Ref. Ares(2025)3226295 - 22/04/2025

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101036458.

Disaggregation Methodology and Working Disaggregation Tool

D3.1

Shruthi Patil, FZJ;
Jörg Verstraete, IMP &
Noah Pflugradt, FZJ

Disclaimer

This report was written as part of the LOCALISED project under EC grant agreement 101036458. The information, documentation and figures available in this deliverable were written by the LOCALISED project consortium and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

How to quote this document

Patil, S.; Verstraete, J.; Pflugradt N. (2024), Disaggregation Methodology and Working Disaggregation Tool (LOCALISED Deliverable 3.1)

This deliverable is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

General information about this Document

Project acronym	LOCALISED
Project full title	Localised decarbonisation pathways for citizens, local administrations and businesses to inform for mitigation and adaptation action
Grant Agreement no	101036458
Deliverable number	D3.1
Deliverable title	Disaggregation Methodology and Working Disaggregation Tool
Deliverable nature	Report
Dissemination level	Public
Work package and Task	WP3 (T3.1)
Contractual delivery date	Month 30 (March 2024)
Actual delivery date	Month 30 (March 2024)
Authors	Shruthi Patil, FZJ; Jörg Verstraete, IMP; Noah Pflugradt, FZJ;
Reviewers	Matthew (FZJ); Enric (IREC)

Revision History

Version	**Date**	**Name**
V1	14.03.2024	First draft
V2	20.03.2024	First internal review
V3	26.03.2024	Second internal review
V4	12.03.2025	Revision after comments from Periodic Technical Report 2

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
DSP	Data Sharing Platform
LAU	Local Administrative Unit
NUTS	Nomenclature of Territorial Units for Statistics (from French: Nomenclature des Unités Territoriales Statistiques)
SDG	Sustainable Development Goals
SECAP	Sustainable Energy and Climate Action Plan
SOI	SDG Oriented Indicator
GDP	Gross Domestic Product
NACE	European Classification of Economic Activities
WP	Work Package
T	Task
D	Deliverable

Executive Summary

The LOCALISED project aims to downscale decarbonisation trajectories consistent with Europe's net-zero targets to local levels to support local authorities, businesses and citizens in speeding up the uptake of mitigation and adaptation actions. To achieve this, a lot of data is necessary at a local level. T3.2 dealt with the collection of the data, properly assessing which data is available and where data may be missing or not available at a sufficiently high spatial resolution. The task also involved the collection of additional data that may be necessary for dealing with these issues. T3.1, of which this deliverable is an outcome, has the objective to properly examine and define procedures to downscale the national target levels from EUCalc (European Calculator, https://www.european-calculator.eu/, Costa, 2022), to smaller regions, and to some extent deal with missing data. It builds on deliverables D3.2 and D3.3 from T3.2 which focus on data collection and links, to a lesser extent, with T2.3 (particularly D2.4 and D2.5), which also concern the collection of spatial data.

The downscaling of spatial datasets collected by LOCALISED, including the national data from EUCalc, requires their spatial disaggregation from a coarse level (e.g. country level) to fine spatial resolutions (e.g. local regions). This process is not straightforward: it requires determining the spatial distribution of the coarse-level data at the fine spatial resolution. T3.1 is aimed at examining the problem and developing a methodology. As the problem deals with many datasets - some at higher resolution than others - the idea arose to exploit the connection between them. Machine learning (in the form of Random Forest) is used to establish the relationship between different variables; additional datasets were collected in T3.2 to serve as proxy-data for this approach. T3.1 also investigates how the quality of datasets can be communicated and defines a quality measure based on the type of data processing performed on the source dataset.

From its conception, the aim of the LOCALISED project was to provide a downscaling to NUTS3 level. However, we learned, through discussions with WP5, that going beyond NUTS3 and down to LAU level would be very beneficial for local authorities to support Sustainable Energy and Climate Action Plans (SECAPs). Downscaling to a LAU level poses additional challenges to consider, relating mainly to the availability of the data and to its spatial differences.

The deliverable D3.1 - Disaggregation Methodology and Working Disaggregation Tool, presents these aspects to complete the methodology for spatially disaggregating datasets in the context of the LOCALISED project.

1 Introduction

1.1 Purpose of the work

This deliverable is the outcome of T3.1 - Extension and Improvement of the Disaggregation Tool. LOCALISED is a data-driven project, which employs - among others - spatial datasets to process data relating to climate, energy, demography, economy, etc., while maintaining the information of the regions where this data is applicable. The data for this purpose was collected in WP2 (for climate related data) and WP3 (for other spatial data) and provided to other WPs. Climate data is used in WP4 to determine possible adaptation measures for each region, in WP5 to support the creation of SECAPs, and in WP6 to disaggregate intersectoral pathways. Furthermore, WP7 uses the data to identify local businesses and industries while the tools developed in WP8 use the data for profiling.

Data was sourced from different databases for WP2 and WP3 (Deliverables D2.5 (Patil et al., 2023) and D3.3 (Verstraete et al., 2023)), respectively and its quality assessed. The datasets are typically available at a specific spatial resolution - Some is provided at the level of the EU member states (a NUTS0 level), while the remainder is available at the level of smaller regions or the municipality level. The different WPs that comprise the LOCALISED project work at different spatial levels and need data on a matching spatial level. The purpose of the developed disaggregation tool is to offer a mechanism for supplying datasets at fine spatial resolutions. As such, this implies the need for so-called spatial disaggregation, which distributes the data defined at a given spatial level across the smaller regions contained within. This is a non-trivial problem which requires an analysis of available data, the collection of additional datasets to support the operation, and the development of methodologies to assess the quality and suitability of the datasets. In addition to data disaggregation, it is also necessary to incorporate mechanisms to address missing data and assess its quality to estimate the reliability of subsequent analyses using the data.

1.2 Spatial hierarchy of statistical regions in the EU

Many collected datasets have an associated spatial resolution. Two obvious examples are population and GDP: which can be considered at the spatial resolution of countries or at different administrative spatial resolutions within the country. Countries have their own administrative divisions at multiple levels, but these differ between countries which complicates comparisons between regions and between countries. To facilitate comparison, the EU has developed the NUTS hierarchy. NUTS stands for Nomenclature of Territorial Units for Statistics - the acronym stems from French: Nomenclature des Unités Territoriales Statistiques - and defines a hierarchy of three spatial subdivisions for the EU member states (Figure 1). The highest level in the hierarchy is the country level (NUTS0), and each proceeding level partitions the previous level into smaller regions, thereby increasing its spatial resolution. The idea behind the NUTS partitioning is that the region definitions to some extent relate to the spatial distribution of the population. As a result of this, the definitions of the regions can change and are consequently revised approximately every four years. For the LOCALISED project, the NUTS definitions for the year 2016 are considered (NUTS - GISCO - Eurostat, 2021). This choice is based on the adoption of this NUTS definition standard in 2018, as the majority of data collected within the project adheres to this standard. The NUTS partitioning tries to follow the countries' administrative divisions, but occasionally it deviates. Poland, for example, is administratively divided in 16 regions, but for the NUTS2 division the Mazowian administrative region is split into two NUTS2 regions, yielding a total of 17 NUTS2 regions.

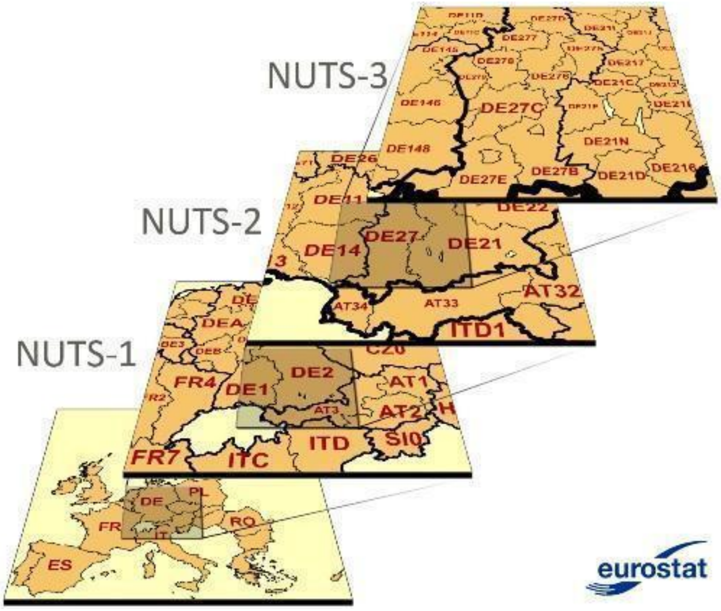


Figure 1 Illustration of NUTS levels (source: Eurostat).

In many countries, the NUT3 level still contains regions larger than the lowest level administrative units. As a result, an additional level is added below NUTS3: the LAU-level (Local Administrative Unit). This level is however more problematic in its definition and can potentially change year to year.

1.3 Spatial levels in the LOCALISED project

In its initial conception, LOCALISED intended to provide a downscaling of the EU pathways from a NUTS0 to a NUTS3 level. WP2 provides decarbonisation pathways on a NUTS0 level using the EUCalc tool (Costa, 2022), which WP3 spatially disaggregates for WP4 and WP5. WP4 is developing the Modular Integrated Decarbonisation Adaptation Solver (MIDAS) model that aims to provide regional mitigation and adaptation measures, which are defined on a NUTS3 level. WP5 aims to support local authorities with indicators oriented towards Sustainable Development Goals (SDGs) and Sustainable Energy and Climate Action Plans (SECAPs). Here it became clear that the LAU level would be much more interesting and provide better support for the local authorities. Table 1 provides a short overview of the spatial levels required by other WPs.

Table 1 Spatial level at which different WPs work.

Work package	Task	Spatial Level
WP2	Provides decarbonisation pathways	NUTS0
WP4	Provides regional mitigation and adaptation measures	NUTS3
WP5	Requires data to fill in SECAPs	LAU
WP7	Requires business-related data	NUTS2

As specified in the proposal, D3.1 deals with the disaggregation of EUCalc decarbonisation pathways from NUTS0 to NUTS3. However, the data requirements in different WPs differ in terms of the acceptable spatial level, and LAU has become a pertinent level. Furthermore, WP5 and WP7 require data that is not part of the EUCalc pathways and was instead collected from different European databases. Examples of such data are "people at risk of income poverty after social transfers" and "gross value added in agriculture, manufacturing and transportation sectors", which occur respectively in WP5 and WP7. Most of such data is however not available at a fine NUTS3 or LAU spatial resolution level. Therefore, such data is as well disaggregated.

In the light of this, the following steps were considered for handling data. First, the data requirements (which data is needed at which spatial level) for the different WPs were determined. Then public databases were searched for the appropriate data. The data was collected at the finest spatial resolution at which it is available, and all datasets ultimately disaggregated to LAU level. The developed Data Sharing Platform (DSP) (preview provided in D3.3 and final version to appear as D3.4) is designed to allow queries and provide data at any spatial resolution. This offers transparent access to the spatial datasets, independent of the resolution of the original source data.

During the data collection process, national databases, such as the Federal Statistical Office of Germany (https://www.destatis.de/EN/Home/_node.html), were also explored. These databases provide data specific to individual countries, leading to inconsistencies in the collected datasets, as similar data may not be available for all EU countries. Additionally, searching each country's national database is challenging due to differences in language and data structures. However, these national sources often contain valuable information. Therefore, at this stage, datasets from select project partner countries—Germany, Poland, and Spain—were examined. This investigation helped collect some datasets for verifying disaggregation methods and provided minor insights into additional data availability across different countries.

Table 2 provides an overview of the amount of data collected at the different spatial resolutions. At a NUTS0 level only 27 regions exist, representing the 27 member states. Apart from the EUCalc pathway data, 332 additional datasets were collected at this level. No datasets were collected at a NUTS1 level. As data commonly exists at a higher spatial resolution or at NUTS0 level, NUTS1 is not a common spatial level for data reporting. The number of regions quickly grows with the NUTS levels, reaching 95314 regions at LAU level. To cope with the potential issues of changing definitions at the LAU level, most data collected at LAU level is point-data that has associated x-y coordinates available. The precise x-y coordinates of the data allow us to overlap the data with LAU regions to obtain data for these regions. If there is a requirement to change LAU definitions in the future, one could overlap the data again to obtain data for new LAU regions.

Table 2 Overview of the number of datasets and amount of data at different spatial levels.

Spatial level	**Number of regions across the EU**	**Number of datasets**
NUTS0	27	332 + EUCalc pathway data
NUTS1	88	-
NUTS2	232	53
NUTS3	1155	160
LAU	95314	124

While offering transparent access to the spatial datasets at different resolutions is useful, it also means that not all data will be of the same quality. The data available at the requested resolution or finer will be of high quality. Spatial disaggregation of data will, inherently, lower the data quality. The methodology to assess the quality level is discussed in Section 2.

Further, data on various topics is collected within the project. Some required data may not be available in some regions or even countries. The method in which missing data is treated is described in Section 3. Last, the disaggregation of the data itself is a stepwise process from the lowest level to the highest, ending with the EUCalc pathway This entire approach is elaborated in Section 4.

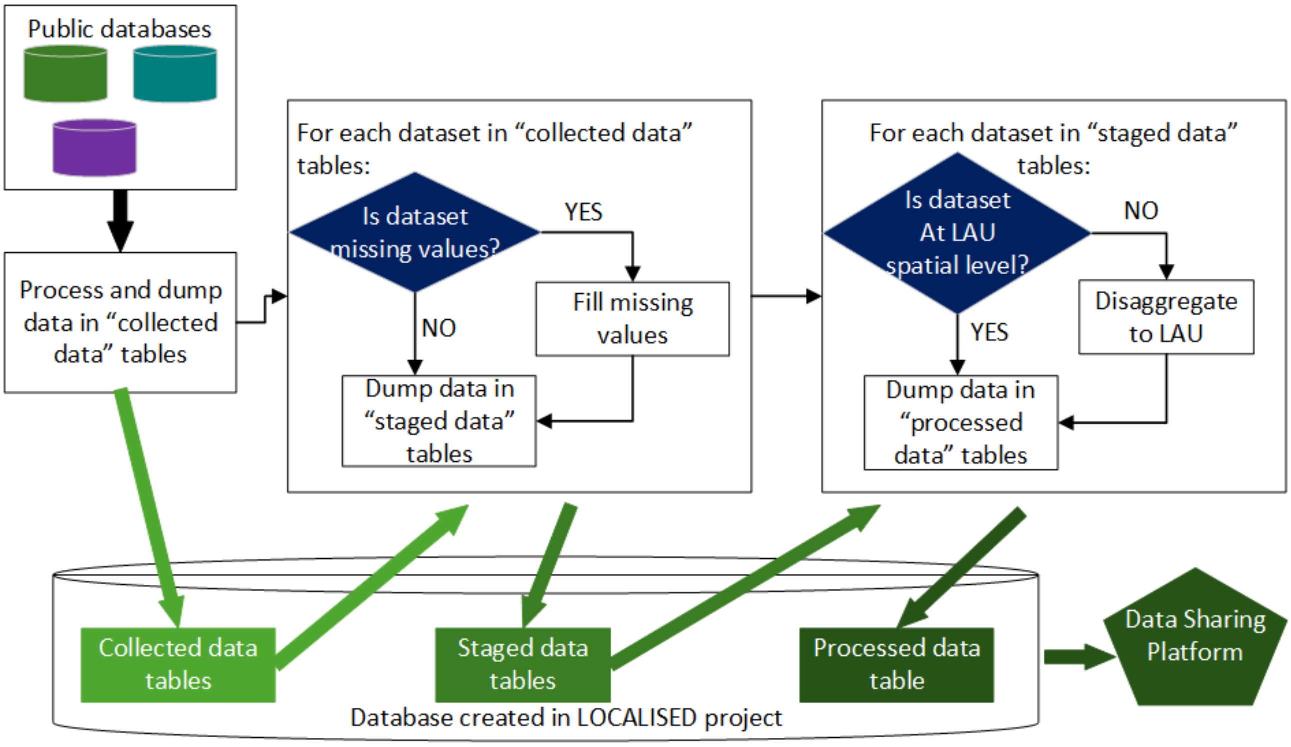

Public databases
Process and dump data in “collected data” tables
For each dataset in “collected data” tables:
Is dataset missing values?
YES
Fill missing values
NO
Dump data in “staged data” tables
For each dataset in “staged data” tables:
Is dataset At LAU spatial level?
NO
Disaggregate to LAU
YES
Dump data in “processed data” tables
Collected data tables
Staged data tables
Processed data table
Database created in LOCALISED project
Data Sharing Platform

Figure 2 Workflow for imputation and disaggregation.

