Irving Fisher Committee on
Central Bank Statistics

BIS

IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks"

Missing values imputation for Central Balance Sheets Data Office (CSBO) accounting data[1]

Iker González Crespo and Pablo Jiménez Segovia,
Bank of Spain

[1] This contribution was prepared for the IFC Satellite Seminar held at the ISI 64th World Statistics Congress, co-organised with the Bank of Canada in Ottawa, Canada, on 15 July 2023. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Canada, the BIS, the IFC or the other central banks and institutions represented at the event.

Missing values imputation for Central Balance Sheet Data Office

Iker González Crespo

Pablo Jiménez Segovia

Abstract

The scope of this paper is to impute missing accounting data in the Central Balance Sheet Data Office (CBSO) questionnaires.

Financial and non-financial companies are obliged to deposit their annual accounts and other administrative records, which are used for statistical purposes. The Bank of Spain receives this information from the non-financial companies through the Commercial Registry. Once this information is gathered an automatic filtering is applied in which missing values are identified so they cannot be used for economic analysis. Therefore, it is proposed to impute this information not to lose these questionnaires, and therefore gain a larger sample (and representativeness of Spanish companies). In order to achieve this goal a series of Machine Learning (ML) algorithms based on random forest have been used to predict these accounting items. Afterwards, an accounting adjustment has been made so that all these imputations balance accounting. Finally, AI explainability techniques have been used to understand the variables that have most influenced when making the imputations.

Keywords: Central Balance Sheet Data Office, CBSO, SME, ML, AI, imputation, miceranger, random forest, R, accounting adjustment, explainability

JEL classification: C53 (Forecasting and Prediction Methods - Simulation Methods), M41 (Accounting).

1. Introduction

The Central Balance Sheet Data Office (CBSO) of the Bank of Spain annually receives and cleans data, primarily accounting data, from nearly one million non-financial companies.

Since the questionnaires through which information is received from the Mercantile Registry sometimes have details of missing information in fields necessary for proper cleansing, and given that this is one of the reasons why approximately 20% of the questionnaires are discarded to date, it was proposed to investigate whether modern statistical techniques could impute those "missing" values. That is the goal of this project.

Different imputation algorithms have been studied, and as mentioned in Section 3.2, the algorithm called "miceranger" was chosen, based on chained random forests.

Once the values have been imputed, in a second iteration, they must balance in an accounting way so that the completed questionnaires pass the subsequent validations carried out in the CBSO. To achieve this, it is necessary to force this reconciliation by applying business rules (the accounting relationships

that exist between the reported items) since no current algorithm, as far as the authors of this article are aware, imputes and reconciles simultaneously.

To carry out this work, a Machine Learning model has been fitted for each information node in which the companies subject to imputation are classified. The nodes are formed as the intersection of each sector of activity (14 groups of activities in the monograph) with each size (according to EU Recommendation). The sizes of the analysed companies are: large, medium, small and micro-enterprise.

In 2019, the Central Balance Sheet Data Office began a proof of concept in collaboration with the Instituto de Ingeniería del Conocimiento (IIC), where a similar imputation was tested, focusing only on four totals (short-term debts, long-term debts, debtors, and creditors). One of the intentions of the current work has been to broaden that approach for four reasons:

- Increase the number of items that can be imputed.
- Attempt to incorporate the accounting logic of relationships between items, beyond their specific totals, to help the algorithm learn. For instance, the "Cash" item could influence another item in a different total, such as "Long-term financial investments."
- This more ambitious path can be extended to include items from the "Income and Expenses" section.
- The accounting explainability of the algorithm and the patterns that can be detected may open doors to deeper insights into accounting relationships.

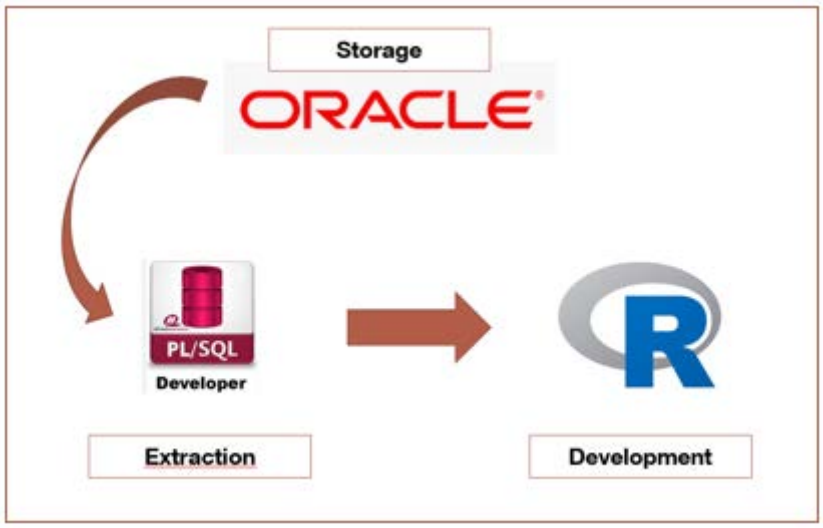
To sum up, this work is the outcome of a combination of data science (extraction, processing, data wrangling, and deployment of Machine Learning models) and business knowledge (understanding of accounting items, questionnaire frequencies, ranges of possible values, and accounting adjustments). After imputing the missing values, an analysis is conducted to assess the correctness of these imputations at an aggregate level, meaning by node, by sector, and by size.

2. The data

A database called GTB (which is the database of questionnaires received in an environment accessible to data scientists) is available for the years 2018 and 2020, hosted in Oracle. This database contains accounting data and assessments of questionnaire quality (reconciliations, checks, reliability of monetary units, cleansing dates, and other metrics analysing the status of questionnaires at each point in time).

This project is conducted for a specific type of questionnaire: those originating from the Mercantile Registry (CBB) and of the "reduced" type (i.e., those used by companies with fewer than 50 employees). In CBB, there is only one questionnaire for each company in each fiscal year.

To connect with the database, the PL/SQL tool is used, and the desired information is extracted, stored in a directory, and accessed using the R programming language. In our case, we select questionnaires that are perfectly filled out (passing all the quality controls used by the Balance Sheet Central, marking the company as consistent and suitable for use in studies). A procedure for removing outliers (an internal process developed in the CBSO) is applied to these questionnaires to discard anomalous cases. Although they comply with all business accounting rules, they have values that exceed the normal value of their node and are eliminated to avoid biases in the studies. These data serve as both the training set (train) and as a reference for algorithm verification (test).

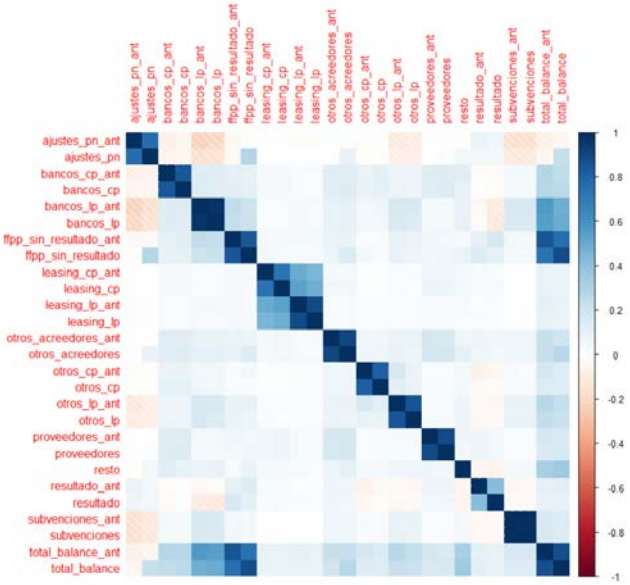


The data to be imputed corresponds exclusively to the balance sheet items. An example is provided to illustrate the type of information we have in the questionnaires for the asset categories of Non-Current Assets and Current Assets:

