Missing Data & Imputation Mthods for ICT-USage in Enterprises

Luis Carlos Castillo

University of Urbino and Universty of Bremen

2024-07-29

Importand considerations

In the process of selecting variables consistent across all six years of the ICT usage in enterprises survey (2014-2019), conducted by ISTAT, several challenges were encountered due to variations in variable codification and measurement across different years. These discrepancies suggest that the survey's methodology evolved over time, possibly to adapt to changes in technology or business practices, which occasionally led to alterations in how certain variables were defined or measured.

It is important for readers to be aware that, although the variable names and codes may vary from year to year, careful review has been conducted to ensure that each selected variable consistently measures the intended construct over the entire period. This preparatory work is critical for the next step of the analysis, which will involve data imputation to address any missing or inconsistent data points, thereby preserving the integrity of the analysis.

List of selected variables

name_EN	ict2014	ict2015	ict2016	ict2017	ict2018	ict2019
company code	Codice	Codice	codice	codice_	codice_	codice_
class of revenues from the sale of goods and services	Ricavi	Ricavi	ricavi	ricavi_cl	ricavi_cl	ricavi_cl
Percentage of employees using the computer out of the total employees	A2	A2	A2	A2_	A3_	A3_
employment of specialists in computer subjects	B1	B1	B1	B1	B1	B1
IT training courses for employees with specialist ict skills	B2a	B2a	B2a	B2a	B2a	B2a
IT training courses for employees without specialist ict skills	B2b	B2b	B2b	B2b	B2b	B2b
Use of internal personnel for ICT infrastructure maintenance	B5a	B5a	B5a	B5a	B5a	B5a
Percentage of employees using computers connected to the internet	C2	C2	C2	C2_	C2_	C2_
Internet download	C4	C4	C4	C4	C4	C4
Enterprise provides mobile devices with mobile connection	C5a	C5	C5	C6	C5	C5
Website	C7	C7	C8	C8	C8	C7
possibility to place orders or reservations online eg online shopping cart	C8a	C8a	C9a	C9a	C9a	C8a
access to product catalogs or price lists	C8c	C8c	C9c	C9c	C9c	C8c
Social network	C9a	C9a	C10a	C10a	C10a	C10a
Social media and multimedia	C9c	C9c	C10c	C10c	C10c	C10c
announcement of vacancies or possibility to apply for employment online	C8g	C8g	C9f	C9f	C9g	C8f
links or references to company profiles on social media	C8h	C8h	C9g	C9g	C9f	C8g
using erp software	E1	E1	E1	D1	D1	D1
use operational crm software	E2b	E2b	E2b	D2b	D2b	D2a
use analytical crm software	E2a	E2a	E2a	D2a	D2a	D2b
web sales through intermediary websites or ecommerce sites marketplaces or apps	J7	I8	H9	G11	I1b	F1B
Size by number of employees	clad4	clad4	clad4	clad3	clad3	clad3
Classification of ICT compnies	dom4	dom4	dom4	dom4	dom4	dom4
Region NUTS-1	rip	rip	rip	rip	rip	rip
Groups of economic sectors	ateco_1	ateco_1	ateco_1	Ateco_1	Ateco_1	Ateco_1

Imputation Methods for Variables with 100% Missing Data (2014-2019)

To address the discontinuities in certain survey questions across different years, stratified imputation methodology was employed. This approach was used for variables that are entirely missing in specific years due to changes in the survey design. A composite strata variable by combining company size, sector, and region was created. For each variable with missing data, we calculated the proportion of responses (e.g., the proportion of UM_E1—Using ERP software—being 1) within each stratum, using data from years where the variable was available. Assuming these proportions remain stable over time, this proportions were used to impute the missing values in the discontinuous years. This involved determining the number of cases to impute for each stratum, calculating the expected counts of 1s and 0s based on the historical proportions, and assigning these values randomly to maintain variability. This robust approach ensures that the imputed data accurately reflects the patterns observed in similar firms, thereby preserving the integrity and continuity of the dataset for comprehensive analysis.

Summary of Variables with 100% Missing Data (2014-2019)

Year	Variables_with_all_Missing_Values
2014	B5a
2015	-
2016	-
2017	B5a, C10a, C10c, C9g, D1, D2a, D2b
2018	C10a, C10c, C9g
2019	B5a

This table provides a concise view of the variables within the ICT usage in enterprises survey that were entirely absent in specific years. Notably, the variable B5a has recurrently been missing across multiple years, pointing to significant discontinuities in the data collection or changes in survey focus. Other variables, particularly C10a, C10c, and C9g, also show repeated instances of complete data absence, primarily in the later years of the survey. This summary assists in identifying patterns of missing data and facilitates targeted approaches for data imputation and analysis continuity.