Figure 2 shows the workflow for the data imputation and disaggregation. Both these stages take place outside of the DSP (Section 5) and can be considered as processing steps between the data acquisition and the data sharing. The collected data is stored in the database, in a set of tables. Missing data is identified and filled (also referred to as data imputation). This data is stored separately in another set of tables in the database. The next step is to perform the spatial disaggregation. The disaggregated data is again stored separately. This last level is made available to the end users through the DSP. The benefit of this approach is that the collection and analysis of data can happen without affecting the content of the DSP, while revisions of or improvements to the data in the DSP can be performed periodically and are traceable. In addition, the lack of additional processing needs at the time of data querying improves the performance of the database.

2 Data Quality Rating

A large amount of data is collected from various sources. As previously explained, all data is collected at the lowest spatial level (i.e. the highest spatial resolution at which it is available), while the DSP provides a transparent unified access method to query data at any spatial level. This implies that not all data will be of the same quality: if we, for example, consider data at NUTS3 level, then some data will have been sourced at NUTS3 or even LAU level, whereas other data may have been available only at NUTS2 or higher. Data queried at a spatial level higher than the one at which it was collected should be of good quality, provided there is no missing data. However, data that is queried at spatial level lower than its source data, must be disaggregated and, as a result, is most likely to be of lower quality. Similarly, while methods were developed to account for missing datasets, the quality of the queried data that refers to this missing data will also be lower.

All data collected and processed in the LOCALISED project is annotated with a five-stage quality rating: VERY LOW, LOW, MEDIUM, HIGH, VERY HIGH. The final quality rating for a value depends on its availability (with respect to spatial level and missing data) and the processing steps performed to obtain its value.

In the first stage, required datasets are collected at their available spatial resolution. At this stage, the values contained in these datasets are assigned the quality rating VERY HIGH.

The next stage is data imputation to account for missing data. The procedure described in Section 3 assigns a quality rating to imputed data based on how the imputed value was determined.

The final stage is spatial disaggregation (for datasets that were not sourced at LAU level). The process of spatial disaggregation generally lowers the quality rating of the input dataset, depending on the level of the source data and the quality of the disaggregation itself. This procedure is explained in Section 4, together with the other details pertaining to spatial disaggregation.

3 Missing Value Imputation

The first processing step when a new dataset is received is missing value imputation and the resulting assignment of quality levels. Dealing with missing values is different for the different spatial levels, due to the size of the datasets. The process will be discussed beginning at the LAU level.

3.1 LAU regions

At the LAU level, most data consist of point data. As an example, consider the power plant locations with their generation capacity sourced from JRC Open Power Plants Database (JRC-PPDB-OPEN) (Kanellopoulos et al., 2019). If there is no power plant in a region, it won't be represented in the dataset since the dataset does not model the absence of power plants. In this case, there will be many LAU regions without data for a power plant, but those values can reliably be set to 0. The quality rating in this case is *VERY HIGH*.

One exception is the survey data collected from EUROSTAT regarding the satisfaction of people with public transport. This data is required byWP5 to fill out the SECAPs but is only available for some LAU regions. Contrary to the previous example, this clearly concerns missing data and imputation is not as straightforward as above. A choice was made to complete the data by assigning a value of 0 to the missing values and assigning a quality rating of *VERY LOW*.

This procedure was applied for the five different datasets that relate to survey variables:

'percentage_of_people_very_satisfied_with_public_transport',
'percentage_of_people_rather_satisfied_with_public_transport',
'percentage_of_people_rather_unsatisfied_with_public_transport',
'percentage_of_people_not_at_all_satisfied_with_public_transport'
'percentage_of_people_with_unknown_satifactory_level_with_public_transport'

3.2 NUTS3 regions

The idea behind the methodology for data imputation is that datasets (variables) are not fully independent and connections between datasets can be made. A machine learning approach aims to uncover such connections and employ them to determine missing data.

At NUTS3 level, there are 1155 regions across the EU (Table 2, Section 1.3). This provides a large enough sample size to employ machine learning algorithms to impute missing values. At this level, there are 160 datasets, with 32 of them suffering from missing values. Figure 3 shows the number of missing values for a sample of the NUTS3 datasets. For some variables the number of missing values is rather small, but for others it is nearly half of the data.

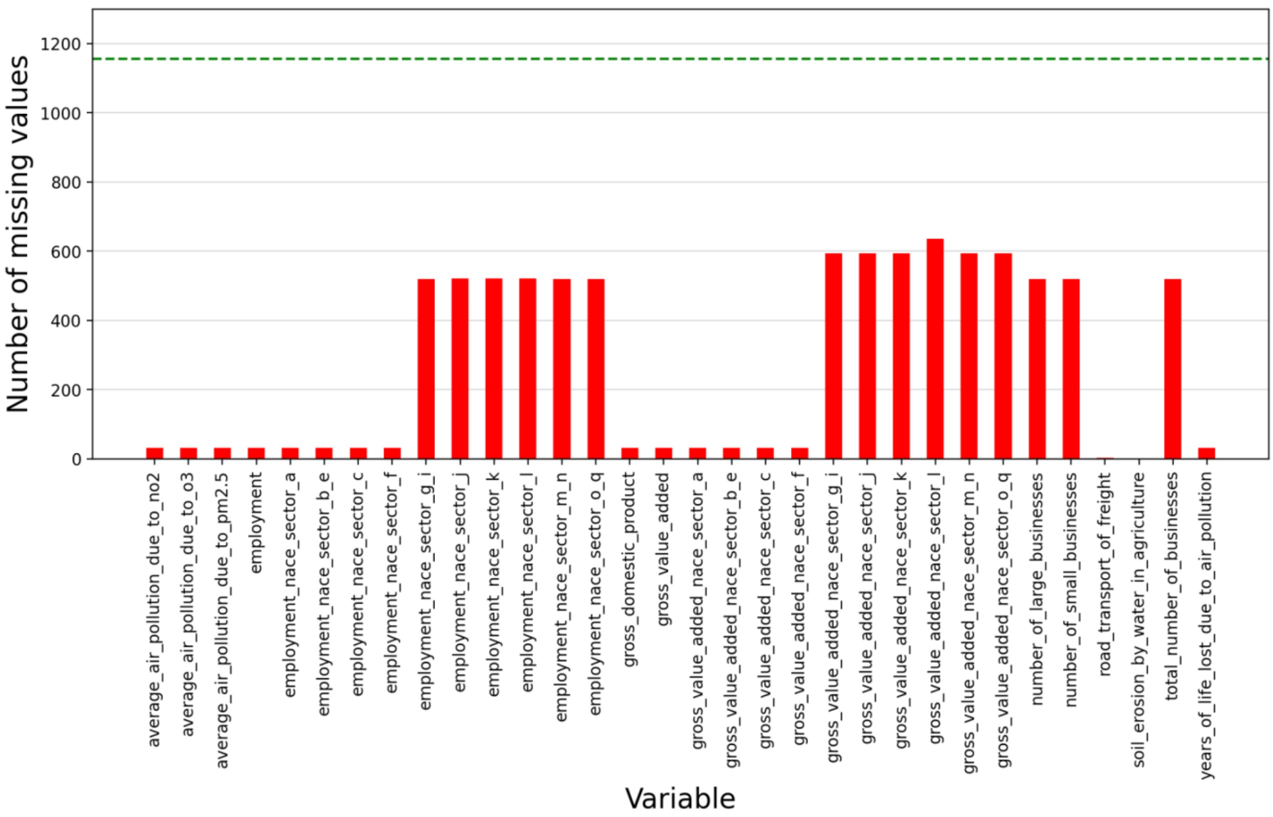


Figure 3 The number of missing values per variable at NUTS3. The green line at the top indicates 1155, the total number of values.

To perform the imputation, training data is necessary. For this purpose, we consider all the NUTS3 levels that do not have missing data. This yields 128 complete datasets with data for each of the 1155 regions. In addition, the datasets at LAU level (excluding the survey datasets, as explained in Section 3.1) are aggregated to NUTS3 level, yielding a complementary 119 complete datasets. In total, this results in 247 datasets that together constitute the predictor set for the model. A single experiment considers all these datasets, using a random selection of 10% of the 1155 regions that do not have missing data for testing and the remaining 90% for training. The iterative imputer is set to perform 10 such experiments - each time picking a different random set of training and testing data - after which the imputed results are verified using the R-squared method.

R-squared score ranges from 1 to negative infinity. A perfect prediction would result in an R-squared score of 1. Subsequently, worse predictions (thus worse imputation) possess lower values. A negative R-squared value indicates that the model's predicted values perform worse than if one were to use the average of existing values as a filler for the missing values. As it has an upper limit, the R-squared score lends itself as a good indicator for the quality of the imputed data.

These imputed values are always floats, which are arithmetically rounded in case the desired values are integers (For example, population in a region should be an integer). This rounding occurs before the calculation of the R-square score. Since 10 different experiments are performed, 10 R-squared scores are obtained. Among these, the lowest R-squared score is considered as an indicator for the quality of the dataset after imputation of the missing data. Table 3 shows how the R-squared scores are mapped to the data-quality labels.

Table 3 Quality ratings for missing value imputation.

R2	**Quality Rating**
> 0.9	*HIGH*
> 0.5 and <=0.9	*MEDIUM*
> 0.2 and <=0.5	*LOW*
<= 0.2	*VERY LOW*

Note that *VERY HIGH* is missing even in case of a perfect prediction (R-squared score equal to 1), we do not assign the label *VERY HIGH*, as this was reserved for data that is directly obtained from the source.

The final R-squared scores (i.e. the minimum of the R-squared scores of the 10 experiments for each variable) for the data imputation of the 32 NUTS3-level sets with missing data are shown on Figure 4. The variables for which the imputation worked best (R-squared score > 0.9) are:

- 'employment',
- 'employment_nace_sector_b_e',[1]
- 'employment_nace_sector_c',
- 'gross_domestic_product',
- 'gross_value_added'.

On the other side of the spectrum, the worst imputation results (R-squared score ≤ 0.5) are seen for the variables:

- 'employment_nace_sector_a',
- 'gross_value_added_nace_sector_c',
- 'gross_value_added_nace_sector_j',
- 'gross_value_added_nace_sector_k',
- 'road_transport_of_freight'.

[1] NACE sectors refer to the statistical classification of economic activities in the European community. The sector descriptions are as follows:

nace_sector_a - agriculture, forestry and fishing
nace_sector_b_e - mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; water supply; sewerage, waste management and remediation activities
nace_sector_c - manufacturing
nace_sector_f - construction
nace_sector_g_i - wholesale and retail trade; repair of motor vehicles and motorcycles; transportation and storage; accommodation and food service activities
nace_sector_j - information and communication
nace_sector_k - financial and insurance activities
nace_sector_l - real estate activities
nace_sector_m_n - professional, scientific and technical activities; administrative and support service activities
nace_sector_o_q - public administration and defence; compulsory social security; education; human health and social work activities

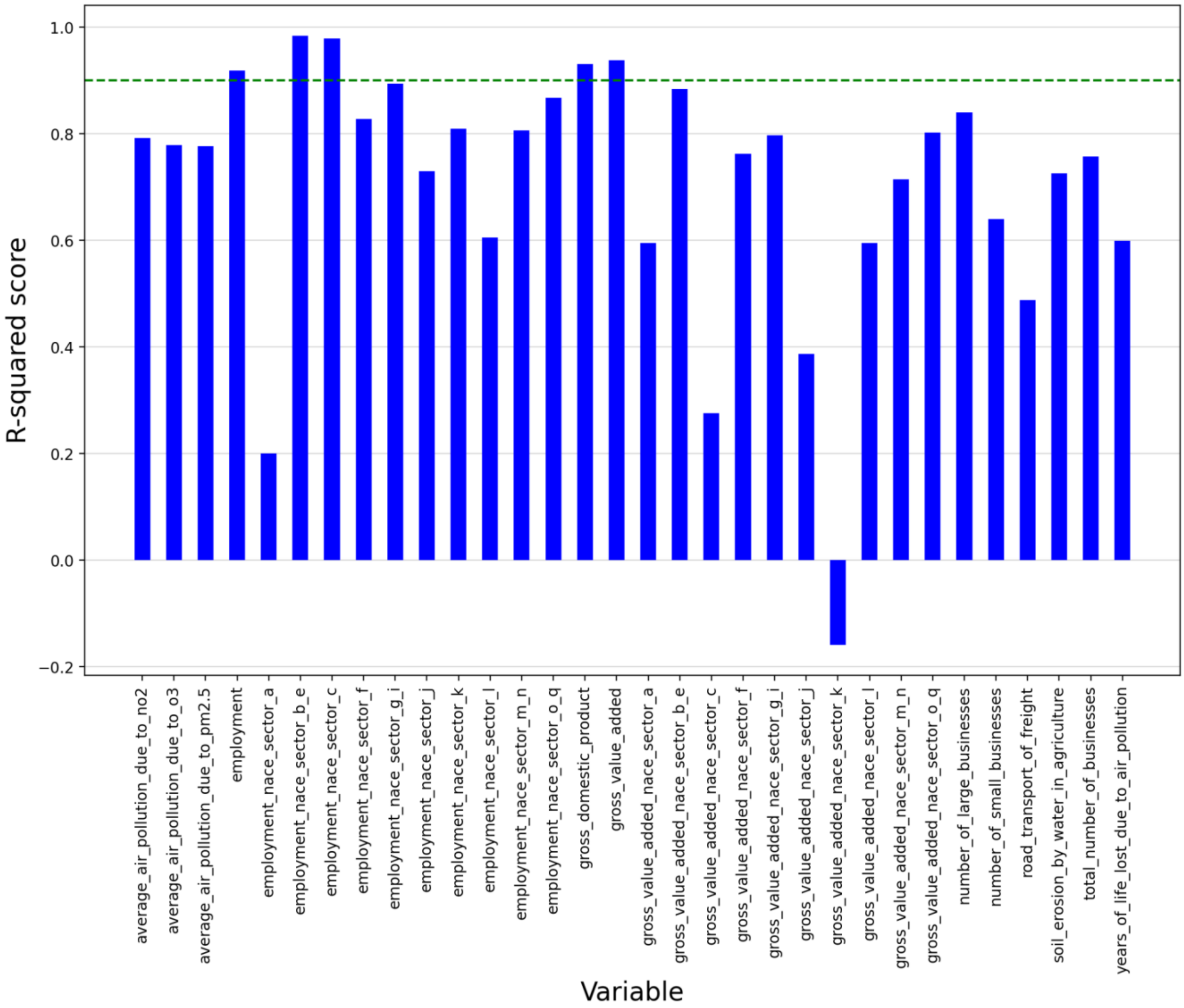


Figure 4 R-squared score for each variable that is missing values. The score shows the minimum of the 10 experiments. The higher the score, the better the imputation; the green light marks the 0.9 limit that matches with HIGH.

