BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Quality checks on granular banking data: an experimental approach based on machine learning

by Fabio Zambuto, Maria Rosaria Buzzi, Giuseppe Costanzo, Marco di Lucido, Barbara La Ganga, Pasquale Maddaloni, Fabio Papale and Emiliano Svezia

March 2020

Number 547

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Quality checks on granular banking data: an experimental approach based on machine learning

by Fabio Zambuto, Maria Rosaria Buzzi, Giuseppe Costanzo, Marco di Lucido, Barbara La Ganga, Pasquale Maddaloni, Fabio Papale and Emiliano Svezia

Number 547 – March 2020

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

QUALITY CHECKS ON GRANULAR BANKING DATA: AN EXPERIMENTAL APPROACH BASED ON MACHINE LEARNING

by Fabio Zambuto[*], Maria Rosaria Buzzi[*], Giuseppe Costanzo[*], Marco di Lucido[*], Barbara La Ganga[*], Pasquale Maddaloni[*], Fabio Papale[†] and Emiliano Svezia[*]

Abstract

We propose a new methodology, based on machine learning algorithms, for the automatic detection of outliers in the data that banks report to the Bank of Italy. Our analysis focuses on granular data gathered within the statistical data collection on payment services, in which the lack of strong ex ante deterministic relationships among the collected variables makes standard diagnostic approaches less powerful. Quantile regression forests are used to derive a region of acceptance for the targeted information. For a given level of probability, plausibility thresholds are obtained on the basis of individual bank characteristics and are automatically updated as new data are reported. The approach was applied to validate semi-annual data on debit card issuance received from reporting agents between December 2016 and June 2018. The algorithm was trained with data reported in previous periods and tested by cross-checking the identified outliers with the reporting agents. The method made it possible to detect, with a high level of precision in term of false positives, new outliers that had not been detected using the standard procedures.

JEL Classification: C18, C81, G21.
Keywords: banking data, data quality management, outlier detection, machine learning, quantile regression, random forests.
DOI: 10.32057/0.QEF.2020.547

Contents

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

[†] Bank of Italy, IT Development Directorate.

1. INTRODUCTION AND MAIN CONCLUSION[*]

Central Banks regularly collect, process and disseminate a wide set of economic and financial statistical data that are used by internal and external users. Ensuring a high quality level for this huge wealth of information is crucial in order to ensure that the central banks' decision-making processes for the various institutional functions (e.g. monetary policy, financial stability, banking supervision and economic research) continue to be based on high-quality information and that high-quality data are disseminated to the public at large.

This requirement has become even more important in recent years due to the significant information gaps revealed by the global financial crisis. This has led to the collection of a growing and more diversified volume of granular data with a view to enhancing the monitoring of the financial system by promptly detecting any vulnerability. In the face of the expanding granularity of statistical data collections, detecting outliers through effective data quality management (DQM) system has become more challenging.

The DQM of large datasets is typically performed by means of automated checks that verify some pre-determined relationships among the data (e.g. accounting, logical and mathematical relationships) collected from reporting agents[2] (RAs). However, there are situations in which such ex ante deterministic relationships are weak or lacking. In these cases the approach that is usually adopted entails plausibility checks, which include the estimation of thresholds and of the related "acceptance regions"; the potential outliers – to be submitted to RAs for cross-checking – are defined as the observations falling outside the acceptance regions.

This approach is not straightforward for various reasons. First, the thresholds must be calibrated according to a robust statistical analysis since, on the one hand, too loose acceptance regions may fail to identify all the anomalies in the data and, on the other hand, if they are too tight this could result in an unnecessarily large number of potential outliers (i.e. too many "false positives"). Second, although the thresholds are typically calibrated on the basis of time series analysis, this approach requires some degree of judgment that goes beyond statistical expertise and involves the experience and knowledge of the data managers responsible for the DQM. Third, the thresholds need to be periodically reviewed, and in some cases updated, to account for the fact that reporting patterns may change over time; if the number of thresholds is large, this activity can be highly time-consuming. In this context, plausibility checks inevitably become more complex to manage when the degree of granularity of the collected information is high and reporting patterns are heterogeneous.

[*] The authors are grateful to Gianluca Cubadda and Roberto Rocci (University of Tor Vergata, Rome) and to participants in the Bank of Italy workshop on "Big Data and Machine Learning" (Rome, 6 June 2018) for useful comments and fruitful discussions on a preliminary draft of the paper. We also wish to thank Guerino Ardizzi, Michele Savini Zangrandi and Elisa Bonifacio (Bank of Italy, Market and Payment System Oversight Directorate) for the useful comments they provided during the construction and analysis of the database. The views expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy.

[2] In our context reporting agents are financial intermediaries subject to reporting regulations.

In order to tackle these issues, the paper explores the use of machine learning techniques to efficiently validate granular datasets collected from RAs[3]. The literature on the use of machine learning techniques within a central bank is rich; however, so far it has mainly focused on economic and financial analysis (Chakraborty and Joseph, 2017). On the contrary, statistical applications for DQM purposes have been relatively unexplored so far; a notable exception concerns the use of classification and regression algorithms to improve DQM of financial and supervisory data (Cagala, 2017, and Farnè and Vouldis, 2018). This paper contributes to this new research area by proposing a method based on quantile regressions to estimate acceptance regions for outlier detection. Specifically, a supervised learning algorithm known as Quantile Regression Forests (Meinshausen, 2006) is adopted in order to identify statistical relationships among the collected data that can be exploited to detect potential outliers, which RAs are then requested to cross-check. From an operational point of view, this approach would appear to improve the current DQM system based on plausibility checks in two ways. First, the new method exploits all the available information to define thresholds that take into account both the heterogeneous reporting patterns and the intrinsic variability observed in the data; in other words, the estimated acceptance regions are tailored to the characteristics of the individual RA (for instance, in terms of the number and type of customers) and to the degree of granularity of the data that are collected. Second, the method reduces the role of expert judgement in the (periodic) calibration of thresholds by providing a more data-driven approach to define dynamic thresholds which are automatically updated over time as new data are collected.

