

1 Detailed methodology for completing data

1.1 Structure-related variable imputation

For variables directly related to the building structure, the imputation strategy was applied in a block-wise manner, treating them as inherently linked. However, different strategies were adopted depending on the building form, based on the amount of available data.

For LR buildings, a "MWTR" strategy (Manual, WHE, Tabula, Region) was applied. Since sufficient data was available, missing values could be completed using external databases and, in only one case, regional imputation. For MR buildings, an additional round of regional imputation and a default value imputation step were added, resulting in the strategy "MWTRRD". For HR buildings, data availability was very limited, to avoid large extrapolations, all values were filled using default values, therefore the final strategy used was "D".

1. Flag M (Manual Imputation):

This first step targets building typologies where only a single variable within the structure-related block is missing. These isolated values are completed manually, based on expert judgment or predefined rules. Also buildings with sloped concrete roofs were modified to flat roofs. Additionally, buildings with sloped concrete roofs were converted to flat roofs, and glass walls were replaced with concrete, as this was deemed more representative for residential buildings in this specific context.

2. Flags W and T (External Databases - WHE and Tabula):

Missing values in structural variables are then filled using external databases. The World Housing Encyclopedia (WHE) and Tabula provide country-specific information. When both sources are available for a country, WHE is prioritized, as it generally contains more detailed information on structural characteristics.

3. Flag R — Regional Imputation (GDUBP Region):

If missing values remain after the previous steps, a regional imputation strategy is applied based on geographical relevance within the GDUBP regions. A priority is given to neighboring countries with larger borders (in case of mainland) or closer (in case of islands), giving more influence to closer and more geographically relevant neighbors. Data from the most relevant country within the same region with available structural data is used to fill the gaps. When no country in the same region had available data, the most relevant country with data was used as a fallback, even if it belonged to a different region. For LR buildings, all GDUBP regions had at least one valid observation. For MR buildings, two GDUBP regions lacked data, limiting regional imputation. For HR buildings, eight regions had insufficient data, so this step was skipped.

4. Flag D — Default Value (MR and HR Only):

Due to the lack of data for HR buildings, no regional strategy could be applied. However, as HR buildings showed low variability in their structural variables, a default value corresponding to the most frequent values was used in all cases. For MR buildings, despite more available data and two rounds of regional imputation, some missing values remained. These were also completed using dominant values. In both

MR and HR categories, the default values corresponded to concrete buildings with light-colored facades and flat roofs.

1.2 Lighting-related variable imputation

For variables pertaining to building lighting characteristics, the variables are treated independently. Each imputation method is described below:

Shading imputation

The original 'shading' variable is converted into three distinct binary (presence/absence) variables, each representing a specific type of shading system:

- Interior shading
- Exterior fixed shading (combining overhangs and facade fixed elements)
- Exterior operable shading

The imputation for 'shading' systems follows a "TWR" strategy (Tabula, WHE, Region) for LR buildings. MR buildings followed a "TWRRD" (additional Region and Default) and HR buildings followed a "WD" strategy, as Tabula was not filled for HR buildings and there was not enough survey data in this typology to enable a regional strategy.

1. Flag T (Tabula Data): Missing values for each of the newly created binary shading variables are imputed using data from the Tabula database. Where shading data is missing in the primary dataset but available in Tabula, its value (presence or absence) is used for imputation.
2. Flag W (WHE Data): Similarly, missing values for the binary shading variables are also imputed using data from the World Housing Encyclopedia (WHE). If shading data is missing in the primary dataset but available in WHE, its value is used.
3. Flag R (Region): For the remaining missing values in the shading variables, a regional imputation approach is applied for LR and MR buildings. Responses from most relevant neighboring countries are used to fill the gaps.
4. Flag D (Default): A default value was applied to HR buildings, which corresponds to the majority of responses in buildings with data. The default value was the presence of interior shading and the absence of other forms of shading. The same values were applied to the MR buildings that remained empty after the R strategy.

