BANCA D'ITALIA
EUROSISTEMA

A decision-making rule to detect insufficient data quality

an application of statistical learning techniques to the non-performing loans banking data

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

Banca d'Italia

Motivation

- It is key to count on **an efficient and effective monitoring of the quality level of the data** transmitted by Reporting Agents (RAs) in order **to provide users with high-reliable data to carry out thorough and robust analyses.**

- **Data Quality Level (DQL) generally follows a positive trend** thanks to subsequent corrections submitted by RAs; however, **a data quality worsening may occur** especially when data production is affected by exogenous and unpredictable events, such as **RAs' IT malfunctions**, **changes in the reporting requirements** or **operative tensions and staff shortage** (also as seen during the pandemic).

- The aim of the study is to define a decision-making rule:
  - to speed up the **detection of DQL worsening;**
  - to provide a **synthetic measure of the DQL**.

Data quality cycle in Banca d'Italia

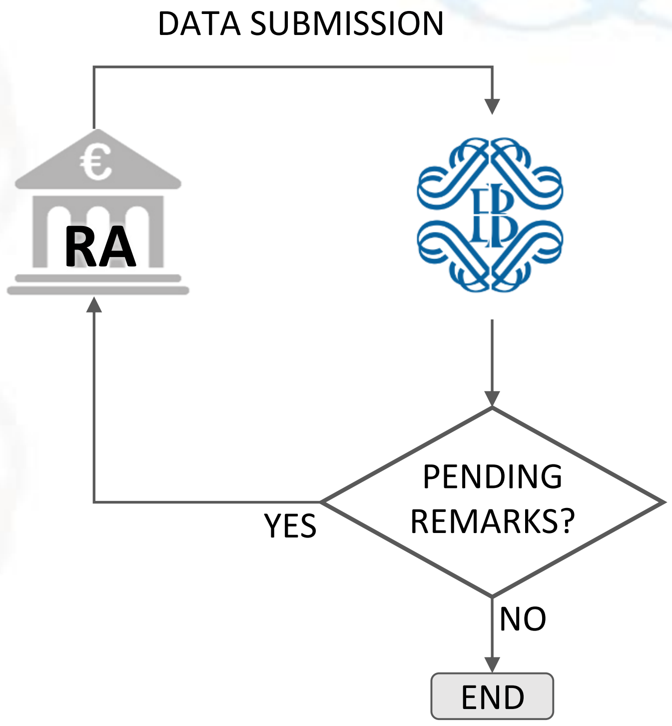
- At each data submission, the reliability of the data is assessed upon arrival by the Banca d'Italia by using a **set of automatic Data Quality Checks** (DQCs).

- A severity level from 0 to 10 is assigned to each DQC.

- When a DQC detects plausible errors (outliers) or deterministic errors, **remarks** are sent to the RA to request for:
  - **corrections** of erroneous data by sending a new data submission

  or

  - **confirmations** of the data. These can be, in turn, accepted or refused by the Data Manager



Current decision rule to release data to users

- Based on the severity level of the DQC, the generated **remarks are classified as “serious” and “non-serious”**. **If at least 1 serious remark is generated, the data submitted are kept on hold to be examined by the Data Manger** (hence not immediately released to the users)

AT EACH DATA SUBMISSION

ONLY NON-SERIOUS REMARKS →	RELEASED
AT LEAST A SERIOUS REMARK →	NOT RELEASED

- Considering 2 subsequent data submissions sent by an RA for a specific reference date, the possible cases are as follows:

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

- **In cases A, C and D**, the Data Manager’s decision is **straightforward;** in case B the *$(k+1)^{th}$* submission may worsen the DQL.
- **The proposed decision-making rule is applied to case B** to detect the unexpected worsening of the DQL.

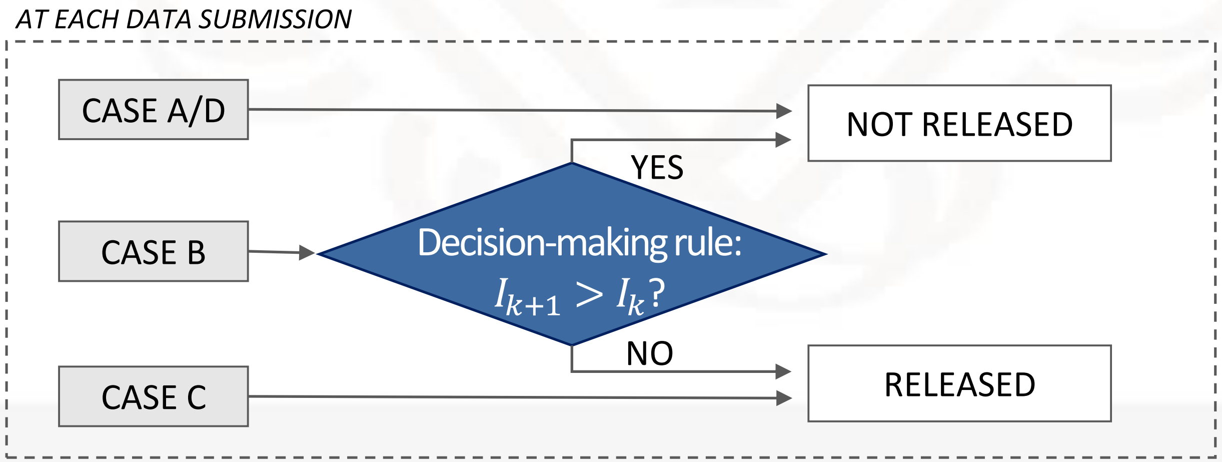
Definition of the proposed decision-making rule