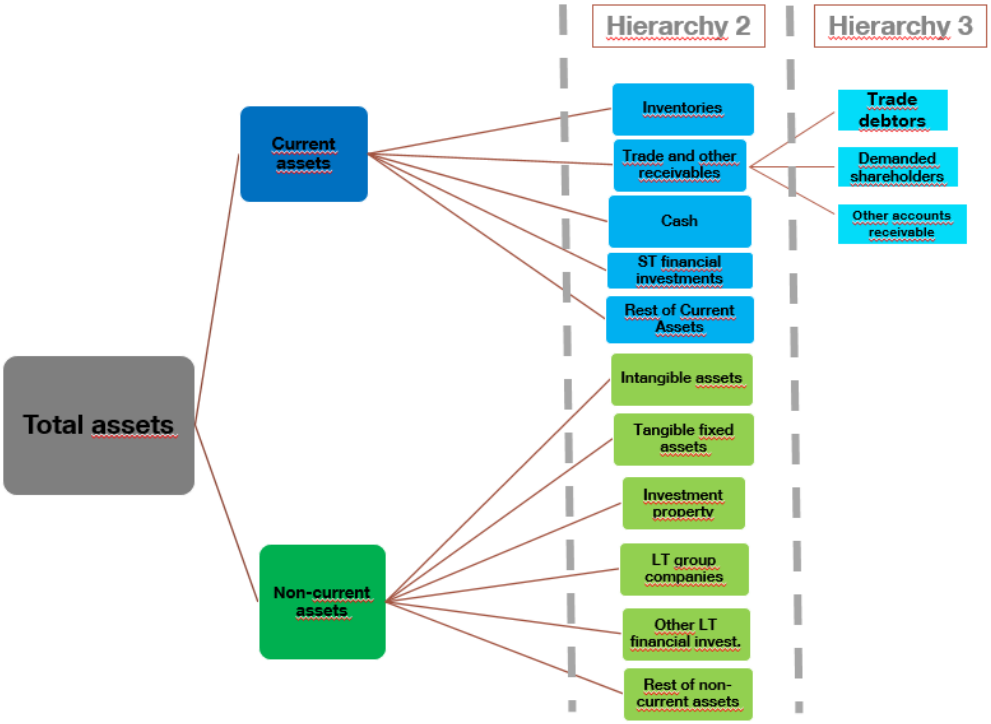
ACTIVO		NOTAS DE LA MEMORIA	EJERCICIO 2020 (2)	EJERCICIO 2019 (3)
A) ACTIVO NO CORRIENTE	11000		424,71460	424,38489
I. Inmovilizado intangible	11100			0,02288
II. Inmovilizado material	11200		356,34557	366,62683
III. Inversiones inmobiliarias	11300			
IV. Inversiones en empresas del grupo y asociadas a largo plazo	11400			
V. Inversiones financieras a largo plazo	11500			
VI. Activos por impuesto diferido	11600		68,36903	57,73518
VII. Deudores comerciales no corrientes	11700			
B) ACTIVO CORRIENTE	12000		1.173,05580	1.068,56633
I. Activos no corrientes mantenidos para la venta	12100			
II. Existencias	12200		527,76339	499,86441
III. Deudores comerciales y otras cuentas a cobrar	12300		501,45542	420,93524
1. Clientes por ventas y prestaciones de servicios	12380		490,71850	396,46022
a) Clientes por ventas y prestaciones de servicios a largo plazo	12381			
b) Clientes por ventas y prestaciones de servicios a corto plazo	12382		490,71850	396,46022
2. Accionistas (socios) por desembolsos exigidos	12370			
3. Otros deudores	12390		10,73692	24,47502
IV. Inversiones en empresas del grupo y asociadas a corto plazo	12400			
V. Inversiones financieras a corto plazo	12500			19,47000
VI. Periodificaciones a corto plazo	12600			
VII. Efectivo y otros activos líquidos equivalentes	12700		143,83699	128,29668
TOTAL ACTIVO (A + B)	10000		1.597,77040	1.492,95122

Note that the value of keys is available for each of the accounting items for the current year (2020 in this example) and the previous year (2019 in this example). Some of the items have been grouped into variables called "Rest" as they are considered too infrequent for algorithms to learn from (having very little variability).

When conducting analyses, data from the previous year's items has been taken into account. This is because having data from the previous year is considered an important variable for imputing values for the current fiscal year, and vice versa. The following image shows the correlation with the previous year's items.



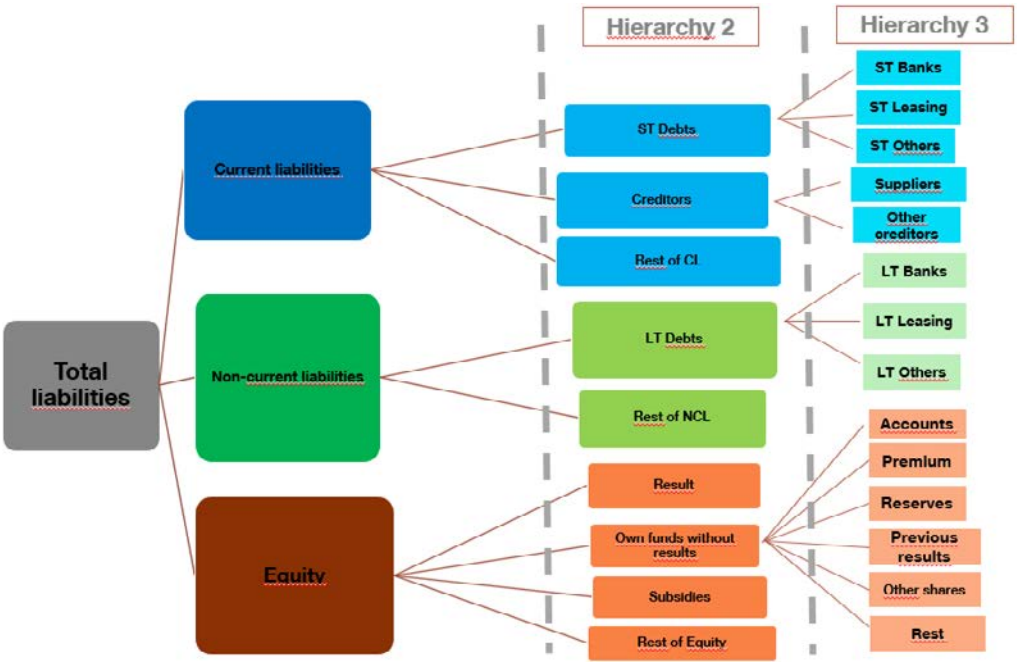
For the current year and the previous year, the following relationships of accounting items for assets exist:



The values of accounting items that are totals (total assets, current assets, and non-current assets) will always be provided due to the technical requirements of the CBSO. The rest of the accounting items are the ones that we will impute and then perform an accounting adjustment to ensure that their sum matches the total of their respective totals.

In the case of assets, we select questionnaires that have all positive accounting items since all their values must be greater than or equal to 0.

On the other hand, liabilities have the following relationships:



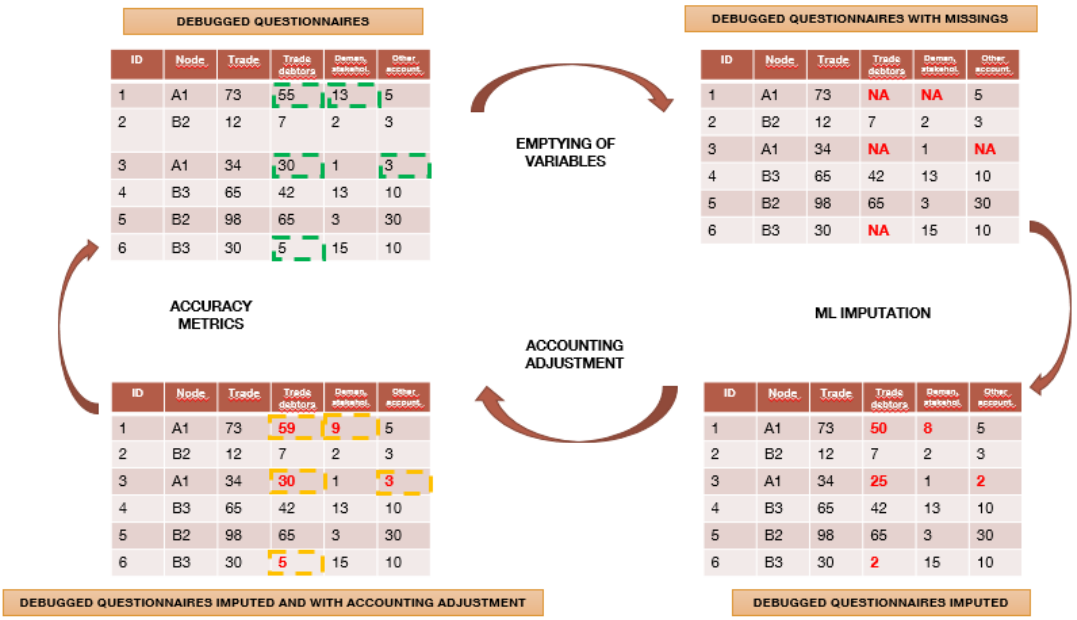
Where the items that are totals, namely, Current Liabilities, Non-Current Liabilities, Equity, will always be reported. The rest of the accounting items are the ones that will be attempted to be imputed. Long-Term Debt, Short-Term Debt, Creditors, and Own funds without results are also total items that must be taken into account when reconciling the corresponding sums.

In the case of liabilities, there are items that must always be positive, while others may be negative at times (such as the result for the year). We only consider questionnaires where those accounting items that must necessarily be positive are positive, with the aim of preventing the algorithm from learning from errors.

Additionally, it was decided not to impute items that are very infrequent, i.e., items that exist in only, for example, 1% of cases. This decision is based on three main reasons:

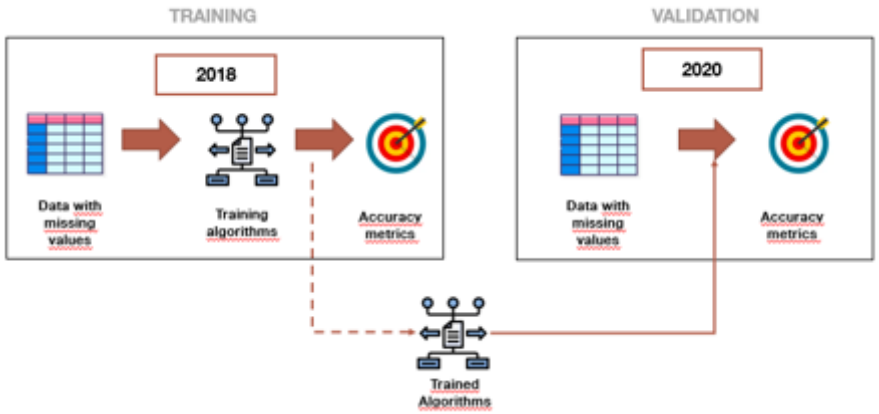
1. Artificial intelligence algorithms do not learn well from variables with little variability. In other words, they struggle to learn patterns effectively. This falls within the realm of "variable selection."
2. Having fewer items reduces the complexity of the development.
3. It seems reasonable not to impute a value for an item that, in practice, is almost always zero.

3. Training and validation method



This imputation scheme is carried out for each node (business sector and company size) as it is believed that the behaviour between each of them may differ. Therefore, an algorithm will be obtained for each node.

Data from the year 2018 is used to train the algorithm and select the one with fewer errors. The data from 2020 serves as validation to check if the final algorithm can generalize well for other years.



