FACULDADE
DE ECONOMIA
UNIVERSIDADE
DO PORTO

Predicting Household Occupancy Status: A Data-Driven Approach Using Machine Learning and GIS

Yara Mwafaq Abed Al Rohman Al suwiti

Internship Report

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

Pedro Campos, PhD

2025

Biographical note

Yara Mwafaq Abed Al Rohman Al-Suwiti earned her Bachelor's degree in Software Engineering from the Jordan University of Science and Technology. She is currently completing the Master's in Data Analytics and Decision Support Systems at the Faculdade de Economia da Universidade do Porto, where she has also fulfilled the requirements for the European Master in Official Statistics (EMOS) designation. Her thesis, "Predicting Household Occupancy Status: A Data-Driven Approach Using Machine Learning and GIS", was conducted within Statistics Portugal (INE) as part of an internship fulfilling the EMOS program requirements. Her academic interests centring on the integration of geographic information systems (GIS) and related software into real-world problem-solving, with a particular emphasis on visualization and its role in evidence-based decision-making. She has also completed professional training in data analysis, software development, quality assurance, and cybersecurity, and aims to advance innovative applications of data and GIS technologies to address complex societal and organizational challenges. Beyond her current fields of expertise, she is committed to continuously expanding her knowledge across diverse disciplines to foster interdisciplinary perspectives and solutions.

Acknowledgments

All praise and thanks are to Allah, the one who, by his blessing and favour, all good works are accomplished.

First and foremost I would like to thank my father, Mwafaq Al Suwiti, my primary source of motivation and courage and my role model, for his unwavering support for me in this long journey, in what was at first an unfamiliar land. Not only did his financial support make it possible for me to have this wonderful opportunity in the first place, but also his immense emotional support and belief in me and my capabilities, which allowed me to see this opportunity through and push myself no matter what. I also thank my immediate family, my mother, my siblings, my best friend, and all those who supported me and have been my rock through it all, and for believing in me and helping me see through my own doubts and remain strong and carry out what needs to be done. I would also like to thank my supervisors Ana M Santos for her invaluable insight and Pedro Campos for his expertise, and I would also like to thank Mimosa Pinho for her constant support and assist with several parts of the theoretical and methodological aspect of this study, all of which encouraged this project to come to life as my initial goal was to learn more about GIS along with applying what I have learned about Data Analysis and Machine Learning to a real-world problem, creating an expandable, reproducible baseline beneficial for future similar issues.

Abstract

This study investigates the feasibility of leveraging administrative tax and utility records, in combination with historical datasets from the National Building Database (BNE) and the Resident Population Base (BPR) of Statistics Portugal, to predict the future occupancy status of dwellings for participation in upcoming surveys. The primary objective is to reduce the operational burden, cost, and time associated with attempting to survey households that are ultimately unavailable or ineligible. The research integrates Geographic Information Systems (GIS) both for spatial representation of the dataset and for the visualization of model predictions. The dataset comprises utility consumption records (e.g., water, electricity) and tax information for dwellings located in the parish of São João da Madeira, covering the period from 2018 to 2025. Data preprocessing involved converting continuous and transactional data into categorical indicators reflecting dwelling status. Multiple supervised learning algorithms were trained and evaluated, with XGBoost achieving the highest performance, yielding an accuracy of 89.75% and a cross-validated weighted F1-score of 89.05% on the test set. The best model was then applied to perform one-step-ahead forecasting, successfully predicting occupancy statuses for the year 2026, validated using domain knowledge tests. The results demonstrate that combining administrative records with machine learning and GIS visualization offers a viable, scalable approach for anticipating dwelling availability, thereby enabling more efficient survey planning and resource allocation.

Keywords: Machine Learning Models, Geographic Information Systems (GIS), Housing Surveys, Administrative Records, Base Nacional de Edifícios (BNE), Base de População Residente (BPR)

Contents

Biographical note . . . iii
Acknowledgements . . . v
Abstract . . . vii

1 Introduction **1**
1.1 Background and Motivation . . . 1
1.2 Institutional Context . . . 1
1.3 Problem Statement . . . 1
1.4 Aim and Objectives . . . 2
1.5 Broader Context and Links to International Practices . . . 3

2 Literature Review **5**
2.1 Sampling Frames and Administrative Data Integration . . . 5
2.2 INSEE: Transitioning from Census-Based to Tax-Based Sampling Frames . . . 7
2.3 Eurostat: Survey Sampling Reference Guidelines . . . 8
2.4 Census Bureau MAF-ARF: Leveraging Administrative Records for Address-Based Sampling . . . 9
2.5 Discussion . . . 10

3 Methodology **13**
3.1 The Data . . . 13
3.2 Data preparation . . . 16
3.3 Machine Learning Models (Theoretical Foundation) . . . 19
3.4 Modeling . . . 20
3.4.1 Initial Model Development and Evaluation . . . 20
3.4.2 Impact of Historical Aggregate Feature on Model Performance . . . 23
3.4.3 Model Selection and Performance Comparison . . . 24
3.4.4 Introduction of New Variables, Re-Modelling and Evaluation . . . 25
3.5 Model Results and Discussion . . . 31
3.5.1 Evaluation Metrics . . . 31
3.5.2 Implementation . . . 31
3.5.3 Visualization of Household Occupancy using GIS . . . 33
3.5.4 Final Model and Comments . . . 34

4 Conclusion **37**
4.1 Final Remarks . . . 37
4.2 Limitations and Future Work . . . 38

A Data Dictionary **i**

B Additional Household Status Heatmaps **ii**

List of Figures

3.1 Data Overview 14
3.2 Percentage of Existing (Non-Missing) Data per column 17
3.3 Top features most associated with target (Pearson & Cramér's V) 18
3.4 Top features most associated with target (Pearson & Cramér's V) After Aggregation 18
3.5 Class Distribution in Target Variable 19
3.6 SHAP summary plot for the XGBoost model without historical aggregate feature 25
3.7 Top features most associated with target (Pearson Cramér's V) With Newly Introduced Data and Aggregate 26
3.8 SHAP analysis for XGBoost 30
3.9 Predicted Dwelling Occupancy by Coordinates (2026) – Excluding Unlocated Buildings 32
3.10 Spatial distribution of occupied and vacant households visualized using ArcGIS Pro. Heatmaps highlight clustering patterns, with *Primary* households concentrated in central urban areas and *Vacant* households more dispersed in peripheral regions. 33

B.1 Additional household status heatmaps generated in ArcGIS Pro. These maps complement the primary and vacant household visualizations in the main text. The first figure representing Households with Status Secondary, and the second figure representing a combination of filters for houselods that are Eliminated, Inaccessable, and Non-existent due to Census error ii
B.2 Further household status heatmaps illustrating spatial distribution for other statuses.The first being Non-existent family accommodation because it was associated with another accommodation, and the second being Family accommodation occupied for other purposes ii

List of Tables

3.1 Model Performance by Oversampling Technique 22
3.2 Model Performance by Oversampling Technique (Excluding Historical Aggregate Feature) 23
3.3 Performance of Selected Models Without the Historical Aggregate 24
3.4 Model performance including historical aggregate and new numerical variables 28
3.5 Model performance excluding historical aggregate 29

A.1 Full Data Dictionary i

Chapter 1

Introduction

1.1 Background and Motivation

Household surveys are a cornerstone of official statistics, providing critical information about demographic, social, and economic conditions. However, the effectiveness of such surveys depends heavily on the quality and accuracy of the underlying sampling frames. Traditional methods of constructing and updating sampling frames, including field visits and manual listing, are increasingly challenged by dynamic population movements, high operational costs, and limited resources. Internationally, statistical agencies are modernizing their methodologies by integrating administrative registers, utility records, and geospatial data, often supported by predictive models and machine learning techniques. These developments aim to reduce inefficiencies, improve coverage, and ensure that statistical outputs remain both timely and relevant.

1.2 Institutional Context

This project was conducted within Statistics Portugal (INE), the central authority responsible for the production and coordination of official statistics in Portugal. INE operates with technical independence, aligned with both national legislation and European standards, and its mission is to produce high-quality, relevant statistical information that serves society's evolving needs. To achieve this, INE promotes methodological innovation, integrated data analysis, and transparent dissemination practices, while safeguarding the confidentiality of data providers. The project was developed at the Porto office of INE as part of an internship, under the joint supervision of the Geographic Information Department (led by Ana M. Santos) and the Methodology Department (led by Pedro Campos), both of whom provided critical guidance.

1.3 Problem Statement

The primary challenge motivating this project arises from the operational constraints faced by household survey teams at Statistics Portugal. These teams rely on pre-assigned survey routes, which include lists of dwellings whose occupancy status is often unknown until a physical visit occurs. Upon arrival, a dwelling may be found occupied by the registered owner, by other residents, vacant, inaccessible, temporarily unavailable, or misclassified in

prior records. This uncertainty directly impacts survey efficiency, as teams expend significant time and resources visiting locations that may ultimately not contribute usable data. Such inefficiencies represent a substantial burden on INE's data collection efforts.

Traditional approaches to mitigating these uncertainties, such as manual verification, repeated field visits, or reliance on outdated records are resource-intensive and cannot scale effectively, particularly in municipalities with high population mobility or rapidly changing housing conditions. The problem is further compounded by the heterogeneity of dwellings, as some units may serve non-residential purposes (e.g., vacation homes, storage facilities) or undergo temporary occupancy changes not reflected in existing administrative records.

This project addresses these challenges by leveraging multiple sources of existing and historical data within INE's databases. Key inputs include prior dwelling status records, utility consumption patterns, tax information, and geospatial location data. Each of these data sources provides complementary insights: previous survey participation indicates historical usage trends; utility consumption serves as a proxy for actual inhabitancy; tax records can identify registered owners or secondary usage patterns; and geographic information allows for spatial analysis of dwelling clusters and accessibility. By integrating these diverse datasets, the project aims to produce a predictive model capable of estimating the likely status of each dwelling prior to field visits.

The predictive approach is designed to classify dwellings according to occupancy probability. For example, moderate to high utility consumption may indicate active residential use, while low or sporadic consumption may suggest intermittent use, vacation occupancy, or vacancy. By setting informed thresholds on these variables, survey teams can prioritize dwellings for inspection, reduce redundant field visits, and allocate resources more efficiently. In addition to immediate operational benefits, this methodology aligns with INE's strategic objectives of modernizing survey practices, promoting data-driven decision-making, and adopting innovative statistical and computational methods.

Ultimately, this project not only seeks to improve the efficiency of household surveys in São João da Madeira but also provides a replicable framework that could inform broader INE practices. By demonstrating the practical value of integrating administrative, geospatial, and historical data with predictive modeling, the work contributes to advancing Portugal's national statistical infrastructure while offering insights applicable to similar statistical agencies facing comparable challenges.

1.4 Aim and Objectives

This project aims to predict the availability of dwellings for household surveys by leveraging existing and historical data stored in INE's databases. These data include prior year dwelling statuses, participation in the Resident Population Database (BPR), tax records, and utility consumption data, complemented by geographic information for a selected municipality (São João da Madeira). The primary objective is to develop a predictive model of dwelling status, where utility consumption is treated as a proxy for "signs of inhabitancy." Thresholds in consumption levels can distinguish between dwellings actively inhabited, used occasionally (e.g., storage or vacation homes), or vacant. By integrating multiple data sources and predictive analytics, the project contributes to INE's strategic goal of modernizing survey methodologies, enhancing efficiency, and ensuring greater accuracy in Portugal's official statistics.

1.5 Broader Context and Links to International Practices

While the immediate motivation for this project is rooted in the operational challenges faced by Statistics Portugal, the initiative also resonates with a wider international trend in the modernization of official statistics. Across Europe and beyond, national statistical institutes are increasingly moving away from traditional, census-based, and field-intensive methods of maintaining household survey frames. Instead, they are turning toward the integration of administrative registers, tax records, and utility data, often complemented by geospatial information systems and advanced analytical techniques.