The empirical analysis carried out in the paper validates granular data on payment services reported by Italian banks, which are characterized by rather weak ex ante relationships among the reported variables. The data analysed are archived in the Bank of Italy's statistical data warehouse (DWH) after they successfully pass through the current DQM system. The machine learning algorithm is tested by cross-checking the identified outliers directly with the RAs. The main result of the empirical exercise is that the proposed methodology detects new outliers, which had not been identified by the current DQM system, with a high level of accuracy in terms of minimizing the number of "false positives". These results confirm that machine learning techniques are worth exploring to enhance the efficiency and effectiveness of statistical DQM processes related to the data that central banks collect from RAs.

The paper is organized as follows. Section 2 illustrates the main features of the Bank of Italy's DQM system that is currently applied to the data transmitted by the RAs. Section 3 describes the characteristics of the dataset concerning payment services and debit cards that we employed in our empirical analysis. Section 4 illustrates the proposed methodology to derive acceptance regions and presents the main results. Section 5 summarizes the main conclusions and outlines future research directions.

[3] For an overview of machine learning techniques, see Hastie, Tibshirani and Friedman (2001) and Hastie, James, Tibshirani and Witten (2013), Bishop (2007).

2. MAIN FEATURES OF THE BANK OF ITALY'S DQM SYSTEM

The Bank of Italy regularly collects a large amount of data from RAs. Data collection relies on a representation model based on the logical framework of a two by two table (matrix model, see Del Vecchio et al., 2007, Froeschl, Grossmann and Del Vecchio, 2003) where the rows indicate the "business facts"[4] and the columns identify the main features of these "business facts"[5] Information is collected at a very detailed and granular level through a number of "surveys", each internally homogeneous in term of the "business facts" being investigated, the type of RAs, the reference date and the reporting deadline.

In order to validate the data collected from RAs, the current DQM system is based on a highly automated two-step process. First, the data are validated via a set of quality checks that are carried out automatically upon receipt of the reports. Second, the identified anomalies are communicated to RAs via automatically generated "remark messages" in order to elicit revisions when necessary. The system also allows back-and-forth interaction between the RAs and Bank of Italy's data managers concerning the "confirmation" of the contested anomalies when these are not reporting errors. At the end of the validation process, the data are released to the DWH and become available to internal users and external dissemination.

The quality checks employed in the current DQM system can be grouped into three macro-categories: formal checks, deterministic checks and plausibility checks.

Formal checks represent the first layer of the DQM system and are designed to ensure that (a) the data format is compliant with the technical standards laid out in the reporting instructions and (b) the metadata used to describe reporting concepts are consistent with the representation model and the Statistical Data Dictionary (SDD) of the Bank of Italy[6].

The purpose of deterministic checks is to assess the internal consistency of the reports by verifying that the linear constraints established ex ante in the reporting rules and intrinsic to the "business facts" are fulfilled (e.g. a balance sheet constraint such as that total assets must equal total liabilities). Other deterministic rules may refer to the consistency between stocks and flows or to the simultaneous presence of data points that are reciprocally related (e.g. the number and the amount of credit transfers brokered by a bank) and must therefore be reported simultaneously. In this respect, deterministic checks represent a form of "hard checks" that in most cases detect incorrect data points with a high degree of precision.

Lastly, plausibility checks evaluate data quality according to statistical rules, by isolating values that "substantially deviate" from some usual or expected pattern. They can be regarded as "soft checks" since they

[4] An example of "business fact" could be the amount of credit transfers brokered by a bank; its qualifying features may include the country of the customer and the corresponding institutional sector.
[5] An extract of the matrix model is presented in the Appendix.
[6] As an example, a reporting template could require the amount of credit transfers brokered by a bank, broken down by the country of the ordering client and its institutional sector (general government, domestic/non-domestic households, domestic/non-domestic non-financial corporations, etc.). In this case, formal checks would serve the purpose of ensuring that, for each feature of a given "business fact" (in this case the country and the institutional sector), RAs consistently use the codes indicated in the reporting instructions (e.g. if the reported institutional sector of the client is "domestic households", then the country must be "Italy").

may be violated by data that, for any reason, legitimately depart from the expected reporting behaviour and are then subsequently confirmed by RAs. Plausibility checks employed in the current DQM system compare absolute and/or percentage changes of a given "business fact" between two pre-determined time periods against pre-defined thresholds. Hence, for this type of controls the definition of the thresholds is crucial. The latter are usually set by the data manager on the basis of the variability observed in the data and of time series analysis; the preliminary estimate is then fine-tuned to take account of the various reporting patterns and avoid the generation of too many anomalies, which would be difficult for RAs to accurately cross-check. These types of checks are normally applied to aggregated data. For example, if the raw data refer to the amount of credit transfers brokered by a bank, broken down by the country of the customer and the institutional sector, thresholds will be defined to assess the plausibility of variations of the total number of credit transfer aggregated at the bank level.

