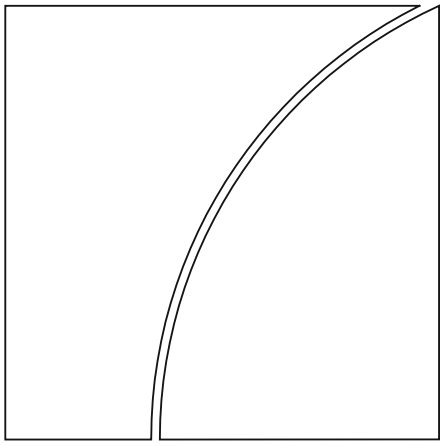
Irving Fisher Committee on Central Bank Statistics



IFC Working Papers

No 22

Statistical matching for anomaly detection in insurance assets granular reporting

by Vittoria La Serra and Emiliano Svezia

October 2022

BANK FOR INTERNATIONAL SETTLEMENTS

IFC Working Papers are written by the staff of member institutions of the Irving Fisher Committee on Central Bank Statistics, and from time to time by, or in cooperation with, economists and statisticians from other institutions. The views expressed in them are those of their authors and not necessarily the views of the IFC, its member institutions or the Bank for International Settlements.

This publication is available on the BIS website (www.bis.org).

ISSN 1991-7511 (online)
ISBN 978-92-9259-599-9 (online)

Statistical matching for anomaly detection in insurance assets granular reporting

Vittoria La Serra, Emiliano Svezia[1]

Abstract

Since 2016, insurance corporations report granular asset data in Solvency II templates on a quarterly basis. Assets are uniquely identified by codes that are required to be kept stable and consistent over time; nevertheless, due to reporting errors, unexpected changes in the codes may occur, causing inconsistencies when compiling insurance statistics. The paper addresses this issue as a statistical matching problem and a supervised classification approach is proposed to detect such anomalies. Test results show the potential benefits of machine learning techniques on data quality management processes and the efficiency gains arising from automation, especially during situations of constraints on human resources, as the ongoing pandemic.

Keywords: insurance data, data quality management, record linkage, statistical matching, machine learning.

JEL classification: C18, C81, G22

Contents

[1] Bank of Italy, Vittoria.Laserra@bancaditalia.it, Emiliano.Svezia@bancaditalia.it, Statistical Data Collection and Processing Directorate. The views expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy.

This paper was presented at the 11th Biennial BIS-IFC Conference on "Post-pandemic landscape for central bank statistics" and received the award for the best paper presented by a young statistician.

Introduction and motivation

In the process of collecting, processing and disseminating statistics, an effective and efficient data quality management (DQM) is of paramount importance in order to ensure the high quality of data. The automation of DQM processes became crucial in presence of increasingly granular databases. Furthermore, since the beginning of the Covid-19 pandemic in 2020, it has become clear the importance of investing resources in making such processes as much automatic as possible, in order to enhance their resilience in presence of situations of human resource constraints.

In the statistical literature, machine learning models are emerging as important tools to approach the DQM on very granular data in an automated way, since they generally outperform traditional modelling approaches in prediction tasks (Chakraborty *et al*., 2017). Restricting the issue to central banks statistics, the Bank of Italy has already applied successfully several machine learning methods to specific DQM processes (see Buzzi *et al*., 2020, Cusano *et al*., 2021, Zambuto *et al*., 2021, Maddaloni *et al*., 2022) and further research in this field is ongoing.

This paper proposes a machine learning approach in a statistical matching framework to solve - in an accurate and efficient automated way - a DQM issue on insurance granular assets data, specifically to check for anomalies in identification codes (ID) reporting. More in detail, the assets' IDs are expected to remain unique and consistent over time, meaning that the IDs assigned by the insurance corporations (ICs) cannot be subject to changes throughout the reporting history of the assets. However, from quarter to quarter, unexpected changes in the code for the same asset can occur. This is either due to annual updates of the requirements, which imply a change for assets' codes, or, more often, it is a consequence of reporting errors. Therefore, an insurance corporation might either consciously revise an asset's code, following requirements from an updated version of the regulation, or erroneously change it to a new, different code than the one used in the previous reporting quarter. Either ways, such changes have important consequences on the work of supervisory authorities and central banks, since they signal a change in the reported assets that in practice has not occurred; this raises DQM issues when analyzing assets' time series and compiling the IC statistics that are then disseminated.