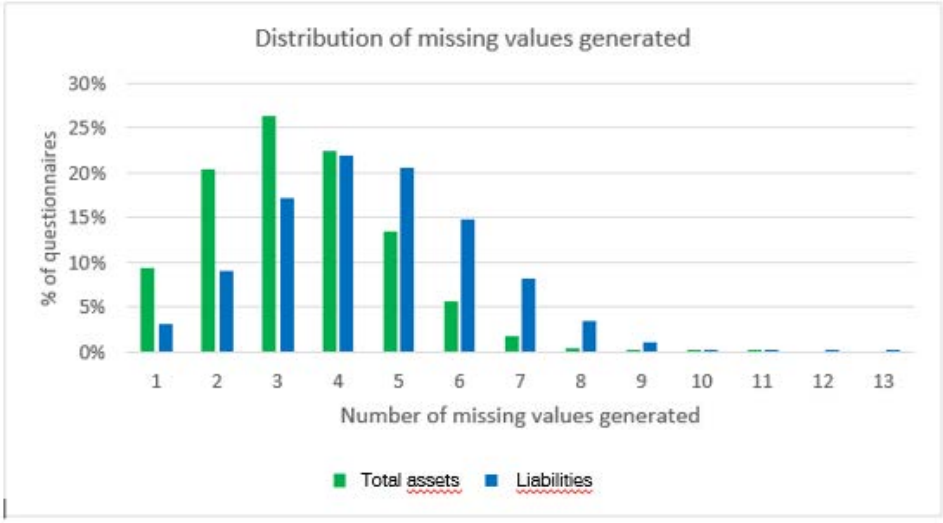
It is important to emphasize that real data from the questionnaires is not modified. In other words, the data reported in the questionnaires is accepted as valid and is not imputed.

Additionally, it is worth mentioning a step performed before imputation: if the "Result" item is reported in the Income Statement and does not appear in the Balance Sheet, that "Result" is not imputed but is transferred directly. This prioritizes the actual data received over imputations.

Going into more detail on each of the steps in the methodology:

3.1 Variable emptying

For both the year 2018 and the year 2020, there are perfectly reconciled and publishable questionnaires for the CBSO. Thirty percent of the items are emptied in a set of over 800,000 questionnaires used in the model. In 2020, the distribution of missing values generated for each questionnaire is as follows:



Where we can see that the distribution of the number of missing values does not concentrate on any specific item, and moreover, the number of blanks generated in the questionnaires is distributed randomly.

This emptying is decided to be done randomly because there is no intention to bias the algorithms, and it is assumed that in production, any combination of missing values may occur. In other words, no hypothesis is made about the patterns of missing values that the algorithm should learn.

3.2 Imputation algorithm

Due to the high computational cost required for this project to read, process, and launch various ML algorithms, a workstation was used to reduce project times.

Different algorithms of various natures were tested to find the one with the best fitting metrics. Among others, the following were tested:

- KNN: a distance-based algorithm. Different metrics and "k" values were tested to obtain the best fit. The distance calculation significantly slowed down the execution of these algorithms. The results were good but below the performance of other algorithms.
- MICE: a classic algorithm for imputation using chained equations. This algorithm proved to be computationally expensive, and the results were also below those of other algorithms.
- Missranger: an iterative algorithm based on random forest. This algorithm achieved very good results but did not explicitly return a model for interpretability. References to this model can be seen here aquí.

- Miceranger: another iterative algorithm based on random forest. The operation of the algorithm is described in the following link vínculo. Two variants of the algorithm were tested (with and without Predictive Mean Matching).

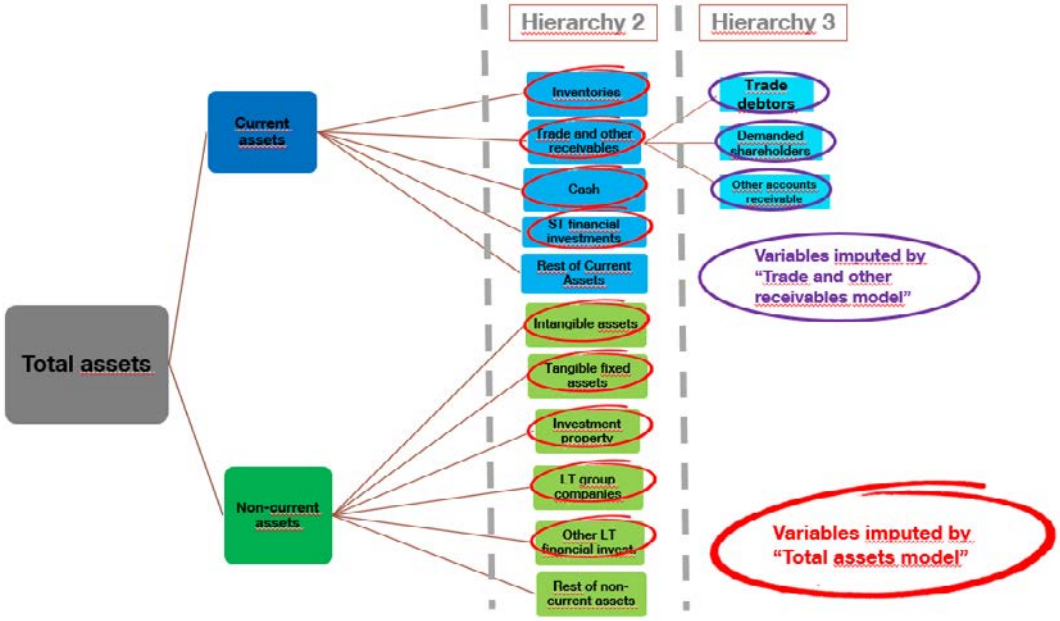
Ultimately, the chosen model was Miceranger without PMM (Predictive Mean Matching) since its fitting metrics were the best among the previously mentioned algorithms. This algorithm works very well because it is parallelizable (faster in execution) and is also an interpretable model (which is necessary for explaining its results). The fitting metrics that have been analysed (some aggregated, some accounting item-specific, and some questionnaire-specific) are as follows:

Type of metric	Metric	Definition
Aggregated	**Total aggregate variation**	sum of all predicted values divided by the sum of all actual values
Aggregated	**Total aggregate variation, broken down by variable**	sum of all predicted values divided by the sum of all actual values, broken down by variable
Accounting item	**Average error**	average of the errors committed in all accounting items
Accounting item	**Average error, broken down by variable**	average of the errors committed in all accounting items, broken down by variable
Accounting item	**Error deviation**	standard deviation of errors made in all accounting items
Accounting item	**Error deviation, broken down by variable**	standard deviation of errors made in all accounting items, broken down by variable
Accounting item	**Maximun error**	maximum error committed in all accounting items
Accounting item	**Correlation**	similarity of the structure of actual items versus imputed items
Accounting item	**% of accounting items with error**	% of accounting items where there has been an imputation error
Questionnaire	**% of questionnaire items with error**	% of questionnaires where there has been an imputation error in at least one accounting item

The training is carried out with questionnaires from 2018. During training, the procedure we use is to train 5 Miceranger models for each node and select the model with the lowest Out Of Bag (OOB) error, a common measure in this type of algorithm based on random forest.

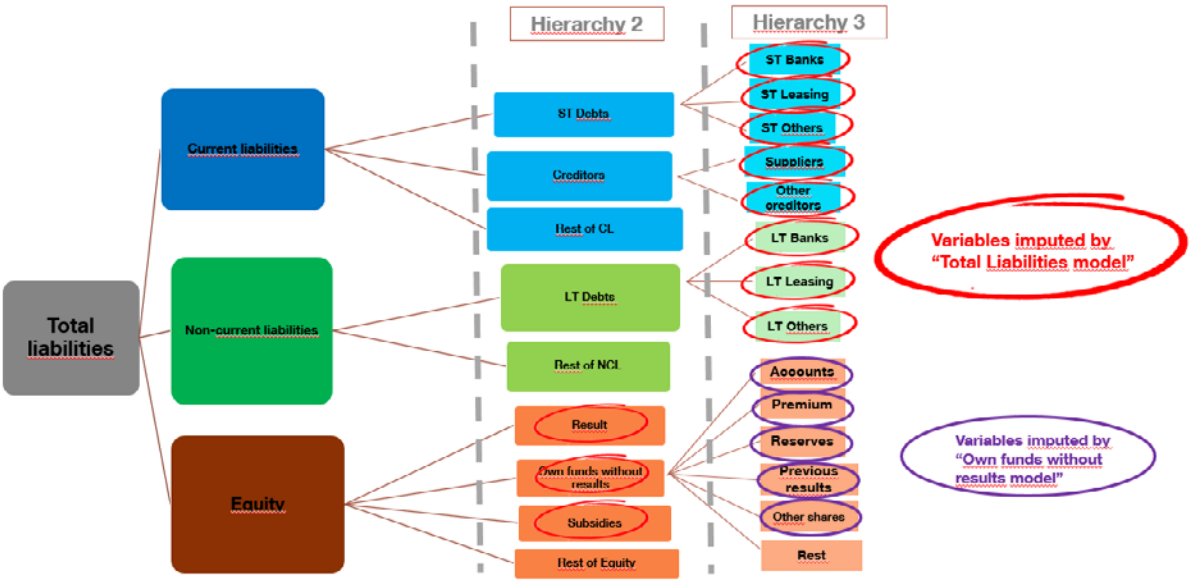
3.3 Training and validation of the algorithm

In the case of questionnaires for asset items, it is considered that there is not enough data for the items indicated in Figure 6 as hierarchy 3 (more detailed items) for the algorithms to learn patterns. Therefore, a model is built for accounting items of hierarchy 2 and another model for variables of hierarchy 3. In summary, for each year (current and previous) in the assets, we have:



In other words, two models are constructed: one for imputing items that fall under Current Assets and Non-Current Assets (hierarchy 2, marked in red), and another for imputing items located under Debtors (hierarchy 3, marked in purple). Accounting items related to "Rest" are not imputed, as they typically appear as zero and are less relevant.