While formal checks are always fully implemented, the number of deterministic and/or plausibility quality checks depends on the nature of the data. In data collections where reported information on the different variables tend to be inter-linked through internal relationships (e.g. based on accounting rules), quality checks can be easily tailored to the reporting rules and therefore deterministic checks turn out to be the most efficient. This contrasts with data collections where ex ante relationships among the reported variables are weak. In this case, deterministic checks cannot be implemented and the calibration of the plausibility thresholds becomes crucial. It is also important to note that deterministic checks are typically applied to raw data as sent by RAs, whereas plausibility checks are typically implemented on aggregations of such raw data. Indeed, the direct implementation of plausibility checks on granular data presents two main drawbacks: first, from an operational point of view, the number of data is very high; second, from a methodological perspective, the calibration of the thresholds can be particularly difficult with granular time series as their volatility is typically more pronounced compared with aggregated time series.

Despite these problems, it is very important to apply a DQM system directly to the granular data transmitted by RAs in order to detect anomalies which would otherwise cancel each other out at aggregated level. This, in fact, may bias subsequent analyses conducted at a more detailed level. For example, checking the total amount of credit transfers brokered by a bank may not highlight the case in which a bank erroneously classifies the entire amount of credit transfers under the household sector and sets to zero the amount ascribable to the non-financial corporate sector.

3. DATA ON PAYMENT SERVICES

The above discussion suggests that managing plausibility checks becomes more complex especially in contexts in which the data collected are highly granular and ex ante deterministic relationships are more limited. In this section, we focus on payment service data, which represent a suitable area for exploring the potential of innovative approaches for the development of new plausibility checks at a very granular level owing to its non-accounting nature and the high level of detail of the information reported. We first discuss the main features

of the data collection and its current DQM and then focus on the portion of the data that are analysed in our empirical exercise.

The characteristics of the statistical data collection on payment services are established by Bank of Italy Circulars No. 272/2008 (for credit institutions) and No. 217/1996 (for payment institutions and electronic money institutions), which incorporate the provisions of Guideline ECB/2014/43.

The data cover customers' use of the payment instruments offered by intermediaries: checks, credit transfers, direct debits, card payments, point-of-sale access, online payments, and information on applicable fees and possible fraud schemes. For the various instruments, data on the number of operations and the corresponding amounts are collected. The amount outstanding at the reference date is requested only for the following assets: number of ATMs, bank offices and cards in circulation; all other data are flow data. Overall, about 300 categories of "business facts" are collected, at various frequencies (quarterly, semi-annual and annual) and for a number of classification attributes (e.g. geographical area and sectoral breakdown of the counterparty).

The collected information is used to compile the dataset sent to the European Central Bank (ECB) and is published by various institutions: the ECB (Bluebook)[7], the Italian Banking Association (ABI), the Bank of Italy (e.g. in the Annual Report, payment system publications and in the Bank's Statistical Database, which are all available to the general public). RAs use the aggregated time series to analyse their individual position against the market average.

3.1 Data Quality Management

As already mentioned, the non-accounting nature of payment service data implies the absence of strong deterministic relationships among the variables being collected. Therefore, DQM is performed mainly through simple trend-based plausibility checks (213 out of a total of 335 checks). Furthermore, due to the large number of reported "business facts" and to the high degree of granularity of the information collected, the implementation and maintenance of such checks becomes complex and time consuming. In order to make this process operationally less demanding, for any "business fact" of the reporting template a specific trend-based check is implemented at an aggregated level – namely on data that do not take into consideration the geographical and the sectoral dimension – and the same thresholds are applied to data reported by all RAs. On the one hand, this approach limits the costs associated with a continuous fine-tuning of plausibility checks by the data managers; on the other hand, it does not allow any heterogeneity in reporting patterns across different RAs. In addition, since such checks are carried out only on relatively aggregated time series, nothing can be inferred regarding the quality of the underlying granular data transmitted by RAs. Moreover, based on past experience, the number of generated remark messages and the incidence of "false positives" is far from trivial: in 2017 approximately 14,000 potential outliers were detected and only 30% of them turned out to be genuine errors which were revised by RAs. An undesired implication of the combination of a systematically large number of anomalies and a relatively small percentage of revisions is that RAs can be led to put less effort in

[7] https://www.ecb.europa.eu/paym/intro/book/html/index.en.html

carefully checking all the remark messages sent by the Bank of Italy, potentially jeopardizing the overall quality of the data released to the users.

The above considerations suggest that the granular data collected in the payment service statistical survey represent an interesting area of reporting within the exploration and implementation of more sophisticated DQM methodologies and one that could deliver substantial gains in terms of efficacy and efficiency, owing to the heavy reliance on plausibility checks in the current DQM framework.

3.2 Debit card data

The reporting of debit cards issuance data is carried out with a semi-annual frequency. The information reported by the issuer banks is very detailed and includes (among the other variables) the amounts outstanding for the cards already issued, the payment schemes offered (distinguishing between "only national" and "national and/or international"), the possibility of using the card on ATM and POS systems, the type of chip technology used, and the residence of the cardholder (reported at the level of the provincial capital, of which there are 110 in total). In our dataset we aggregated the elementary observations at the bank-province-semester level: each observation thus indicates the number of debit cards issued by the bank *i*, at the end of the reporting semester *t*, for a given province *p*.

The data falling under the scope of our analysis are extracted from the Bank of Italy's statistical DWH and include data from 2014.H2 to 2018.H1.[8] Our dataset includes all bank-province pairs for which an average of at least 1,000 debit cards was reported over the considered time period. The final sample includes 18,000 observations, reported by 213 banks, accounting for about 97 percent of the total number of debit cards issued by the Italian banking system[9].