The paper is organized as follows. Section 1 describes the data from which the dataset used in the analysis is derived, presenting its structure and details on the Italian case. In Section 2, a record linkage approach based on machine learning models for classification is proposed; different models are tested on an Italian dataset and a robust and high-performance random forest approach is chosen, for whom results are presented. Section 3 summarizes the main conclusions, showing the advantages of the proposed approach and opening new ground for future research.

1. Data description

Since 2016 European insurance corporations (ICs) report to their national supervisory authority, according to the *Implementing Technical Standards* (ITS)[2] drawn by EIOPA, and to the national central banks, quarterly data on their individual balance sheets. The data is organised in templates according to the *Solvency II Directive*[3]. They provide very granular and highly valuable information especially with template S.06.02 which contains asset-by-asset information on the single holdings of insurance corporations, showing the investments in debt securities, equity and investment fund shares, as well as loans, deposits and properties.

Template S.06.02 allows, on the one hand, supervisory authorities to perform a comprehensive and detailed risk assessment upon insurance undertakings and, on the other, central banks to compile statistics on insurance sector, useful to analyse its interconnections within the financial system and to gather knowledge on households' wealth and income from insurance policies. This template is used both for supervisory and statistical purposes, enriched with specific for the latter.

More in details, template S.06.02 comprises quantitative information on each position held, such as the market and nominal value, quantity and accrued interest of the asset, along with qualitative features, which include – wherever applicable – the type of insurance undertaking[4], the type of asset[5], the issuer and/or counterparty sector, the issuer and/or counterparty area, the currency, the issue and maturity dates, the name of the issuer, the description of the asset.

Each asset in the template is reported with an identification code. Asset codes are standardized in most cases (e.g. ISIN codes for securities), although in some cases insurance corporations can assign internal codes and report them (CAU, Code Attributed by the Undertaking).

The dataset used in the paper consists of Italian data from the S.06.02 template and it is also integrated with attributes from the European Central Bank's Centralised Securities Database (CSDB) - the European harmonised security registry. In detail, the population of Italian ICs is composed of around 100 entities; overall, reported data comprises almost 30 reporting quarters and around 70,000 assets at each period.

On average, in each quarter, there is a turnover of 8% in number and 4% in market value share for the reported assets[6]. In line with the reporting instructions, asset ID codes that are only reported in one of two adjacent quarters should consist in new purchased or sold assets. However, these also include the cases of changes in the codes, whose exact percentage in the data is unknown. Therefore, 8% can be taken as the maximum expected percentage of cases of anomaly, which highlights

[2] *Commission Implementing Regulation* (EU) 2015/2450 of 2 December 2015 and following amendments, laying down implementing technical standards with regard to the templates for the submission of information to the supervisory authorities, according to Directive 2009/138/EC of the European Parliament and of the Council.

[3] Directive 2009/138/EC of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance.

[4] Life, non-life, composite and reinsurance.

[5] As from now CIC (Category Identification Code).

[6] Between two adjacent quarters, the turnover of the assets is the percentage of brand new reported codes (not reported in the first quarter) over the total reported assets in the second quarter.

how, even having a limited impact on the general data quality, errors in ID codes cannot be neglected.

2. The proposed approach

2.1 A record linkage problem

Statistical matching techniques, as described in D'Orazio *et al*. (2006), have the objective to draw information from two (or more) different datasets by linking them with respect to some common observed variables. Such techniques were originally proposed with the aim of data integration, i.e. to link two (or more) datasets coming from independent surveys and build a richer dataset containing information from both (Okner, 1972).

A specific case of statistical matching is record linkage, which is applied when the statistical units in two datasets are supposed to be at least partially overlapping (D'Orazio *et al*., 2006) and the objective of the analysis is to identify the list of common units between the two.

The topic was first introduced and formalized by Fellegi *et al*. (1969) and a classical approach was then proposed by Jaro (1989). Since mid-Nineties, many applications of record linkage have concerned the issue of linking historical census data, as in the works of Ferrie (1996), Rosenwaike *et al*. (1998) and Ruggle (2002). Different methodologies for performing record linkage have later been proposed, such as mixture models (Larsen *et al*., 2001) and Bayesian approaches (Fortini *et al*., 2001; Tancredi *et al*., 2011).