To understand the differences in performance, we checked the Pearson correlation between these variables, the top 3 correlated variables for each, and the number of missing values in each case. If the variables are highly correlated with other datasets, the correlated variables could be used by the random forest model for prediction. The results for the best performing variables are shown in Figure 5.

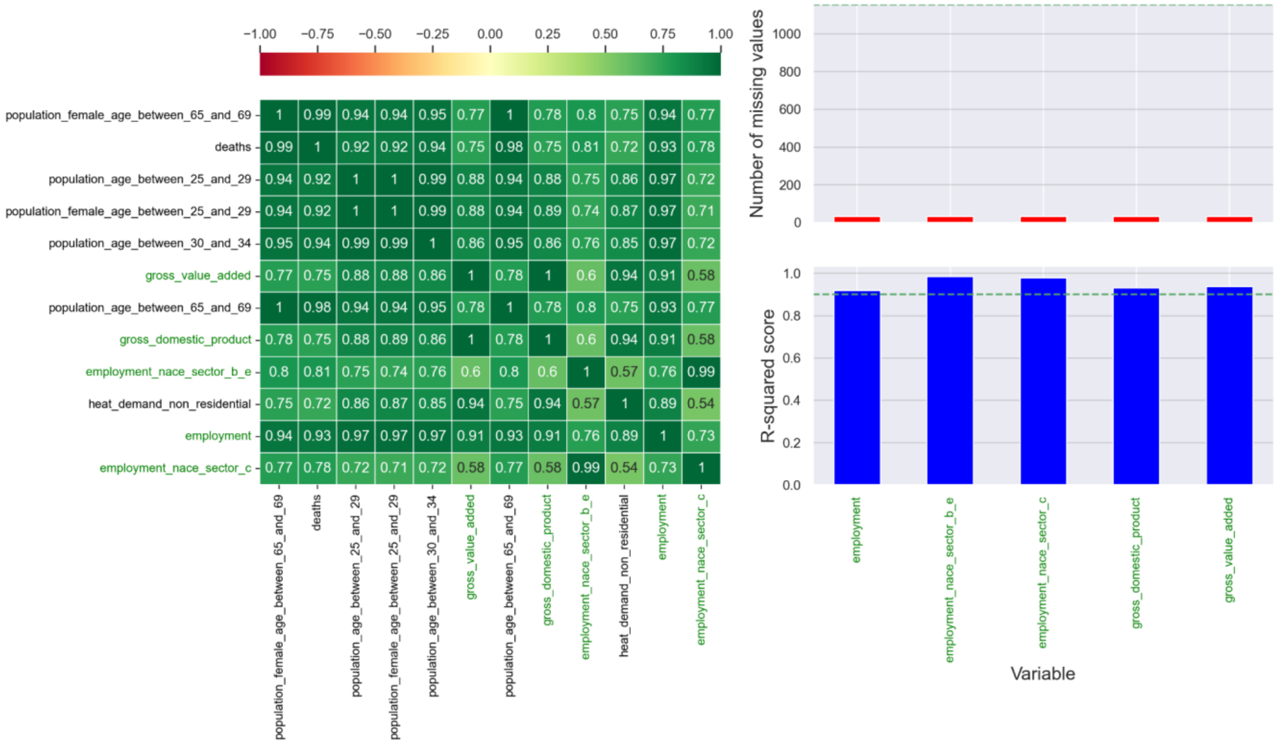
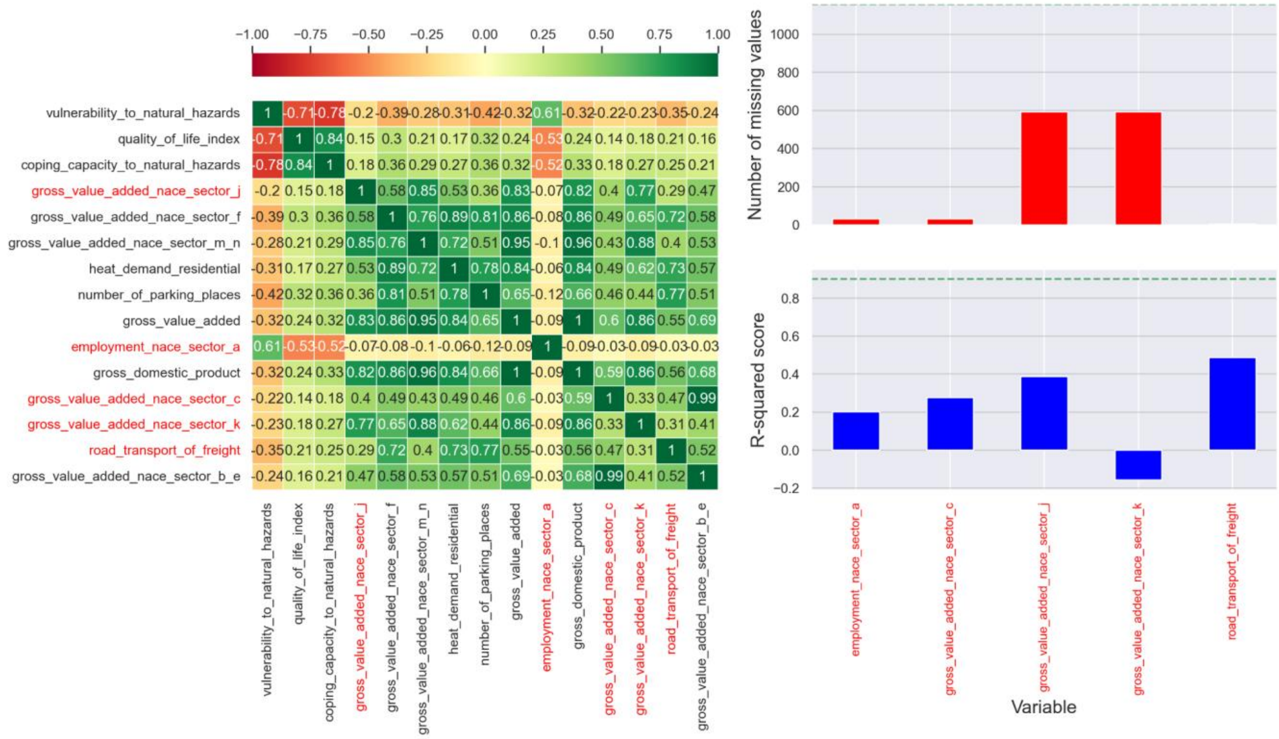


Figure 5 Correlation matrix for the best performing variables at NUTS3. These variables are highlighted in green. [top-right] Number of missing values in case of each best performing variable. [bottom-right] The R-squared scores in case of each best performing variable.

From Figure 5, it is evident that very few records are missing for the best performing variables (marked in green). Additionally, the figure shows that they highly correlate with some of the predictor variables. This creates the perfect combination and leads to a good imputation performance of the imputer.

A similar figure (Figure 6) provides insights into the worst performing variables. Here, different explanations exist for the different variables. In the case of 'gross_value_added_nace_sector_j' and 'gross_value_added_nace_sector_k', the number of missing values is quite high, leading to poor data imputation. However, while 'employment_nace_sector_a' and ‘road_transport_of_freight' have a low number of missing values, they exhibit only low correlations with any of the possible predictor variables. Of note, 'gross_value_added_nace_sector_c' has a relatively low number of missing values but only correlates highly with the variable 'gross_value_added_nace_sector_b_e'. As it turns out, this variable also misses data in the same regions as 'gross_value_added_nace_sector_c' and cannot serve as a predictor. This in turn also leads to poor data imputation.



	vulnerability_to_natural_hazards	quality_of_life_index	coping_capacity_to_natural_hazards	gross_value_added_nace_sector_j	gross_value_added_nace_sector_f	gross_value_added_nace_sector_m_n	heat_demand_residential	number_of_parking_places	gross_value_added	employment_nace_sector_a	gross_domestic_product	gross_value_added_nace_sector_c	gross_value_added_nace_sector_k	road_transport_of_freight	gross_value_added_nace_sector_b_e
vulnerability_to_natural_hazards	1	-0.71	-0.78	-0.2	-0.39	-0.28	-0.31	-0.42	-0.32	0.61	-0.32	-0.22	-0.23	-0.35	-0.24
quality_of_life_index	-0.71	1	0.84	0.15	0.3	0.21	0.17	0.32	0.24	-0.53	0.24	0.14	0.18	0.21	0.16
coping_capacity_to_natural_hazards	-0.78	0.84	1	0.18	0.36	0.29	0.27	0.36	0.32	-0.52	0.33	0.18	0.27	0.25	0.21
gross_value_added_nace_sector_j	-0.2	0.15	0.18	1	0.58	0.85	0.53	0.36	0.83	-0.07	0.82	0.4	0.77	0.29	0.47
gross_value_added_nace_sector_f	-0.39	0.3	0.36	0.58	1	0.76	0.89	0.81	0.86	-0.08	0.86	0.49	0.65	0.72	0.58
gross_value_added_nace_sector_m_n	-0.28	0.21	0.29	0.85	0.76	1	0.72	0.51	0.95	-0.1	0.96	0.43	0.88	0.4	0.53
heat_demand_residential	-0.31	0.17	0.27	0.53	0.89	0.72	1	0.78	0.84	-0.06	0.84	0.49	0.62	0.73	0.57
number_of_parking_places	-0.42	0.32	0.36	0.36	0.81	0.51	0.78	1	0.65	-0.12	0.66	0.46	0.44	0.77	0.51
gross_value_added	-0.32	0.24	0.32	0.83	0.86	0.95	0.84	0.65	1	-0.09	1	0.6	0.86	0.55	0.69
employment_nace_sector_a	0.61	-0.53	-0.52	-0.07	-0.08	-0.1	-0.06	-0.12	-0.09	1	-0.09	-0.03	-0.09	-0.03	-0.03
gross_domestic_product	-0.32	0.24	0.33	0.82	0.86	0.96	0.84	0.66	1	-0.09	1	0.59	0.86	0.56	0.68
gross_value_added_nace_sector_c	-0.22	0.14	0.18	0.4	0.49	0.43	0.49	0.46	0.6	-0.03	0.59	1	0.33	0.47	0.99
gross_value_added_nace_sector_k	-0.23	0.18	0.27	0.77	0.65	0.88	0.62	0.44	0.86	-0.09	0.86	0.33	1	0.31	0.41
road_transport_of_freight	-0.35	0.21	0.25	0.29	0.72	0.4	0.73	0.77	0.55	-0.03	0.56	0.47	0.31	1	0.52
gross_value_added_nace_sector_b_e	-0.24	0.16	0.21	0.47	0.58	0.53	0.57	0.51	0.69	-0.03	0.68	0.99	0.41	0.52	1

Figure 6 Correlation matrix for the worst performing variables at NUTS3. The worst variables are highlighted in red. [top-right] The number of missing values in case of each worst performing variable. [bottom-right] The R-squared scores in case of each worst performing variable.

3.3 NUTS2 regions

The methodology for data imputation of missing data for NUTS2 regions is similar to that for NUTS3. However, with only 232 regions at NUTS2 level (Table 2, Section 1.3), the dataset is on the smaller side, which may result in a poor performance of any algorithm that requires training.

The project collected 53 datasets at this level, 40 of which suffer from missing data. Figure 7 shows the number of missing values for a sample of these 40 NUTS2 datasets. The number of missing values for some is rather small, but for others well over 50%.

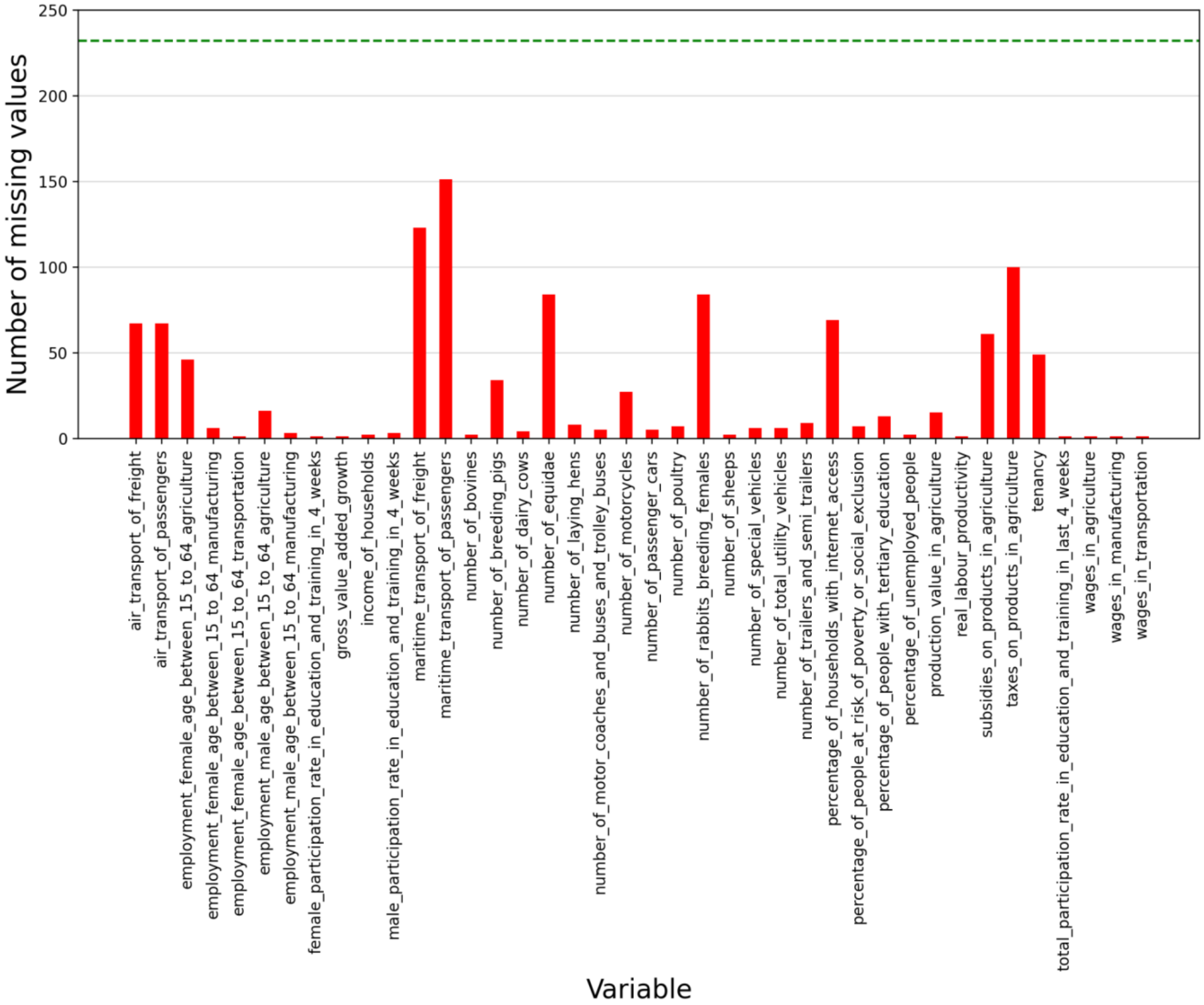


Figure 7 The number of missing values per variable at NUTS2. The green line at the top indicates 232, the total number of values.

Similarly to how NUTS3 data imputation benefited from the aggregated datasets at LAU level, NUTS2 data imputation benefits from both the aggregated LAU datasets (124 datasets) and NUTS3 datasets (128 that did not have missing data).

Figure 8 shows the final R-squared values (the minimum of the 10 experiments) for the NUTS2 datasets that were missing data. Compared to the NUTS3 level, the imputation does not work so well for a larger number of datasets.