When evaluating the quality of the debit card data transmitted by the RAs, it is important to note that this assessment should be carried not only with respect to the percentage or absolute changes in individual data over time, as is done in the current trend-based checks, but also with respect to additional information specific to the reporting entity and which could affect the number of debit cards issued, Examples of such information are the number of bank customers, the type of accounts they hold, the existence of other payment instruments and/or services offered, and the geographical area. Indeed, this piece of information is important in order to capture various potential sources of heterogeneity among the RAs in the observed amounts of debit cards. More specifically, from the Bank of Italy DWH we extracted the following variables. To account for the time varying characteristics of the bank customer base in a given province, we included the variable $depositors_{ipt}$, defined as the log of the total number of depositors resident in a given province, and $perc\_ca_{ipt}$, denoting the percentage of depositors who hold a current account. In order to capture the relevance of a firm's overall

[8] This period was selected to avoid structural changes in the statistical data collection due to changes in the relevant European or national regulations.

[9] Our empirical exercise is finalized by cross-checking the potential outliers with banks. At present, this process is carried out by email and phone. Restricting the focus on the most relevant issuers mitigates the cost associated with such validation while ensuring a high impact on the quality of related statistics.

activity in the payment service market, we included the variable $size_{it}$, defined as the log of the total amounts transacted (both as an issuer and as an acquirer of payment services related to payment cards). In addition, we accounted for the balance between issuing and acquiring services by including the variable $iss\_acq\_ratio_{it}$, computed as the difference between the total amount transacted as an issuer and the total amount transacted as an acquirer, divided by the total amounts transacted; it can take values ranging from -1 (a bank focusing only on acquiring activity) and +1 (a bank focusing only on issuing activity). We also controlled for time effects by including a seasonal dummy indicating the specific semester of observation (*sem*), as well as a trend variable (*trend*) computed as the number of semesters starting from the first period in our dataset (2014.H2). Lastly, we included bank-specific (α_i) and province-specific (μ_p) fixed effects to account for time invariant characteristics of banks and geographical areas.

4. MACHINE LEARNING-BASED CHECKS FOR DEBIT CARD DATA

In this section we illustrate an approach to enhance the DQM of debit card issuance data by improving the flexibility and the efficiency of plausibility checks. We resort to machine learning methodologies to exploit complex statistical relationships in the reported data in order to estimate and update the thresholds to be used to detect the outliers according to purely statistical criteria, without any expert judgment on the part of the data managers.

It is important to emphasize that in the empirical analysis our data have already gone through the current DQM system that, as mentioned in Section 3, is applied on relatively aggregated data. Conversely, the approach described below is intended for implementation on granular data; it then becomes complementary to our current DQM and leads to the identification of additional (potential) outliers.

4.1 Outlier Detection Procedure

According to the definition of outliers originally proposed by Tukey (1977), in the following we will refer to them as data points that lie outside an interval identified by some estimated thresholds. This interval must (a) be robust to the presence of anomalous data points in the data, (b) take into account the characteristics of the reporting entity, (c) adapt to the level of disaggregation of the reported data. In order to define thresholds exhibiting all of these properties, we estimate quantiles of the conditional distribution of the target variable (Koenker and Bassett, 1978; Koenker, 2017). Accordingly, we employ quantile regression to estimate the following general model for quantiles *q* of the target variable *y*, namely the number of debit cards issued:

$$\boldsymbol{q}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{X}) + \boldsymbol{e}$$

where X is an $n \times k$ matrix of observed variables that capture bank features as well as the multilevel structure of our dataset, plus a random disturbance term e. The model is used to derive thresholds that can be used as benchmarks to detect outlier candidates (Meinshausen, 2006).

Quantile regression makes it possible to estimate conditional quantiles of the target variable for a given level of probability τ falling into the interval (0,1):

$$q_\tau(y|X = x) = q_\tau(x) = sup\{y: F(y|X = x) \leq \tau\} \quad (1)$$

where $F(y|X = x)$ is the conditional cumulative distribution function of *y*. Under this definition, $q_\tau(x)$ verifies the following equation:

$$P(y < q_\tau(x)) = F(q_\tau(x)) = \tau. \quad (2)$$

By estimating appropriate pairs of conditional quantile functions it is then possible to create suitable prediction intervals that will include the value of new observations for the target variable; each prediction interval is associated with a given level of probability. The width of the intervals is a function of the predictors and typically reflects the degree of variability of the phenomenon. Hence, such approach provides useful benchmarks when the interest is on the "extreme" values of a distribution rather than on the expected ones. Accordingly, in our approach, control thresholds correspond to suitable prediction intervals for the target variable.

In our analysis, different quantile regression models have been estimated to compute the thresholds: Linear Quantile Regression (LQR), Linear Quantile Regression with Fixed-Effects (LQR FE), and Quantile Regression Forest Model (QRF).

Linear quantile regression models (LQR and LQR FE) can be obtained as solutions, for chosen levels of the τ parameter, of the following minimization problem regarding an appropriate loss function (Koenker and Hallock, 2001; Koenker, 2004; Meinshausen, 2006):

$$\underset{q}{\operatorname{argmin}} \sum_i \rho_\tau \left(|y_i - q_\tau(x_i)|\right) \quad (3)$$

where the subscript *i* denotes the observation and $\rho_\tau(|y_i - q_\tau(x_i)|)$ is the check function defined as

$$\rho_\tau = \begin{cases} \tau * |y_i - q_\tau(x_i)|, & if\ y_i > q_\tau(x_i) \\ (1-\tau) * |y_i - q_\tau(x_i)|, & if\ y_i \leq q_\tau(x_i) \end{cases}$$

The choice of the absolute loss function in (3) implies that the estimation of the quantile functions $q_\tau(x)$ is more robust to the presence of outliers in the target variable than the traditional linear regression function.