Imputation Methods for Variables with Low Levels of Missing Data (2014-2019)

To handle the low missing values in the dataset, the mice package for multiple imputation was used. This process involved generating multiple imputed datasets to account for the uncertainty associated with missing data. Depending on the nature of the variable categorical or continuous different imputation methods were used, such as Predictive Mean Matching (pmm), Random Forest (rf), Classification and Regression Trees (cart), logistic regression (logreg) and random sample from observed values (sample). For each method, five imputed datasets with fifteen iterations to ensure convergence were created. Following the imputation, regression analyses to evaluate the performance of each method was conducted, using accuracy metrics to determine the best model. The method with the highest accuracy was then selected, and the corresponding imputed dataset was integrated into the original dataframe. This systematic approach ensured a robust handling of missing data, enhancing the reliability and interpretability of our subsequent analyses.

Summary of Variables with Low Levels of Missing Data (2014-2019)

Year	Variables_with_Low_Missing_Data
2014	C3a, C9c
2015	C8g, C8h, C9c
2016	C9a, C9c, C9f, C9g
2017	C8g, C9f
2018	C9a, C9c, C9f
2019	C8a, C8c, C8f, C8g

This table outlines the variables within the ICT usage in enterprises survey that exhibited low but notable missing data percentages, approximately between 21% and 26%, across different survey years. It illustrates that certain variables, such as C8g and C9c, frequently appear with missing data, suggesting recurring issues with these specific survey items. Understanding the patterns of these less significantly missing variables helps in applying consistent and appropriate imputation techniques to address these gaps, thus enhancing the overall data quality and reliability for subsequent analysis.

Missing values for 2014

The missing data chart for 2014 highlights some significant challenges in data continuity for certain variables in the ICT usage in enterprises survey. Notably, the variable B5a exhibits a complete absence of data for this year, while other variables such as C8a, C8c, C8g, C8h have missing values up to 26%. The gap in B5a is the result of changes in the survey's measurement approaches and modifications in variable codification over different survey iterations, while the other variables are missing at random.

To manage these inconsistencies and uphold the analytical integrity of the study, a methodical approach to data imputation is being employed. For the variable B5a, with its absence in multiple years, we are utilizing interpolation techniques to estimate missing values based on available data from the years the variable is present. This ensures that the reconstructed data aligns with the observed trends over the survey period. For the other variables the mice package will be used for imputation considering the type of missingness.

Missing chart 2014

Missing values for 2015

The 2015 missing data chart for the ICT usage in enterprises survey indicates various levels of missingness among the variables, with some showing significant gaps. Notably, variable C8a, C8c, C8g, and C8h displays approximately 24% missing data. Other variables exhibit lower missing data percentages but are still noteworthy for the purposes of this study.

To address these data gaps, we will employ imputation methods tailored to the specific characteristics and missing data patterns of each variable.

Missing chart 2015

Missing values for 2016

The 2016 missing data chart for the ICT usage in enterprises survey indicates various levels of missingness among the variables, with some showing significant gaps. Notably, variable C9a, C9c, C9f, and C9g displays approximately 22% missing data. Other variables exhibit lower missing data percentages but are still noteworthy for the purposes of this study.

To address these data gaps, we will employ imputation methods tailored to the specific characteristics and missing data patterns of each variable.