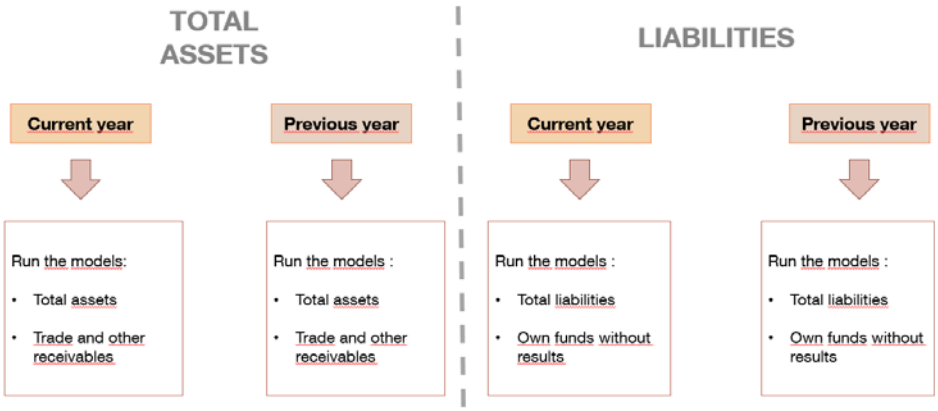
In the case of liabilities, questionnaires have more information for the more detailed items (hierarchy 3), and that's why the modelling approach is different. For each year (current and previous), the following scheme is used for liabilities:



The process began by creating a model to impute the variables marked in red (total liability model) because it was considered sufficient for the algorithm to learn, separating the Result item due to its accounting economic relevance. Later, for production purposes and to recover all the items in the questionnaire, it was decided that it was necessary to impute those belonging to the rest of the Net Equity items marked in purple (Net Equity model without result).

Similar to the approach for assets, accounting items related to "Rest" are not imputed, as they typically appear as zero and are less relevant.

In summary, for each year (current or previous) and each balance sheet category (assets or liabilities), the following models are executed:



3.4 Accounting adjustment

Once the imputations are completed, it is necessary to reconcile them accounting-wise, ensuring that the accounting items match their respective totals. For this, in each questionnaire, an adjustment factor is created for each total, defined as:

$$adjustment\ factor = \frac{\text{total} - \sum non - imputed\ variables}{\sum imputed\ variables}$$

This factor will multiply each of the imputations made by our algorithm. Depending on whether the items and totals can be negative or not, calculations had to be adapted to reconcile. If all items are positive, the corresponding adjustment factor for that total can be directly applied. However, when there are potentially negative items, it was necessary to adapt the programming to subtract those items before applying the adjustment factor.

This part has been particularly complex and has consumed a significant portion of the project's time due to the large number of cases involved. The assurance provided by this strategy is that ultimately, all accounting items are reconciled.

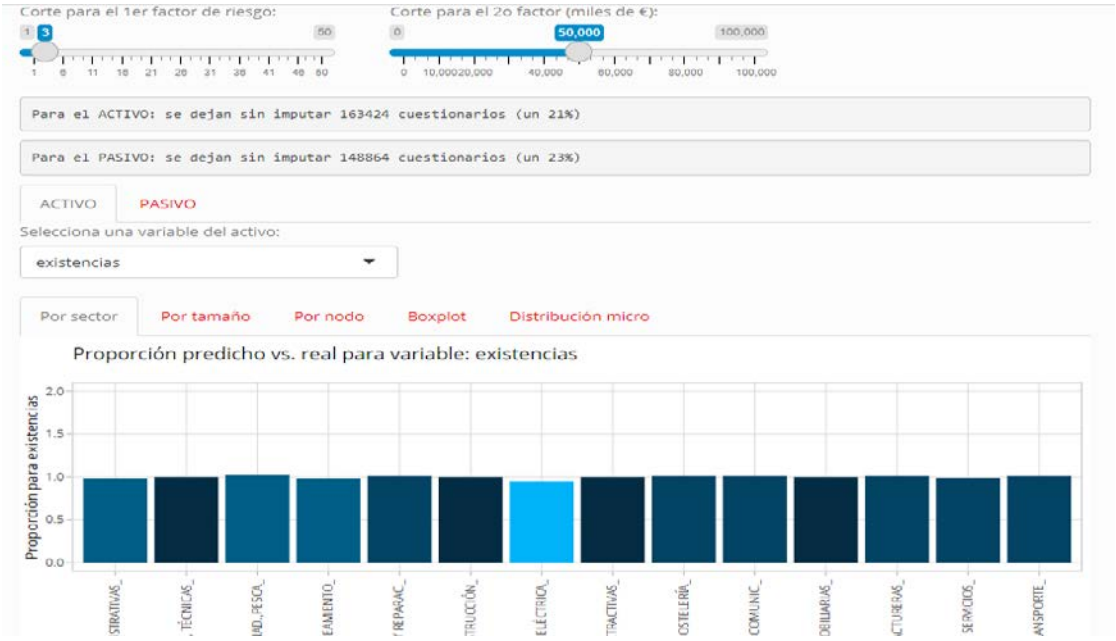
4. Risk factors

Taking into account these accounting adjustments, two risk factors have been identified when the imputation by the algorithms is not correct:

1. Adjustment factor: the quantity by which we need to multiply our imputations to reconcile the items with their totals. Note that if the algorithm imputes quantities that do not follow the actual proportion of accounting items and is multiplied by a high accounting adjustment, then errors will increase considerably.
2. Amount pending imputation: the difference between the total and the actual values (provided by the companies and therefore not imputed) associated with that total.

It is about balancing the risk associated with having a poor imputation with the attempt to impute the maximum number of questionnaires possible. Taking a conservative stance, a threshold of 3 was set for the first factor and 100,000 euros for the second factor. It is required that at least one of the two factors does not exceed the established threshold (in this case, that the first factor is less than 3 or the second factor is less than 100,000 euros) to be able to impute the questionnaire.

Simultaneously, a web application was developed using Shiny to perform quality control of aggregates (sector, size, and node) and make decisions about these thresholds. Additionally, this app serves to verify that the imputations are reliable for the data of the year 2020:



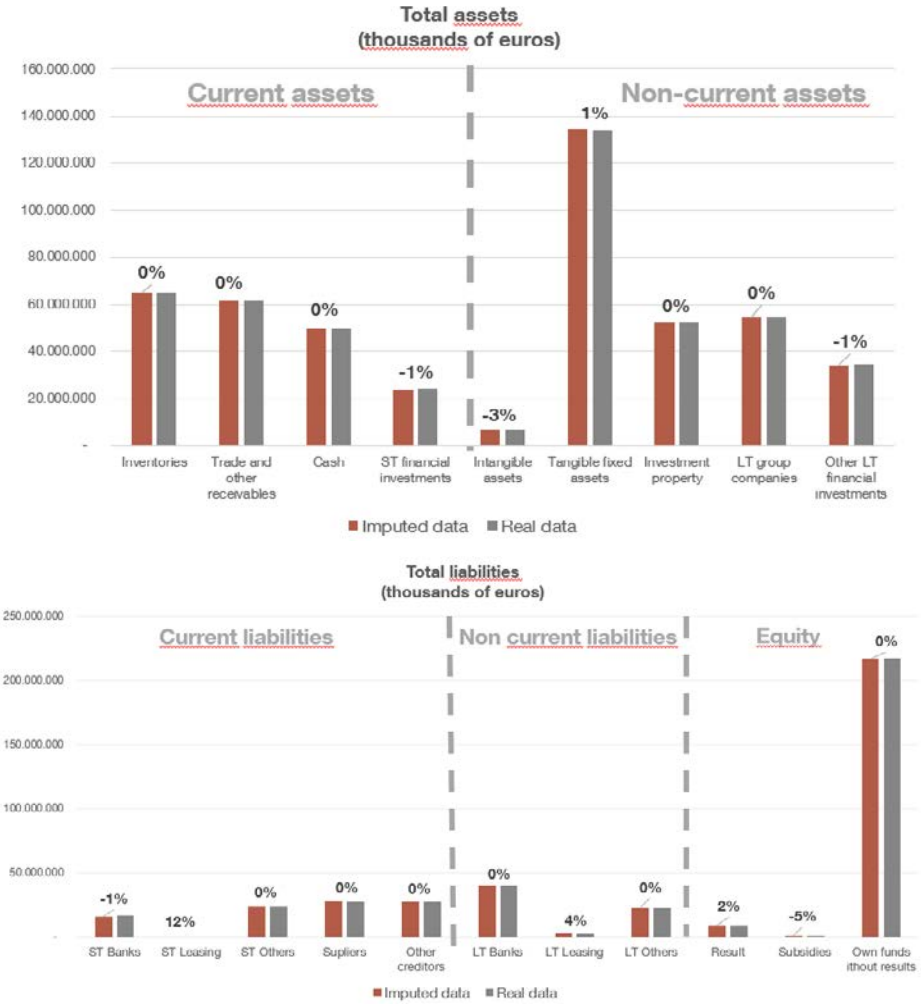
This interactive app will be made available to researchers so that they can consider which nodes have more certainty and which ones have less.

5. Results

As mentioned earlier, the primary goal is to try to analyze if the imputations are reliable at an aggregate level (for the statistics that the CBSO publishes in its reports). The data used for the results are those related to the year 2020, as our models were trained on 2018, and in 2020, we have real data to compare the imputed values with the actual values.

5.1 Aggregated imputations

In aggregate terms, the difference between the actual and imputed accounting items is minimal, as observed in the following graph:



Where the differences between the actual and imputed values are very low percentages.