A QRF model uses the same steps followed in regression random forests to grow trees (Meinshausen, 2006). However, at each leaf node, it retains all *y* values instead of only the mean of *y* values. Therefore, it keeps a raw distribution of *y* values at each leaf node. Since the conditional cumulative distribution function (CDF) of *y* can be viewed as an expected value

$$F(y|X = x) = P(Y < y|X = x) = E\left(1_{\{Y<y\}}\middle|X = x\right)$$

and the random forest regression approximates the conditional mean by a weighted mean over the observations of the target variable, then the quantile function can be similarly estimated as

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)\, 1_{\{Y<y\}} \quad (4)$$

using the same weights $w_i(x)$ as for random forests. Weights are functions of the number of leaves in each of the trees of random forest (Breiman, 2001).

Our procedure leverages the useful properties of quantile regression in order to identify potential outlier data points in debit card data. More specifically, the process of deriving and identifying the potential outliers consists of the following steps:

- first, we divide the observations in our dataset $[(y_1, x_1), \dots, (y_n, x_n)]$ into a "training" and a "test" set (see below);
- second, a set of conditional quantiles functions $q_\tau(x)$ for the target variable is estimated on the training set, for different values of τ and for the different regression models mentioned above. At this stage, a model selection step is also performed via cross-validation to identify the most suitable model to be applied in the following steps;
- third, for the "new" data points included in the test set, the observed covariates x_{new} are plugged into the estimated quantile functions $\hat{q}_\tau(x_{new})$ and the estimated quantiles values are then combined to define various types of prediction intervals for the observed value of debit cards y_{new} (see section 4.3). The intervals are defined as functions of the set of estimated conditional quantiles, i.e.:

$$[f_1(\hat{q}_{\tau 1}(x_{new}), \dots, \hat{q}_{\tau k}(x_{new}))\, , f_2(\hat{q}_{\tau 1}(x_{new}), \dots, \hat{q}_{\tau k}(x_{new}))];$$

- lastly, potential outliers are identified as test observations whose value for y_{new} falls outside the estimated prediction intervals; the identified anomalies are then cross-checked directly with RAs.

The above procedure can be repeated using a rolling window to take account of new reported data and promptly capture changes in banks' business characteristics and in market conditions that could affect reporting patterns. In particular, in our empirical exercise we followed an "incremental" approach such that the dataset was refreshed twice by taking two snapshots of the data available in the DWH at subsequent points in time (see Table 1). The first snapshot was taken at the end of September 2017 and included the data reported for the reference period December 2014-June 2017. This initial dataset was split into a training and a test set: the former comprised observations spanning from December 2014 to June 2016 (9,177 observations) and was employed for model selection; the latter included observations reported in the period December 2016–June 2017 (4,704 observations) and was employed to compute prediction intervals with the selected model and identify the outliers to be cross-checked with the RAs[10].

[10] Although quantile regression is robust to the presence of outliers in the target variable, our estimation procedure may be affected by observations representing outliers with respect to the predictors of our models. Such outliers are unlikely to occur in our dataset since most of the predictors are related to accounting data that are typically characterized by high

After the first round of data validation, a second snapshot of the data was carried out at the end of September 2018 by including new observations for the period December 2017–June 2018. At this point, the new training set included the observations for the timespan December 2014-June 2017[11] (13,771 observations) and was employed to re-train the model previously selected. The updated model was used to validate the observations in the new test set comprising the observations for the two additional semesters (4,229 observations).

The proposed approach has two main advantages: first, the estimated thresholds, or equivalently the prediction intervals, are "tailored" to the characteristics of the RAs; second, the thresholds "adapt automatically" over time as new values for the characteristics of RAs are observed.

Table 1. Sample observations (training and test set)

	First snapshot (December 2014-June 2017)			Second snapshot (December 2014-June 2018)		
	Train	Test	Total	Train	Test	Total
N. observations	9,177	4,704	13,881	13,771	4,229	18,000
Percentage	66%	34%	100%	77%	23%	100%

4.2 Estimation and model selection

As explained in the previous Section, our analysis aims at estimating prediction intervals for the number of debit cards reported by each entity, in a given province, at a specific point in time. In the training phase, the quantile functions used to define the intervals are computed by estimating the following general quantile regression models:

$$q_\tau(x_{ipt}) = f(depositors_{ipt}, perc\_ca_{ipt}, size_{it}, iss\_acq\_ratio_{it}, trend, sem, \alpha_i, \mu_p) \quad (5)$$

where $q_\tau(x_{ipt})$ denotes the τ-quantile of the target variable conditioned on a set of attributes x_{ipt} observed for bank *i*, in province *p*, at time *t* (see Section 3.2 for a detailed description of the attributes). To compute prediction intervals associated to different probability levels, we estimated conditional quantiles corresponding to different τ, namely: 0.01, 0.25, 0.75, and 0.99.

As illustrated in section 4.1, the general model (5) was estimated for each τ following various approaches. Firstly, we considered traditional parametric linear quantile regression models (LQR and LQR FE) where each conditional quantile is obtained by minimizing the loss function (3) (Koenker and Basset, 1978).

reliability. For example, customer accounts represents the basis to generate accounting reports that are monitored through deterministic rules in the current DQM system. Furthermore, this information needs to be strictly monitored by reporting entities themselves for anti-laundry purposes.