More recent contributions to this topic make use of machine learning techniques (Feigenbaum, 2016, Rijpma *et al*., 2020), which is the framework of the current work. In fact, the issue of unexpected changes of ID codes in insurance data introduced above can be approached as a record linkage problem.

As described in Section 1, each reported asset in a quarter is identified by a unique ID code and comes with a set of reported features. If a change in an asset's ID code occurs, so that an insurance corporation I reports asset "a" in quarter Q_t and recodes it as "b" in quarter Q_{t+1}, it is expected for the reported features of the two apparently different assets "a" and "b" to be the same[7], since they actually refer to the same asset. Comparing the reported features of the two assets is therefore necessary to assess whether their difference in ID code is in fact an anomaly, stemming from an unexpected change that has taken place.

As in a record linkage framework, two datasets of assets, each referring to two adjacent reporting quarters $\{Q_t, Q_{t+1}\}$, can be compared to assess whether there are common units between the two datasets, where each unit is an asset, identified by its ID code and its reporting IC.

[7] The features selected in the dataset are "structural". For this reason, they should remain stable except for limited statistical reclassifications that can occur.

Assets in the two quarters are compared with respect to the observed features and such comparisons are carried out using distance measures, one for each feature's type, whether categorical (nominal or ordinal), numerical or textual.

Nominal variables, such as reporting/counterparty sector or issuer/counterparty area, are compared with an overlap measure (Boriah *et al.*, 2008), taking 0 as a measure of minimum distance if the reported values are equal and 1 otherwise; ordinal variables, such as the categorized maturity date, are compared via the Manhattan distance, while numerical variables, as the assets' market value, are compared using a Euclidean distance; lastly, textual variables as the assets' description are compared using a Levenshtein measure for strings. All distance measures are normalized to take values in the interval $[0, 1]$.

Each pair of assets, one reported in Q_t and one reported in Q_{t+1}, can either be a "match" if the two assets share the same ID code or a "non-match" if they do not.

A comparison matrix is built, as reported in Table 1. Each row in the matrix refers to a pair of assets from the two adjacent quarters and each column refers to an observed feature, either nominal, ordinal, numerical or textual, for which a distance measure $d(.)$ is chosen. The $d_f(a, b)$ distance measure for two assets a and b on a feature f is a value calculated in $[0,1]$, where the endpoints respectively indicate minimum and maximum distance between the two observed values for feature f.

The features in the comparison matrix are used as input (covariates) to supervised statistical models together with the status of each pair that is a binary target variable with values "match" or "non-match" to be predicted.

Input to supervised models: the comparison matrix and the target variable Table 1

			COMPARISON MATRIX			
Asset codes		TARGET VARIABLE	Nominal	Ordinal	Numerical	Textual
Q_t	Q_{t+1}	Status	$i \in \{1 \dots n_i\}$	$j \in \{1 \dots n_j\}$	$k \in \{1 \dots n_k\}$	$w \in \{1 \dots n_w\}$
a	a	Match	$d_i^1(a,a) \dots d_i^{n_i}(a,a)$	$d_j^1(a,a) \dots d_j^{n_j}(a,a)$	$d_k^1(a,a) \dots d_k^{n_k}(a,a)$	$d_w^1(a,a) \dots d_w^{n_w}(a,a)$
a	b	Non-match	$d_i^1(a,b) \dots d_i^{n_i}(a,b)$	$d_j^1(a,b) \dots d_j^{n_j}(a,b)$	$d_k^1(a,b) \dots d_k^{n_k}(a,b)$	$d_w^1(a,b) \dots d_w^{n_w}(a,b)$
b	b	Match	$d_i^1(b,b) \dots d_i^{n_i}(b,b)$	$d_j^1(b,b) \dots d_j^{n_j}(b,b)$	$d_k^1(b,b) \dots d_k^{n_k}(b,b)$	$d_w^1(b,b) \dots d_w^{n_w}(b,b)$
…	…	…	…	…	…	…

For the application, two subsets are selected from the whole Italian database, including all reported assets from the two subsequent quarters 2021-Q1 and 2021-Q2, and the observed arrays of features on pairs of reported assets are compared by building a comparison matrix; with the goal of detecting the changes in ID codes that ICs have reported between the two quarters, comparison is only made for pairs referring to the same ICs. Even with this constraint, the number of rows in the matrix, i.e. the number of compared pairs, approaches 150 million units. Given the size of the dataset, it would be impossible for data analysts to manually check all pairs of assets.