Window share imputation

The window share imputation strategy is identical to that for shading systems in LR buildings and largely similar for MR buildings.

For MR and HR buildings, a manual step was added: 'Glass buildings' were converted to 'Bay/Large windows'. This maintains coherence with the previous choice in structural variables to avoid entire glass walls.

The strategy for HR buildings diverged from that applied in shading imputation. For HR, it was considered reasonable to apply the window share values from MR building forms (Flag F). This step provided sufficient data for window share in HR buildings, enabling the performance of the R strategy.

Finally, for any remaining missing window share values in MR and HR residential buildings (Flag D), these are filled with the category 'Many windows'. This imputation is based on observed trends and a majority of available data for these building types, suggesting this category as a reasonable default.

The final imputation strategy for HR buildings was "TWMFRRD" (Tabula, WHE, Form, Region, Region, Default), while for MR buildings, a similar strategy was applied, excluding the F (Form) step.

1.3 Thermal-related variable imputation

For thermal-related variables initially the presence of insulation was defined prior to other variables. This sequencing is methodologically strategic: it facilitates the identification and recovery of missing data related to insulation and informs the selection of default values for the remaining thermal variables. It is thus established first due to its foundational role in the subsequent processing of other thermal variables.

The other variables are treated individually and independently, although the same treatment is given to both variables describing insulation, given their similar nature. The methodology followed is described below:

Presence of insulation imputation

The overarching imputation strategy for this variable, denoted as "TFCR" (Tabula, Form, Climate, Region):

1. Flag T: Completion of missing data on the presence of insulation in European countries based on Tabula.
2. Flag F: Combination of the presence of insulation variable among different building forms. Whenever a conflict is present among the forms, the 'Yes' is prevalent over the 'No' values.
3. Flag C: Definition of the country's climate as cold, hot or neither hot or cold. In cold climates define 'Yes' while in the hot define 'No'.
4. Flag R: For countries with neither hot nor cold climate and, the missing data are completed with the data from the closest country with a common climate.

Window type imputation

The overarching strategy is denoted by "IM" (Insulation rule and Manual).

1. Flag I: An exploratory analysis is performed to understand the distribution of window types in relation to other variables such as insulation presence and climate. It suggests that insulation presence is a strong predictor for window type. For this reason missing values in insulated buildings have been set to "Double glazing," and uninsulated ones to "Single glazing".

2. Flag M (Extreme Case Checks): A manual review is performed for "Triple glazing" entries. To address any potential data inconsistencies four values were modified for 'Double glazing' considering they are insulated: LR in Poland, MR in Armenia, HR in Kosovo and Turkey. One value was modified to 'Single glazing' considering it is uninsulated: HR in Malaysia.

Temperature system imputation

The overarching strategy is denoted by "TVFIRM" (Tabula, Values from other variables, Form, Insulation rule, Regional, Manual).