However, since there are many combinations of accounting items and types of aggregation (size, sector, and node), in the aforementioned Shiny web application, results can be analysed interactively. The metric displayed in the application is a ratio of the actual value to the imputed value. In other words, the closer it is to 1, the closer the imputed and actual values are.

This is illustrated with three examples for the accounting item "Cash" for the three types of aggregation:

5.2 Questionnaire imputations

Despite not being the main goal of the project, the imputations by questionnaire yield acceptable results. To verify these results, an error metric has been created for each item i and questionnaire j as follows:

$$error_{ij} = abs(real_{ij} - imputed_{ij})$$

The absolute value is taken to avoid cancelling positive and negative errors when calculating aggregate metrics for this error.

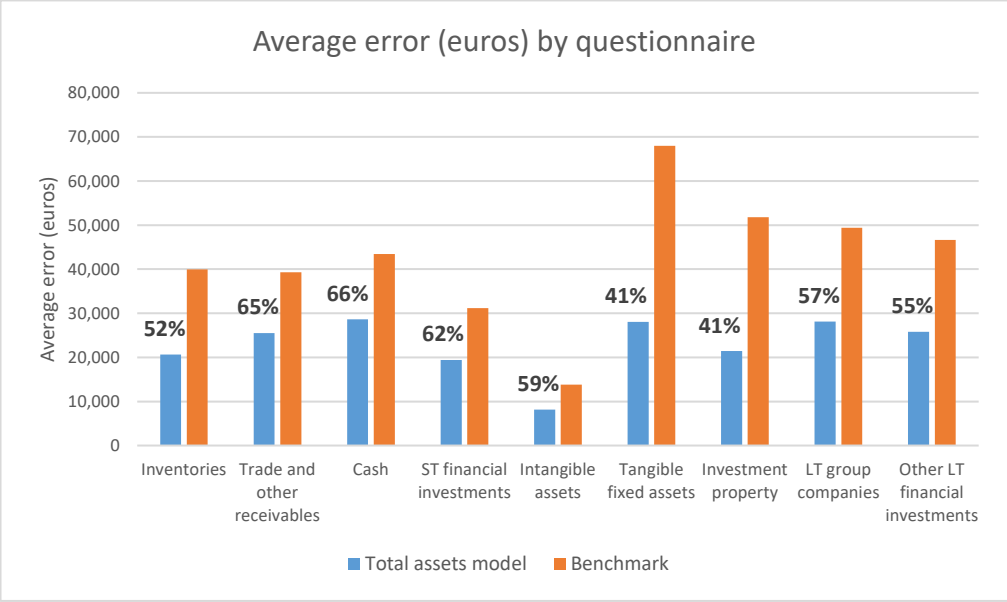
This way, we have the total error of the imputations made for each of the questionnaires in 2020.

5.2.1 Benchmark

When deciding whether an algorithm makes good predictions or not, it is common to use a "naive" model as a baseline or benchmark. In this case, the following model has been constructed as a benchmark:

- Calculate the mean for each accounting item for each of the nodes.
- Impute missing values with the mean of the node and the respective accounting item. This is precisely the system used by the Central de Balances when it wanted to recover specific companies: estimating the missing value based on the normal mean data of its aggregate (crossing business sector and size).
- Adjust those predictions to reconcile accounting-wise.

Comparisons have been made between the different models adjusted with Miceranger and this "benchmark" model. For example, for the Total Assets model, the following comparisons of the previously defined error (calculating the average error per questionnaire) have been obtained for each accounting item:



Where the percentages represent the percentage difference in error of our Total Assets model (obtained through ML) compared to the benchmark model (previously mentioned baseline model). In this graph, it is observed that for each accounting item, we obtain a significantly lower average error for our model compared to the benchmark model, using 610,103 imputed questionnaires for this comparison.

5.2.2 Distribution of errors

The distribution of errors (defined in section 5.2) for each of the models is as follows:

		Distribution of errors		
Model	**Imputed questionnaires**	**Mean**	**Median**	**P99**
Total assets	610.103	53.000 €	1.000 €	835.000 €
Trade and other receivables model	529.322	5.000 €	0€	75.000 €
Total liabilities	512.342	37.500 €	1.000 €	594.000 €
Own funds without results	484.093	33.500 €	0€	458.000 €

Finally, observing the P99 (99th percentile), we see that errors for the vast majority of questionnaires are not high. However, if we were to obtain the maximum errors for each model, they would be high amounts, and this is due to an incorrect imputation where the algorithm does not make a mistake in the amount but rather in the accounting item where it imputes it. This could be mitigated by lowering the threshold in the second risk factor (Section 4).

In other words, the distributions of these errors indicate that, for the vast majority of questionnaires, the models perform a good imputation.

5.3 Algorithm weight in respect to the accounting adjustment

In addition to the results, an analysis has been conducted on how much weight the algorithm has had compared to the adjustment in the final imputation for the year 2020. This metric has only been created for the asset models (total assets and receivables), as all items are positive, making it easier to create this metric and drastically reducing its possible scenarios.

1. For questionnaires that have any imputation, the sum of the breakdowns of the total (for Total Assets, the breakdowns would be current assets and non-current assets) is obtained. This sum is calculated for the algorithm's own predictions (SumPredP) and the predictions after the accounting adjustment (SumPredT).
2. Once these sums are obtained, depending on the scenario faced, the metric is calculated in one way or another:
    a. If SumPredP < SumPredT, then:

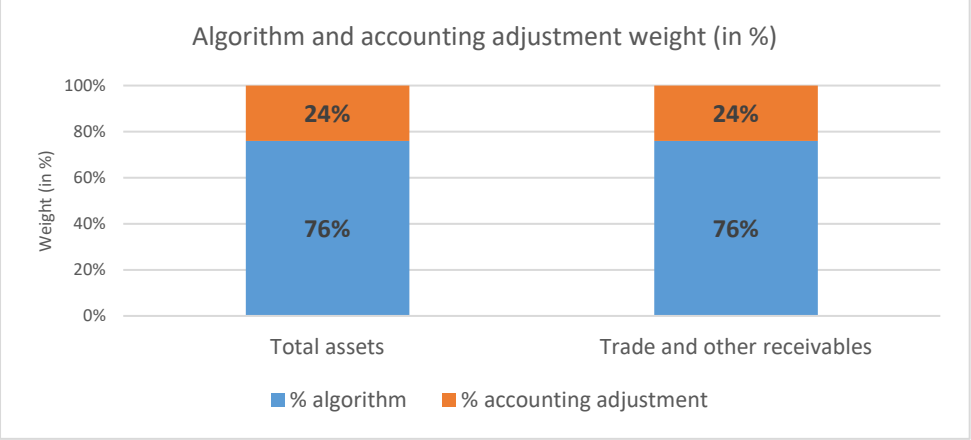
$$algorithm\ weight\ (\%) = \frac{SumPredP}{SumPredT}$$

    b. If SumPredP>=SumPredT, then:

$$algorithm\ weight\ (\%) = \frac{SumPredT}{SumPredP}$$

3. Finally, the average weight of the algorithm is calculated for each of the questionnaires.

Taking into account this described metric, the results for the Total Assets and Receivables models (items always positive) are as follows:



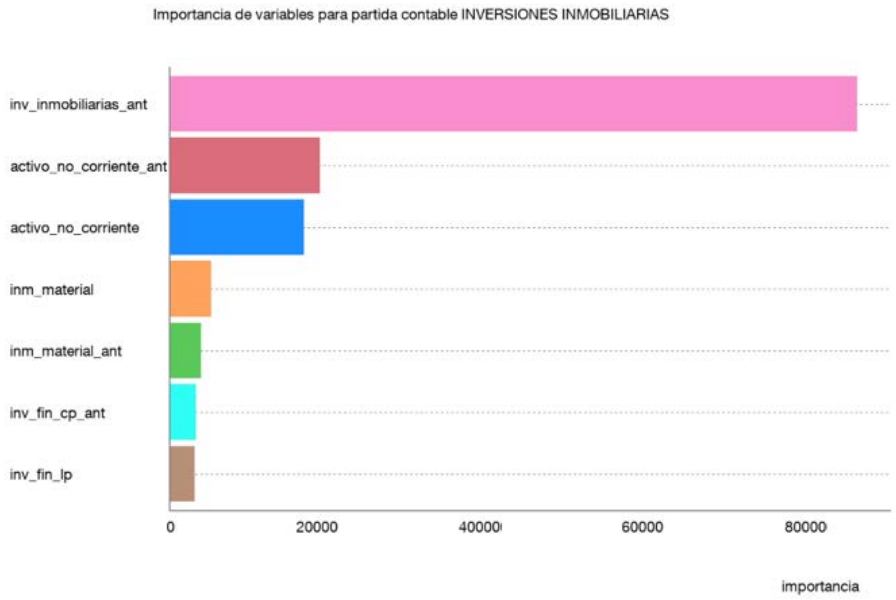
In conclusion, the results of the imputations at the aggregate level are very good. At the questionnaire level, the algorithms impute considerably well for the vast majority of questionnaires, although there are significant errors for a very small percentage of them.

6. Explainability

Once the final imputations have been obtained for each of the questionnaires, it is of interest to know which accounting items have contributed the most (and to what extent) when deciding the imputation. Since our Miceranger algorithm returns a random forest for each of the imputed items, explainability has been studied both globally and locally.

