing Data Quality', Volume 2001 Issue 025 International Monetary Fund.

Cornell-Farrow, S., Garrard, R. (2020), 'Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia, Communications in Statistics: Case Studies, Data Analysis and Applications', 6:2, 228-246 (https://doi.org/10.1080/23737484.2020.1752849).

Cule, E., De Iorio, M. (2012), 'A semi-automatic method to guide the choice of ridge parameter in ridge regression', Imperial College London and University College London.

Cule, E., De Iorio, M. (2013), 'Ridge regression in prediction problems: automatic choice of the ridge parameter', Genetic epidemiology, 37(7), 704-714.

Damia, V., Aguilar, C. P. (2006), 'Quantitative quality indicators for statistics an application to euro area balance of payment statistics'; Occasional Papers series n. 54, European Central Bank.

European Commission and Eurostat (2019), 'Quality Assurance Framework of the European Statistical System (V.2.0)'.

Gonzalez-Garcia J. R., Pastor G. (2009), 'Benford's Law and Macroeconomic Data Quality' Volume 2009: Issue 010 International Monetary Fund.

Hastie, T., Tibshirani, R., Friedman, J. (2009), 'The elements of statistical learning: data mining, inference, and prediction', Springer Science & Business Media.

Hoerl, A., Kennard, R. (1970), 'Ridge Regression: Biased Estimation for Nonorthogonal Problems'. Technometrics, 12(1), 55-67 (https://doi.org/10.1080/00401706.1970.10488634).

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), 'An introduction to statistical learning with applications in R', Springer Science & Business Media.

Le Cessie, S., Van Houwelingen, J. (1992), 'Ridge Estimators in Logistic Regression'. Journal of the Royal Statistical Society, Series C (Applied Statistics), 41(1), 191-201.

Pereira, J. M., Basto, M., da Silva, A. F. (2016), 'The logistic lasso and ridge regression in predicting corporate failure', Procedia Economics and Finance, 39, 634-641.

Pipino, L. L., Lee, Y. W., Wang, R. Y. (2002), 'Data Quality Assessment', Communications of the ACM Vol. 45 (pages 211-218), No. 4ve.

Rajaguru, H., Chakravarthy, S. (2019), 'Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer', Asian Pacific journal of cancer prevention: APJCP. 20. 3777-3781. 10.31557/APJCP.2019.20.12.3777.

Schaefer, R. L., Roi, L. D., Wolfe, R.A. (1984), 'A ridge logistic estimator, Communications in Statistics - Theory and Methods', 13:1, 99-113 (https://doi.org/10.1080/03610928408828664).

Van Wieringen, W. N. (2020), 'Lecture notes on ridge regression' Vrije Universiteit Amsterdam.

Appendix A - Additional descriptive analysis and results

Figure A.1. Box-plot of the number of remarks sent generated by DQCs

a) Training set **b) Validation set**

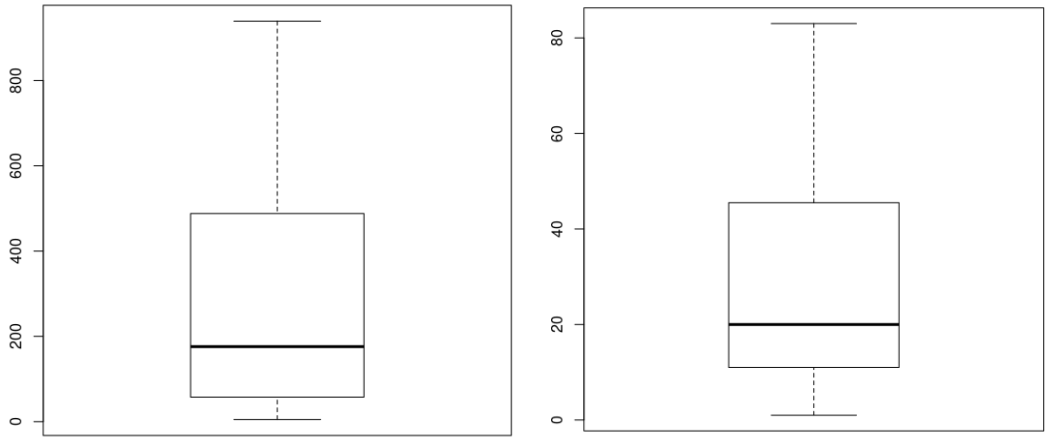


Table A.1: Validation set – summary statistics

Variable	*Min*	*Max*	*Median*	*Mean*	*SD*
Imbalance	*-10.61 bn*	*9.08 bn*	*-0.66 mio*	*-0.09 bn*	*1.28 bn*
Log_Imbalance	*0.00*	*23.09*	*15.40*	*13.71*	*6.30*
Log_Imbalance2	*0.00*	*532.92*	*237.14*	*227.45*	*140.79*
Records	*0.001 K*	*645.10 K*	*0.89 K*	*30.08 K*	*95.88 K*
Log_Records	*0.00*	*13.38*	*6.79*	*6.98*	*2.88*