Missing chart 2016

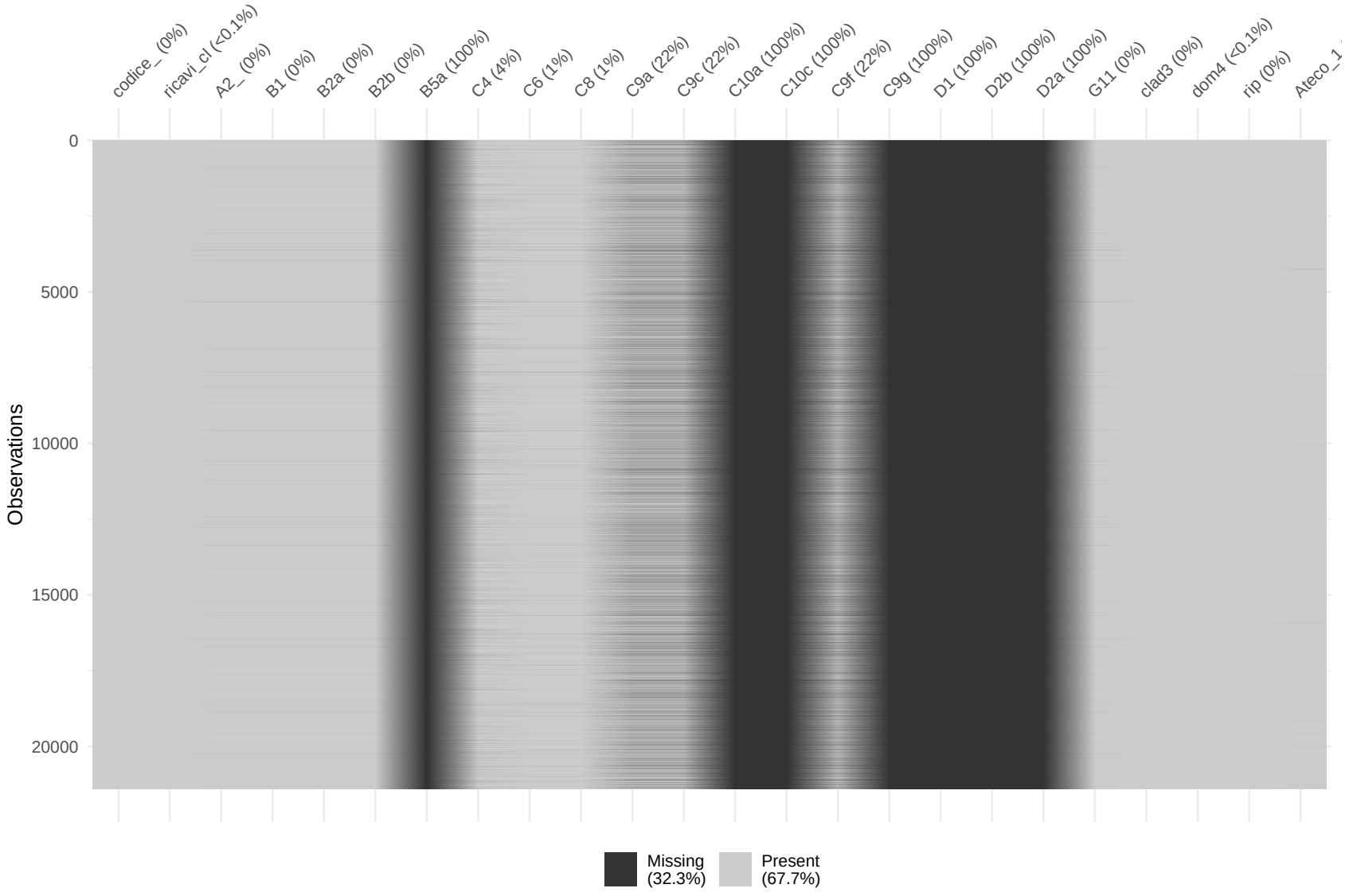
Missing values for 2017

The 2017 missing data chart for the ICT usage in enterprises survey reveals significant gaps in data availability due to changes in the questionnaire, leading to the discontinuation of several variables including B5a, C10a, C10c, C9g, D1, D2a, and D2b, each showing 100% missing data. These substantial data omissions are the result of modifications to the survey's focus and question set.

To mitigate the impact of these missing values, we will apply interpolation techniques for the variables that have been intermittently recorded over the years. This method will allow us to estimate missing values by leveraging trends and patterns from available data in other years where these variables were included.

Additionally, for variables such as C8g and C9f, which exhibit up to 22% missing data, we will employ imputation methods that utilize the existing data within the same year. These approaches are designed to enhance data completeness and ensure the robustness of subsequent analyses, thereby maintaining the integrity of the study despite the survey's evolving structure.