6.1 Global

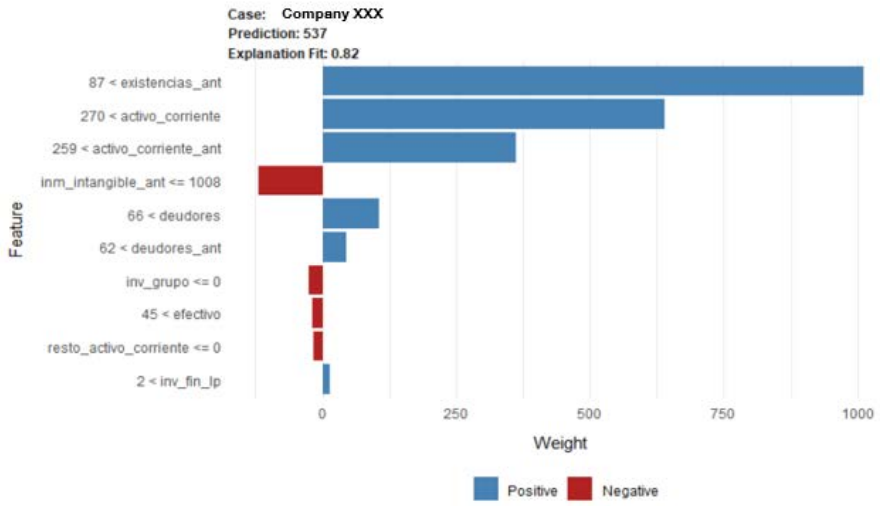
In this case, the accounting items that have contributed the most in a general sense to each of the imputations of the different accounting items have been analysed. For this purpose, variable importance has been used. To illustrate this, an example is provided for the random forest resulting from the accounting item "Inversiones Inmobiliarias" (Investment properties):



Where, in this example, it can be observed that the "Inversiones Inmobiliarias" (Investment properties) item depends on the real estate investments from the previous year and also on items closely related to it.

6.2 Local

In the case of local explainability, LIME (Local Interpretable Model-Agnostic Explanations) has been used to analyse, at the questionnaire level, which accounting items have had the most impact on that imputation and in what direction. As an illustration, the following example is provided:



In this case, it is a small-sized company where it can be observed that the high inventory from the previous year contributes very positively to the imputation of current inventory for the current year. Similarly, a higher volume of current assets (for the current year and the previous year) also contributes to a higher imputation of inventory. Therefore, it seems that the algorithm provides reasonable local explanations about which items influence and in what direction.

7. Production

The last part of this project is the deployment of this entire procedure to automatically carry out this imputation in the internal applications of the Bank of Spain.

7.1 Method

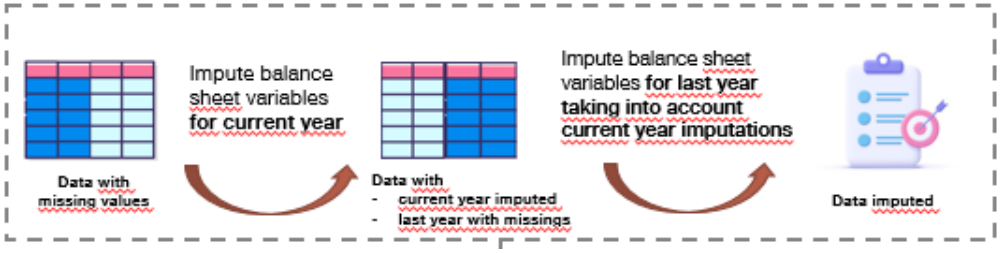
In the database, there is no distinction between a value reported as zero and an empty value. Therefore, the first step is to transform those empties corresponding to square summations into zeros since any imputation would lead to a mismatch.

Subsequently, the saved models are applied to the current year's data. This should be done following the order in which the flow has been designed for both assets and liabilities (the process where imputations are chained):

1. First, impute current year's data:
    - a. Impute and reconcile hierarchy 2 items (total assets model for assets/total liabilities model for liabilities).
    - b. Considering the imputations made in a., impute and reconcile hierarchy 3 items (debtors model for assets/ equity without result model for liabilities).
2. With the imputed information from the current year, impute the previous year's data:
    1. a. Impute and reconcile hierarchy 2 items (total assets model for assets/total liabilities model for liabilities).

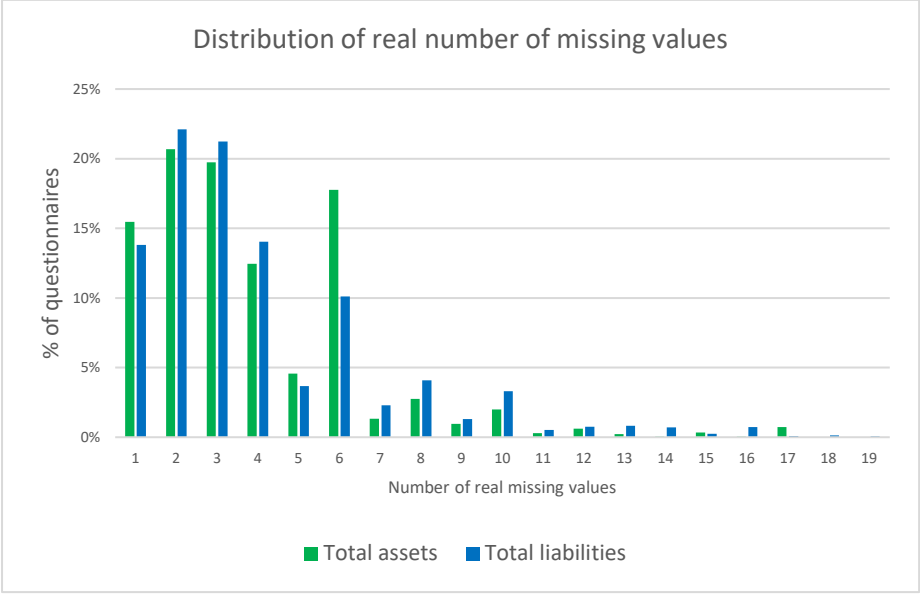
2. b. Considering the imputations made in a., impute and reconcile hierarchy 3 items (debtors model for assets/ equity without result model for liabilities).

Finally, once the imputations for the entire questionnaire are obtained, the output format is modified to fit the requirements of the Bank of Spain's accounting database (GTC). It can then follow its usual course of subsequent validations and quality control. This format is of a key-value type.



7.2 Real data into production

Having the data from 2020, the steps of the method described earlier have been followed. Firstly, taking into account the validations from the Accounting Treatment Manager of the Bank of Spain (GTC), those items that are empty and whose sum does not match have been emptied (if the item is empty but the sum matches, it is understood that the item is zero). In this way, the distribution of the real number of missing values is as follows:



The data we have after applying these GTC validations is as follows:

- 1,899 questionnaires to be imputed for the assets (7,562 entries to be imputed)
- 3,627 questionnaires to be imputed for the liabilities (14,849 entries to be imputed)

Once the data is obtained, values are predicted using the pre-trained algorithms, and an accounting adjustment is performed for both asset and liability entries. By applying the two mentioned risk factors, 94% of the questionnaires for assets and 90% of the questionnaires for liabilities are imputed. In other words, the imputation results are as follows:

- 1,791 questionnaires imputed for assets (7,141 entries imputed)
- 3,268 questionnaires imputed for liabilities (13,015 entries imputed)

The final output format (key-value) for recording in internal applications is as follows:

id	clave	valor
14029130257	123009	47
14029130257	127009	89
14029130257	125009	68
14029130257	111009	0
14029130257	112009	194
14029130257	113009	441
14029130257	114009	0
14029130257	115009	86
14030078636	122009	9
14030078636	125009	9
14028856637	122009	40
14029120474	122009	1

With the idea of verifying how the algorithms perform imputations, the business team has analysed a series of questionnaires (randomly selected). These analysts have checked that, in general, the imputations made by the algorithms make sense and are based on logical criteria (values of entries from the previous year, logical relationships, etc.).

Once this output is defined, collaboration with the Information Systems Department is needed to launch this process on internal servers periodically (BATCH).

8. Next steps

The lines of research that can be continued are as follows:

- Deployment on the internal servers of the Bank of Spain.

- Search for better algorithms: Clearly, the more accurate the imputation algorithm is, and therefore the better it imputes microdata, the better the entire procedure will be. The current pace of development in the field of algorithms and data science requires staying vigilant for innovations that could enhance any of the projects already developed.

- Apply these techniques to the Income Statement questionnaires.

BANCO DE ESPAÑA
Eurosistema

MISSING VALUES IMPUTATION FOR CENTRAL BALANCE SHEET DATA OFFICE (CBSO) ACCOUNTING DATA

64TH WORLD STATISTICS CONGRESS

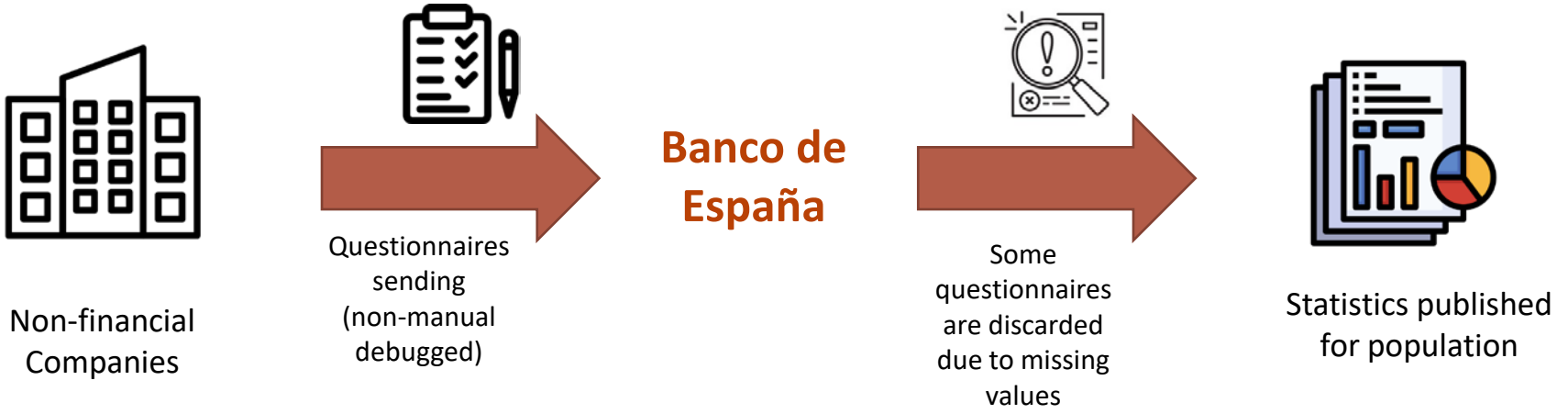
Iker González Crespo
Data Scientist at Central Balance Sheet Data Office
Statistics Department