[11] The outliers detected in the first round that were not revised at the time of the second snapshot were dropped from the refreshed training set.

Secondly, quantile functions were estimated by employing the QRF algorithm, which is an extension of the random forest algorithm that allows a robust, non-linear and non-parametric estimation of empirical conditional quantiles[12] (Meinshausen, 2006). Similarly to random forest models, QRF requires the specification of both the number of trees in the forest and a parameter (*mtry*) indicating the number of variables to be considered at each split during the construction of the trees. The QRF model was estimated by setting the number of trees at 500 and considering a grid of possible values for the *mtry* parameter.

To implement model selection, in the first snapshot of our procedure we estimated quantile functions for various τ-quantiles of interest using data in the training sample and comparing the performance of the various models by computing the average of the empirical loss in (2) via a ten-fold cross-validation. To preserve the longitudinal structure of our dataset – which includes data for each bank, in each capital province, for various semesters – we performed cross-validation block-wise. Each fold was obtained initially by sampling unique bank-province pairs; the final folds were then obtained by including all the original observations corresponding to the selected pairs. The model with the lowest average loss across the various reference quantiles was selected and employed to estimate prediction intervals in the test sample.

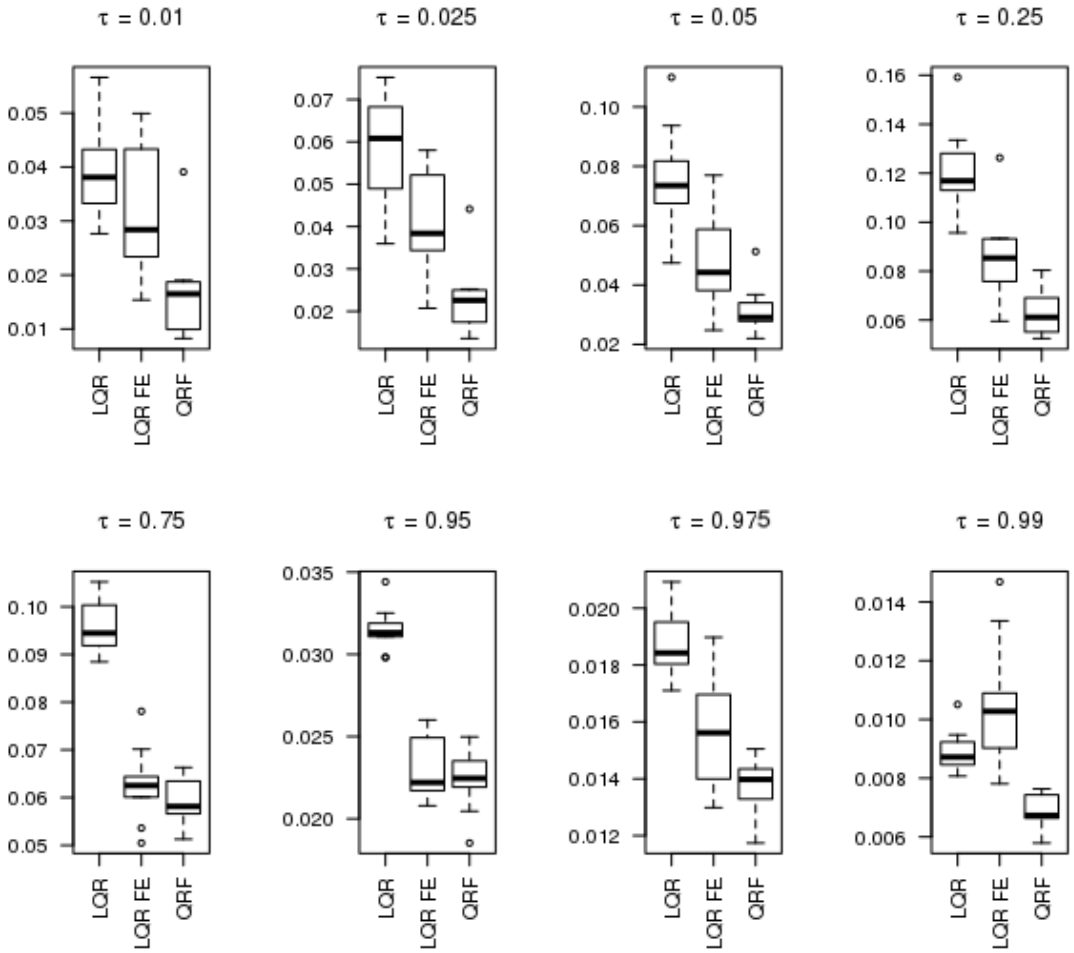
The results of the cross-validation performed in step one are presented in Figure 1. For various quantiles a box plot for the average loss computed for the observations iteratively left out is reported for the two estimation approaches described above, namely linear quantile regression with fixed effects (LQR FE) and quantile regression forest (QRF)[13]. For benchmark purposes, the results for a (simple) baseline model consisting of a parametric quantile regression without fixed-effects (LQR) are also reported.

Both the LQR FE and the QRF model show lower average loss compared with the benchmark, suggesting that the inclusion of fixed effects improves the predictive performance. However, for the 0.99 quantile, the performance of LQR FE model is worse on average compared with the benchmark. The QRF model is the one that reproduces the empirical distribution of observed data more closely, especially with respect to the lower tails of the distributions; instead, for the upper quantiles, the average performance of the two approaches is rather similar. Furthermore, LQR FE appears to be characterized by a greater variability in performance especially for the extreme tails of the distribution (τ equal to 0.01, 0.025, 0.975, 0.99), those of greatest interest to our analysis. Based on these measures of performance, the QRF model was selected and employed to estimate prediction intervals for data validation.