Agencies such as INSEE in France, Eurostat at the European level, and the U.S. Census Bureau have already undertaken substantial reforms to reduce the dependence on resource-heavy enumeration and to improve both efficiency and coverage through administrative data linkage. These experiences provide valuable reference points for understanding the challenges and opportunities that accompany such transitions, from issues of data quality and harmonization to questions of institutional capacity and legal safeguards.

By positioning itself within this international trajectory, Statistics Portugal not only addresses immediate operational needs but also contributes to a broader methodological conversation about the future of survey design. The present project can thus be seen as part of this continuum: a localized case study in São João da Madeira that draws on administrative and geospatial data to anticipate dwelling occupancy, while also aligning with global practices in the pursuit of more efficient, accurate, and sustainable approaches to household survey management.

Chapter 2

Literature Review

The literature on household survey methodology demonstrates a clear shift toward modernized sampling frames that integrate administrative, geospatial, and auxiliary data. Traditional field-based methods; such as manual dwelling listings or census-based frames, are increasingly challenged by dynamic population patterns, high operational costs, and the need for timely statistical output. In response, statistical agencies worldwide have explored approaches that leverage existing administrative records, utility consumption data, and advanced predictive techniques to improve survey efficiency, coverage, and accuracy.

International examples illustrate diverse strategies in this modernization process. INSEE has gradually transitioned from census-based to tax-based sampling frames, demonstrating how administrative data can enhance both precision and operational feasibility. Eurostat provides methodological guidance emphasizing harmonization and best practices across member states, reinforcing principles of probability-based sampling, quality assurance, and model-assisted estimation. Meanwhile, the U.S. Census Bureau's MAF–ARF initiative showcases the integration of address-based systems with auxiliary administrative records—including tax, utility, and program data—to produce comprehensive, adaptable sampling frames.

These experiences highlight the growing importance of auxiliary and spatial information in survey design, not only for improving coverage but also for supporting predictive modelling. In particular, the use of historical and administrative data to anticipate dwelling occupancy directly informs the approach taken in this thesis. By synthesizing these lessons from international practice, the chapter frames the methodology for predicting household occupancy using INE's administrative and geospatial datasets, situating the work within a broader context of global statistical modernization.

2.1 Sampling Frames and Administrative Data Integration

Accurate sampling frames are a vital cornerstone of statistical surveys, yet traditional methods for constructing and maintaining them are prone to errors. Outdated contact information, unoccupied dwellings, and omitted new constructions can reduce the representativeness and quality of survey data (Santos, A. M., Portugal, A., Marcelo, C., Magalhães, J., Cruz, P., Campos, P., Leitão, P., & Mina, S., 2024) . Moreover, updating sampling frames through conventional field methods is often resource-intensive and slow, making them insufficient for rapidly changing socio-economic environments or resource-constrained regions.

Administrative Data Integration

Integrating administrative datasets into sampling frames can significantly enhance both accuracy and timeliness. For instance, population registries and housing databases provide complementary sources of information, including sociodemographic characteristics, dwelling occupancy, and geographic distribution (DeWaard et al., 2022; Foley, 2020; Sullivan & Genadek, 2024). By cross-referencing these datasets, inconsistencies can be identified and corrected—for example, a dwelling marked as vacant in one dataset may be confirmed as occupied based on utility consumption records. Together, these sources provide a robust foundation for sampling frames, ensuring both household- and dwelling-level attributes are accurately represented.

Geospatial Technologies: GIS and Satellite Imagery

Geographic Information Systems (GIS) and satellite imagery further support the construction and maintenance of sampling frames by visualizing population distributions and detecting demographic changes over space (Goodchild, 2007a; Langford, 2013; Langford & Unwin, 1994). These technologies allow statistical agencies to monitor population movements, dynamically update sampling frames, and identify dwellings in remote or otherwise inaccessible areas, which would be difficult to capture through traditional field surveys.

Machine Learning for Sampling Optimization

Machine learning models can enhance sampling efficiency by analyzing patterns in administrative and geospatial data to predict occupancy status and prioritize regions for survey updates (Kontokosta & Tull, 2017). Such predictive approaches enable targeted sampling based on socio-economic or demographic characteristics, potentially reducing costs while improving coverage. Nonetheless, challenges remain, including harmonizing disparate datasets, complying with privacy regulations, and ensuring adequate computational resources and expertise.

Limitations and Challenges

Despite these advances, several limitations persist. Administrative datasets often contain gaps or inconsistencies, investments in infrastructure and software are substantial, and machine learning applications require careful oversight to maintain data quality and compliance (DeWaard et al., 2022; Merly, Alpa, & Sillard, 2019). The integration of these diverse data sources necessitates careful management to fully realize the benefits in accuracy, efficiency, and adaptability.

In summary, combining administrative data, GIS, satellite imagery, and predictive modelling offers substantial improvements in the construction and maintenance of sampling frames. By leveraging complementary datasets and advanced analytical techniques, statistical agencies can reduce errors, enhance representativeness, and respond more effectively to evolving socio-economic and geographic conditions (Favre-Martinoz, 2015; Grafström & Tillé, 2013; Kontokosta & Tull, 2017).

2.2 INSEE: Transitioning from Census-Based to Tax-Based Sampling Frames

INSEE has progressively modernized its household survey methodology by transitioning from decennial census-based sampling frames to tax-based sources, complemented by spatial balancing techniques. These innovations directly inform the thesis aim of predicting dwelling occupancy by demonstrating how multi-source and spatially optimized data can enhance survey representativeness.

Evolution of Sampling Frames

Historically, INSEE relied on population censuses to construct sampling frames. The decennial rhythm and methodological changes created representation challenges (Favre-Martinoz, 2015). Tax data emerged as a viable alternative due to its near-complete coverage and precise geolocation. The Fidéli system (2016) further refined tax data by addressing double-counting and aligning definitions with census concepts (Merly-Alpa & Sillard, 2019).

Methodological Innovations

To optimize survey accuracy and efficiency, INSEE adopted spatial balancing algorithms for primary unit selection, reducing variance by up to 20% for spatially correlated variables (Favre, Martinoz, Merly, & Alpa, 2016; Grafström & Tillé, 2013). Primary units were defined as adjacent municipality clusters ($\geq 2,500$ dwellings), with a traveling salesman algorithm optimizing geographic compactness. This reduced the average unit extent by 25% compared to prior samples.

Coordination of Samples

INSEE coordinated household surveys with the Labour Force Survey (LFS) by defining coordination units ($\geq 10,000$ dwellings) and applying indirect sampling methods to ensure consistent weights (Deville & Lavallée, 2006). Mixed-mode surveys, including internet and phone data collection, are increasingly integrated alongside traditional face-to-face interviews (Cotton & Dubois, 2019).

Regulatory and Practical Drivers

EU regulations, particularly the Integrated European Social Statistics framework, set specific accuracy thresholds for indicators such as unemployment rates, shaping sample size and stratification (Cases, 2019). Operational efficiency was improved through the NAUTILE application, which streamlined sample selection and household marking, supported by agile project management.

Outcomes and Implications

The updated master sample improved first-stage accuracy five- to six-fold, leveraging Fidéli income variables and spatial balancing, while reducing respondent burden through synchronized surveys (Tillé, 2019). These developments highlight the importance of methodological rigor, regulatory compliance, and operational pragmatism in official statistics.

2.3 Eurostat: Survey Sampling Reference Guidelines

The methodological advancements implemented by INSEE demonstrate how national statistical offices leverage administrative and spatial data to enhance survey representativeness and precision. While these innovations are tailored to the French context, Eurostat's guidelines provide a complementary, pan-European perspective on survey design. In particular, Eurostat emphasizes probability-based sampling, quality assurance, and model-assisted estimation techniques, which inform best practices applicable across diverse contexts. Together, these frameworks underscore the importance of integrating multiple data sources and optimizing sampling methodologies—principles directly relevant to developing predictive models of dwelling occupancy in this thesis.

Eurostat's guidelines provide a foundational framework for designing high-quality surveys across the EU, emphasizing **probability-based designs**, methodological rigor, and quality assurance (Cochran, 1977; Kish, 1965).

Foundational Concepts

The guidelines distinguish descriptive surveys, which estimate population parameters (e.g., means, totals), from analytical surveys, which examine relationships between variables. Social surveys are voluntary and diverse, while business surveys are mandatory with stricter sampling protocols. Probability sampling is emphasized as the standard method to minimize bias.

Survey Design and Quality Assurance

Multiple sampling frames are discussed:

- **Simple Random Sampling (SRS)**: Baseline approach for efficiency benchmarking.
- **Systematic Sampling**: More efficient than SRS when logical ordering exists.
- **Stratified Sampling**: Divides populations into homogeneous strata with proportional or Neyman allocation for precision optimization.
- **Probability Proportional to Size (PPS)**: Leverages size-specific data to improve sampling accuracy.

Quality metrics such as design effect and effective sample size help assess efficiency losses in complex designs.

Advanced Methodologies

Model-assisted estimation uses auxiliary data, including regression and post-stratification, to improve estimates. Real-world challenges are addressed:

- **Nonresponse**: Unit nonresponse is mitigated through reweighting; item nonresponse through imputation techniques (hot-deck, regression).
- **Frame Imperfections**: Multiple frame approaches handle over- or under-coverage.

Practical Applications

Examples include the EU Labour Force Survey, demonstrating stratified sampling with regional precision requirements, and PISA educational surveys, illustrating clustering effects on effective sample size. Software implementations (SAS, SPSS, R) support practical applications (VLISS Virtual Laboratory, 2020). Eurostat's guidelines provide a rigorous framework for survey design, ensuring probability-based sampling, quality assurance, and adaptability to modern statistical demands. They complement national initiatives such as INSEE's methodological advancements, highlighting international best practices in official statistics.

2.4 Census Bureau MAF-ARF: Leveraging Administrative Records for Address-Based Sampling

Building on the principles outlined by INSEE and Eurostat, the U.S. Census Bureau's Master Address File–Administrative Records Frame (MAF-ARF) exemplifies the integration of administrative data into high-quality sampling frames (DeWaard et al., 2022; Foley, 2020; Sullivan & Genadek, 2024). The MAF-ARF approach is highly relevant to this thesis, illustrating how combining multiple administrative sources can support predictive modelling of dwelling occupancy.

Structure and Sources

The MAF-ARF draws on multiple administrative data sources to construct a comprehensive, up-to-date address frame:

- **Tax records**: Provide near-universal coverage of households and businesses, with annual updates that capture changes in occupancy and property ownership.
- **Utility records**: Offer additional validation for dwelling existence and occupancy status, including active service accounts and geographic coordinates.
- **Social program participation lists**: Include information from programs like SNAP or Medicaid, which help identify households not captured through tax or utility data.

Regular updating, deduplication, and validation procedures minimize coverage errors and enhance temporal accuracy (Keith, Finlay, & Genadek, 2021). The MAF-ARF framework is also designed to handle missing or conflicting data through hierarchical rules that prioritize sources by reliability.

Methodological Advantages

Administrative data integration provides several methodological benefits:

- **Near-complete coverage**: The frame captures most residential addresses, reducing the risk of undercoverage inherent in traditional census enumeration.
- **Accurate geolocation**: Addresses are geocoded, enabling spatially explicit survey design and small-area estimation.

- **Support for complex designs**: The frame allows stratified, cluster, and probability-proportional-to-size sampling, improving precision for diverse survey objectives.

- **Reduction of nonresponse bias**: Linking multiple administrative sources helps detect households that may otherwise be missed, improving representativeness.

Machine learning and model-assisted estimation techniques are applied to predict characteristics of households and support adaptive sampling designs, optimizing data collection efficiency (DeWaard et al., 2022; Foley, 2020).

Practical Implementation

The MAF-ARF is used operationally to:

- Coordinate the selection of households across multiple surveys, including the American Community Survey and the Current Population Survey.

- Ensure consistent sampling weights and coverage across surveys to support longitudinal and cross-sectional analysis.

- Support small-area estimation and spatially explicit population estimates by combining administrative data with survey responses.