The best performing variables (R-squared score > 0.9) are

- 'female_participation_rate_in_education_and_training_in_4_weeks',
- 'number_of_passenger_cars',
- 'tenancy',
- 'total_participation_rate_in_education_and_training_in_last_4_weeks'

The correlation graph (Figure 9) suffices for explaining the good performance: each of these variables has at least one predictor that correlates very well.

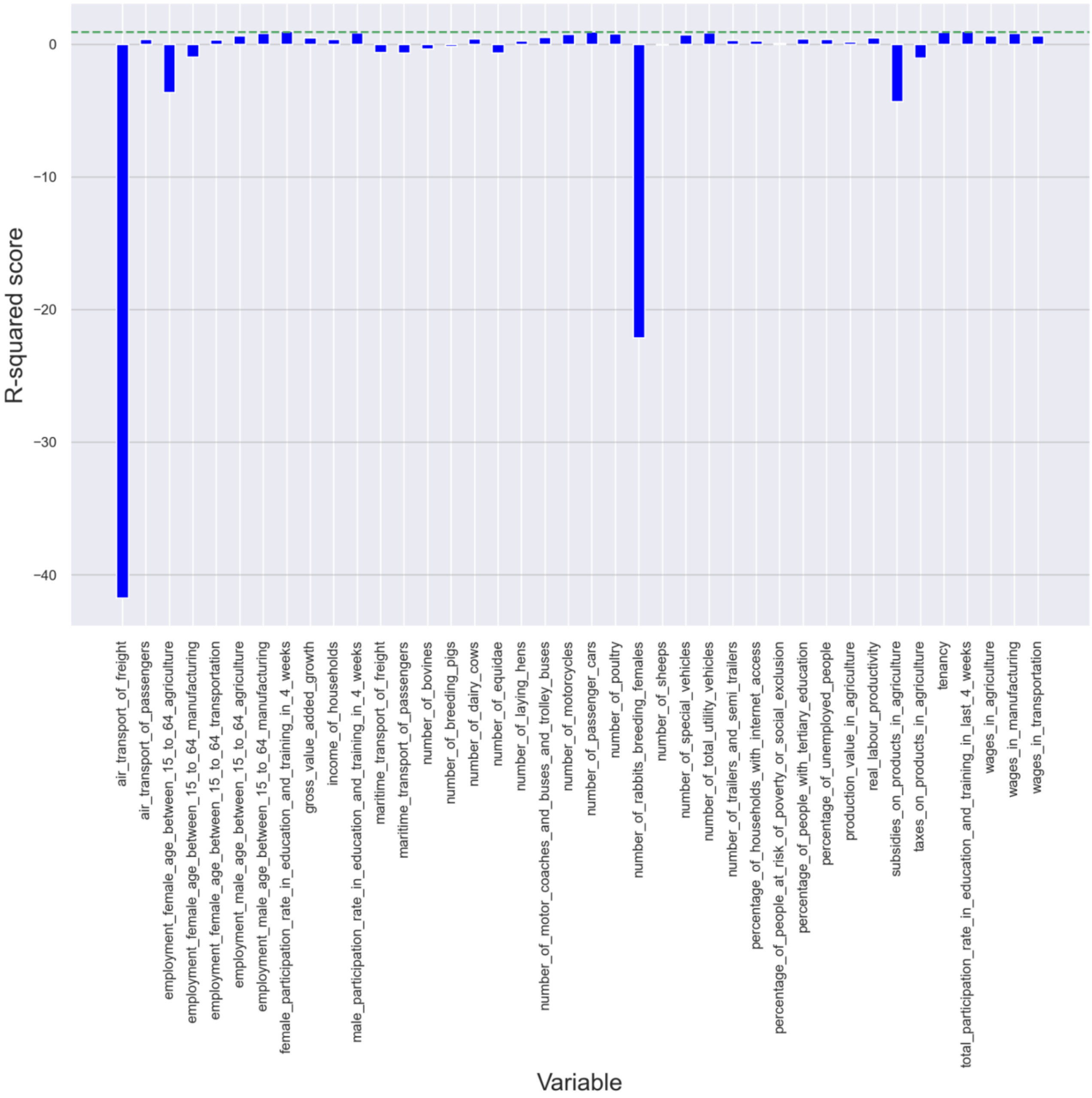
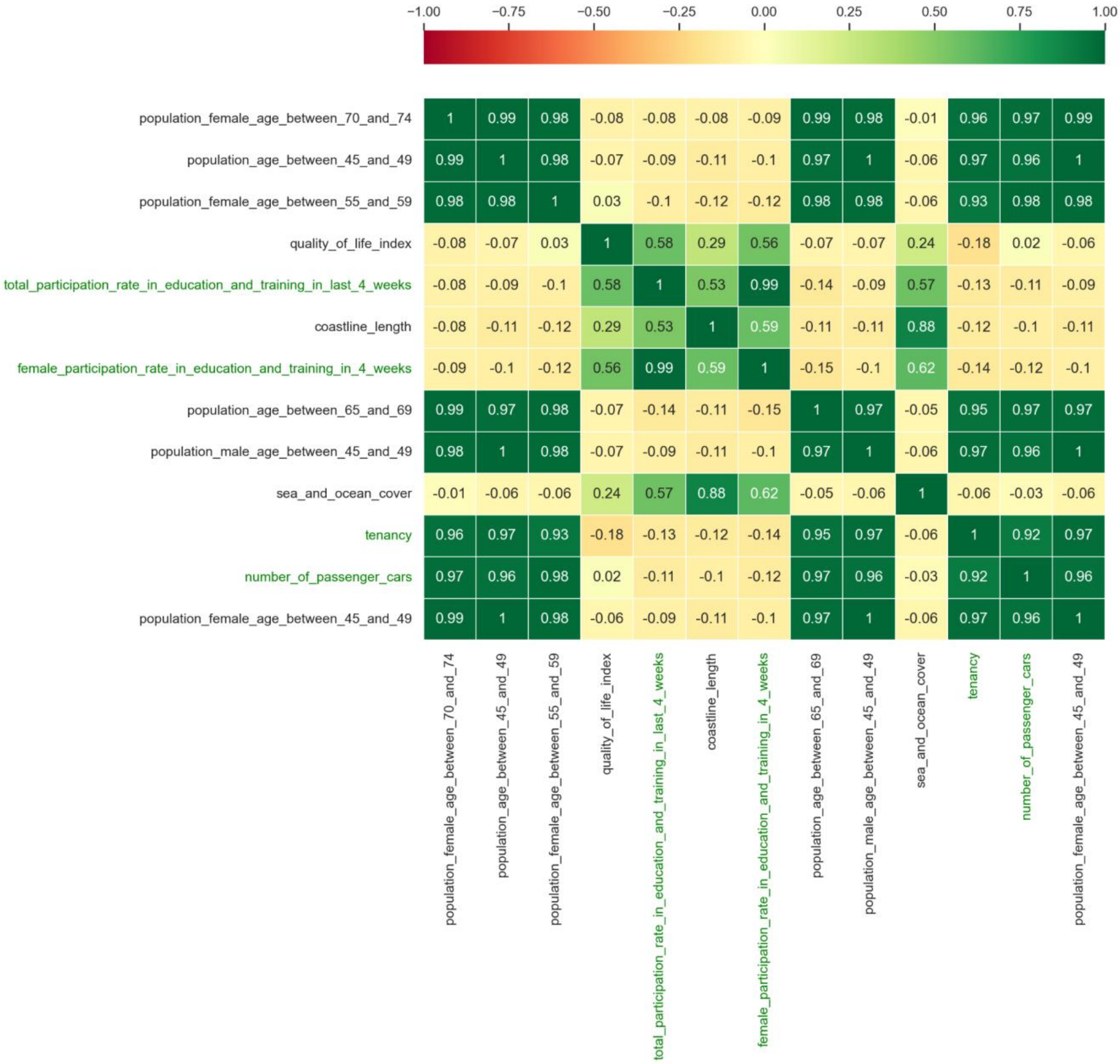


Figure 8 R-squared score for each variable that is missing values. The score shows the minimum of the 10 experiments. The higher the score, the better the imputation. The green line shows 0.9 that indicated HIGH quality rating.



	population_female_age_between_70_and_74	population_age_between_45_and_49	population_female_age_between_55_and_59	quality_of_life_index	total_participation_rate_in_education_and_training_in_last_4_weeks	coastline_length	female_participation_rate_in_education_and_training_in_4_weeks	population_age_between_65_and_69	population_male_age_between_45_and_49	sea_and_ocean_cover	tenancy	number_of_passenger_cars	population_female_age_between_45_and_49
population_female_age_between_70_and_74	1	0.99	0.98	-0.08	-0.08	-0.08	-0.09	0.99	0.98	-0.01	0.96	0.97	0.99
population_age_between_45_and_49	0.99	1	0.98	-0.07	-0.09	-0.11	-0.1	0.97	1	-0.06	0.97	0.96	1
population_female_age_between_55_and_59	0.98	0.98	1	0.03	-0.1	-0.12	-0.12	0.98	0.98	-0.06	0.93	0.98	0.98
quality_of_life_index	-0.08	-0.07	0.03	1	0.58	0.29	0.56	-0.07	-0.07	0.24	-0.18	0.02	-0.06
total_participation_rate_in_education_and_training_in_last_4_weeks	-0.08	-0.09	-0.1	0.58	1	0.53	0.99	-0.14	-0.09	0.57	-0.13	-0.11	-0.09
coastline_length	-0.08	-0.11	-0.12	0.29	0.53	1	0.59	-0.11	-0.11	0.88	-0.12	-0.1	-0.11
female_participation_rate_in_education_and_training_in_4_weeks	-0.09	-0.1	-0.12	0.56	0.99	0.59	1	-0.15	-0.1	0.62	-0.14	-0.12	-0.1
population_age_between_65_and_69	0.99	0.97	0.98	-0.07	-0.14	-0.11	-0.15	1	0.97	-0.05	0.95	0.97	0.97
population_male_age_between_45_and_49	0.98	1	0.98	-0.07	-0.09	-0.11	-0.1	0.97	1	-0.06	0.97	0.96	1
sea_and_ocean_cover	-0.01	-0.06	-0.06	0.24	0.57	0.88	0.62	-0.05	-0.06	1	-0.06	-0.03	-0.06
tenancy	0.96	0.97	0.93	-0.18	-0.13	-0.12	-0.14	0.95	0.97	-0.06	1	0.92	0.97
number_of_passenger_cars	0.97	0.96	0.98	0.02	-0.11	-0.1	-0.12	0.97	0.96	-0.03	0.92	1	0.96
population_female_age_between_45_and_49	0.99	1	0.98	-0.06	-0.09	-0.11	-0.1	0.97	1	-0.06	0.97	0.96	1

Figure 9 Correlation matrix for the best performing variables at NUTS2. The best performing variables are highlighted in green.

The variables for which the imputation provided poor results (R-squared score ≤ 0.5) are

- 'air_transport_of_freight',
- 'air_transport_of_passengers',
- 'employment_female_age_between_15_to_64_agriculture',
- 'employment_female_age_between_15_to_64_manufacturing',
- 'employment_female_age_between_15_to_64_transportation',
- 'gross_value_added_growth',
- 'income_of_households',
- 'maritime_transport_of_freight',
- 'maritime_transport_of_passengers',
- 'number_of_bovines',
- 'number_of_breeding_pigs',
- 'number_of_dairy_cows',
- 'number_of_equidae',
- 'number_of_laying_hens',
- 'number_of_rabbits_breeding_females',
- 'number_of_sheep',
- 'number_of_trailers_and_semi_trailers',
- 'percentage_of_households_with_internet_access',
- 'percentage_of_people_at_risk_of_poverty_or_social_exclusion',
- 'percentage_of_people_with_tertiary_education',
- 'percentage_of_unemployed_people',
- 'production_value_in_agriculture',
- 'real_labour_productivity',
- 'subsidies_on_products_in_agriculture',
- 'Gaxes_on_products_in_agriculture'.

Figure 10 shows the correlations between these variables and their top 3 correlating variables. Here as well, the imputation does not work well either due to low correlations with other variables or the correlating variables are missing data themselves.

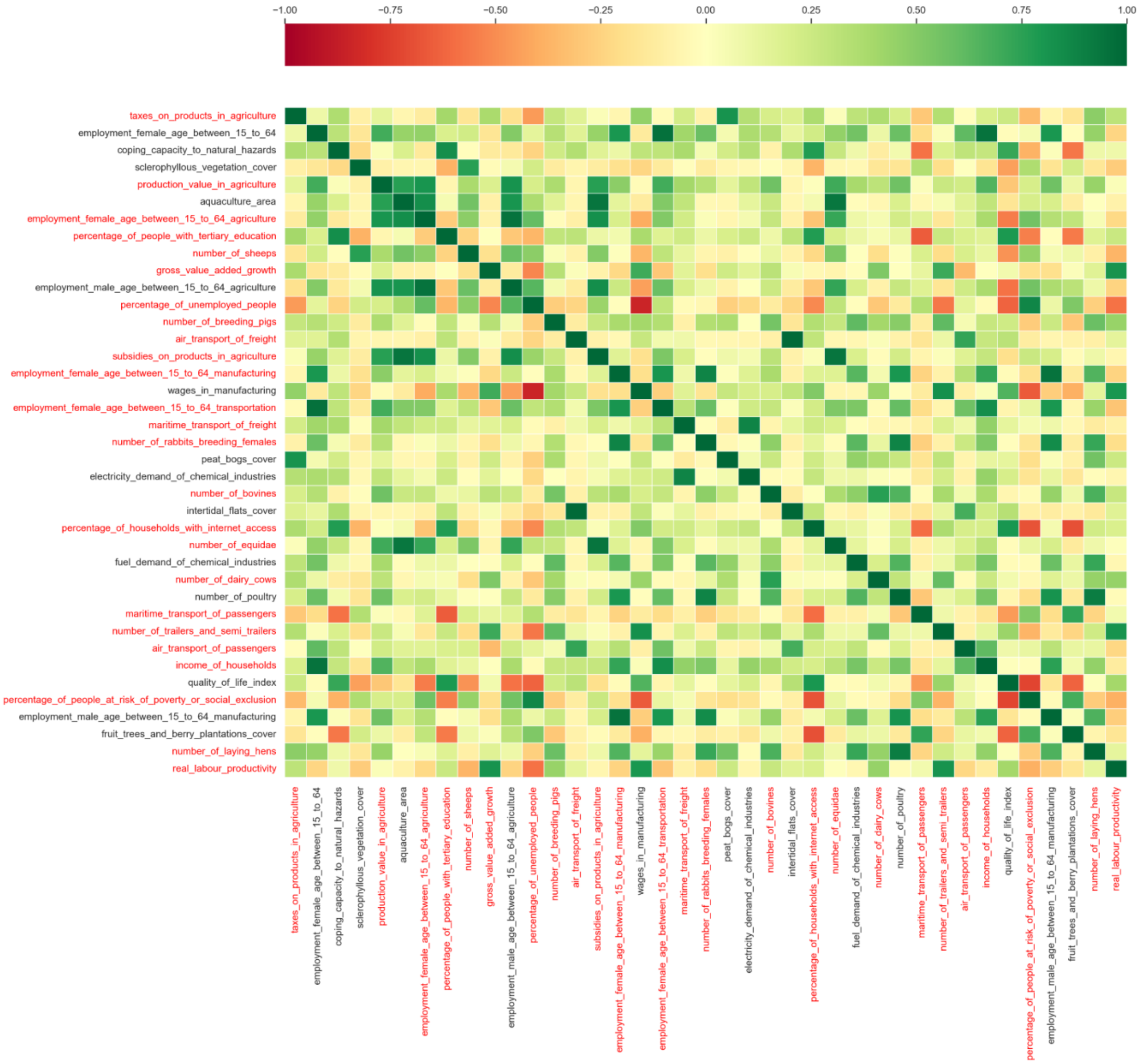


Figure 10 Correlation matrix for the worst performing variables at NUTS2. The worst performing variables are highlighted in red.