The comparison matrix is built and afterwards split into a "training set" and a "test set", respectively including 80% and 20% of the data. Moreover, both datasets are proportionally stratified with respect to the "asset type"[8] feature, in order to obtain a representative dataset.

The two datasets are respectively used to train and test supervised classification models, to predict the status variable in the comparison matrix from the computed distances between features.

2.2 Model selection

Four widely used supervised classification models are considered: logit, bagging, random forests and neural networks.

The logit model is adopted as a benchmark, being a high-performing yet easy-to-interpret classifier (Feigenbaum, 2016). Among machine learning classification models, bagging (Breiman, 1996), random forests (Breiman, 2001) and neural networks (Bishop, 1995) are considered and respective hyperparameters are tuned for each model: number of bootstrap samples for bagging, numbers of trees and variables randomly sampled as candidates at each split (*mtry*) for random forest, number of nodes in the hidden layer for neural networks.

In order to assess robustness, the models are trained and tested multiple times on differently unbalanced data with respect to the binary target variable. More in detail, the models are trained and tested on comparison matrices having $p\%$ cases of match and $(1-p)\%$ cases of non-match, with p ranging from 1% (extreme unbalance) to 50% (perfect balance).

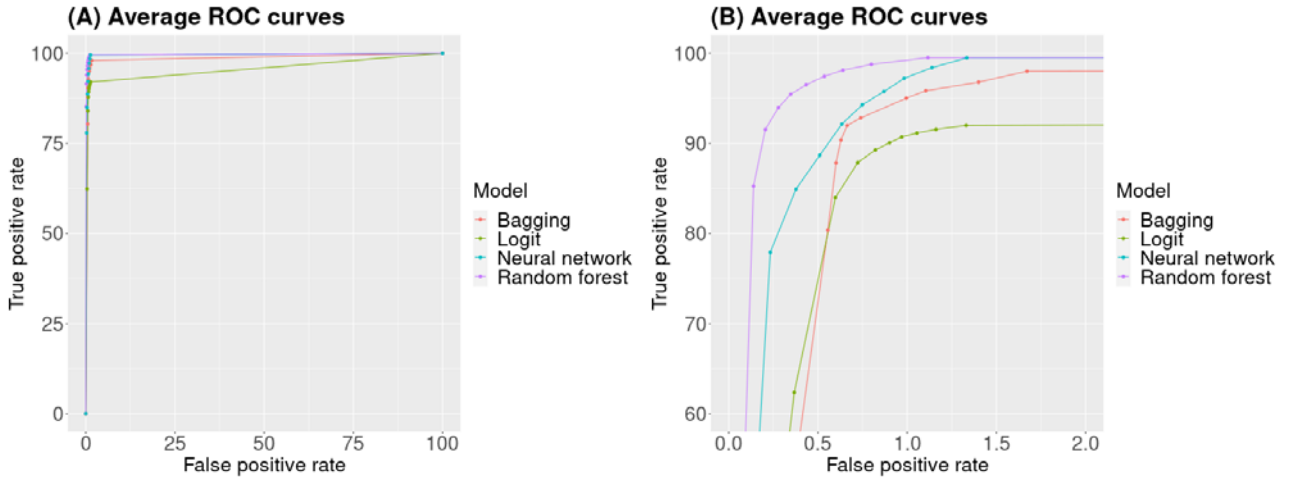
The test results are presented averaging over the results obtained on test sets unbalanced with different p.

Average Receiver Operating Characteristics (ROC) curves are shown in Figure 1, built through the computation of average false positive and true positive rates, varying with the probability threshold for classification.

[8] Third digit of the CIC code. Ten classes are considered.

Average ROC curves for the four tested models

Figure 1



Panels A and B both show the average ROC curves for the four models; panel A shows the whole curves, while panel B focuses on a smaller range, to better spot the differences in the models.

Area Under Curve (AUC) indexes for the average ROC curves are presented in Table 2; this index is the chosen criterion to assess each model's best hyperparameters combination[9].