[12] The joint estimation of the full conditional distribution implies that QRF has the advantage of preventing the problem of quantile crossing that may instead affect approaches that estimate each quantile function separately.

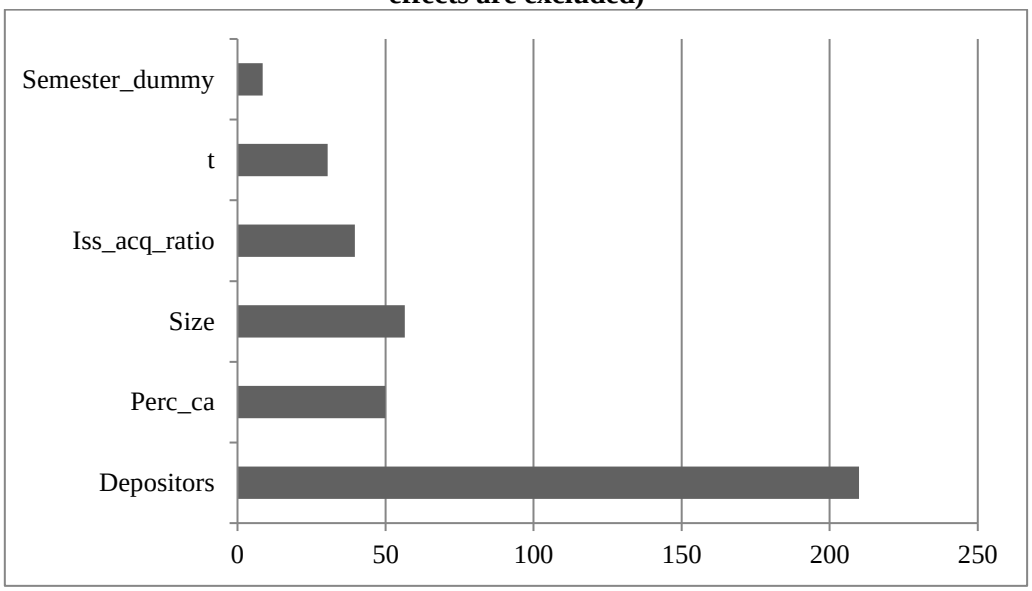
[13] For the QRF model results correspond to a value of the *mtry* parameter equal to the square root of the number of predictors. This value was selected over a grid of alternative values based on the average cross-validation loss.

Figure 1. Box plots for the average cross-validation empirical loss for different quantiles and models



The relative importance of the variables included in the selected QRF model is reported in Figure 2. More specifically, for each variable in the model (excluding dummies for bank and province fixed effects) we report the average increase in mean squared error recorded over all the trees when the values of that variable are permuted, normalized by the standard deviation of all differences. As expected, the geographical distribution of bank customers seems to be by far the most important predictor, followed by the variables accounting for customer characteristics and the bank business model. Not surprisingly, variables related to time effects are less relevant; this could be due to the fact that the time span of our analysis is relatively short (only eight semesters).

Figure 2. Measure of the importance of the variables included in the quantile regression forest (fixed effects are excluded)



Variable importance for the QRF model.

4.3 Identification and validation of potential outliers

After training the selected model, prediction intervals were estimated for the observations in the test set in order to perform data validation. Specifically, acceptance thresholds corresponding to the following prediction intervals were computed as:

$$I_1(x) = [q_{0.01}(x), q_{0.99}(x)] \quad (6)$$

$$I_2(x) = [q_{0.025}(x), q_{0.975}(x)] \quad (7)$$

$$I_3(x) = [q_{0.25}(x) - 1.5 \cdot (q_{0.75}(x) - q_{0.25}(x)), q_{0.75}(x) + 1.5 \cdot (q_{0.75}(x) - q_{0.25}(x))] \quad (8)$$

The potential outliers were cross-checked by directly contacting RAs; Table 2 summarizes the results.

Table 2. Cross-checking of potential outliers with the RAs

Prediction intervals	**I1**	**I2**	**I3**
a-Total number of potential outliers	**373**	**489**	**457**
b-Anomalies detected and revised ("true positives")	289	312	292
c-Confirmed observations ("false positives")	84	177	165
d-Precision (=b/a) (percentage)	**77.5%**	**63.8%**	**63.9%**

As for the interval defined in equation (6), 373 potential outliers were identified, 289 of which turned out to be incorrect data ("true positives") and, as such, were revised by the RAs; the associated precision, measured by the ratio of the number of true positives and the sum of true and false positives (i.e. the number of detected outliers whose values were confirmed by banks and that were not revised), is equal to 77.5 percent. Moving to

the narrower interval defined in equation (7), the number of both potential outliers and true positives increased to 489 and 312, respectively, although precision declined to 63.8%; similar results in terms of precision were obtained on the basis of the interval defined in equation (8), based on a combination of quartile functions. These results show that the choice of the reference quantiles to construct prediction intervals is an important aspect of the proposed methodology because it has an impact on the total number of anomalies, true positives and false positives.

In brief, the choice of the optimal width for the prediction intervals is context-specific and should take into account the costs of accidentally validating an anomaly compared with the benefits of having an additional outlier revised in the dataset. Moreover, it is important to note that in the context of our application we do not have any information on the anomalies that are not detected by the model and therefore we are not able to compute performance measures based on false negatives.