3.4 NUTS1 and NUTS0 regions

At NUTS1 level, no data is collected, as the required variables (datasets) were available either at a higher spatial resolution or at NUTS0 level.

For NUTS0 regions, a single variable (dataset) only contains 27 samples - coinciding with the 27 member states. At NUTS0 level, there are 332 datasets with only 11 of them suffering missing values. Figure 11 shows the number of missing values in each of those 11 datasets. In addition to those 321 complete NUTS0 datasets, all the complete datasets and higher NUTS level can be aggregated to complete the NUTS0 data. This large number of datasets however does not change the fact that the sample size is too small for performing missing data imputation using machine learning in the same way as for NUTS2 and NUTS3 regions.

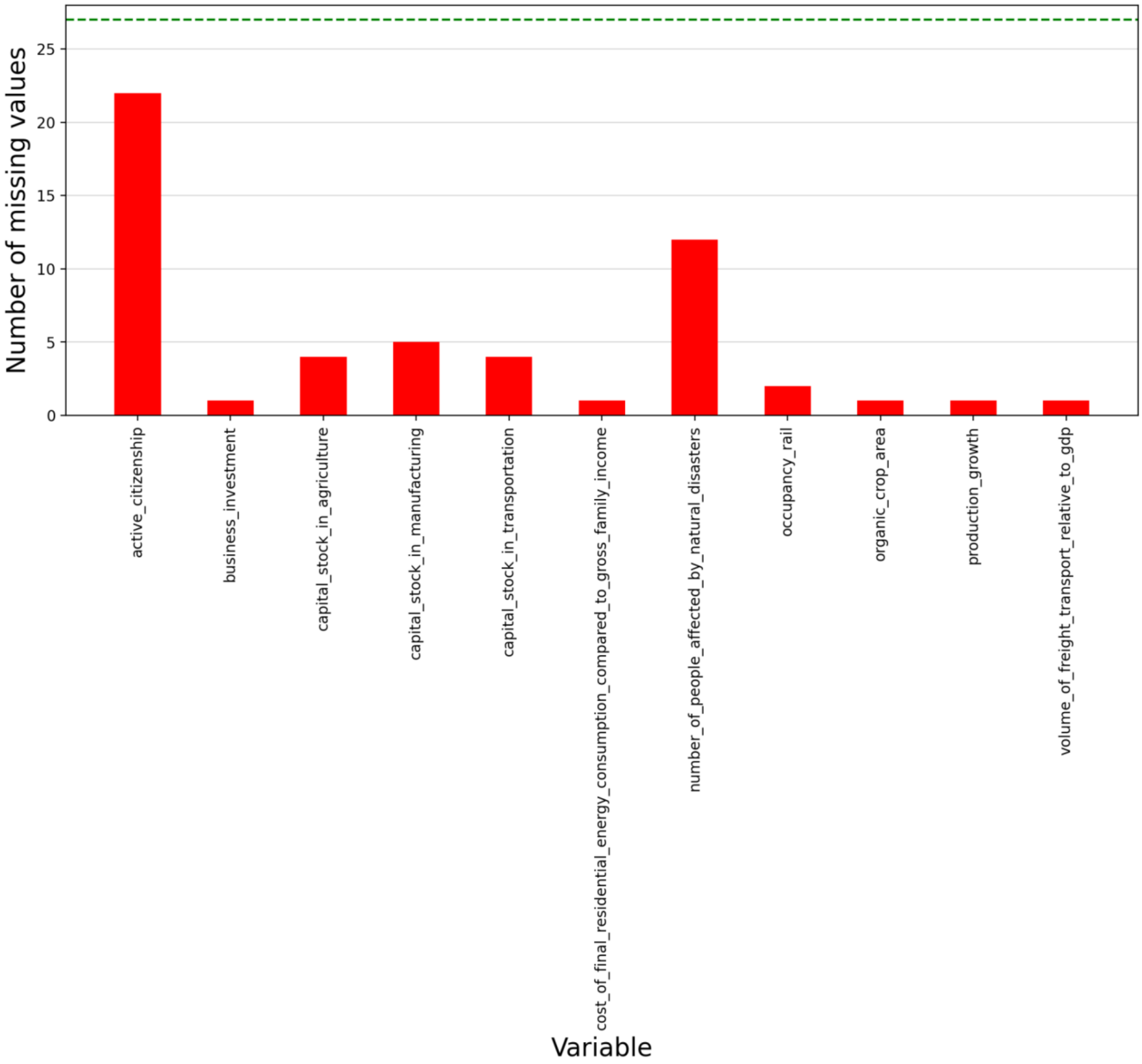


Figure 11 Overview of the number of missing values for the 11 NUTS0 datasets that have missing values.

The same approach was performed as for the survey data at LAU level: regions at NUTS0 that did not have a value are assigned a value of 0 in order to have a full dataset, but the quality of the value is marked as *VERY LOW*.

4 Disaggregation Methodology

4.1 Introduction

Spatial disaggregation is an approach to increasing the spatial resolution of the data. It should be noted that this is not a transformation in the sense of e.g. a coordinate transformation but rather constitutes a remapping of the contained data.

Different approaches have been considered in literature, ranging from statistical to machine learning methods (Monteiro, 2018). Common to all these approaches is that additional knowledge, that sheds light on how the data should be distributed, is required. This can come in the form of statistical knowledge, expert-provided input, or through the use of other datasets. The idea behind using additional data is presented in Figure 12

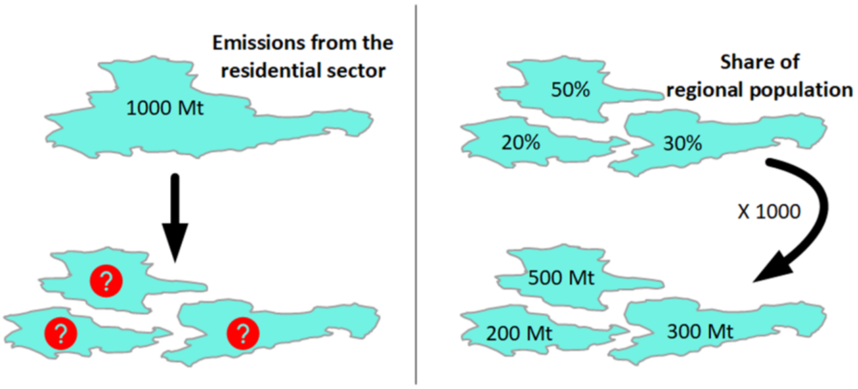

Emissions from the residential sector
1000 Mt
?
?
?
Share of regional population
50%
20%
30%
X 1000
500 Mt
200 Mt
300 Mt

Figure 12 Conceptual example of spatial disaggregation using proxy data.

The example in Figure 12 aims to disaggregate emissions from the residential sector over three sub-regions. While this data is not known at the level of the sub-regions, it stands to reason that the emissions from the residential sector are correlated with the population. As the population in this example is known at the level of the sub-regions, the distribution of the population can be used as an indicator to disaggregate the residential sector emissions among the sub-regions accordingly. This example illustrates the need for a dataset that exhibits a good correlation with the data to be disaggregated and is available at the higher spatial resolution. This data that helps steer the disaggregation is referred to as proxy data.

Given the variation in size (NUTS0 has 27 regions, whereas LAU has 95314 regions), different approaches are considered for different levels. However, all levels make use of proxy data. The need for spatial disaggregation was clear from the start of the project. The procedures for identifying suitable proxy datasets were developed in WP3 and detailed in D3.3. WP3 considered the collection of the proxy datasets in parallel with the collection of the datasets that had immediate relevance with the project.

The approach for involving proxy data bears some resemblance to the methodology for data imputation. The total number of regions at the NUTS3 and NUTS2 levels allow for machine learning to aid spatial disaggregation. Consider, as an example, the disaggregation of "employment" data from a NUTS3 to a LAU level. It is first necessary to determine the relationship between "employment" and the data we have at LAU level. The candidate proxy data at LAU level can be aggregated to match the NUTS3 level - this is a more trivial operation. With all datasets now at a NUTS3 spatial resolution, it is possible to establish how the values of the candidate proxy datasets connect to the "employment" data.

As was the case for data imputation (Section 3), a random forest model is used to determine the relation between the NUTS3 dataset and the aggregated candidate proxy datasets. This established relation is then used to predict the data at LAU level. There is one caveat: the predictor will predict all values at LAU level but has no knowledge of constraints. The values of "employment" in sub-regions at LAU level should sum up to the known value of the containing region at NUTS3 level, but the predictor potentially returns values that do not meet this constraint. As this constraint has to be met, the outcome of the predictor is rescaled so the calculated values correctly sum up to the NUTS3 values.

The disaggregation of NUTS0 dataset suffers - similarly as its data imputation - from too small a sample size. This prevents automatically identifying proxy datasets. So, for the spatial disaggregation of NUTS0 data, proxy datasets were manually specified.The quality of the data is dependent on how good the random forest model managed to predict the values. This is in turn dependent on the quality of the datasets that were used and the quality of their connection. The quality is assessed on a scale of five levels (Section 2), and is determined based on the R-squared score of the predictions. The mapping of the R-squared score to the quality levels is the same as for the data imputation (Table 3 Section 3.2).

Table 4 provides a summary of the mechanisms for spatial disaggregation and for assessing the quality of the resulting datasets.

Table 4 Methodology for disaggregation and quality assessment by spatial level.

Spatial level	**Disaggregation approach**	**Disaggregation quality rating**
NUTS3	- Random Forest model - LAU data as predictors - Some predictions are overridden by manual proxies if required	Based on prediction scores/manual
NUTS2	- Random Forest model - LAU + NUTS3 data as predictors - Some predictions are overridden by manual proxies if required	Based on prediction scores/manual
NUTS1	- (No data collected at this level)	-
NUTS0	- Manual proxy assignment based on the learnings from above	Manual

In the next sections, an analysis of the performance of the spatial disaggregation for the different NUTS levels is provided.

4.2 Spatial disaggregation of NUTS3 datasets

All the data collected at the NUTS3 level undergoes a disaggregation process to achieve a finer spatial resolution at the LAU level. These datasets are systematically categorized into three distinct groups: **general statistics**, **economic indicators**, and **other indicators**. The third category, which encompasses various environmental and societal factors, includes variables related to pollution levels, exposure to natural hazards, and other relevant aspects. To ensure a comprehensive evaluation, the performance of the data disaggregation process is analyzed separately for each of these three categories. The following subsections provide a detailed discussion of the effectiveness, challenges, and key observations associated with disaggregating data in each category.

4.2.1 General statistics

In this section, the performance of the spatial disaggregation for general statistical data (such as population, deaths, etc.) from NUTS3 level to LAU level is discussed. Figure 13 shows the target variables, i.e. the NUTS3 level variables that need to be disaggregated, on the X-axis and the predictor variables, i.e. the candidate proxy data that is available at LAU level on the Y-axis. At each intersection point of a target variable and a predictor variable is an ellipse. The colour of the ellipses match with the assigned quality label (Table 3, Section 3.2), while the size is an indicator for the importance of the predictor in this model (calculated using permutation importance). Note that the cumulative size of the ellipses for a single target variable should be the same for all variables. Only predictor variables with an importance > 0.05 are shown in the plot as the importance is too small for the other variables to be significantly represented in the figure.

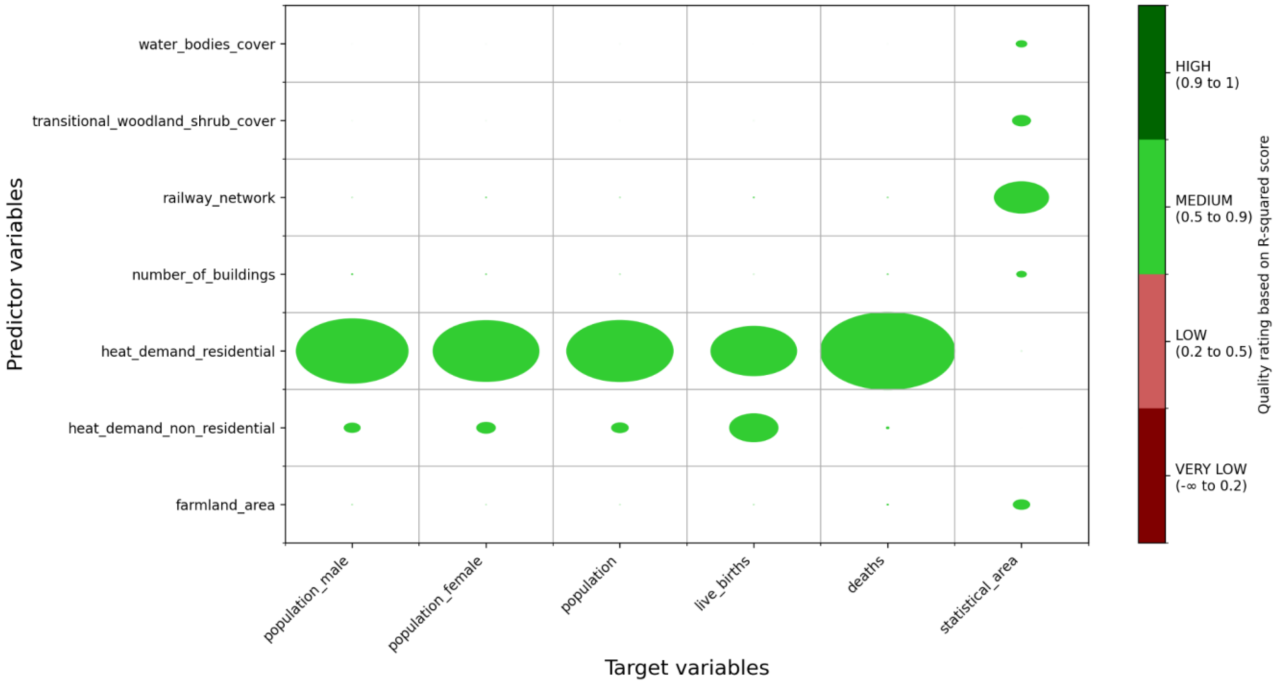


Figure 13 R-squared score and importance of different predictor variables for general statistical variables at NUTS3 (limited to variables with importance > 0.05).

Figure 13 clearly shows that for target variables such as "*population*", the variable "*residential heat demand*" is a good predictor: the R-squared score evaluates to MEDIUM and the predictor is deemed most important for the model. Similar result are obtained for "*male population*", "*female population*", "*live births*" and "*deaths*". Similarly, it is clear that for "*statistical area*", the different land cover variables such as "*railway network*", "*water bodies*", etc. are evaluated as good predictors.

All of these target variables are however disaggregated with a quality rating "MEDIUM", indicating that there is room for improvement. However, improvement is dependent on additional datasets and considering that the LAU level is the smallest statistical unit, a lot of data simply does not exist at this spatial resolution.

4.2.2 Economic indicators

The quality of the spatial disaggregation of the economic indicators that need to be disaggregated from NUTS3 level to LAU level is depicted in Figure 14.