- 1) Flag T (Tabula Data): Missing temperature systems values are initially imputed by leveraging data from the Tabula database for specific building types (LR and MR).
- 2) Flag V (Values from other variables - NaN to None): This rule addresses a common issue in surveys where non-mandatory questions can lead to ambiguity. Specifically, for the temperature system variable, it's unclear whether a missing value truly signifies an unknown response or if it implicitly means "None" (no temperature system present). This ambiguity arises because respondents might have left the temperature system question blank if they believed no such system existed in their building, rather than explicitly selecting a "None" option, since no respondent chose the option "None". To resolve this, we make the hypothesis that if a respondent provided information for other related building systems, but left the temperature system blank, then it is interpreted as "None". This rule analyzes the two questions preceding it: ventilation system and air infiltration. If the temperature system is missing, but any of the previous questions are answered, we infer that the respondent actively provided information for other systems. Therefore, the absence of a value for the temperature system in such cases is modified to 'None', indicating the implicit absence of a temperature system.
- 3) Flag F (From Other Forms - First Pass): This strategy addresses missing temperature system values by looking at data available for other building forms within the same country. For LR buildings the order of choice is first MR and second HR data. Similar cross-imputation logic is applied for MR (from HR then LR) and HR (from MR then LR).
- 4) Flag I (Insulation Rule): The initial distribution of temperature systems across countries and its relationship with insulation presence was studied showing a prevalence of 'Heating' systems in insulated buildings. For this reason, Missing temperature system values are set to 'Heating' in insulated countries.
- 5) Flag R (Regional Neighbors): For any temperature system values still missing, a regional imputation strategy is employed. If a neighboring country has a defined temperature system and shares a similar climate, its temperature system value is used to impute the missing value.
- 6) Flag M (Manual Checks and Corrections): Following automated imputation, a final manual review and correction phase addresses specific outlier or remaining missing cases.
 - After all the above treatment, specific countries like 'Cuba' and 'Guantanamo Bay' are still missing data and are manually assigned 'None' across all building forms.

- Tabula only provides data on the presence of heating systems, but not about cooling systems. For Greece and Spain, the values from Tabula were manually corrected to 'Both'.
- In 'Guam' and 'Senegal' buildings were described to have no window glazing and cooling systems, these last values were modified to 'None'.

Insulation thickness imputation

An exploratory analysis of insulation thickness variables demonstrated a substantial number of missing values, even within countries known to have insulated buildings. The analysis also revealed that insulated countries are heavily concentrated in Europe. This observation highlights the considerable potential of Tabula, which provides detailed data for European countries, as a valuable source for this imputation.

The Tabula database sometimes provides insulation thickness directly. However, when only the U-value of building elements is available, the insulation thickness was calculated from the thermal transmittance using the following formula:

$$e = \lambda \times \left(\frac{1}{U_{value}} - \frac{1}{U_0} \right)$$

Where:

- e is the thickness of the insulation [m]
- λ is the thermal conductivity [W/mK]
- Uvalue is the thermal transmittance of the insulated wall [W/m²K]
- U0 is the thermal transmittance of the uninsulated wall [W/m²K]

A thermal conductivity of 0.04 was used, which represents an approximation value for most common insulation materials. The values of U0 used were based on the provided values in the French 3CL method for Energy Performance Certificates (Légifrance, 2006) for the material described for the walls or roofs. When multiple buildings are described for the same building typology, only the walls and roof types described in the majority were considered.

The imputation strategy for insulation thickness, applied uniformly to both wall and roof insulation thickness, varied based on the insulation status. For uninsulated countries, the "SI" (Survey, Insulation rule) strategy was employed. Conversely, for insulated countries, the "VTFRD" (Values from other variables, Tabula, Form, Regional, Default) strategy was utilized. Both strategies are detailed below:

- Uninsulated countries
 - 1) Flag S (Survey - Not Insulated): This specifically handles cases where the survey explicitly stated no insulation. The thickness was then set to '0cm'.
 - 2) Flag I (Insulation Rule - Not Insulated): This rule extends the '0cm' imputation based on the presence of insulation status regardless of its source.
- Insulated countries
 - 1) Flag V (Values from other variables - Insulated): This strategy attempts to derive insulation thickness from available insulation period data. It prioritizes the following specific historical periods in order to infer thickness, '90s', '2000s', '70s-80s'.

- 2) Flag T (Tabula data - Insulated): Missing insulation thickness values are imputed by data from the Tabula database.
- 3) Flag F (Form - Insulated): This strategy addresses remaining missing insulation thickness values by looking at data available for other building forms within the same country.
- 4) Flag R (Regional neighbors - Insulated): For any insulation thickness values still missing, a regional imputation strategy is employed prioritizing neighboring climatically similar countries.
- 5) Flag D (Default - Insulated): For any insulation thickness values still missing, a default value is imputed using the most frequent value for the insulated countries in the studied building typology.