- The proposed rule is based on a synthetic data quality indicator computed through past evidence from the Data Quality Management (DQM) activity:
  - **number of remarks** (**R**) generated by the **DQC *c***
  - **severity level** (***τ***)
  - **number of confirmations** (**Conf**)

- Definition of **a synthetic data quality indicator $\mathbf{I_k}$** for the *k^th^* data submission sent by an RA for a specific reference date:

$$I_k = \sum_c \tau_c \cdot \left(R_{c,k} - Conf_{c,k}\right)$$

- If the DQC detects deterministic errors precisely (non-confirmable DQCs), $Conf_{c,k}$ is by construction equal to 0.
- **The higher the value of $\mathbf{I_k}$, the lower the DQL** of the *k^th^* data submission.
- The proposed decision-making rule is defined as follows:



What quantities are available for the calculation of I_k and I_{k+1}?

- Let us assume we want to compare the DQL of the *(k+1)th* data submission with the DQL of the *kth*. Once the *(k+1)th* data submission is received, the **availability of the information for the calculation of I_k and I_{k+1}** is as follows:

	I_k	I_{k+1}
Number of remarks	✓	✓
severity level	✓	✓
Number of confirmations	✓	✗

- The **number of confirmations related to remarks, generated by the confirmable DQC c for the *(k+1)th* data submission, is estimated**:

$$\widehat{\text{Conf}}_{c,k+1} = \sum_{r=1}^{R_{c,k+1}} \widehat{\text{Conf}}_{c,k+1,\,r} \qquad where: \quad \widehat{\text{Conf}}_{c,k+1,\,r} = \begin{cases} 1, & if\, p\left(Conf_{c,k+1,\,r}\right) > cut-off \\ 0, & otherwise \end{cases}$$

- ***cut-off*** is a threshold lying within (0, 1) assessed with a cross-validation method
- The **estimation of the probability *p(Conf)*** is derived applying **machine learning techniques** to a dataset including remarks generated by confirmable DQCs actually observed in the previous reference dates.

Dataset and Model selection

- **Dataset: Banks Non-performing loans dataset (NPL)**, collected by Banca d'Italia on a biannual basis
  - over 17 million of records between 30*th* June 2017 and 30*th* June 2019
  - about 65K remarks generated, of which 5,083 by confirmable DQCs
  - 15 dummy variables for DQCs and 415 for RAs
  - numeric variables: differences among quantitative aggregates of remarks, number of records sent and reference dates
- **Model selection: the logistic regression model outperforms.**

	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 • *from June 2017 to December 2018* • *4,643 remarks*	**Accuracy**	0.83	0.83	0.73	0.73	0.73	0.73	0.50
	Recall	0.95	0.92	0.99	0.94	0.99	0.98	0.39
	Precision	0.83	0.86	0.73	0.75	0.73	0.73	0.83
	Negative predictive value	0.80	0.73	0.55	0.50	0.54	0.54	0.33
Validation set • *June 2019* • *440 remarks*	**Accuracy**	0.81	0.78	0.76	0.75	0.78	0.75	0.78
	Recall	0.94	0.85	0.97	0.90	0.99	0.94	1.00
	Precision	0.83	0.87	0.78	0.80	0.78	0.79	0.78
	Negative predictive value	0.62	0.51	0.18	0.39	0.50	0.33	NA

Sources: NPL dataset – Banca d'Italia

Application of the decision-making rule

- Considering the subsequent submissions of the case B, the decision-making rule allows the Data Manager to **automatically and promptly identify cases where the DQL decreases** and it **prevents the users to use non-fit-for-use data**.

Reference dates between years 2017 and 2018

		$(k+1)^{th}$ submission		
		Not-released	Released	Total
k^{th} submission	Not-released	269	407	696
	Released	51	275 (Case B)	326
	Total	320	682	1,002

Results of the decision rule for the *$(k+1)^{th}$* submissions of Case B	Percentage
Released submissions	89%
Additional Not-released submissions	11%

Reference date of June 2019

		$(k+1)^{th}$ submission		
		Not-released	Released	Total
k^{th} submission	Not-released	15	23	38
	Released	1	14 (Case B)	15
	Total	16	37	53

Results of the decision rule for the *$(k+1)^{th}$* submissions of Case B	Percentage
Released submissions	93%
Additional Not-released submissions	7%

Sources: NPL dataset – Banca d'Italia

Conclusions

- The proposed decision-making rule **improves the current DQL monitoring** by promptly detecting additional cases of DQL worsening.

- The synthetic data quality indicator I_k provide a **synthetic measure of the overall quality of data** transmitted by the RAs.

- **The decision-making rule is accurate**. It was assessed by comparing its results with the outcome resulting from an application of the decision-making rule based on the real status of the remarks confirmability: in 97% of cases the conclusions coincide.

- The proposed method can be **flexibly applicable to various data collections**.

- For the NPL dataset, the **implementation of the decision-making rule** in the Banca d'Italia's collection system is **ongoing**.

BANCA D'ITALIA
EUROSISTEMA

Thank you for your attention!

barbara.laganga@bancaditalia.it