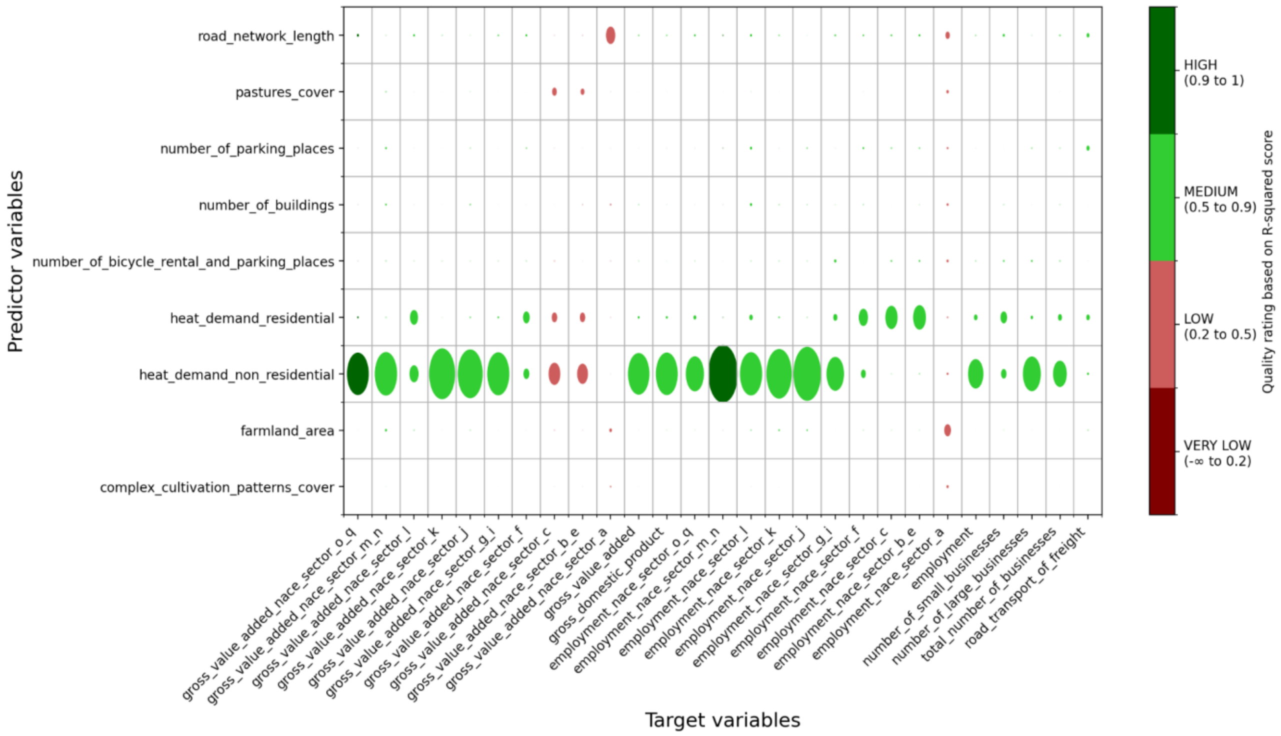


Figure 14 R-squared score and importance of different predictor variables for economic variables at NUTS3 (limited to variables with importance > 0.05).

For some cases, such as "*employment_in_nace_sector_a*", the prediction quality is poor and the predictor is not deemed important. Nace sector A is the agriculture sector, and it seems that none of the predictor variables are important and none even stand out. In some cases, such as "gross_value_added_nace_sector_c", the quality is considered low but the importance of the predictor "*non-residential heat demand*" is rather high. This indicates that, while this predictor variable is deemed the most important one for the model to determine the disaggregation, the model cannot disaggregate with good results. Both cases are indicative of the issue that there is simply not enough data at LAU level to perform a spatial disaggregation of sufficient quality.

4.2.3 Other indicators

The quality of other indicators, which mainly relate to pollution, vulnerability and risks that are available at NUTS3 and for which a disaggregation to LAU is depicted in Figure 15.

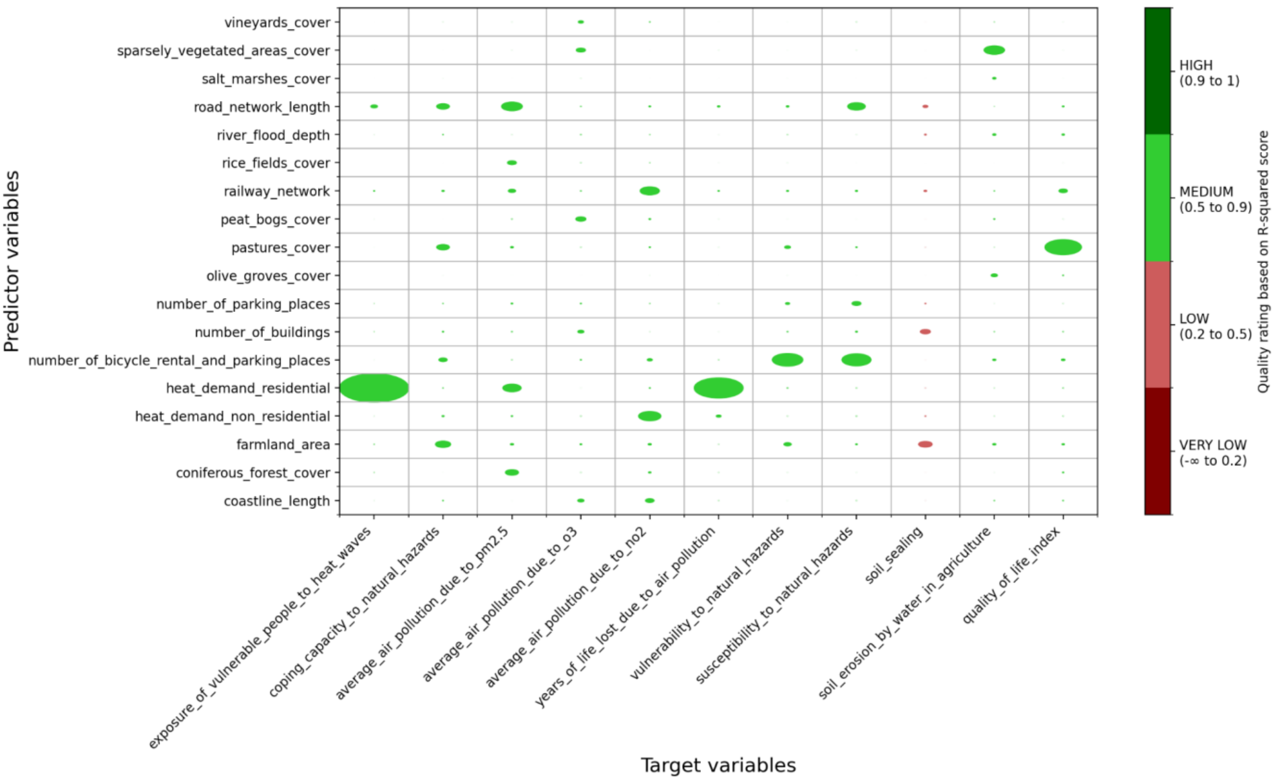


Figure 15 R-squared score and importance of different predictor variables for other variables at NUTS3 (limited to variables with importance > 0.05).

It is very visible on Figure 15 that a lot of the air pollution related data disaggregates quite well using different predictor variables relating to greenery cover, infrastructure, such as road network, and population density in terms of residential heat demand. Given the different pollutants and knowledge on the source of these pollutants - some more connected to road traffic, others with heating - this result is within expectations.

The algorithm cannot find decent predictors for "*soil sealing*": the quality is low and none of the predictor values is deemed important. Additional LAU level data would be needed to improve this prediction.

We also performed a test with other target variables such as "*exposure of vulnerable people to heat waves*", "*coping capacity to natural hazards*", "*vulnerability to natural hazards*", "*susceptibility to natural hazards*" and "*quality of life index*". These indicators are calculated based on several other indicators and are made available in the ESPON Database Portal (https://database.espon.eu/) at NUTS3 level. While there appears to be a good prediction "*exposure of vulnerable people to heat waves*" using "*residential heat demand*", one must be careful with the interpretation. A high residential heat demand implies a high population; the exposure of vulnerable people to heat waves first and foremost requires people. However, the number of people is not a determining factor for the risk of exposure. The data may show some correlation which the model catches, but spatial disaggregation of such indicators is not meaningful. The random forest model is therefore discarded in this case. The source values should be provided at LAU level and, since we do not have this data at LAU, the assigned values at LAU level are the same as the parent NUTS3 region. The quality rating for this dataset at LAU level is set to LOW (the dataset at NUTS3 level is VERY HIGH as that is the level of the source data).

4.2.4 Geographical distribution of data - further analysis

To have verifiable results, first a prediction of values at NUTS3 level will be considered and the predicted NUTS3 level should be similar in distribution as the target distribution. This is shown in Figure 16, which shows the variable "*employment*" and its prediction using "*non-residential heat demand*". The top row shows the data at NUTS3 level; from left to right it shows the known data for "employment", the predictor variable "*non-residential heat demand*" (aggregated from LAU) and the predicted spatial distribution at NUTS3 level. Each map also contains a violin plot that visualises the distribution of values. The spatial correlation between the target variable at NUTS3 level (top left) and the aggregated predictor variable at NUTS3 (top middle) is immediately obvious from the maps. The predicted values at NUTS3 level (top right) are slightly underestimated as confirmed on the violin plot but the geographics spread is very similar to the original values.

On the second row of Figure 16, the predictor variable is shown at LAU level (bottom middle), alongside the prediction at LAU level for the "*employment*" (bottom right). Here it is evident that the high heat demand values are very concentrated in the big cities, which creates a large difference with the small regions that have relatively small values in comparison. As a result, this causes the predictor to also assign very low values for "*employment*" in those regions and effectively resulting in a disproportionately big employment in the cities. The low values in the small regions are very likely too small and not realistic, despite the fact that the disaggregation to LAU level of "*employment*" using "*non-residential heat demand*" is evaluated as MEDIUM based on R-squared (Figure 14). An important data-aspect is that non-residential heat demand data is collected from the Hotmaps project (https://zenodo.org/records/4687026), which provides a geo-tiff image. In a geo-tiff, the spatial regions are regular shaped regions that do not necessarily overlap well with the LAU regions. However, at LAU level, the pixel size of the geo-tiff is potentially too big and suffers from partial overlaps making it potentially less suitable to work at LAU level.

The example using "*non-residential heat demand*" illustrates that an indicator may appear sufficient to reach the initial goal of this deliverable (NUTS3), however, downscaling further to LAU levels may need more data and verification. This investigation will continue within the future work of WP3.

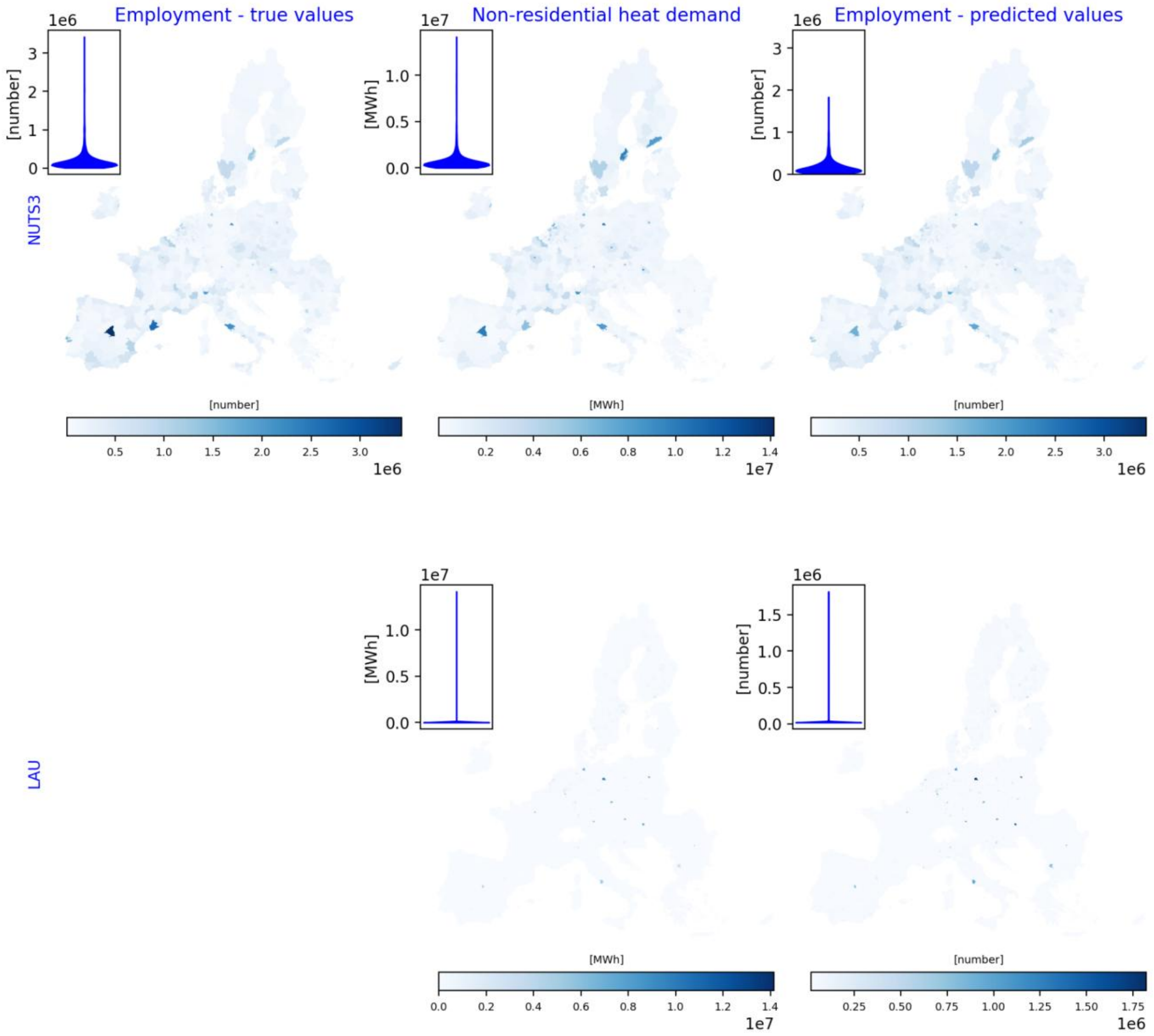


Figure 16 Comparison of the spatial distribution of "employment" and "non-residential heat demand".

4.3 Spatial disaggregation of NUTS2 datasets

The spatial disaggregation of NUTS2 datasets follows the same approach as the spatial disaggregation for NUTS3 datasets (Section 4.3). The result of the analysis of the candidate proxy data sets are shown on Figures 17-21. There are a few datasets with good predictor variables, e.g. "*male employment in agriculture age 15-64*" is well predicted by "employment in NACE sector a" (Figure 17). This is a rather trivial case as NACE sector a is the agricultural sector, but it is helpful to see such a connection confirmed in the methodology.