The Census Bureau also implements rigorous quality assurance, including frame audits, duplicate detection, and error modelling, which directly inform survey operations and reduce variance in estimates. The MAF-ARF illustrates state-of-the-art integration of administrative records for survey sampling. It provides practical guidance for operationalizing multi-source frames, methodological inspiration for spatially optimized sampling, and a benchmark for predictive modelling applications, all of which inform the design and validation of the datasets used in this thesis.

2.5 Discussion

The theoretical foundation reviewed demonstrates how international statistical agencies are converging on a common direction: the modernization of sampling frames through administrative integration, geospatial technologies, and advanced modelling techniques. Across contexts, a consistent theme emerges: traditional field-based approaches are increasingly insufficient for capturing dynamic socio-economic realities.

The Portuguese case (Santos et al., 2024) highlights the role of administrative records, GIS, and machine learning in enhancing both the accuracy and efficiency of sampling frames. This is complemented by INSEE's transition to tax-based frames and the application of spatial balancing methods (Favre-Martinoz, 2015; Grafström & Tillé, 2013; Merly-Alpa & Sillard, 2019), which demonstrate how multiple data sources and algorithmic techniques can improve representativeness while reducing costs and respondent burden. At a broader level, Eurostat's methodological guidelines (Cochran, 1977; Kish, 1965; Lohr, 2019) reinforce the importance of probability-based sampling, survey quality assurance, and model-assisted estimation, offering a harmonized framework that transcends national contexts. Finally, the U.S. Census

Bureau's MAF-ARF (DeWaard et al., 2022; Foley, 2020; Sullivan & Genadek, 2024) exemplifies how large-scale integration of tax, utility, and program data, coupled with geospatial validation and predictive analytics, can deliver high-quality, operational sampling frames.

Together, these diverse but complementary approaches directly inform the aim of this thesis: the predictive modelling of dwelling occupancy. The reviewed literature underscores several methodological principles crucial for this objective: the integration of heterogeneous administrative datasets (Wallgren & Wallgren, 2014; Zhang, 2012), the leveraging of spatial and geostatistical techniques (Deville & Tillé, 2004; Goodchild, 2007a), the adoption of probability-based and quality-assured sampling designs, and the use of machine learning to enhance efficiency and adaptability (Kreuter & Peng, 2020; Rao & Molina, 2015). By synthesizing these lessons, the thesis situates its predictive modelling framework within a broader international trajectory of statistical modernization, aligning local implementation with globally recognized best practices.

Chapter 3

Methodology

Before discussing the data and its sources in detail, it is important to provide an overview of the methodological approach adopted in this study. The overarching objective of this research was to develop predictive models of household occupancy by integrating multiple sources of information and combining them with advanced machine learning techniques. To achieve this, the methodology followed a sequential structure which began with data integration and preparation, followed by model development, evaluation, and spatial visualization of results.

The study utilized heterogeneous datasets that included administrative records, energy consumption data, and geospatial attributes. These sources were combined to create a comprehensive feature set capable of capturing demographic, behavioural, and locational characteristics relevant to occupancy outcomes. After preprocessing and feature engineering, several machine learning algorithms were applied, with a particular focus on tree-based methods such as Random Forest and XGBoost, due to their robustness, interpretability, and capacity to handle high-dimensional data. Model performance was assessed through comparative evaluation using standard classification metrics.

Beyond predictive modelling, the methodology also incorporated a spatial analysis component. The results of the best-performing models were projected onto geographic information systems (GIS) to capture spatial dependencies, highlight clustering effects, and provide visual insights into the distribution of occupancy predictions across regions. This step allowed the study to bridge predictive analytics with spatial representation, thereby enhancing both interpretability and applicability of the findings.

This chapter is therefore structured to reflect the methodological workflow: the data sources are first described, followed by the data preparation process, the theoretical foundations of the chosen algorithms, the iterative model development, and the evaluation of results. The chapter concludes with a discussion of implementation details and the visualization of household occupancy predictions within a GIS framework.

3.1 The Data

The initial data was imported into and handles within Jupyter Notebook, with 10,000 rows of raw data available for use. The data sprouted from two different sources; Base Nacional de Edifícios (BNE), and Base de População Residente (BPR), which are internal databases exclusive to INE.

The Base Nacional de Edifícios, also known as the National Building Database (BNE), was first created in 2021 by Statistics Portugal to serve as a centralized geographic database to serve as a centralized geographic database of buildings accross the country. It is a high value cartographic internal database exclusive for statistics Portugal and aims at supporting statistical operations, public policies and decision-making processies (Instituto Nacional de Estatística. (2021). *Base Nacional de Edifícios (BNE)*. Statistics Portugal. https://www.ine.pt). The BNE contains georeferenced information of buildings, where they are identified for statistical purposes by a unique identification number and an identifying point in cartographic coordinates (location data consisting of longitude and latitude), as well as a list of attributes regarding those buildings (*Base de Georeferenciação de Edifícios: documento metodológico. Lisboa, INE, 2003*).

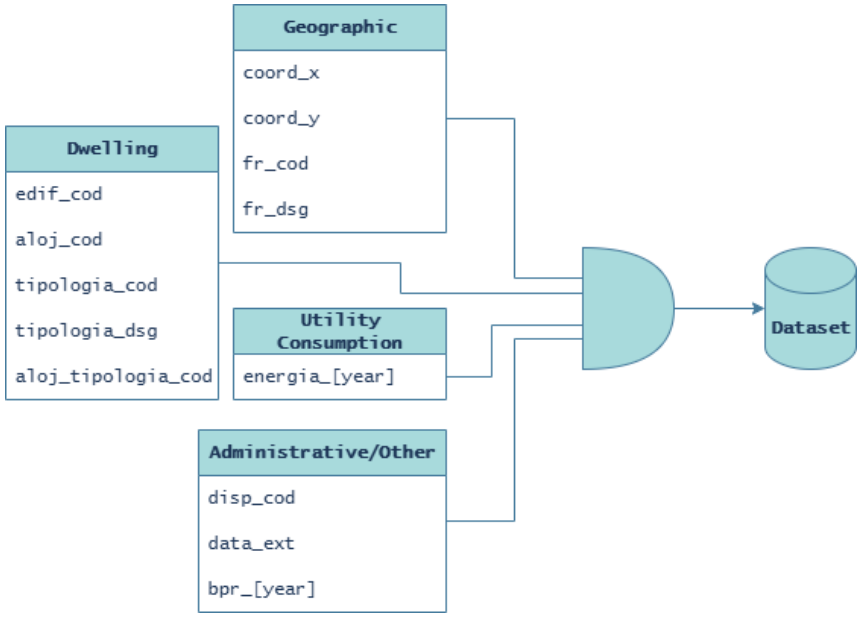


Figure 3.1: Data Overview

The dataset[1] classifies the buildings by use into four categories:

1. Exclusively residential building (100%): any building where the entirety of the usable area is exclusively allocated for housing or complementary uses, such as parking, storage, or social uses.

2. Mainly residential building (from 50% to 99%): any building in which 50% or more of the total usable area is allocated to housing or complementary uses, such as parking, storage, or social uses.

3. Mainly non-residential building (until 49%): any building in which 50% or more of the total usable area is allocated to non-residential purposes.

4. Non-residential building: any building where the entirety of the usable area is allocated non-residential or commercial uses, such as warehouses, shops, or for manufacturing.

On the other hand, The Base de População Residente (BPR), or Resident Population Base, was first established in 2017 as part of an effort to modernize population statistics through the integration of administrative data sources. These sources include social security

[1]The full list of variables and their descriptions will be provided in the appendix in Table A.1.

databases, tax authority records, education system registries, and civil registration systems, among others. The BPR provides granular, annually updated, georeferenced estimates of the resident population in Portugal. It covers multiple territorial levels, from national scales (e.g., NUTS I, II, and III) down to local administrative divisions such as Freguesias (Parishes). The database contains individual-level records encompassing demographic characteristics, residential locations, household structures, socioeconomic attributes, and population counts. Its primary purpose is to support statistical production, territorial planning, and policy evaluation by offering a reliable and up-to-date portrait of the population.

The integration of the Base Nacional de Edifícios (BNE) and the Base de População Residente (BPR) into one repository represents a major step toward the modernization of the household surveys process in Portugal. When used together, these two databases significantly enhance the efficiency, accuracy, and spatial resolution of household surveys operations. The BNE provides a detailed, georeferenced map of all buildings in the country, allowing authorities to precisely locate and classify physical structures, while the BPR offers up-to-date information on who lives where, based on harmonized administrative sources. This synergy supports a more targeted and resource-efficient enumeration, reduces duplication or omission of addresses, and helps identify coverage gaps. Furthermore, the combined use of these registries enables pre-survey planning, improves sampling frames, and facilitates real-time validation of population data against existing residential structures. Ultimately, this integrated approach lays the groundwork for a register-based household survey model, minimizing reliance on exhaustive fieldwork while maintaining improving data quality, timeliness, and policy relevance.

This is where the dataset at hand sprouted from, as mentioned before, combining some of the most relevant attributes from both BNE and BPR to aid in generating a predictive model first that would be of benefit to the survey teams by enhancing and decreasing the list of buildings on the survey route to only viable ones, thus saving time and effort during fieldwork, and secondly providing precise location data of viable candidate buildings, visualized on a map using Geographic Information Software (such as ArcGIS Pro) and mapping techniques, to illustrate the distribution of the selected candidates and help draw a clearer, more efficient route of survey collection, and better concentrate the efforts of the field workers on viable cases to save time and effort.

The dataset, illustrated in the appendex in Table A.1, was split into an 80/20 split due to the restricted sample of records provided, and to allocate more data for training to better assist in the accuracy of the predictive models. The accuracy measure used was the F1-score, which is the harmonic mean of precision and recall, providing a balanced measure of a model's accuracy when dealing with imbalanced classes. The F1-score ranges from 0 to 1, where a higher score indicates better performance in correctly identifying the positive class while minimizing false positives and false negatives. Initial inspection of the data revealed heavy class imbalances that required several attempts to deal with, although their existence is justified by the fact that a subset of variables within the dataset were the same variable but with values of previous years, and since one of those variables (aloj_sit_cod_2025) was used as the target variable to test the models before moving forward to predicting one step ahead, it made sense that the variables of previous years would contribute the most to its prediction as they are technically leading into it, thus constituting a heavy class imbalance, which called for the use of F1-score as an evaluation metric to provide a more meaningful assessment of the model's ability to correctly identify the minority classes.

The dataset contains pre-processed historical categorical data of dwellings describing the status/situation of the dwellings for the years 2018 to 2025 under the variables pertained to Situação do Alojamento Familiar, or Family Housing Situation (aloj_sit_cod_xxxx) which was derived from the BNE, with approximately 13.6% of total missing values across these categories, the variables contained the information about the status of a dwelling in a numerical format with status code referring to its respective categorical status as defined in a document provided by INE, where the definition of each variable and code as well as their meanings is described. For Family Housing Situation, the numerical categories represented the following:

1: 'Alojamento familiar de residência habitual', 2 for 'Alojamento familiar de residência secundária', 3 for 'Alojamento familiar vago', 5 for 'Alojamento familiar para demolição', 6 for 'Alojamento familiar inexistente por erro dos Censos', 7 for 'Alojamento familiar inexistente por ter sido associado a outro alojamento', 10 for 'Alojamento familiar ocupado para outros fins', 11 for 'Alojamento familiar não localizável', 12 for 'Alojamento familiar inacessível', 96 for 'Alojamento familiar demolido', 97 for 'Alojamento familiar substituído', 98 for 'Alojamento familiar eliminado', and finally 99 for 'Alojamento familiar indefinido'.