15th July 2023

BANCO DE ESPAÑA
Eurosistema

INDEX

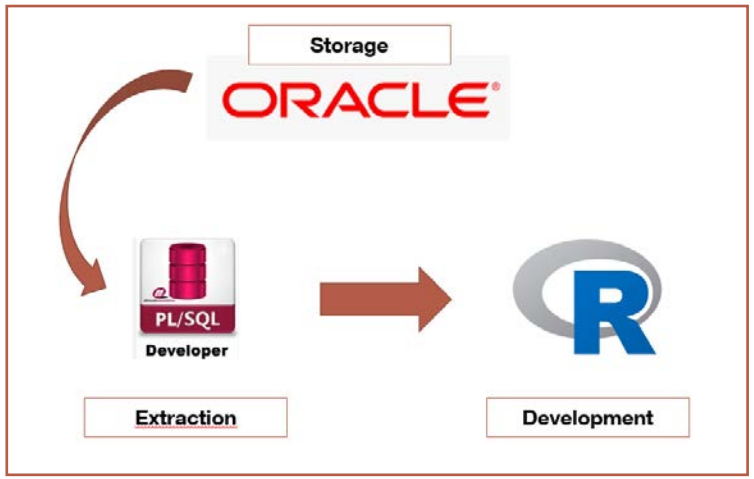
Banco de España receives almost 1 million of accounting questionnaries from non-financial companies. This information needs to be debugged by Central Balance Sheet Office (CBSO) in order to publish statistics for the spanish population.



What if we were able to impute these missing values, and thus:

- Enrich our databases
- Publish much more accurate statistics

GOAL OF THE PROJECT



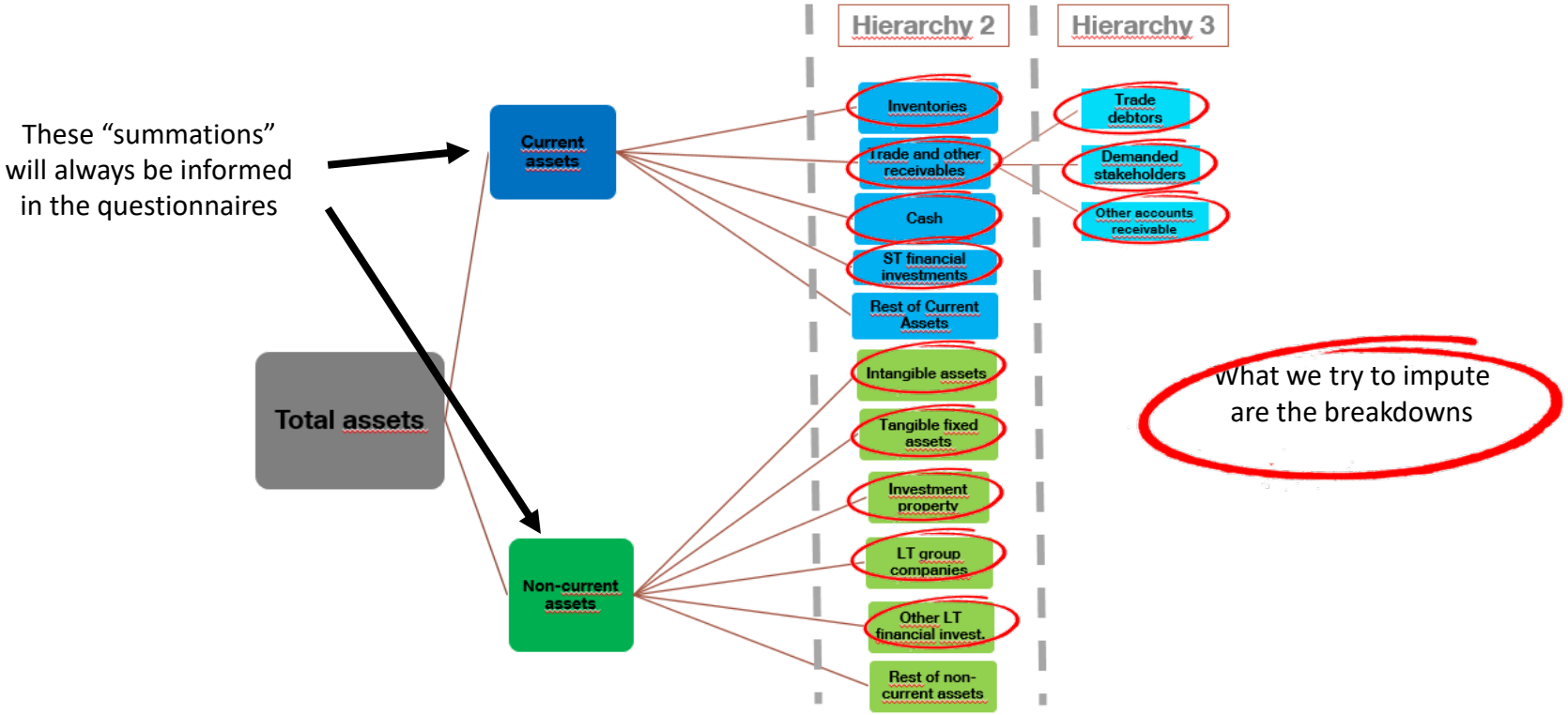
- The databases are stored in Oracle
- Information of questionnaires for years 2018 and 2020
- Usage of a Workstation due to computational cost
- Development has been conducted in R

- Focus in one specific type of questionnaire (only for SMEs)
- Select the balance variables (total assets, total liabilities and equity)
- There are 2 years to impute (current and last)

Example of questionnaire

ACTIVO		NOTAS DE LA MEMORIA	EJERCICIO 2020 (2)	EJERCICIO 2019 (3)
A) ACTIVO NO CORRIENTE	11000		424,71460	424,38489
I. Inmovilizado intangible	11100			0,02288
II. Inmovilizado material	11200		356,34557	366,62683
III. Inversiones inmobiliarias	11300			
IV. Inversiones en empresas del grupo y asociadas a largo plazo	11400			
V. Inversiones financieras a largo plazo	11500			
VI. Activos por impuesto diferido	11600		68,36903	57,73518
VII. Deudores comerciales no corrientes	11700			
B) ACTIVO CORRIENTE	12000		1.173,05586	1.068,56633
I. Activos no corrientes mantenidos para la venta	12100			
II. Existencias	12200		527,76339	499,86441
III. Deudores comerciales y otras cuentas a cobrar	12300		501,45542	420,93524
1. Clientes por ventas y prestaciones de servicios	12380		490,71850	396,46022
a) Clientes por ventas y prestaciones de servicios a largo plazo	12381			
b) Clientes por ventas y prestaciones de servicios a corto plazo	12382		490,71850	396,46022
2. Accionistas (socios) por desembolsos exigidos	12370			
3. Otros deudores	12390		10,73692	24,47502
IV. Inversiones en empresas del grupo y asociadas a corto plazo	12400			
V. Inversiones financieras a corto plazo	12500			19,47000
VI. Periodificaciones a corto plazo	12600			
VII. Efectivo y otros activos líquidos equivalentes	12700		143,83699	128,29668
TOTAL ACTIVO (A + B)	10000		1.597,77040	1.492,95122

The aim in this project is to impute all the balance sheet variables with missing values. However, to be illustrative, it is only shown the relations between assets variables:



We get the debugged questionnaires in order to let our algorithms learn relations between all the variables in a proper way.

Debugged questionnaires

ID	Node	Trade	Trade debtors	Deman, stakehol.	Other account.
1	A1	73	55	13	5
2	B2	12	7	2	3
3	A1	34	30	1	3
4	B3	65	42	13	10
5	B2	98	65	3	30
6	B3	30	5	15	10

Generate 30% of missing values (randomly) for each accounting variable

Debugged questionnaires with missing values

ID	Node	Trade	Trade debtors	Deman, stakehol.	Other account.
1	A1	73	NA	NA	5
2	B2	12	7	2	3
3	A1	34	NA	1	NA
4	B3	65	42	13	10
5	B2	98	65	3	30
6	B3	30	NA	15	10

For each node (sector and size), **fit a ML model to impute values**

Debugged questionnaires imputed

ID	Node	Trade	Trade debtors	Deman, stakehol.	Other account.
1	A1	73	50	8	5
2	B2	12	7	2	3
3	A1	34	25	1	2
4	B3	65	42	13	10
5	B2	98	65	3	30
6	B3	30	2	15	10

Accounting adjustment (to match all the sums)

Debugged questionnaires imputed and with accounting adjustment

ID	Node	Trade	Trade debtors	Deman, stakehol.	Other account.
1	A1	73	59	9	5
2	B2	12	7	2	3
3	A1	34	30	1	3
4	B3	65	42	13	10
5	B2	98	65	3	30
6	B3	30	5	15	10