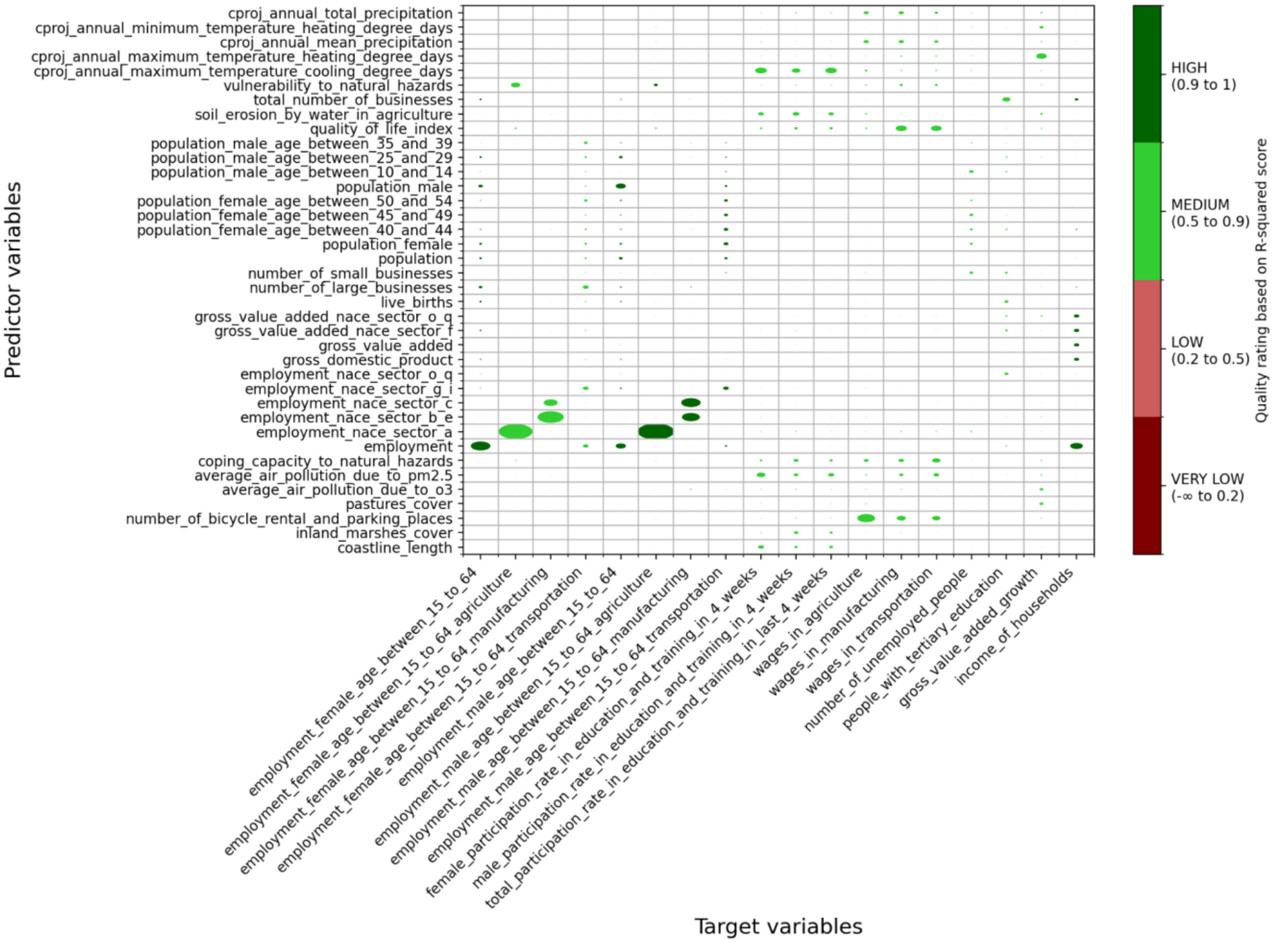


Figure 17 R-squared score and importance of different predictor variables at NUTS2 (limited to variables with importance > 0.05).

Figure 18 shows, using variables for transport, that it is possible for predictor variables to be considered important, but their use still does not yield a disaggregation with a good R-squared value. The explanation is that the two assessments (*R-squared* and *importance permutation*) consider different aspects. The importance permutation provides insight into which of the predictor variables has the biggest impact in performing the disaggregation, but this does not necessarily imply a good disaggregation - which is what R-squared aims to assess. The number of refineries logically connects with the maritime transport of freight (reflected by the high importance permutation), but the spatial distribution of *maritime freight transport* is not really a variable that lends itself to a spatial distribution within countries or large regions. Similarly, air transport of freight will only be possible at airports: while the transported goods may connect to "gross_value_added_nace_sector_k", the spatial distribution of this predictor variable may not very well match the locations of airports.

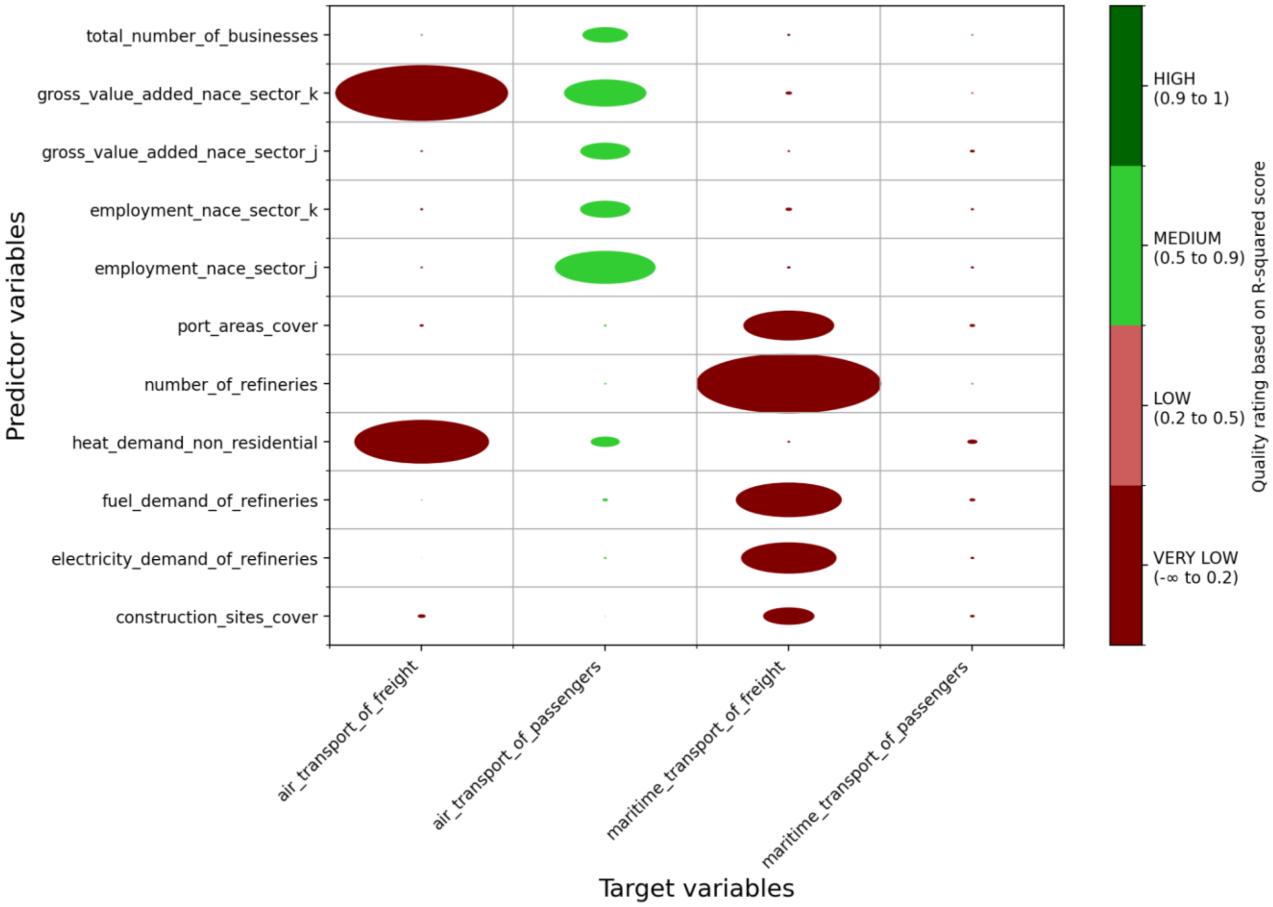


Figure 18 R-squared score and importance of different predictor variables for transport variables at NUTS2 (limited to variables with importance > 0.05).

Similarly, Figure 19, Figure 20 and Figure 21 show the predictor variables for economic variables and some general statistics, for variables relating to animal population, for data on vehicle stock, respectively.

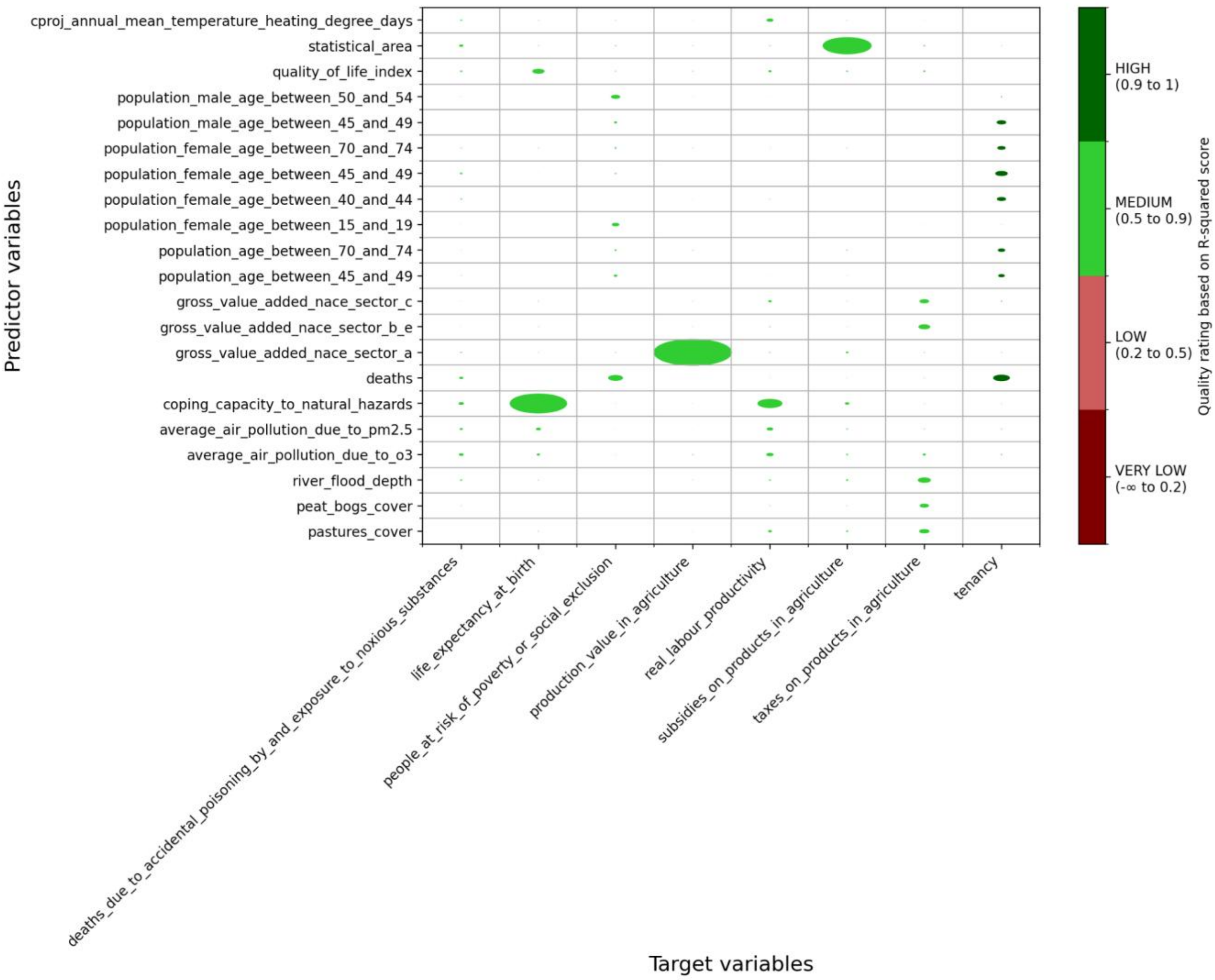


Figure 19 R-squared score and importance of different predictor variables for economic variables and some general statistics at NUTS2 (limited to variables with importance > 0.05).

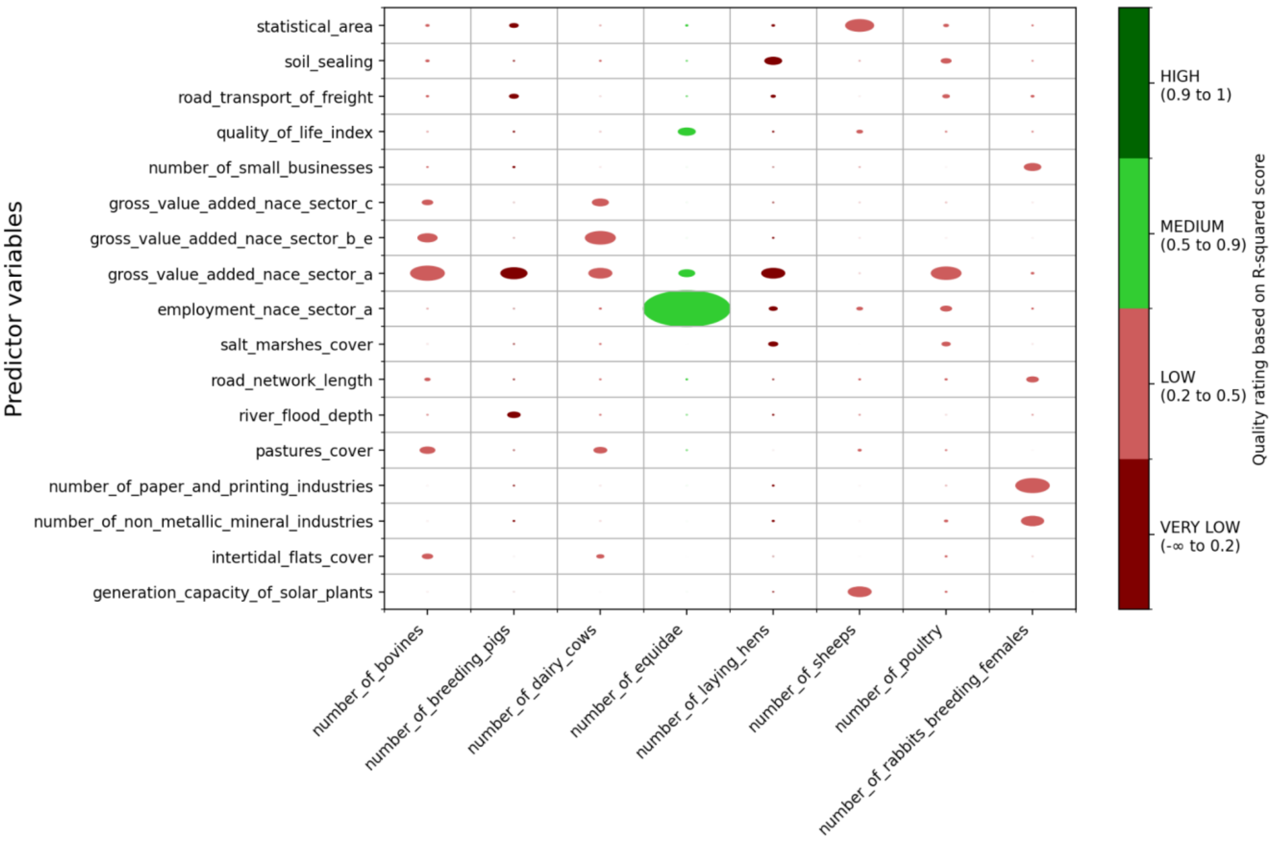


Figure 20 R-squared score and importance of different predictor variables for animal population at NUTS2 (limited to variables with importance > 0.05).