The data also contains the variables "edif_cod" for building code, and "aloj_cod" for individual unit code, which when combined create a unique identifier for each individual dwelling within a complex or suburban. Other variables included within the dataset were variables derived from the BPR named (bpr_xxxx) for years 2020 to 2023, which describe whether a selected dwelling was included in the BPR survey/study for a selected year in which the cases that correspond to "yes" for dwelling participation were given the value X, and NaN for the rest of the dwellings which there are no information on. These variables were mostly empty, with the percentage of missing values across them coming to approximately 63% as they hadn't been updated/imputed yet for the purpose of this study. Additionally, the variable "disp_per_sigla" exists to essentially identify the external entity that provided information on the respective building where applies, "fr_cod" is a variable that represents the fregusia (parish) code while "fr_dsg" contains the name of that parish. "tipologia_cod" with "tipologia_dsg" which represent the code of typology of a building and the description of it respectively. Additionally, "disp_cod" indicates the approximate availability of selected dwelling to participate in the survey. "aloj_tipologia_cod" indicates the typology of a dwelling with values: family dwelling, collective, indefinite. And finally, "data_ext" represents data extraction date.

For the better part of the project timeline, the empirical part was limited to work with historical data related to the status of a dwelling to produce the initial results test, without including tax and utility data, or BPR variables until they are updated. Each of the categorical values that were represented numerically were mapped to their respective categorical values as defined in *FNA_4B_Tabelas_Descodificacao_20130228_V2.7* to avoid them being treated as inherently numerical values where the larger numbers would be considered as a "better" status.

3.2 Data preparation

The first step is to clean the variable/column names from white spaces, check data types and non-null counts, generate descriptive statistics for all columns (including non-numeric), and check for missing values in each column. After which a status map for dwelling status was created for readability and better interpretation of the data by the models. Next, the

study moves to handling the missing data in the columns included in the initial model which are all categorical data represented numerically, wherein a map is applied to these categorical variables.

When investigating the Family Housing Situation variables, they exhibited a low number of missing data, and were handled as follows: replaced the null and missing values with status 3 'Alojamento familiar vago', the reasoning for that was that it is known that the dwelling exists based on existing historical data but no data regarding that dwelling for a specific year is present. Thus, as it is a matter of fact that this specific dwelling exists, the most appropriate action that was settled on is to set the status to vague/vacant and not indefinite. It is also worth mentioning that considering the amount of missing data per variable is significantly low, if this approach introduces any bias or inaccuracy in the model it would vary by small percentages.

Illustrated below is the amount of existing (non-missing) data per variable in the initial tests:

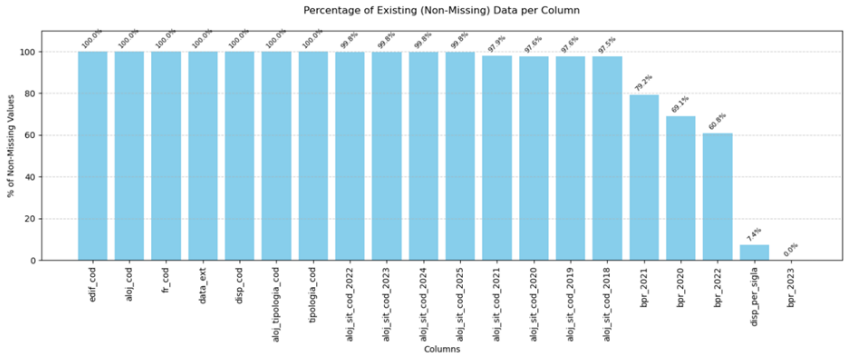


Figure 3.2: Percentage of Existing (Non-Missing) Data per column

Furthermore, due to the significant amount of missing data and ambiguity of the variable "disp_per_sigla", it was excluded from the model. Although the BPR variables also had a significant amount of missing values and a level of ambiguity, upon further inspection and referring to INE's official definition of BPR along with some sample values, it was safe to conclude that the missing values signified that the dwelling was not included or has not participated in the BPR of a specific year. After consulting with the department of Methodology, the study concluded that it is safe to handle the missing data using the categorical value "Unknown", as the existing data in those variables was a simple "X" indicating the participation/inclusion of the dwelling in this BPR variable. Thus, the variables content after handling the missing data would be one of two categories: either "X" for true cases of inclusion, or simply "Unknown" for cases where information is just not available. Label Encoder is applied to all variables except the target variable: "aloj_sit_cod_2025", to explore their correlation with the target variable and point out significant features that aid in the feature selection process. Some variables exhibited low positive or negative correlation, while some exhibited a complete lack of correlation to the target variable, those being identifiers that would normally not contribute to the models learning process. The variables with 0 correlation were excluded from further steps. Initially, chi^2 test is applied to find significant features (feature selection) and selecting features with p-value < 0.05, and a threshold of 90% to highlight variables that could lead to possible leakage within the models (Figure 3.3). However, the choice of the chi^2 test is discarded as it fails to capture the strength or magnitude of the association between variables;

it only indicates whether a statistically significant relationship exists. Therefore, in this study, Cramér's V and Pearson's correlation are opted for instead, as they provide a standardized measure of association strength, allow for direct comparison across variables, as well as being less sensitive to large sample sizes, making them more informative for feature selection and leakage detection. By doing this, the study aimed to reduce the dimensionality of the dataset by keeping only the features most relevant to predicting the target, thereby simplifying the model and potentially improving its interpretability and performance. Previous values of the "aloj_sit_cod_xxxx" showed high correlation to the target variable when using Cramer's V (see Figure 4.2). This is explained by the fact that they are essentially past values of the same variable. However, keeping them as they are in the dataset would cause leakage issues as they far surpassed the pre-set 90 percent threshold. To remedy this, the study continues by aggregating them into a single categorical variable using most-frequent as a method of selection to reduce dimensionality and redundancy of highly correlated historical data into just one feature (Figure 3.4).

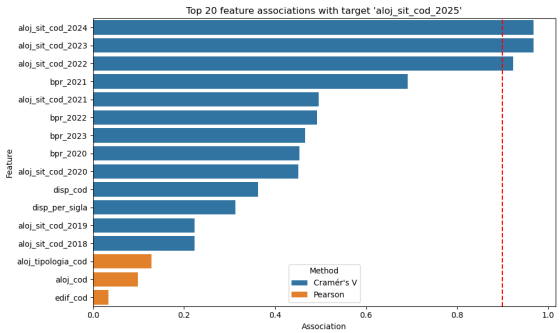


Figure 3.3: Top features most associated with target (Pearson & Cramér's V)

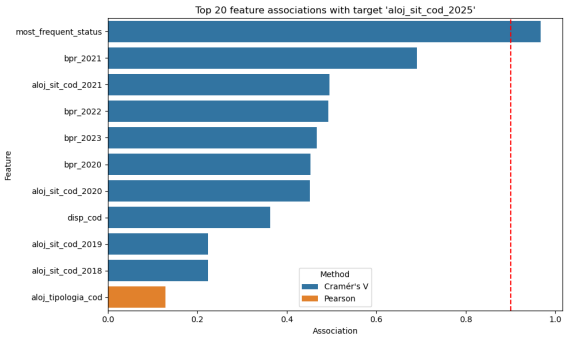


Figure 3.4: Top features most associated with target (Pearson & Cramér's V) After Aggregation

The dataset exhibited substantial class imbalance, with class 0 (Alojamento familiar de residência habitual) representing approximately 82% of the dataset, and several minority classes containing fewer than 1% of samples. This imbalance was visualized using frequency histograms (Figure 3.5).

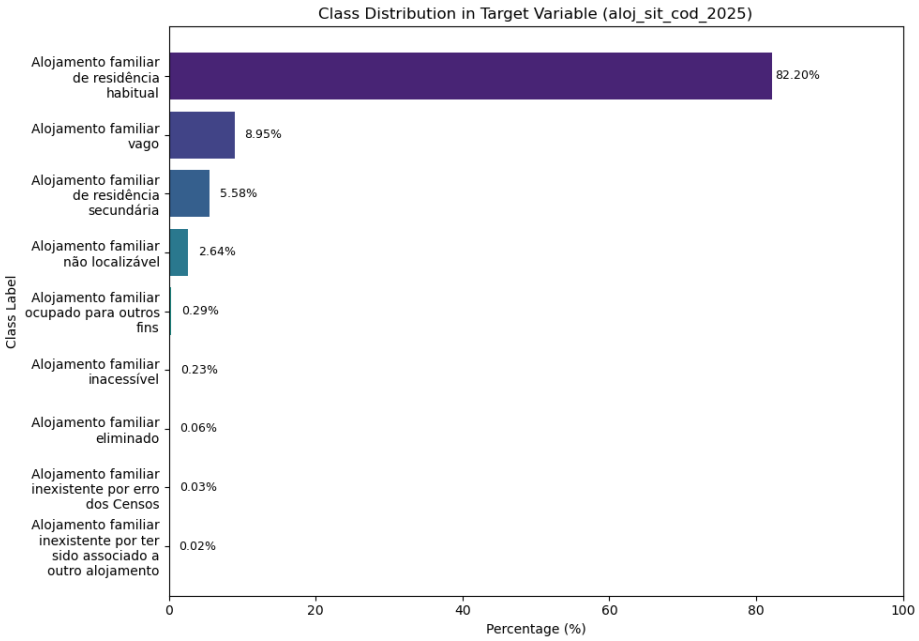


Figure 3.5: Class Distribution in Target Variable

Given the imbalanced distribution, special attention was given to mitigating bias during model training, which is discussed in the modelling section. Finally, the dataset was split into an 80/20 split, with the larger part going to training the models to ensure sufficient amounts of data for the model to learn considering the limited amount of data at hand.

3.3 Machine Learning Models (Theoretical Foundation)

In order to predict household occupancy status, this study employs a set of machine learning models that balance interpretability, robustness, and predictive power. The choice of models—Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—was motivated by their complementary strengths in handling heterogeneous datasets that include spatial, temporal, and categorical information. Applying multiple algorithms allows for a more comprehensive evaluation of model performance and ensures that findings are not limited to a single methodological approach.

Logistic Regression (LR) represents the most classical and interpretable model considered in this study. As a generalized linear model, LR estimates the probability of class membership by fitting a logistic function to a linear combination of input features. Its simplicity enables straightforward interpretation of coefficients, making it particularly valuable for establishing baseline performance and for understanding direct feature–outcome associations. However, LR assumes linearity in the log-odds of the target variable, which may limit its ability to capture complex, non-linear patterns present in occupancy-related data (Hosmer, Lemeshow, & Sturdivant, 2013)

Random Forest (RF) is an ensemble learning method based on decision trees. By aggregating predictions from multiple trees trained on bootstrapped samples and random subsets of features, RF reduces variance and enhances generalization. This makes it well-suited for handling high-dimensional, noisy, and mixed-type datasets such as those employed in this

research. Moreover, RF naturally provides measures of feature importance, offering valuable insights into which variables—spatial coordinates, energy consumption indicators, or administrative codes—contribute most to predicting occupancy status (Breiman, 2001).

Extreme Gradient Boosting (XGBoost) extends the ensemble approach by implementing gradient boosting with advanced regularization techniques. Unlike RF, which combines trees in parallel, XGBoost builds trees sequentially, with each iteration aiming to correct errors from previous ones. This results in a highly efficient and flexible model capable of capturing non-linear interactions and subtle dependencies between features. Its scalability and strong predictive performance have made XGBoost one of the leading algorithms for structured tabular data tasks. Importantly for this study, XGBoost can effectively model the complex interplay between spatial and energy-related variables, thereby improving classification accuracy across multiple occupancy classes (Chen & Guestrin, 2016).

By incorporating models with varying levels of complexity and interpretability, this study leverages the strengths of both traditional statistical approaches and modern ensemble learning techniques. This comparative framework enables a robust assessment of predictive performance while also providing interpretable insights into the factors shaping occupancy status, which is essential for informing spatial analysis and potential policy applications.

3.4 Modeling

Building on these theoretical foundations, the following section details the implementation of the selected models within the predictive framework, including the preprocessing steps, training procedures, and evaluation criteria, including: F1-score, accuracy, cross-validation score, and precision. The selection of these algorithms was motivated by the predominantly categorical nature of the dataset, as each is well-suited to handling categorical variables, either natively or with minimal preprocessing. Logistic Regression was chosen for its simplicity and interpretability, particularly in modelling linear relationships between features and the target variable, although it may be limited in capturing complex, non-linear interactions. Random Forest, as an ensemble method based on decision trees, offers robustness to noise and strong performance with categorical data, but at the expense of reduced interpretability and potential overfitting. XGBoost, a gradient boosting framework, is known for its high predictive accuracy and efficiency with large datasets, while also accommodating categorical and numerical data. Despite its higher computational cost and complexity, its inclusion was justified by its widespread use and success in structured data prediction tasks. The combination of these three models thus provides a comprehensive approach that balances interpretability, generalizability, and predictive power, aligning with the characteristics and demands of the data.