5. CONCLUSIONS

This paper investigates whether machine learning techniques can enhance the current approach to DQM for statistical data collections. In the literature on the application of machine learning techniques within central banks, this research area is still relatively unexplored. In principle, it could foster major improvements in the efficiency and effectiveness of DQM, in particular when applied to granular databases characterized by weak pre-determined relationships (accounting, logical, mathematical) among the collected data.

From a methodological point of view, we implemented a supervised learning algorithm capable of learning complex data patterns from past reporting behaviour and we calculated output prediction intervals for the target variable, with these intervals representing acceptance thresholds for the reported data. This approach yields flexible control thresholds that are tailored to the characteristics of individual reporting entities (such as the customer base, the geographical location or the business model) and are automatically updated as new data are reported.

The above methodology has been applied to validate debit card issuance data reported by Italian banks and included in the semi-annual payment services data collection. The algorithm was trained with data reported in previous periods and tested by validating the identified outliers directly with the reporting entities.

The results show that the machine learning procedure was able to detect additional anomalies compared with those identified by the current plausibility checks. Overall, the two approaches to DQM can be regarded as complementary: the current approach is being carried out on relatively aggregated data while the new proposed approach can be applied to granular data. Moreover, although a formal comparison between the results of the two approaches is inappropriate (see below), it is worth mentioning that the level of precision (given by the number of “true” outliers as a percentage of the total number of potential outliers detected by the procedure) reached by the new approach is much higher than the one observed for the existing trend-based checks. Finally, the new methodology is a purely statistical and automated one, and this makes the overall process less time-consuming and more suitable for the DQM of large and granular datasets.

The agenda for future research is quite rich. First, the robustness of the current approach will be thoroughly assessed by monitoring the performance of the algorithm over time. Second, when relying on the automated validation process, a new approach is required for communicating the resulting remark messages to the reporting entities. In fact, in the context of the current DQM procedure, the execution of plausibility checks for a given RA is based exclusively on data included in the same statistical data collection and referred to the individual reporter, and this makes the communication relatively straightforward. In contrast, the new approach isolates outliers that are inconsistent with the reporting patterns observed for the population of RAs and includes data reported in multiple data collections.

From a methodological standpoint, the complexity of the problem and the richness of possibilities offered by the variety of machine learning techniques leave room for further developments to be addressed in the future. First, the procedure described in the paper leverages a supervised learning approach designed to predict plausible ranges for a given target variable. A possible extension is to adopt unsupervised learning approaches capable of learning more complex data patterns and to identify observations that are outliers in the multi-dimensional space defined by the different variables representing different "business facts". Second, another possible approach could be to exploit information on the characteristics of outliers provided by the current DQM system in order to build a classification algorithm that is able to discriminate between plausible and outlier data points. A final note by way of conclusion: in our analysis we accounted for the hierarchical structure of the data by adopting a dummy variable approach; future extensions of our work should employ more sophisticated methods specifically designed to model spatial and individual effects (e.g. mixed-effects regression models or quantile boosting).

REFERENCES

Bishop, C.M. (2007), "Pattern Recognition and Machine Learning", Springer.

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

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

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

Del Vecchio, V., Di Giovanni, F., Pambianco, S. (2007), "The Matrix Model. Unified model for statistical data representation and processing". Available at https://www.bancaditalia.it/statistiche/raccolta-dati/sistema-informativo-statistico/.

Farnè, M, Vouldis, A.T. (2018), "A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks' supervisory data", ECB Working Paper No. 2171.

Froeschl K.A., Grossmann W, Del Vecchio V., (2003), "The Concept of Statistical Metadata", European Commission Information Society Technologies Programme, MetaNet Project, Deliverable 5.

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

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

Koenker, R. (2004), "Quantile regression for longitudinal data", *Journal of Multivariate Analysis*, 91, 74–89.

Koenker, R. (2017), "Quantile regression 40 years on", *Annual Review of Economics*, 9, 155-176.

Koenker, R., Bassett, G. (1978), "Regression Quantiles", *Econometrica*, 46 (1), 33-50.

Koenker, R., Hallock K.F. (2001), "Quantile Regression", *Journal of Economic Perspectives*, 15 (4), 143-156.

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

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

APPENDIX

The Bank of Italy's statistical information system (SIS) is based on two main conceptualizations (Del Vecchio et al., 2007): the first describes the general architecture of an SIS (conceived as a hierarchy of models), the second is related to the generic model devoted to defining the statistical data and the operations to be performed on the overall data, which define the so called Matrix Model.[14]

In the following table an excerpt of the matrix model for the data referred to issued debit cards is reported.

2.3 Payment services: debit cards in circulation	Code	Subcode	Maturity	Currency	Residence of counterparty	Province of counterparty	Chip technology	Frequency
ATM on national circuits:	58620	02	3	1	1	x	x	S
ATM and POS on national circuits:	58620	04	3	1	1	x	x	S
ATM on both national and international circuits:	58620	06	3	1	1	x	x	S
ATM and POS on both national and international circuits:	58620	08	3	1	1	x	x	S

The rows indicate the "business fact" to be reported, while the columns indicate the breakdowns (in terms of variables) applicable to each specific collected item (as described in the rows). The "x" indicates that a specific breakdown is applied to a given row.

A description of the values corresponding to each column is reported below:

Maturity: 3= "not defined or irrelevant"

Currency: 1= euro (this value is reported only for measures that are expressed as amounts)

Residence of counterparty: 1= "Owner resident in Italy"

Province of counterparty: it is the entry for Italy in ISO 3166-2, which defines codes for the names of the provinces.

Chip technology: "micro.chip", "other".

Frequency: S= "semi-annual".

[14] The model gets its name from the graphic representation used to define the data structure, which is a matrix of rows and columns.