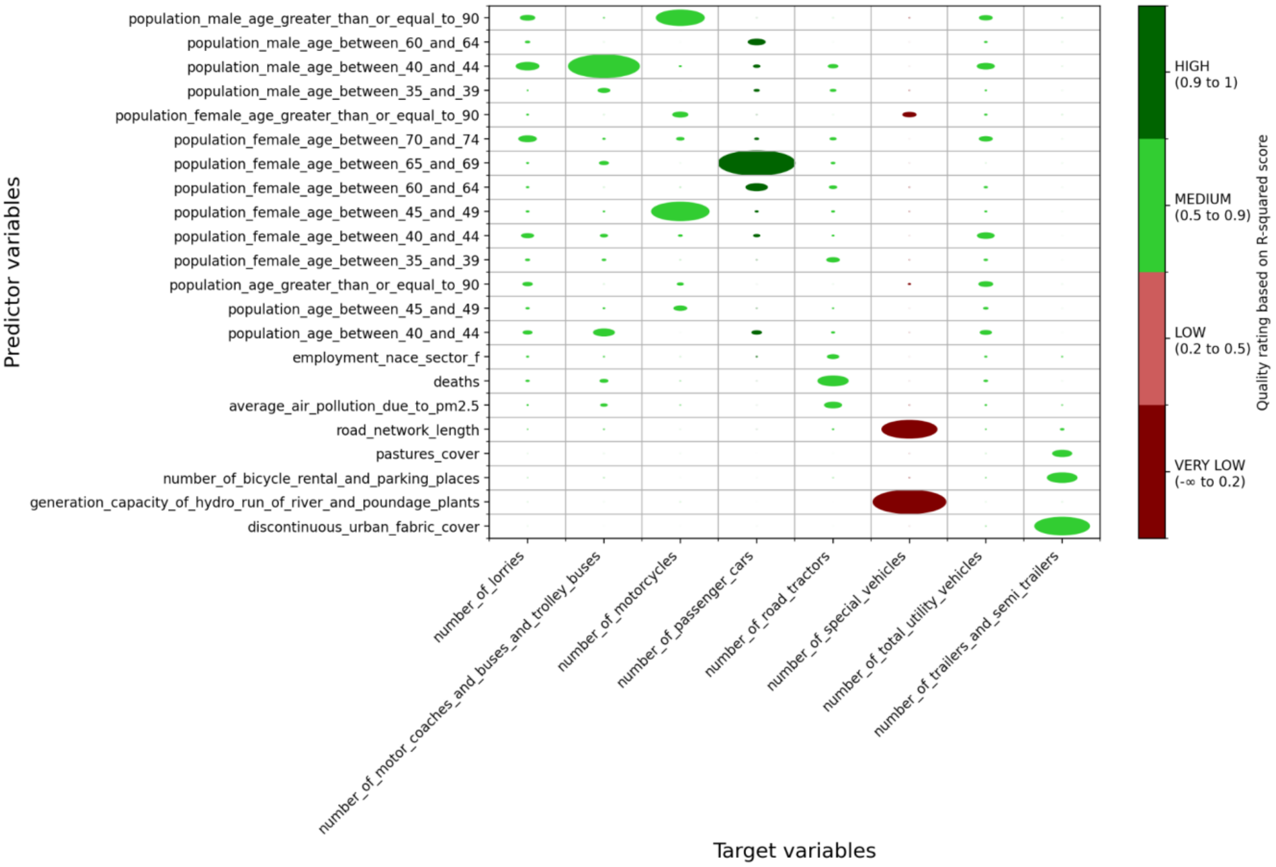


Figure 21 R-squared score and importance of different predictor variables for vehicle stock at NUTS2 (limited to variables with importance > 0.05).

4.4 Spatial disaggregation of NUTS1 and NUTS0 datasets

No datasets at NUTS1 level were collected. At NUTS0 level are all the dataset that originate from the EUCalc Pathways, as well as 332 additionally collected datasets. This is the level at which we have the largest number of datasets, as much data is supplied or estimated at country level. However, with 27 member states, each of these datasets has just 27 datapoints. This is far too little to train any automatic system.

Here, it was necessary to resort to a manual assignment of the predictor variables, based on the relationships seen between target and proxy datasets at NUTS3 and NUTS2 levels, in combination with an understanding of the data. Table 5 shows the proxies assigned to some variables, collected at NUTS0 level. All further variables can be found in the DSP.

Table 5 Examples of proxies assigned to variables at NUTS0.

Variable	Proxy
production_growth	gross_value_added
emission_factor_of_gasoline	(no proxy, same value all regions)
occupancy_bus	(no proxy, same value all regions)
paper_and_cardboard_waste	population + dump_sites_cover
housing_cost_overburden_rate_by_poverty_status	income_of_households / population
energy_demand_in_agriculture_and_forestry_and_fisheries_from_electricity	gross_value_added_in_nace_sector_a
energy_demand_of_chemical_and_petrochemical_industries_from_electricity	electricity_demand_of_chemical_industries
energy_demand_of_chemical_and_petrochemical_industries_from_natural_gas	fuel_demand_of_chemical_industries
residential_energy_demand_space_cooling	annual_mean_temperature_cooling_degree_days + income_of_households + population

Similarly, Table 6 shows the proxies assigned to some of the EUCalc variables.

Table 6 Examples of proxies assigned to EUCalc variables.

Variable	Proxy
eucalc_agr_domestic_production_afw_cereal	non_irrigated_arable_land_cover
eucalc_agr_emissions_ch4_liv_enteric_abp_dairy_milk	number_of_dairy_cows
eucalc_bld_emissions_co2_residential_sfh_non_elec_hw	heat_demand_residential

eucalc_dhg_energy_demand_heat_district	(heat_demand_residential+heat_demand_non_residential) / statistical_area
eucalc_ind_emissions_co2e_chemicals	Electricity_demand_of_chemical_industries + fuel_demand_of_chemical_industries
eucalc_tra_energy_demand_freight_aviation	air_transport_of_freight
eucalc_wat_water_consumption_household	population
eucalc_tra_vehicle_fleet_freight_marine_bev	maritime_transport_of_freight
eucalc_ind_material_production_glass	electricity_demand_of_non_metallic_mineral_industries+fuel_demand_of_non_metallic_mineral_industries

5 Working Disaggregation Tool

The disaggregation workflow is implemented in the Python programming language. It is published on the GitHub account of Forschungszentrum Jülich. It can be found under https://github.com/FZJ-IEK3-VSA/ETHOS.zoomin.

The workflow reads the collected data from a local database, along with details regarding proxies or random forest model to be used to disaggregate the data to LAU level. Based on this input, the data is disaggregated. The disaggregated data is written to the final table in the database i.e. processed data table (refer to Figure 2).

In a subsequent step, the national decarbonisation pathway data from EUClac is disaggregated in a similar way.

6 Data Access

6.1 Data sharing platform

The processed data table is accessible through the DSP. This platform is being developed in T3.3. An initial version was presented in Deliverables D3.2 and D2.5. The goal of the DSP is to provide a single access point for all the spatial data. It allows access to datasets for a specified region, at the specified resolution - which is independent of the resolution of the source.

The DSP is an Application Programming Interface (API) that provides access to the databases; the format of the API URL is described below. The URL can be constructed with the desired set of parameters and pasted in any web browser to access the corresponding data.

API URL Format: http://data.localised-project.eu/dsp/v1/region_data/?api_key=api_key&region=region

The parts highlighted in red are the parameters. These are described in Table 7.

Table 7 API parameters.

Parameter	Description	Options
api_key	The confidential API key.[2]	-
region	The region code corresponding to the country's region for which the data is to be queried.	Any region code. Note: a list of regions, corresponding to the specified resolution and country can be queried in the following manner: http://data.localised-project.eu/dsp/v1/region_metadata/?api_key=api_key&resolution=resolution&country=country

Example API query:

Suppose the query is to be made for Berlin in Germany. The NUTS3 region code of Berlin is DE300. The API URL would be:

[2] The DSP is not yet fully ready for public sharing. It is due in September, 2024. In the meantime, the confidential API key is only shared within the consortium.

http://data.localised-project.eu/dsp/v1/region_data/?api_key=XXXXXXX&region=DE300

The result would look as follows:

```
HTTP 200 OK

Allow: GET

Content-Type: application/json

Vary: Accept

{

    "count": 1,

    "next": null,

    "previous": null,

    "results": [

        {

            "value": 3669491.0,

            "year": 2020,

            "var_name": "population",

            "var_description": null,

            "var_unit": "number",

            "var_aggregation_method": "sum",

            "taggings": [

                {

                    "tag_dimension": "sector",

                    "tag_name": "general stat"

                },

                {

                    "tag_dimension": "type",

                    "tag_name": "stock"

                },

                {

                    "tag_dimension": "commodity",

                    "tag_name": "not applicable"

                },
                {
                    "tag_dimension": "resource",
                    "tag_name": "not applicable"
                },
                {
                    "tag_dimension": "link",
                    "tag_name": "not applicable"
                },
                {
                    "tag_dimension": "other",
                    "tag_name": "total"
                }
            ],
"var_source_name": "Eurostat",
"var_source_link": "https://ec.europa.eu/eurostat/databrowser/view/DEMO_R_PJANGRP3/default/table?lang=en",
"var_source_citation": "Eurostat, “Population on 1 January by age group, sex and NUTS 3 region.” https://ec.europa.eu/eurostat/databrowser/view/DEMO_R_PJANGRP3/default/table?lang=en (accessed Jun. 29, 2023).",
"var_source_license": "Creative Commons Attribution 4.0 International (CC BY 4.0)",
"original_resolution": "NUTS3",
"disaggregation_binary_criteria": null,
"disaggregation_proxy": "Disaggregation using random forest model. Top 3 important variables in population prediction: heat_demand_residential, heat_demand_non_residential, road_network_length",
"calculation_equation": null,
"pathway_file_name": null,
"climate_experiment": null,
"quality_rating": "VERY HIGH"}]}
```

Please note that only one variable, population, is shown in the above example result. The actual query provides all the data collected and curated for the region. The table below describes the variable fields:

Table 8 Variable fields in the API query output.

Field	Description
value	The value of the variable.
year	The year for which the data is collected.
var_name	Name of the variable.
var_description	A detailed description of the variable is provided, where required. It is left blank, otherwise.
var_unit	The unit in which the value is expressed.
var_aggregation_method	When the data is queried at higher level, the data from lower level is aggregated. The type of aggregation performed is indicated here. This could, for example, be sum or mean.
taggings	This is a list of keywords that helps describe and categorise the variable further.
var_source_name, var_source_link, var_source_citation	The information regarding data source is specified here.
original_resolution	The resolution at which the data was available and therefore, collected is specified here.
disaggregation_binary_criteria[3]	During disaggregation of some datasets, the proxy value in certain LAU regions are preset to 0 based on a criteria. This criteria could, for example, be population density greater than a particular threshold. Consequently, the disaggregation of values from a parent region to its child LAU regions only considers those LAU regions with non-zero values.

[3] This is an experimental feature that needs further testing.

disaggregation_proxy	This field specifies the disaggregation method used to downscale the values from its original resolution to LAU resolution.
calculation_equation	If a dataset is not collected but calculated based on other datasets, the equation is shown in this field.
pathway_file_name	This field is specific to EUCalc national decarbonisation pathway variables that are downscaled here to local regions. They describe the pathways.
climate_experiment	This field is specific to climate indicators. They describe which climate scenario, or the Representative Concentration Pathway (RCP) considered to arrive at the climate projections.
quality_rating	Specifies the quality of the data.

API documentation:

Further information regarding the API can be found in the official documentation under: http://data.localised-project.eu/dsp/v1/docs/

6.2 API Client

An API client provides ready-to-run scripts that allow one to query the API data with minimal effort. These scripts not only allow the user to query the data, but also to save it in a desired format, such as .csv or .json, in their local machine. Such an API client is developed for the API and is currently hosted on GitHub. This can be found under: https://github.com/FZJ-IEK3-VSA/LOCALISED-Datasharing-API-Client

This client includes functions written in the Python language that allow users to make queries and save the regional data. This helps avoid querying the data each time. The usage of these functions is described through an example script. This can be found in a Jupyter Notebook under: https://github.com/FZJ-IEK3-VSA/LOCALISED-Datasharing-API-Client/blob/master/examples/single_region_all_variables.ipynb

7 Conclusion

Deliverable D3.1 is aimed at providing the spatial disaggregation tool in order to have complete datasets at all resolutions. The concept was extended from disaggregating from NUTS3 (as per the proposal) to LAU level (due to increased interests and applicability of the outcomes). The spatial disaggregation tool aims to use proxy data, i.e. data which exhibits a connection to the dataset to be disaggregated, and for which a similar spatial distribution can be reasonably assumed.

The use of proxy data implies the need for additional datasets, which were collected in the context of D2.5 and D3.3. The "*simple*" disaggregation where the spatial distribution of the proxy dataset is directly applied was investigated. The aim of the research was to automatically find proxy datasets and automatically assess the quality of the resulting disaggregations and to improve on this simple approach using machine learning techniques. The reasons for the employment of machine learning techniques are:

1. To identify complex relationships between proxy datasets and the data to be disaggregated.
2. To be able to quickly adjust to updated and improved datasets.

For datasets at NUTS0 level, the automation possibilities are limited as every NUTS0 datasets only has 27 data points. The methodology works well to disaggregate to NUTS3 level, provided suitable proxy data is available and identified. The availability and suitability as such became key for the functioning of the tool. For the disaggregation to LAU level, the problem of data availability is worse as the general lack of proxy data at this resolution limits the possibility of improving over the "simple" disaggregation, even with manually selected proxy datasets.

The accuracy of the disaggregated data depends on the availability and quality of local-level proxy data. To ensure transparency for both other WPs within the project and external users, each value is accompanied by a quality rating.. Furthermore, the users of these tools will be able to adjust all values for their region. So even if the LOCALISED estimate for a particular region is incorrect due to lack of quality data, this will not hurt the usefulness of the tools developed.

Together with the spatial disaggregation, the methodology also provided an approach to estimated data for regions where data are missing. This approach also uses additional datasets to help to determine the values of missing data. There can be different reasons why data are *missing*, which relates to different interpretations of the data. This was taken into account, as explained in this deliverable.

As the data gathering and processing is an ongoing process throughout the project, the results of the disaggregations (and thus the quality of some datasets) can increase when more suitable proxy data becomes available.

8 References

Patil, S.; Verstraete, J.; Pflugradt, N.; Seydeswitz, T.; Costa, L.; Radziszewska, W. (2023), Climate change database and other spatial data for 3 EU countries (LOCALISED Deliverable 2.4)

Verstraete, J.; Patil, S.; Pflugradt N., Radziszewska W. (2023), Database for 3 EU countries with relevant data for the year 2020 (LOCALISED Deliverable 3.2)

James G., Witten D., Hastie T., Tibshirani R., Taylor J. (2023). Statistical Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-38747-0_2.

Monteiro J., Martins B., Pires J.M. (2018). A hybrid approach for the spatial disaggregation of socio-economic indicators. Int J Data Sci Anal 5, 189–211. https://doi.org/10.1007/s41060-017-0080-z.

Costa L., (2022), First library of model outputs at EU and MS level (LOCALISED Deliverable 2.2)

GISCO - Eurostat, 'NUTS - GISCO - Eurostat', 2021. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts (accessed Nov. 23, 2022).

Kanellopoulos K., De Felice M., Hidalgo Gonzalez I., and Bocin A., 'JRC Open Power Plants Database (JRC-PPDB-OPEN)'. Zenodo, Dec. 13, 2019. doi: 10.5281/zenodo.3574566.

ESPON Database Portal, 'Indicator: Territorial coping capacity to natural hazards | ESPON Database Portal', 2022. https://database.espon.eu/indicator/2224/ (accessed Nov. 23, 2022).

Patil, S.; Verstraete, J.; Pflugradt, N.; Seydeswitz, T.; Radziszewska, W. (2023), Climate change database and other spatial data (LOCALISED Deliverable 2.5).

Verstraete, J.; Patil, S.; Pflugradt N., Radziszewska W. (2023), Database with all relevant data for the year 2020 (LOCALISED Deliverable 3.3)

LOCALISED

www.localised-project.eu