3.4.1 Initial Model Development and Evaluation

Handling Class Imbalance, Tuning and Evaluating with Multiple Samplers (First Results)

The dataset analyzed in this study exhibits pronounced class imbalance, primarily due to its historical origin and the application of data pre-processing methods such as missing value imputation and aggregation of correlated variables using a most-frequent value strategy. This imbalance presented significant challenges for model training, particularly for

minority classes with very limited samples. To address the issue of class imbalance within the training dataset, multiple synthetic oversampling techniques were incorporated into the model development pipeline. Specifically, the methods employed included Synthetic Minority Over-sampling Technique (SMOTE; (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)), BorderlineSMOTE ((Han, Wang, & Mao, 2005)), Adaptive Synthetic Sampling (ADASYN; (He, Bai, Garcia, & Li, 2008)), and KMeansSMOTE ((Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009)). These techniques function by generating synthetic samples of the minority class, thereby promoting a more balanced class distribution, which is critical for mitigating bias in predictive modelling and enhancing the classifier's ability to generalize to underrepresented classes (Fernández et al., 2018). To facilitate an efficient and systematic evaluation of these sampling methods in conjunction with hyperparameter optimization, a custom-designed function,
tune_and_evaluate_multiple_samplers_random(), was developed. This function integrates the application of oversampling techniques with the hyperparameter tuning process within a unified workflow. Hyperparameter optimization is performed using the *RandomizedSearchCV* method, which allows for a randomized and computationally efficient exploration of the parameter space. The weighted F1-score was selected as the primary evaluation metric due to its suitability for imbalanced classification problems, as it harmonizes the trade-off between precision and recall across classes of varying prevalence. Implementation leveraged the imblearn pipeline, enabling each sampling technique to be combined with the classifier in a streamlined and reproducible manner. This pipeline structure ensures that synthetic oversampling occurs exclusively within training folds during cross-validation, thus preventing data leakage and preserving the integrity of model evaluation. For each sampler-model combination, stratified k-fold cross-validation was employed to maintain class proportion consistency across folds. Following hyperparameter tuning, the optimized model was evaluated on an independent test set to assess generalization performance. The evaluation comprised multiple classification metrics, including accuracy, precision, recall, and a detailed classification report, providing a comprehensive understanding of model performance. Results from each configuration, including hyperparameters, evaluation metrics, and sampler characteristics, were systematically recorded to enable direct and rigorous comparison, and adjustments and runtime actions of the code were logged into a custom-made log file with time stamps for further revisions. This integrated approach supersedes previously utilized standalone tuning and evaluation procedures by consolidating data preprocessing, hyperparameter optimization, and performance assessment into a single, modular, and reproducible framework. Such a methodology enhances code efficiency and clarity, while ensuring that the efficacy of various sampling strategies in addressing class imbalance is evaluated under consistent and controlled experimental conditions.

Table 3.1: Model Performance by Oversampling Technique

Oversampling	Logistic Regression	Random Forest	XGBoost
None	0.9980 / 0.9973	0.9990 / 0.9985	0.9990 / 0.9990
SMOTE	0.9985 / nan	0.9980 / nan	0.9995 / nan
BorderlineSMOTE	0.9960 / nan	0.9990 / nan	0.9990 / nan
ADASYN	0.9875 / nan	0.9990 / nan	0.9995 / nan
SMOTEN	0.9975 / nan	0.9980 / nan	0.9985 / nan
KMeansSMOTE	– (error)	– (error)	– (error)

Note: Values shown as Test Accuracy / CV Weighted F1-score. "nan" indicates unavailable CV score. "– (error)" indicates sampler failed to run.

Handling Class Imbalance, Tuning and Evaluating with Multiple Samplers (Second Results)

At one point during model development, a derived historical feature, originally constructed as an aggregate of three highly correlated variables, was accidentally excluded. Its removal resulted in improved cross-validated performance, particularly for Random Forest and XGBoost models using BorderlineSMOTE, where accuracy reached up to 80%. This suggests that this feature, despite it being a core part of the original dataset, may have introduced redundancy or overfitting, and its exclusion was ultimately retained to improve generalization. The results demonstrated limited improvement from advanced resampling methods, with several technical challenges encountered during model training. For instance, Logistic Regression combined with BorderlineSMOTE achieved a cross-validation weighted F1-score near 0.75, yet exhibited poor recall on minority classes, indicating struggles to correctly identify rare categories. KMeansSMOTE consistently failed due to insufficient clustering among infrequent classes, preventing effective synthetic sample generation. Other oversampling approaches, including SMOTE and ADASYN, resulted in unstable cross-validation metrics with frequent 'nan' values and corresponding drops in test accuracy, often falling in the 63–65% range. Conversely, Logistic Regression without oversampling—but with class-weight balancing—produced more reliable and stable performance, with a weighted F1-score around 0.73 and test accuracy close to 73%. Similar patterns emerged with Random Forest and XGBoost classifiers. While these tree-based models generally achieved higher test accuracies—often exceeding 99%—and better handling of class imbalance, oversampling methods such as BorderlineSMOTE still introduced instability in cross-validation scores, frequently yielding 'nan' values. Notably, the exclusion of a historical aggregate feature in prior runs led to more consistent performance around 77–80% accuracy with BorderlineSMOTE, suggesting feature engineering decisions can significantly affect the interplay between resampling techniques and model robustness.

In terms of resampling methods, the expected benefits of advanced oversampling techniques were limited and often inconsistent. For example, Logistic Regression paired with BorderlineSMOTE achieved relatively strong performance during cross-validation, with a weighted F1-score near 0.75. However, its performance deteriorated significantly on the minority classes, highlighting a trade-off between overall accuracy and the ability to detect rare outcomes. Other techniques such as KMeansSMOTE repeatedly failed to generate mean-

ingful synthetic samples due to the absence of clear cluster structure among the infrequent classes. Meanwhile, SMOTE and ADASYN produced unstable metrics, frequently returning NaN values during cross-validation, and test accuracies declined to the 63–65% range in several iterations. These outcomes suggest that oversampling in multi-class imbalanced settings—especially when class overlap or feature noise is present—can be more challenging than anticipated.

Interestingly, models trained without oversampling, but with built-in class-weight adjustments, yielded more stable and interpretable results. Logistic Regression with balanced class weights produced consistent performance, reaching both test accuracy and weighted F1-scores around 73%. Similar stability was observed for ensemble models such as Random Forest and XGBoost. In cases where oversampling was avoided, test set accuracy consistently exceeded 99%, though this metric alone must be interpreted with caution due to class imbalance. When oversampling was applied to these tree-based models, especially with methods like BorderlineSMOTE, performance became less reliable across validation folds. The most robust and reproducible improvements occurred only when the historical aggregate was removed, suggesting that the interaction between oversampling techniques and correlated features can significantly affect model stability and generalization. These findings underscore the importance of both careful feature engineering and critical evaluation of resampling strategies when working with imbalanced multi-class data.

Table 3.2: Model Performance by Oversampling Technique (Excluding Historical Aggregate Feature)

Oversampling	Logistic Regression	Random Forest	XGBoost
None	72.95% / 0.7371	83.50% / 0.8269	87.25% / 0.8620
SMOTE	63.05% / nan	72.35% / nan	72.75% / nan
BorderlineSMOTE	78.45% / 0.7529	77.70% / 0.8189	80.85% / 0.7967
ADASYN	65.10% / nan	73.35% / nan	73.20% / nan
SMOTEN	72.55% / nan	69.30% / nan	68.90% / nan
KMeansSMOTE	– (error)	– (error)	– (error)

Note: Values shown as Test Accuracy / CV Weighted F1-score. "nan" indicates unavailable CV score. "– (error)" indicates sampler failed to run.

3.4.2 Impact of Historical Aggregate Feature on Model Performance

The predictive modelling results presented in the previous sections were obtained excluding the historical aggregate feature. This decision aimed to assess baseline model performance without potentially confounding information that might lead to data leakage. Among the models and oversampling techniques tested, the **XGBoost classifier without oversampling** achieved the highest test accuracy (87.25%) and the best weighted F1-score (0.8620). This suggests that the model effectively captures the underlying patterns in the data even without relying on aggregated historical metrics. Excluding the historical aggregate feature is justified given its potential to contain information from future time points, which would not be available in a real-world prediction context. This exclusion promotes model generalizability and avoids overly optimistic performance estimates. Moreover, the limited benefits observed from oversampling methods and the failure of some samplers indicate that the current feature set—absent the aggregate variable—already provides substantial predictive power. Con-

sequently, oversampling was not essential for improving performance in this phase. It is important to note, however, that these results are not the final outcome of the study. Subsequent analyses will introduce additional variables, including spatial (location) data, which are expected to further improve model performance and predictive accuracy. The models will be re-evaluated accordingly in later sections. Future work could also explore incorporating the historical aggregate feature within a rigorous temporal validation framework to mitigate leakage risks, as well as alternative strategies for addressing class imbalance.

In summary, these initial findings demonstrate that the XGBoost model, without oversampling and excluding the historical aggregate feature, provides a stable and efficient baseline for further model development.

3.4.3 Model Selection and Performance Comparison

Table 3.3 presents the test accuracy and cross-validated weighted F1-scores for the evaluated classifiers combined with various oversampling techniques. These metrics provide insight into both the overall predictive accuracy and the balance of classification performance across classes.

Table 3.3: Performance of Selected Models Without the Historical Aggregate

Model + Oversampling	Aggregate	Test Accuracy	CV Weighted F1-score
None + XGBoost	No	87.25%	0.8620
BorderlineSMOTE + Random Forest	No	77.70%	0.8189
BorderlineSMOTE + Logistic Regression	No	78.45%	0.7529

Note. Only models trained without the historical aggregate are shown, as models with the aggregate exhibited inflated performance and unreliable cross-validation scores due to suspected data leakage.

Among the different configurations, the **XGBoost classifier without any oversampling** achieved the highest test accuracy of 87.25% and the best weighted F1-score of 0.8620, indicating superior performance and robustness in predicting the target variable. This suggests that, for this dataset, the native class imbalance did not substantially hinder the model's effectiveness. The combination of **BorderlineSMOTE with Random Forest** also demonstrated competitive results, with a test accuracy of 77.70% and a weighted F1-score of 0.8189. This indicates that selective oversampling can provide benefits in some cases, particularly when used with certain classifiers. Other oversampling methods such as SMOTE, ADASYN, and SMOTEN either yielded lower performance metrics or resulted in fitting errors, as was the case with KMeansSMOTE, which failed to run successfully across multiple attempts. Given the comparative results, the **XGBoost model without oversampling** was selected for the final model deployment. This choice is supported not only by its superior test accuracy and cross-validated F1-score but also by its relative stability and efficiency during training. Additionally, the lack of improvement from oversampling techniques suggests that alternative strategies—such as adjusting class weights or further feature engineering—could be explored in future work to address any residual class imbalance. Overall, this evaluation confirms the effectiveness of the XGBoost classifier for the problem at hand and guides subsequent modelling and interpretation steps. To better interpret the predictive behavior of the selected XGBoost model, SHAP summary plots were generated (Figure 3.6). These plots illustrate

the contribution of each feature to model output, providing insight into feature importance and the directionality of their effects.