Table A.2: Validation set - correlation matrix

Variable	**Imbalance**	**Log_Imbalance**	**Log_Imbalance 2**	**Records**	**Log_Records**
Imbalance	1.00	-0.08	-0.11	-0.17	-0.11
Log_Imbalance	-0.08	1.00	0.96	0.17	0.34
Log_Imbalance2	-0.11	0.96	1.00	0.27	0.46
Records	-0.17	0.17	0.27	1.00	0.57
Log_Records	-0.11	0.34	0.46	0.57	1.00

Table A.3: Verification of the decision rule: comparison with the benchmark decision rule that considers the observed *Conf*

Reference dates between years 2017 and 2018		**Benchmark decision on the *(k+1)th* submission (using *Conf*)**		
		Released	Not-released	Tot.
Decision on the *(k+1)th* submission considering $\widehat{Conf}$	Released	244	7	251
	Not-released	2	22	24
	Total	246	29	275

Accuracy = 0.9673

Reference date 2019		**Benchmark decision on the *(k+1)th* submission (using *Conf*)**		
		Released	Not-released	Tot.
Decision on the *(k+1)th* submission considering $\widehat{Conf}$	Released	13	0	13
	Not-released	0	1	1
	Total	13	1	14

Accuracy = 1

Appendix B - Application of the decision-making rule in cases A, C and D

In Section 7.2, the decision-making algorithm has been assessed by considering two consecutive data submissions that are classified as released by the current method (named 'case B' in the Introduction of the paper).

The results of the application of the decision rule to the cases A, C and D, discussed in this Appendix, are consistent with the rationale behind the current method that considers the overall data quality lower when at least one serious error affects the data transmitted (not-released cases) than when the submission only presents non-serious issues (released cases).

In fact, as described in Table B.1, in case C where the previous and the further submissions are classified respectively as not-released and released cases by the current method, as expected, the proposed decision rule suggests the overall quality of the $(k+1)^{th}$ submissions is better than the k^{th} reports (99% for reference dates from 2017 to 2018; 100% for June 2019).

Vice versa, according to the decision-making algorithm applied to case A, the overall quality of further submissions is mostly lower than the corresponding previous reports (Table B.2). In this latter case, the proposed decision rule shows that the 41% of further submissions with reference dates from 2017 to 2018 shows an increase in the overall data quality: even though the RA has corrected most of the previously data errors, at least one serious error occurred.

Although case D is made up of the k^{th} and the $(k+1)^{th}$ submissions classified as not-released, the decision rule shows an increase in the overall data quality of 90% and 100% of further submissions with reference dates respectively from 2017 to 2018 and June 2019 (Table B.3). This result is consistent to the RA's data reporting process whose aim is the improvement of the data transmitted reliability.

Table B.1: Application of the decision-making algorithm to 'case C'
(number and percentage in brackets)

DQL of the $(k+1)^{th}$ submission compared to the k^{th} report	***Reference dates between 2017 and 2018***	***Reference date of June 2019***
Higher	401 (99%)	23 (100%)
Lower	6 (1%)	0 (0%)
Total	407 (100%)	23 (100%)

Table B.2: Application of the decision-making algorithm to 'case A'
(number and percentage in brackets)

DQL of the (k+1)th submission compared to the kth report	*Reference dates between 2017 and 2018*	*Reference date of June 2019*
Higher	21 (41%)	0 (0%)
Lower	30 (59%)	1 (100%)
Total	51 (100%)	1 (100%)

Table B.3: Application of the decision-making algorithm to 'case D'
(number and percentage in brackets)

DQL of the (k+1)th submission compared to the kth report	*Reference dates between 2017 and 2018*	*Reference date of June 2019*
Higher	242 (90%)	14 (100%)
Lower	27 (10%)	1 (0%)
Total	269 (100%)	15 (100%)