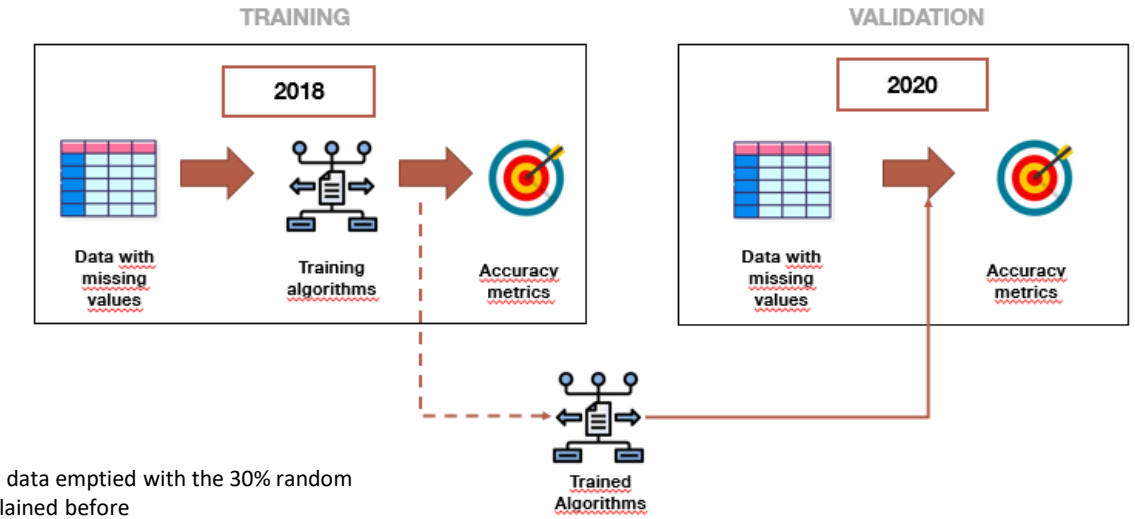
Accuracy metrics between imputed and real variables

The ML algorithm used is **miceranger (link).** This is a technique based on ensembles of random forests. Some advantages of this model are:

- Fits very well for any kind of data (high accuracy metrics)
- Runs in parallel (speed up the process)
- Interpretable model (explain the imputations)

We **fit one model per node** (sector and size) **and for each year** (current and last).

Data from 2018 is used to train the best algorithms and save them. Afterwards, these models are used to predict data from 2020 (and check if they still fit well).



(*) Data with missing values: data emptied with the 30% random missing values technique explained before

BANCO DE ESPAÑA
Eurosistema

Once all the imputations are set, there is a need to force all the sums to fit.
For every "summation" we created a factor to adjust the imputations:

$$factor = \frac{total - \sum non\ imputed\ variables}{\sum imputed\ variables} = \frac{difference}{\sum imputed\ variables}$$

Example

ID	Trade	Trade debtors	Demanded stakehol.	Other acc.	Difference	Factor
1	50	2	2	10	48	4
2	30	5	15	8	15	1,15

Real data Algorithm imputations

We **use this factor as a metric to balance the risk** we want to take when imputing questionnaires

Sensitivity analysis with interactive app to set the optimal factor

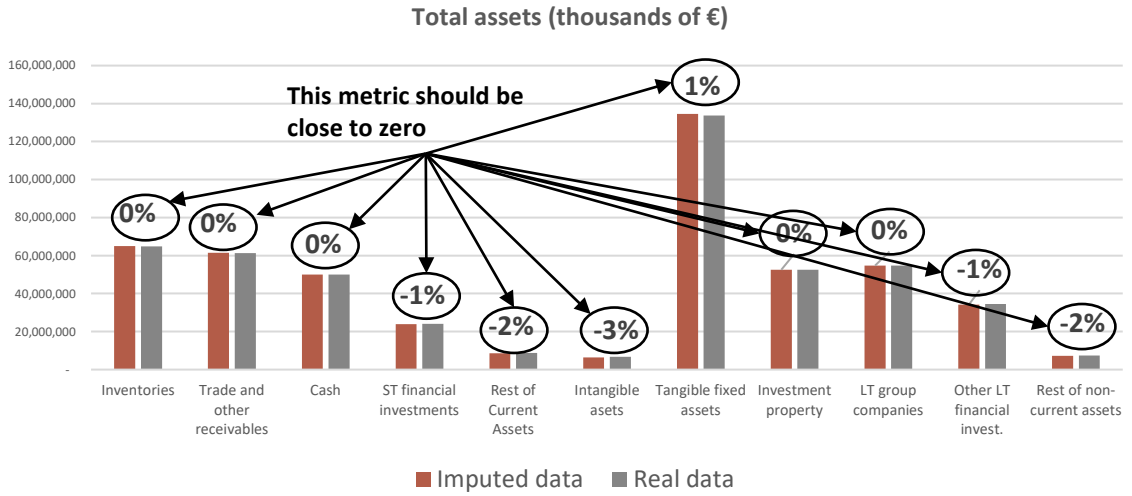
Note that depending on the sign of the variables, this metric can suffer different approaches. In the case of assets, all variables must be positive, so is easier.

A simulation has been conducted to analyze how well the statistics will hold after adding the imputations (and thus gaining sample).

The **results are significantly accurate** by many aggregations:

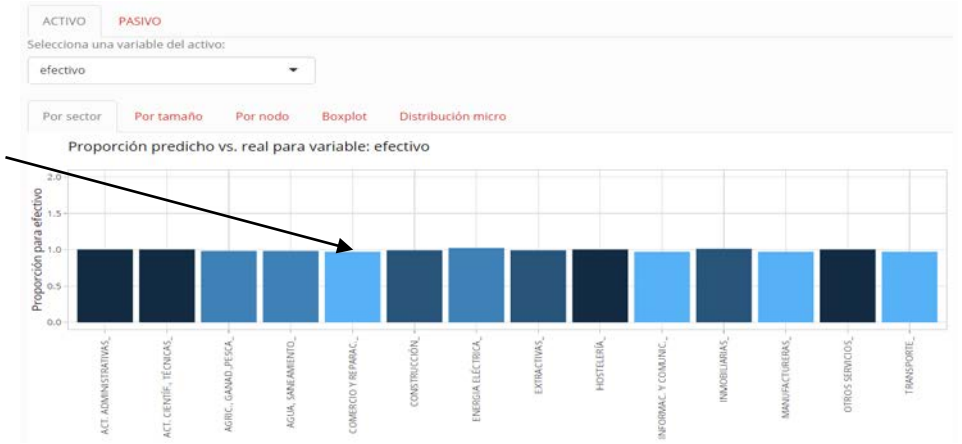
- Variable
- Size
- Sector
- Node (sector and size)

Variable-level

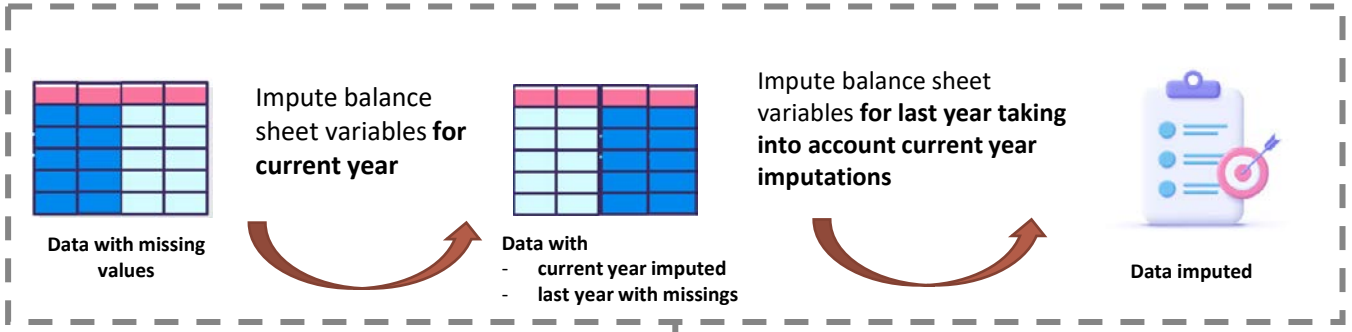


Size-level

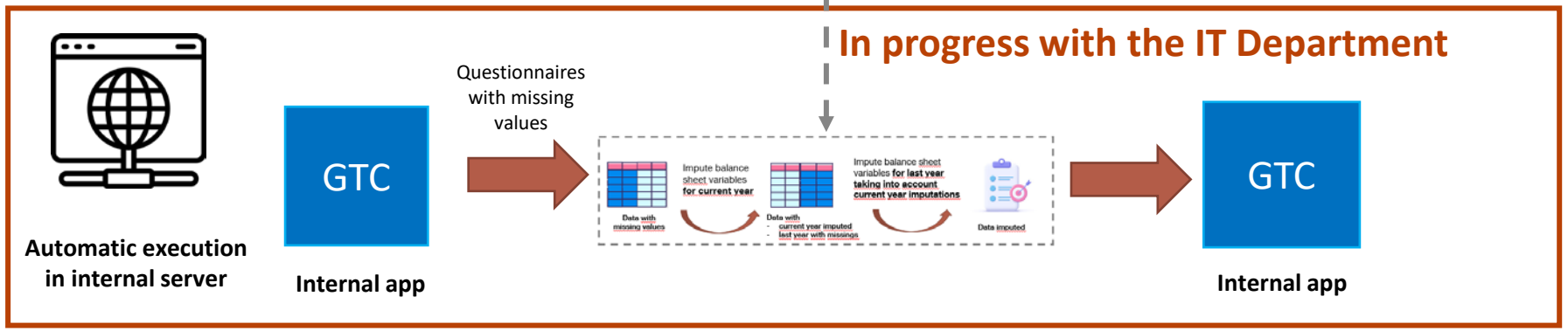
Sector-level



Dealing with streaming real data, we need to link these models (current and last year) as a way to impute the whole variables of the questionnaires:



This development should be linked to the internal processes and applications of Banco de España, in order to communicate one with each other:



BANCO DE ESPAÑA
Eurosistema

THANKS FOR YOUR ATTENTION

STATISTICS DEPARTAMENT