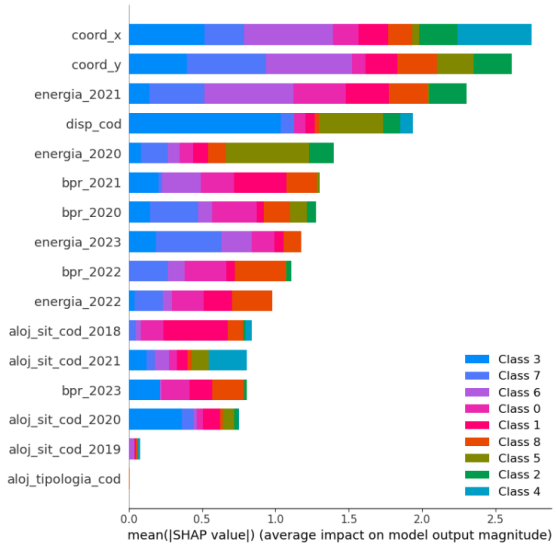


Figure 3.6: SHAP summary plot for the XGBoost model without historical aggregate feature

Note: The results presented prior to this point correspond to models trained without the inclusion of spatial (location) data and additional variables introduced later in this study. Subsequent sections will detail the incorporation of these new features, re-training of models, and a comparative evaluation of their impact on predictive performance.

3.4.4 Introduction of New Variables, Re-Modelling and Evaluation

In the late stages of the study, and before moving on to the semi-final steps of one-step-ahead prediction, as well as visualizing the results using maps, new data became available to be introduced into the dataset that requires re-modelling and re-evaluating the models, with the possibility of this new data enhancing the performances of the predictive algorithms by providing more context aiding in predicting future values with more accurately. The new data introduced consisted of spatial data in the form of X/Y coordinates, and utility data categorized by type and year.

First step was to merge the data from the three available sources:

1. The primary dataset (BNE-BPR)
2. Special data files (location data)
3. Utility records (Water and Electricity)

To ensure a clean and thorough approach, this section of the study reproduces the processes of the initial one all the way from the beginning. From importing the raw data, filtering and cleaning white spaces, to EDA and inspecting data types and percentage of missing values per variable and deciding on the best approach to handle them, especially now that we have introduced numerical and special data that require different handling techniques than categorical data would.

The special data exhibited more records than the primary dataset, with the number of records exceeding 25,000 records which lead to the merged data increasing in length but also leaving a significant amount of blanks in all other variables, decidedly, the data was first filtered by "edif_cod" upon merge by filtering out blank cells, as this variable will be a key variable come the stage of representing the final results using maps. However, after matching the data in the special source with the primary data source, the special data exhibited a small number of missing values where some dwellings did not have a registered location, with 202 dwelling, approximately 2%, missing their respective spatial data, which raises the question of which is the best way of handling it? Is it to change the status of the dwellings with no location data to "not found" or remove the entries corresponding to the missing values?

Ultimately, the study moved on by keeping "edif_cod" as the primary filtering method, as it is unclear if the data was missing at random or due to unavailability, and handled the missing special data using a centroid assignment approach by considering Junta de Freguesia de S. João da Madeira as a central landmark.

Next, the utility records, consisting of water consumption values and electricity consumption values weas originally stored as such: "ano" for the corresponding year of a value, "edif_code" and "aloj_cod" as common variables, "tipo_consumo" for consumption type which contained the two categories (Agua e Electricidade) in a single variable, "VT_medio" for the corresponding consumption value, and finally "aloj_princ_bne" which indicates the category of the corresponding dwelling on whether it is a primary residence or secondary. Upon inspecting the last variable, it was clear that the data provided the utility records was exclusively data pertained to primary residences, with "principal" code = 1, which would further restrict and limit the amount of available data only to dwellings registered as primary residences thus contrasting with the goal of exploring all types of dwellings, therefore the variable was not given high priority when merging the datasets as to not cause imbalances in the dwelling cases/statuses. Furthermore, the variables "tipo_consumo" and "VT_medio" were segmented and labeled by the year they correspond to using the "ano" variable as a label. The complete structure of the data is illustrated in Table A.1.

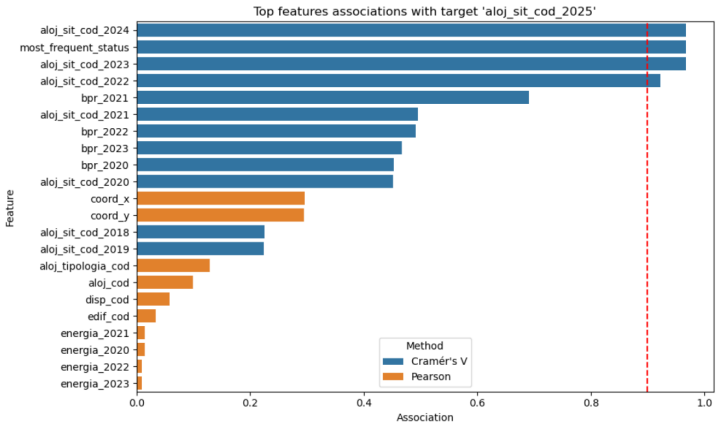


Figure 3.7: Top features most associated with target (Pearson Cramér's V) With Newly Introduced Data and Aggregate

Regardless of the utility data exhibiting very low association to the model., they were included in the initial pipeline for the updated model, as they are part of the core of this study. on the other hand, the newly added spacial data exhibited moderate correlation to

the target variable, using Pearson's correlation as an evaluation method, and thus remained included throughout the model as they are essential for the final step of visualization with GIS programs. Like in the previous process, the historical aggregate was included in the initial test and removed later to compare the performance of the models once again.

Tuning and Evaluating with New Data (First Results)

This part of the experiment relatively follows the same steps as the first modelling attempt that was done with the originally available dataset, where a historical aggregate of the top associated variables was included in the split, only difference here being the newly added energy and spacial data. Logistic regression, which had previously shown unrealistically high results, failed upon the addition of the new numerical data. In its original form, the dataset's categorical features (after encoding) effectively created a high-dimensional binary feature space in which the model could exploit discrete boundaries between classes. This often resulted in artificially strong separation, especially when categories were highly correlated with the target variable.

With the introduction of numerical variables, the geometry of the feature space changed substantially.Logistic regression, which had previously produced unrealistically high scores in the earlier, fully categorical dataset, experienced a severe decline in performance following the inclusion of new numerical variables. In its original form, the dataset's categorical features (after encoding) effectively created a high-dimensional binary feature space in which the model could exploit discrete boundaries between classes. This often resulted in artificially strong separation, particularly when categories were strongly correlated with the target variable.

With the introduction of numerical variables, the geometry of the feature space changed substantially. Logistic regression is a linear classifier in the transformed log-odds space and assumes that the classes can be separated by a single linear decision boundary. Many of the newly added continuous variables are unlikely to have a strictly linear relationship with the log-odds of the target, meaning the decision boundary required to separate the classes may now be non-linear or highly curved. Without transformations (e.g., polynomial terms, splines) or scaling to adjust for different ranges, these features distorted the separation, reduced margin clarity, and amplified multicollinearity between predictors—leading to model instability.

The results confirm this effect: the model consistently returned very low test accuracies, and cross-validation weighted F1-scores were unavailable (NaN) in most oversampling configurations due to convergence issues or failure to learn minority class structure. The relatively better score with SMOTEN (43.05% accuracy) suggests that the algorithm's compatibility with categorical variables preserved some predictive signal, while oversamplers designed for continuous spaces (e.g., SMOTE, ADASYN) offered little benefit, further implicating the new numerical variables as a primary driver of degradation.

In contrast to the performance collapse observed with logistic regression—where the inclusion of the new numerical variables disrupted the model's linear separability assumptions—tree-based ensemble models such as Random Forest and XGBoost reacted in the opposite manner. While logistic regression suffered from reduced predictive capacity due to increased complexity and non-linearity in the feature space, Random Forest and XGBoost produced excessively high accuracy values. This inflation was not indicative of superior modelling capability, but rather a consequence of data leakage introduced by the inclusion of a highly correlated historical aggregate. Owing to their capacity to capture complex, non-linear interactions, these models fully exploited the leaked information, resulting in near-perfect scores

that lack generalisability and cannot be considered reliable indicators of model performance.

Random Forest and XGBoost both exhibited near-identical behaviour in this run, with inflated and unrealistic performance metrics due to the presence of the leaked historical aggregate variable. This feature acted as a proxy for the target, enabling the models to achieve near-perfect accuracy across all oversampling configurations (up to 99.95%). Such results are a textbook symptom of overfitting caused by leakage—where the model memorises rather than learning generalisable patterns.

Tree-based models are particularly susceptible to this effect because they can exploit high-cardinality and strongly predictive features without relying on any assumptions of linearity. When leakage is present in both training and test data, these models can perfectly separate classes without capturing any transferable signal. The cross-validation weighted F1-scores confirm this lack of robustness: in most cases they were NaN, indicating instability or failure to correctly predict minority class instances once the leakage signal was weakened or absent in some CV folds. The only exceptions were BorderlineSMOTE and the un-sampled runs, where the CV F1-scores remained extremely high (≥99.8%), again reflecting the influence of the leaked feature rather than genuine predictive performance. The complete results for this run are presented in Table 3.4.

Table 3.4: Model performance including historical aggregate and new numerical variables

Oversampler / Model	Logistic Regression	Random Forest	XGBoost
SMOTE	28.50% / NaN	99.80% / NaN	99.90% / NaN
SMOTEN	43.05% / NaN	99.85% / NaN	99.95% / NaN
BorderlineSMOTE	15.45% / 31.83%	99.85% / 99.80%	99.95% / 99.89%
KMeansSMOTE	–	–	–
ADASYN	12.70% / NaN	99.85% / NaN	99.90% / NaN
None	20.25% / 36.90%	99.85% / 99.86%	99.90% / 99.89%

Note: The results are displayed as Test Accuracy / CV Weighted F1-score. "nan" indicates unavailable CV score. "– (error)" indicates sampler failed to run.

These results illustrate how the same dataset modification can produce two extreme and undesirable modelling behaviours: underfitting, as seen with logistic regression, and severe overfitting, as seen with Random Forest and XGBoost. The inclusion of the highly correlated historical aggregate fundamentally altered the learning dynamics of all models, either by violating their assumptions or by enabling them to exploit spurious correlations. This underscores the critical importance of rigorous feature auditing prior to model training, particularly when incorporating aggregated or historical variables. Given the strong evidence of data leakage in this run, the reported metrics cannot be regarded as representative of real-world performance. To address this, a second modelling attempt was undertaken with the exclusion of the leakage-prone feature, enabling a more reliable evaluation of model capabilities.

Tuning and Evaluating with New Data (Second Results)

The second modelling attempt applied a 0.90 association threshold to exclude the historical aggregate and any other variables that might cause leakage. Notably, the energy-related variables exhibited very low association with the target variable, failing to even meet a 0.01

threshold that would justify their exclusion under the same criteria. In contrast, the spatial coordinates demonstrated moderate association: the X coordinate was moderately positively correlated with the target, while the Y coordinate was moderately negatively correlated. The inclusion of these spatial features contributed noticeably to the models' predictive performance.

Following the removal of the leakage-prone variables, the performance of all models decreased substantially compared to the first attempt (see Table 3.5). This decline reflects the elimination of artificially inflated scores and offers a more realistic assessment of each model's true predictive capacity. Logistic regression once again underperformed across all oversampling techniques, returning low test accuracies and unavailable CV weighted F1-scores in most cases. This continued weakness reinforces the earlier observation that the model struggles to capture the non-linear relationships present in the revised dataset.

By contrast, Random Forest and XGBoost delivered substantially more stable and interpretable results than in the leakage-affected run. While their test accuracies ranged between 85% and 89.75% depending on the oversampling method, the cross-validation weighted F1-scores—when available—were closely aligned with the test results, indicating improved generalisability. The best-performing configuration for Random Forest was with BorderlineSMOTE (88.48% CV weighted F1-score), whereas XGBoost achieved its strongest result without oversampling (89.05% CV weighted F1-score).