Missing chart 2017

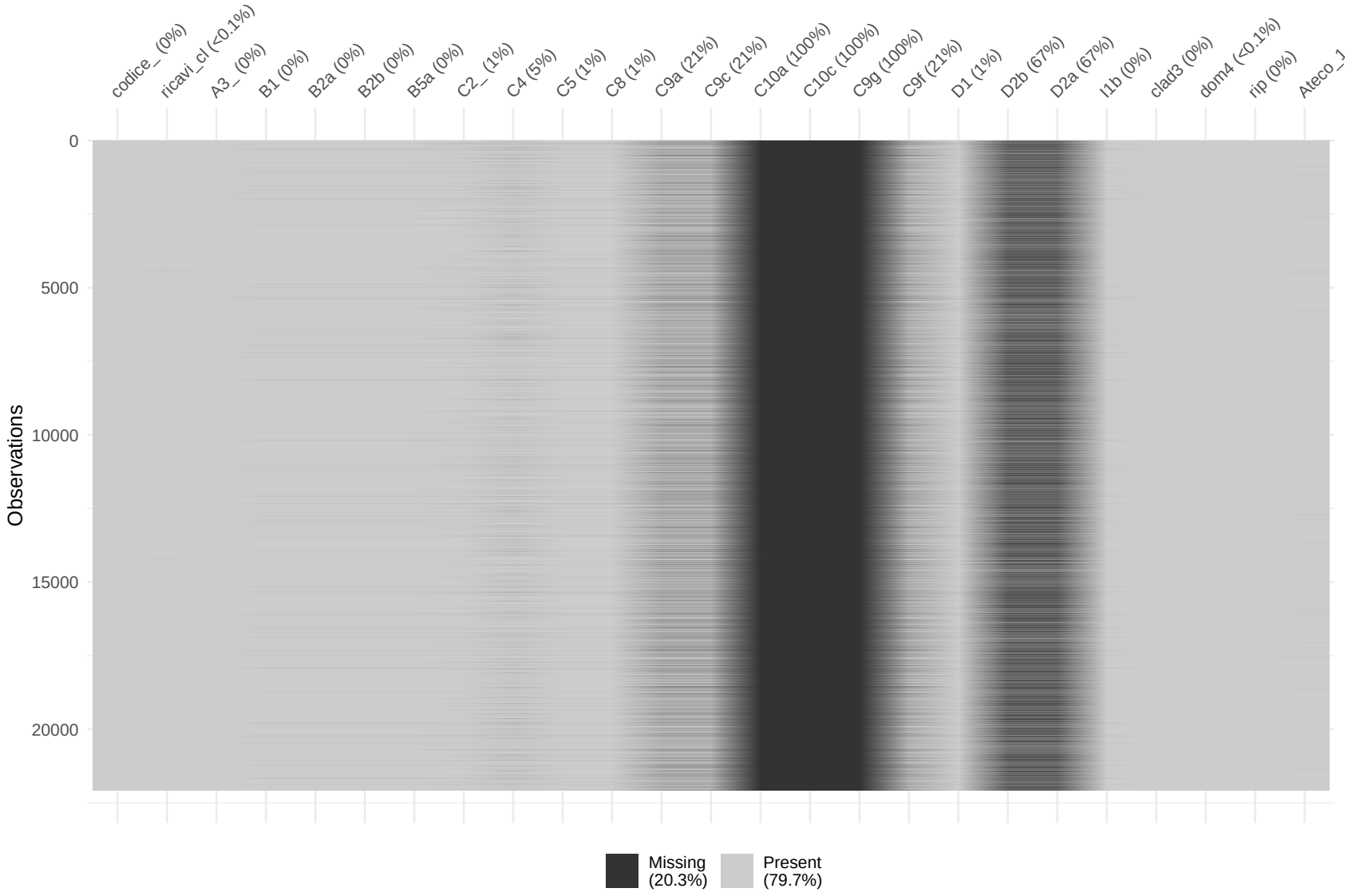


Missing values for 2018

The 2018 missing data chart reveals continued issues with data continuity, particularly for variables C10a, C10c, and C9g, each persisting at 100% missing. This pattern highlights the ongoing challenges related to survey changes over the years.

Additionally, variables such as C9a, C9c, and C9f display 21% missing data, suggesting recurrent gaps. To address these issues, interpolation will be used for variables with complete data gaps, leveraging trends from available years. For those with partial missing data, imputation will utilize existing data from the same year to ensure comprehensive and accurate analysis across the dataset.

Missing chart 2018



Missing values for 2019

The 2019 missing data chart for the ICT usage in enterprises survey shows that variable B5a continues to have 100% missing data, highlighting a consistent absence across several years. Additionally, variables C8a, C8c, C8f, and C8g each report 21% missing data.

To address these gaps, B5a will undergo interpolation using related data from other years where it was recorded, ensuring a more complete dataset for analysis. For the variables with 21% missing data, we will apply imputation techniques that leverage the available data within the same year to fill these gaps, thereby maintaining the analytical integrity of the survey data.

Missing chart 2019