AUC indexes for the tested models

Table 2

Model	AUC index for average ROC curves
Logit (benchmark)	95.66%
Bagging	98.62%
Random forest	99.64%
Neural network	99.52%

The results reported in Table 2 show that all four models perform well on the tested data, with a minimum AUC measure of 95.66% observed for the logit model and a maximum AUC of 99.64% observed for the random forest. As expected (Breiman, 2001), among the random-trees-based models, random forest outperforms bagging.

The superiority of the random forest model against the others can be clearly observed in Figure 1 - panel B, where its ROC curve always shows larger true positive rates over false positive rates, for all probability thresholds for classification, with respect to the other models. For instance, with a 1% false positive rate, the random

[9] The best hyperparameters combination for the three machine learning models are 100 bootstrap samples for bagging, 200 trees and 7 *mtry* for random forest, 30 neurons in the hidden layer for neural networks.

forest can achieve a 99% true positive rate, while the neural network, bagging and logit models respectively achieve 98%, 95% and around 91% rates.

2.3 The results

The random forest model is selected among the tested ones, in relation of its superiority in the robustness analysis held in Subsection 2.2.

More results for the selected model are shown in this Section for an unbalance proportion p of 5%, chosen for illustrative purpose, being smaller than 8%, namely the maximum expected unbalance proportion in the dataset (see Section 1).

Performance measures for the model are presented in Table 3, varying with the probability threshold for classification. The presented indexes are accuracy, balanced accuracy, true positive rate (TPR), true negative rate (TNR), false discovery rate (FDR) and the difference between true positive rate and false discovery rate.

Model performance indexes for the random forest selected model ($p = 5\%$)

(percentage values)

Table 3

	Probability threshold for classification									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Average
Accuracy	99.05	99.41	99.51	99.55	99.56	99.54	99.46	99.35	99.16	99.4
Balanced accuracy	99.27	98.99	98.71	98.36	97.98	97.52	95.91	94.38	92.27	97.04
TPR	99.52	98.53	97.82	97.05	96.23	95.27	91.96	88.86	84.62	94.43
TNR	99.02	99.45	99.59	99.68	99.73	99.77	99.86	99.9	99.93	99.66
FDR	15.72	9.52	7.3	5.87	5.02	4.43	2.82	2.07	1.53	6.03
TPR-FDR	83.79	89.02	90.53	91.17	91.21	90.83	89.13	86.78	83.09	88.39

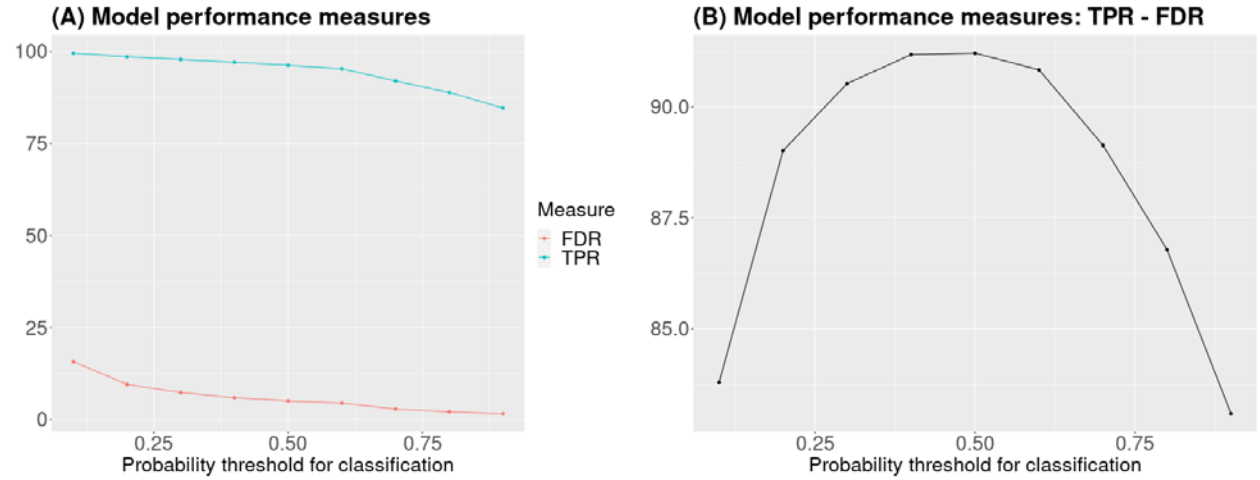
General model performance measures such as accuracy and balanced accuracy show that the model correctly identifies true cases of match and true cases of non-match with high frequencies, for all probability thresholds. Mean values for the two measures are respectively 99.40% and 97.04%.