Table 3.5: Model performance excluding historical aggregate

Oversampler / Model	Logistic Regression	Random Forest	XGBoost
SMOTE	18.60% / NaN	88.55% / NaN	88.90% / NaN
SMOTEN	38.70% / NaN	87.50% / NaN	85.10% / NaN
BorderlineSMOTE	13.75% / 20.45%	**89.65% / 88.48%**	89.30% / 88.29%
KMeansSMOTE	–	–	–
ADASYN	19.30% / NaN	88.40% / NaN	88.80% / NaN
None	23.25% / 31.02%	88.20% / 88.36%	**89.75% / 89.05%**

Note: The results are displayed as Test Accuracy / CV Weighted F1-score. "nan" indicates unavailable CV score. "– (error)" indicates sampler failed to run.

This study also used SHAP (SHapley Additive exPlanations) to gain detailed insights into feature contributions at both global and local levels(Lundberg & Lee, 2017). . SHAP provides a consistent and theoretically grounded method to explain individual predictions by attributing them to input features, helping to uncover the underlying relationships learned by the model on a deeper level complementing the analysis of F1-scores. SHAP (SHapley Additive exPlanations) is a unified framework for interpreting machine learning models by quantifying the contribution of each feature to the prediction output. Based on concepts from cooperative game theory, SHAP values represent the average marginal contribution of a feature across all possible combinations of features (Lundberg & Lee, 2017). This approach allows for consistent, locally accurate, and globally interpretable explanations of complex models. In the context of this predictive model, SHAP analysis revealed that although energy-related features initially showed low correlation wi-th the target variable, they had substantial impact on the model's decisions. These features, together with spatial data, consistently exhibited the highest average SHAP values, indicating their strong influence on the prediction out-

comes. Thus, SHAP not only helped validate the importance of these new data sources but also enhanced the interpretability of the model, guiding further analysis and potential policy implications.

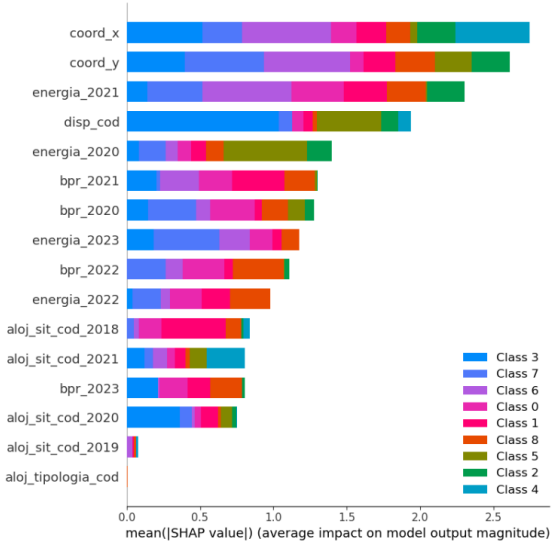


Figure 3.8: SHAP analysis for XGBoost

The SHAP summary plot (Figure 3.8) highlights the relative importance of the input features across all predicted classes. The two spatial coordinates (coord_x and coord_y) emerge as the most influential variables, indicating that locational information is a primary determinant of occupancy prediction. This suggests that spatial patterns, possibly linked to urban density, infrastructure availability, or socio-economic clustering, strongly shape the model's classification outcomes. Among the non-spatial features, energy consumption variables (energia_2020–energia_2023) consistently rank highly, reinforcing the hypothesis that energy usage acts as a reliable proxy for occupancy status. The prominence of energia_2021, in particular, demonstrates its strong contribution across multiple classes. This aligns with the expectation that households with distinct energy usage patterns can be more accurately classified in terms of their occupancy.

Administrative and categorical variables, such as building situation codes (aloj_sit_cod) and typology (aloj_tipologia_cod), also contributed meaningfully, though their average impact was lower than that of spatial and energy-related features. Their influence, however, is non-negligible, particularly in distinguishing between specific classes where locational and energy indicators alone may not suffice.

The balanced presence of SHAP contributions across classes (as reflected by the color distribution within each bar) further indicates that these features are not only globally important but also play differentiated roles in predicting specific occupancy categories. This highlights the ability of the model to leverage heterogeneous data sources—spatial, energetic, and administrative—in complementary ways.

Overall, the SHAP analysis validates the importance of integrating both spatial and energy dimensions in the modelling framework. It further demonstrates that while correlations at the descriptive level may appear weak, the predictive model is able to extract non-linear and multi-class relevant patterns from these features, thereby improving classification performance and interpretability.

Additionally, the absence of extreme accuracy inflation and the alignment of test and cross-validation scores indicate that the removal of the leakage-prone feature successfully mitigated over-fitting in the tree-based models. Although logistic regression remained unsuitable for the task, both Random Forest and XGBoost demonstrated the ability to model the underlying patterns in the dataset effectively. Ultimately, the best-performing configurations were Random Forest with BorderlineSMOTE and XGBoost without oversampling, with XGBoost achieving a marginally higher CV Weighted F1-score, indicating better generalization in cross-validation.

3.5 Model Results and Discussion

3.5.1 Evaluation Metrics

To assess model performance, two primary metrics were used: *test accuracy* and the *cross-validated weighted F1-score*. Test accuracy measures the overall proportion of correctly classified instances on unseen data, providing a straightforward indication of predictive effectiveness. However, accuracy alone can be misleading in imbalanced datasets where the majority class dominates. The weighted F1-score addresses this limitation by accounting for both precision and recall across all classes, weighted by their support. This metric balances false positives and false negatives and is particularly suitable when class distributions are skewed, as in this case. Together, these metrics provide a comprehensive evaluation framework, capturing both general predictive power and balanced class-wise performance.

3.5.2 Implementation

Using the previously trained XGBoost model, one-step forecasts were generated for the 2026 dwelling availability (aloj_sit_cod_2026p). This model leverages historical utility consumption, tax records, structural attributes, and spatial information to estimate which dwellings are likely to be viable for inclusion in upcoming survey operations. Although the true 2026 occupancy labels are not yet available, the model's prior evaluation demonstrated a **test accuracy of 89.75% and a cross-validated weighted F1-score of 89.05%**, indicating strong generalization and reliability in predicting occupancy status.

The predictions cover all dwellings in the study area, each linked with its building and unit identifiers (edif_cod, aloj_cod) and spatial coordinates (lat, long) recorded using ETRS89, the official format used by INE. ETRS89 is a geodetic coordinate system for Europe that provides a consistent reference frame for mapping and surveying by aligning with the stable part of the Eurasian tectonic plate. It is a metric system, meaning all coordinates are expressed in meters, which allows for precise distance and area measurements. Preliminary examination of the predicted data reveals distinct spatial clusters, where dwellings with higher predicted availability are concentrated in specific neighbourhoods, while lower predicted availability is observed in others. These patterns likely reflect underlying neighbourhood-level trends captured by historical utility, tax, and structural features.

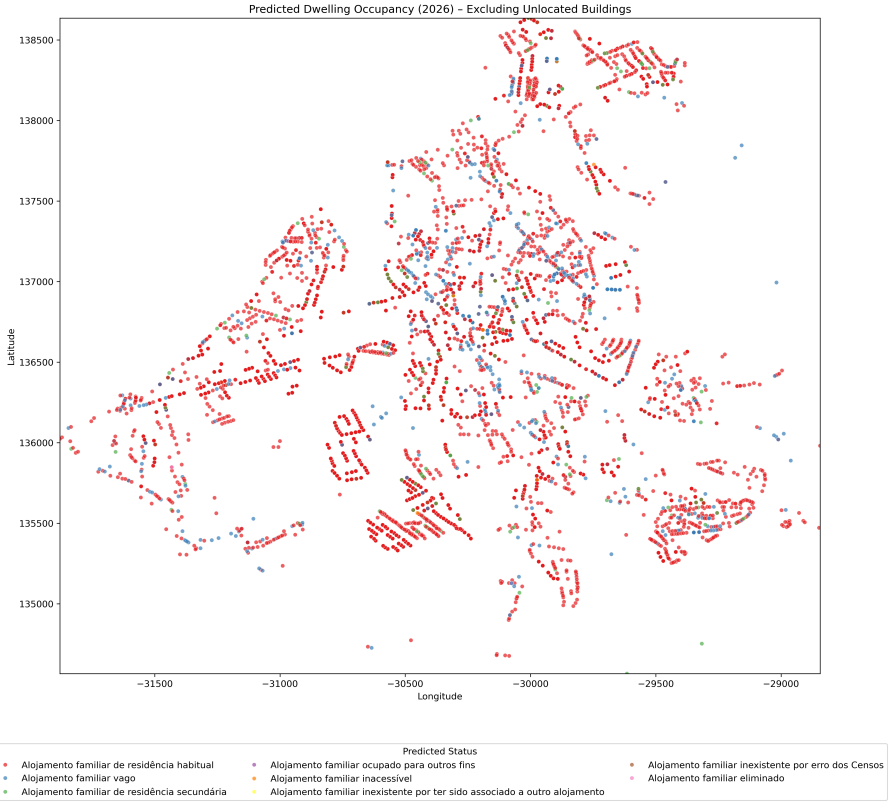


Figure 3.9: Predicted Dwelling Occupancy by Coordinates (2026) – Excluding Unlocated Buildings

Note: Scatterplot shows predicted occupancy statuses for all dwellings with available spatial coordinates. Color coding indicates the predicted status of each dwelling. Buildings with missing spatial data ("unlocated") are excluded to prevent scale distortion.

A domain-specific validation was performed using buildings with missing location data. In the original dataset, approximately 202 buildings lacked precise spatial coordinates and their locations were imputed with the general municipal coordinates according to INE's database conventions. In these cases, the respective housing statuses were recorded as 'unlocated.' The predictions generated by the XGBoost model respected this pattern, classifying these buildings consistently with the original data. Given that the spatial features used are the most recently available and the model exhibits high accuracy and cross-validated performance, this alignment provides additional confidence in the relative accuracy of the predicted occupancy statuses.

The model predictions are driven by historical patterns in the data rather than dwelling-specific identifiers, which were excluded to prevent possible leakage. Key predictors include energy consumption metrics, structural attributes of the building, and spatial variables. The robustness of the XGBoost model is evident in its insensitivity to the inclusion or exclusion of unit-specific identifiers (edif_cod and aloj_cod), confirming that predictions are based on meaningful patterns rather than memorization of individual dwellings. The predictions generated provide actionable information for upcoming survey operations. By focusing survey efforts on dwellings with higher predicted availability, survey planners can reduce unnecessary field visits, optimize resource allocation, and improve data collection efficiency. In addition,

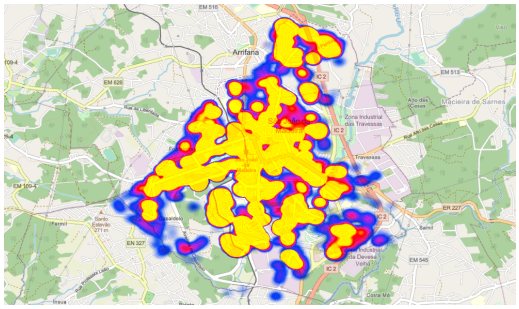
the approach is scalable and can be applied to future years with updated historical data to continuously improve predictive accuracy and operational planning.

In summary, the one-step forecasts for 2026 demonstrate the practical utility of machine learning in survey planning. The combination of XGBoost's predictive power, historical data, and spatial variables enables informed decision-making, ensuring that survey operations are efficient and data-driven. Domain-specific validation, such as the treatment of unlocated buildings, further supports the reliability of the predictions. These results provide a foundation for further analysis, including visualization of spatial clusters and integration with GIS-based operational planning. In addition to scatter plots generated in Python, the predicted 2026 dwelling occupancy statuses were imported into ArcGIS to create high-resolution spatial heat maps and cluster analyses. These maps visualize areas of higher and lower predicted dwelling availability across the study region, highlighting spatial trends and clusters that may not be immediately evident from point-based scatter plots. By overlaying predictions with municipality boundaries and other relevant geospatial layers, these heat maps provide actionable insights for targeted survey operations.