True negative rate (TNR) shows very good results, remaining stably around a 99% value for all thresholds. Making 95% of the unbalanced test set, the cases of non-match are naturally easier to be detected by the classification model.

With the aim of choosing the best probability threshold to consider in the model, true positive rates (TPR), false discovery rates (FDR) and the difference between the two are also presented, both in Table 3 and in Figure 2.

Benjamini *et al*. (2001) define FDR as the "*expected proportion of false discoveries among the discoveries*". Given a binary confusion matrix, FDR is the percentage of negative cases that the model incorrectly classifies as positive cases; in the current analysis, FDR coincides with percentage of non-matching assets that are erroneously classified as matches by the model and it is therefore an interesting cost measure to minimize.

Model performance measures (TPR, FDR, TPR-FDR) Figure 2



As observable in Table 3 and Figure 2 - panel A, both the true positive rate and false discovery rate slowly decrease with the probability threshold.

For instance, taking the lowest threshold for classification, the selected model ensures a 99.52% rate of correctly classified cases of match, with the consequence of a 15.72% incorrectly classified cases of non-match; instead, taking the highest threshold, less than 2% false discoveries are made but only 84.62% cases of match are correctly identified.

Therefore, a trade-off between the two indexes must be found in order to choose an appropriate probability threshold for the model. However, the two rates do not have the same weight in the current analysis: although false discovery rate is a cost measure to minimize, maximizing the true positive rate is considered as a priority for the model effectiveness. Detecting most of the true cases of match is in fact the goal of the analysis and it is therefore desirable to select a lower probability threshold for classification which would ensure to reach the goal, even if that implies that some cases of non-match are erroneously classified as matches.

To assess for the best threshold, the difference between TPR and FDR is calculated and reported in Table 3 and Figure 2 - panel B. The maximum value for the index is reached on a 0.5 threshold; however, the maximum increase in the index is observed when switching from threshold 0.1 to 0.2. The latter ensures a 98.53% true positive rate, with the cost of a 9.52% false discovery rate. Therefore, in the current analysis the threshold value chosen for the model is 0.2.

In light of the presented results, with the goal of identifying the anomalous cases of changes in assets' ID codes between two quarters, assuming that the percentage of such changes in a quarter is 5% of all reported assets, a random forest model results as the best choice. In fact, among the tested models, the random forest ensures large accuracy and balanced accuracy. Moreover, selecting a probability threshold for classification of 0.2, the best model provides a true positive rate around 99% and a cost of a 9% in terms of false discovery rate that is considered acceptable in the DQM process.

3. Conclusions and further developments

Annual updates of the requirements or errors in insurance reporting can cause unexpected and undesirable changes in the reported asset codes, from quarter to quarter. An automated method to detect such changes is necessary to improve the data quality of insurance statistics, which are published at international level, given the size of the available data, the level of granularity on the single assets and the unneglectable impact that such changes have on compiled statistics.

A record linkage approach is proposed to reach the goal, making use of supervised machine learning classification models.

Real Italian data from 2021 are used for the application; four models are considered, i.e. logit model as a benchmark, bagging and random forests as random-tree-based machine learning models and neural networks. Robust results are presented, testing the models on differently sampled data, stratified on varying percentages of cases of changes in the codes in two adjacent quarters, since the true proportion of such anomalies in the data is currently unknown with precision.

The tested models show good performance in terms of average AUC and results show the superiority of the random forest model to approach the problem with respect to the other tested classifiers.

Assuming a 5% proportion of anomalies in the data and taking a 0.2 probability threshold for classification, the selected random forest model shows good performance for all measures of interest, both in terms of effectiveness and efficiency, ensuring large accuracy and balanced accuracy, with around 99% rate of correctly identified cases of changes in ID codes (TPR), accepting the cost of a false discovery rate that approaches 9%.

The presented test results give a robust estimate of the improvement in data quality that would derive from running the selected model on production data, with the goal of successfully identifying cases of unexpected and unwanted changes in the ID codes. However, the actual performance in production of the proposed methodology must be validated through the cross-check with the insurance corporations of the estimated cases of changes in a quarter during a real data production round; in that occasion TPR and FDR could decrease and increase, respectively, from the test results presented in the paper.

In the future, the model training phase might be improved by considering all the available Italian data, not only focusing on two reporting quarters but analyzing all couples of subsequent reporting quarters since 2016, in order to gain a larger amount of information on the assets. This possibility has not been explored yet due to the computational effort needed to elaborate all historical data available.

Moreover, further analyses might be conducted to evaluate the performance of the classifiers on different "asset type", since the results might vary depending on that feature and potentially might be improved by training different models for specific category of assets.

Finally, in the future, the presented approach might be extended to a symmetrical data issue on the ID codes, such as the reuse of the same code for two different assets in two subsequent quarters, which is again an unexpected behavior in the reporting of the assets' ID codes that can attempt at the quality of the data.

References

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., Golani, I. (2001) "Controlling the false discovery rate in behavior genetics research", *Behavioural Brain Research* 125, 279–284.

Bishop, C. M. (1995), "Neural Networks for Pattern Recognition", *Oxford University Press*.

Boriah, S., Chandola, V., Kumar, V. (2008), "Similarity Measures for Categorical Data: A Comparative Evaluation", SIAM International Conference on Data Mining.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45, 5-32.

Breiman, L. Bagging (1996), "Predictors", *Machine Learning*, 24, 123-140.

Buzzi, M. R., Costanzo, G., Di Lucido, M., La Ganga, B., Maddaloni, P., Svezia, E., Zambuto, F., Papale, F. (2020), "Quality checks on granular banking data: an experimental approach based on machine learning", *Questioni di Economia e Finanza* 547.

Chakraborty C., Joseph A. (2017), "Machine Learning at Central Banks", Bank of England Staff Working Paper, No. 674.

Cusano, F., Marinelli, F., Piermattei, S. (2021), "Learning from revisions: a tool for detecting potential errors in banks' balance sheet statistical reporting", *Questioni di Economia e Finanza* 611.

D'Orazio, M., Di Zio, M., Scanu, M. (2006), "Statistical Matching, Theory and Practice", Wiley.

Feigenbaum, J. (2016), "A Machine Learning Approach to Census Record Linkage", Working paper.

Fellegi, I., Sunter, A. (1969), "A theory for record linkage", *Dominion Bureau of Statistics*.

Ferrie, J. P. (1996), "A new sample of males linked from the public use micro sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census manuscript schedule", *Historical methods: a journal of quantitative and interdisciplinary history*, Vol. 29, 141-156.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001), "On Bayesian record linkage", *Research in Official Statistics*, 4: 185–198.

Jaro, M. (1989), "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida", *Journal of the American Statistical Association*, 84: 414-420.

Larsen, M. D., Rubin, D. B. (2001), "Iterative automated record linkage using mixture models", *Journal of the American statistical association*. 96:453, 31-41.

Maddaloni, P., Continanza, D. N., del Monaco, A., Figoli, D., di Lucido, M., Quarta, F., Turturiello, G. (2022), "Stacking machine-learning models for anomaly detection: comparing AnaCredit to other banking datasets", *Questioni di Economia e Finanza*, n. 689.

Okner B. (1972), "Constructing a New Data Base from Existing Microdata", *Annals of Economic and Social Measurement*, Volume 1, number 3.

Rijpma, A., Cilliers, J., Fourie, J. (2020), "Record linkage in the Cape of Good Hope Panel", *Historical methods*, vol. 53, no. 2, 112-129.

Rosenwaike I., Hill, M. E., Preston, S. H., Elo, I. T. (1998), "Linking death certificates to early census records: the African American matched records sample", *Historical methods: a journal of quantitative and interdisciplinary history*. Vol. 31, 65-74.

Ruggles, S. (2002), "Linking historical censuses: a new approach", *History and computing*, vol. 14, 213-224 (2002).

Tancredi, A. and Liseo, B. (2011), "A hierarchical Bayesian approach to record linkage and population size problems", *Annals of Applied Statistics*, 5: 1553–1585.

Zambuto, F., Arcuti, S., Sabatini, R., Zambuto, D. (2021) "Application of classification algorithms for the assessment of confirmation to quality remarks", *Questioni di Economia e Finanza* 631.