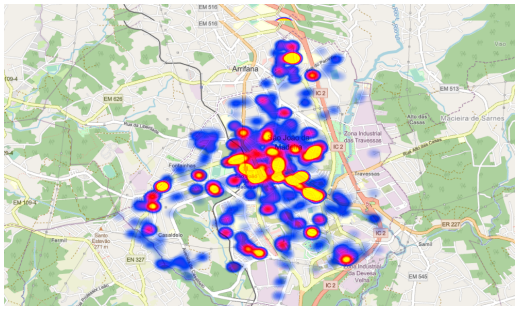
3.5.3 Visualization of Household Occupancy using GIS

The integration of GIS-based data provides the opportunity to capture locational dependencies, as spatial clustering and neighbourhood effects often influence occupancy outcomes. The role of volunteered and spatial data as key resources in modern geographic analysis has been widely discussed (Goodchild, 2007b). These results align with previous work showing that occupancy can be inferred from behavioral or proxy data, particularly through energy use and spatial indicators (H. Dong & Lam, 2012).

To visualize spatial patterns of household occupancy, density-based heatmaps were generated using ArcGIS Pro for each household status. For clarity, the main analysis focuses on two representative statuses: *Primary* (occupied) and *Vacant* (unoccupied), which represent the extremes of occupancy outcomes. Heatmaps for other statuses are provided in Appendix B.



(a) Heatmap of households with status *Primary*.



(b) Heatmap of households with status *Vacant*.

Figure 3.10: Spatial distribution of occupied and vacant households visualized using ArcGIS Pro. Heatmaps highlight clustering patterns, with *Primary* households concentrated in central urban areas and *Vacant* households more dispersed in peripheral regions.

The heatmap for *Primary* households (Figure 3.10a) shows significant clustering in central urban regions, reflecting areas with high residential density and likely access to infrastructure and services (Anselin, 1995; H. Dong & Lam, 2012). In contrast, the *Vacant* house-

hold heatmap (Figure 3.10b) demonstrates more dispersed patterns, with localized clusters in peripheral or transitional zones. This dichotomy supports previous findings that vacancy patterns often indicate urban decay or underutilization of housing (Hedman et al., 2016; Holloway, 2020).

By placing the occupied and vacant heatmaps side by side, it is possible to directly compare the contrasting spatial patterns. The distribution of *Primary* households suggests a strong neighborhood effect, whereas *Vacant* households appear concentrated in less desirable areas, highlighting locational dependencies in occupancy. ArcGIS Pro facilitated this analysis by allowing adjustable heatmap radii and transparency, enabling clear visualization of spatial intensity while retaining geographic context.

Additional heatmaps for other household statuses, including secondary and temporary residences, are presented in Appendix B. These supporting visualizations confirm the general patterns observed in the primary comparison: clustering is prevalent among active households, while non-primary or vacant households are more dispersed. Together, the predictive modelling results and GIS-based visualizations provide a comprehensive understanding of occupancy dynamics across the study area, bridging quantitative predictions with spatial insights for urban planning and policy applications.

These results align with previous work showing that occupancy can be inferred from behavioural or proxy data, particularly through energy use and spatial indicators (B. Dong & Lam, 2012).

3.5.4 Final Model and Comments

After thorough evaluation of multiple machine learning algorithms and oversampling methods, **XGBoost** was selected as the final predictive model for this study. It consistently achieved the highest performance across experiments, with a **test accuracy of 89.75% and cross-validated weighted F1-score of 89.05%**, demonstrating strong generalization and reliability for predicting dwelling availability.

Preliminary tests indicated that including building and unit identifiers (edif_cod, aloj_cod) could skew results in some models, particularly **Logistic Regression**, which tended to memorize apartment-specific patterns rather than learn meaningful predictive relationships. In contrast, XGBoost was robust to the inclusion or exclusion of these identifiers, confirming that its predictions rely on genuine patterns in historical utility, tax, and spatial data.

Hyperparameter Grid and Justification

The final XGBoost pipeline incorporated hyperparameter tuning via *RandomizedSearchCV* within a cross-validated framework. The hyperparameters explored were carefully chosen based on the characteristics of the dataset and the study objectives:

- **n_estimators = [100]**: Defines the number of trees in the ensemble. Each tree sequentially corrects errors from prior trees. A value of 100 was selected to balance computational efficiency with sufficient capacity to capture meaningful patterns in historical occupancy data. Preliminary experiments indicated that increasing the number of trees beyond 100 did not significantly improve performance.

- **max_depth = [3, 5]**: Controls the maximum depth of individual trees. Depth 3 limits over-fitting and promotes generalization across dwellings and years, while depth 5

allows the model to capture slightly more complex interactions, such as the interplay between building-level variables and unit-specific historical occupancy, without creating overly complex trees that memorize the data.

- **reg_alpha| = [0, 0.01, 0.1]**: L1 regularization on leaf weights, encouraging sparsity and reducing the influence of less informative or noisy features. Small non-zero values help the model focus on historically predictive patterns, while 0 allows verification of whether regularization is necessary.

- **reg_lambda = [1, 2]**: L2 regularization on leaf weights, smoothing the model and preventing extreme weight variations. This stabilizes learning across heterogeneous features and mitigates over-fitting to idiosyncratic patterns in historical occupancy.

This configuration enabled XGBoost to effectively leverage historical utility, tax, and spatial data, while mitigating risks of over-fitting or leakage. XGBoost will be employed to estimate the one-step-ahead dwelling availability status (aloj_sit_cod_2026p). By leveraging past data, it can accurately predict which dwellings are likely to be viable for inclusion in upcoming surveys, helping to reduce resource expenditure and improve the efficiency and accuracy of data collection. Having validated the model's robustness and optimal configuration on historical occupancy data, the following section applies XGBoost to predict one-step-ahead dwelling availability for the upcoming year, enabling practical assessment of potential participation in future survey operations.

Chapter 4

Conclusion

4.1 Final Remarks

This study explored the prediction of dwelling occupancy by integrating multiple sources of data, including administrative records, energy consumption, and geospatial information. By combining these datasets with advanced predictive modeling techniques, including XGBoost and Random Forest classifiers, the study demonstrated the potential for high-quality, multi-source data to produce robust and interpretable predictions of occupancy outcomes.

The inclusion of electricity consumption data played a pivotal role in model performance, providing strong predictive signals that enhanced accuracy across most occupancy categories. In particular, rare occupancy types remained challenging to predict, but the models performed well for the majority of units, demonstrating the value of energy usage as a behavioral proxy. The analysis of feature importance further highlighted which variables were most influential in predicting occupancy, offering practical insights for policymakers and urban planners.

Spatial analysis using GIS complemented the predictive modeling, revealing locational patterns and clustering effects that informed the interpretation of model results. These visualizations highlighted areas with higher probabilities of specific occupancy outcomes and demonstrated neighborhood-level dependencies consistent with prior literature. By visualizing the spatial distribution of predicted occupancy, this study enabled a more nuanced understanding of local patterns and supported the validation of model outputs.

The methodological approach aligns with principles discussed in the literature review, including frameworks from INSEE, Eurostat, and the U.S. Census Bureau's MAF-ARF. Each of these initiatives emphasizes the importance of integrating multi-source data, optimizing sample frames, and accounting for spatial structure—principles that were operationalized in this study through the combination of administrative, energy, and geospatial datasets.

Overall, this study confirms that data-driven predictive modeling, when combined with spatial visualization techniques, can generate actionable insights for urban planning, resource allocation, and occupancy management. The findings underscore the importance of multi-source data integration and suggest that predictive modeling can be a valuable complement to traditional survey-based approaches.

4.2 Limitations and Future Work

Despite its contributions, the study faced several limitations. Computational constraints restricted the exploration of more complex or larger-scale models, such as deep learning architectures, which may have improved performance on rare occupancy categories. Temporal discrepancies and incomplete coverage in some administrative and energy datasets could have introduced bias or affected model generalizability.

A point worth noting was the absence of water consumption data. While electricity usage provided significant predictive power, the lack of water consumption data prevented its inclusion in the models. Given the strong influence of electricity data on model outcomes, incorporating water usage—either fully or partially—could have further enhanced model accuracy and provided additional behavioral insights.

Spatial visualizations were limited to static snapshots of occupancy probabilities. Future research could integrate high-resolution or real-time geospatial data, allowing dynamic updates and supporting more responsive urban management. Additionally, the exploration of ensemble approaches or hybrid models combining multiple algorithms may improve predictions for rare categories and further reduce error rates.

Future directions for research include expanding the study to multiple cities or regions to assess model generalizability, integrating additional administrative sources for temporal and demographic coverage, and embedding predictive models within GIS-based decision support systems for practical planning applications. Comparative studies across different jurisdictions would also provide insight into the scalability and transferability of the proposed approach. Finally, the inclusion of additional behavioral proxies—such as water consumption or mobility data—could further improve model accuracy and interpretability.

In conclusion, this thesis demonstrates that the integration of administrative, energy, and geospatial data, coupled with advanced predictive modeling and spatial visualization, offers a powerful framework for understanding and predicting dwelling occupancy. The methodological insights and practical outcomes presented here provide a foundation for future research, policy applications, and the development of data-driven strategies for urban management.

Bibliography

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical Analysis*, *27*(2), 93–115.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining (pakdd 2009)* (pp. 475–482). Springer.

Cases, B. (2019). *Eu iess regulation and sample accuracy thresholds.*

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Cotton, P., & Dubois, F. (2019). *Mixed-mode surveys in national statistical offices.*

Deville, J.-C., & Lavallée, P. (2006). *Calibration and weighting in survey sampling.* Springer.

Deville, J.-C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, *91*(4), 893–912.

DeWaard, B., et al. (2022). *Maf-arf methodology for address-based sampling* (Technical Report). U.S. Census Bureau. Retrieved from https://www.census.gov/...

Dong, B., & Lam, K. P. (2012). A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and prediction. *Building and Environment*, *49*, 1–13.

Dong, H., & Lam, J. (2012). Inferring occupancy from behavioral and energy use data in buildings. *Energy and Buildings*, *54*, 100–108.

Favre, J., Martinoz, P., Merly, G., & Alpa, S. (2016). Spatially optimized sampling strategies in insee surveys. *Survey Methodology Journal*, *42*(2), 125–146.

Favre-Martinoz, J. (2015). Evolution of household survey sampling frames at insee. *Revue de Statistique*.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets.* Cham: Springer.

Foley, M. (2020). *Using administrative records for survey sampling: Lessons from the maf-arf* (Technical Report). U.S. Census Bureau. Retrieved from https://www.census.gov/...

Goodchild, M. F. (2007a). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Goodchild, M. F. (2007b). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Grafström, A., & Tillé, Y. (2013). Doubly balanced spatial sampling. *Journal of Official Statistics*.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in intelligent computing (icic 2005)* (pp. 878–887). Springer.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328).

Hedman, R., et al. (2016). Mapping urban vacancy using spatial analytics. *Urban Studies*, *53*(12), 2541–2558.

Holloway, D. (2020). Patterns and causes of housing vacancy in urban contexts. *Housing Policy Debate*, *30*(4), 623–644.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). New York: Wiley.

Keith, M., Finlay, K., & Genadek, K. (2021). *Administrative records and the master address file: Methodological insights* (Technical Report). U.S. Census Bureau. Retrieved from https://www.census.gov/...

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, *197*, 303–317. doi: 10.1016/j.apenergy.2017.04.005

Kreuter, F., & Peng, R. (2020). *Big data and social science: A practical guide*. Chapman and Hall/CRC.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis*, *45*(3), 324–344. doi: 10.1111/gean.12012

Langford, M., & Unwin, D. J. (1994). Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal*, *31*(1), 21–26. doi: 10.3138/W836-5K13-1432-4480

Lohr, S. L. (2019). *Sampling: Design and analysis* (2nd ed.). Chapman and Hall/CRC.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems (neurips)* (Vol. 30).

Merly, F., Alpa, E., & Sillard, P. (2019). Administrative data integration for improved sampling frames. *Journal of Official Statistics*, *35*(2), 301–325. doi: 10.2478/jos-2019-0015

Merly-Alpa, C., & Sillard, P. (2019). *The fidéli system for tax-based sampling at insee*.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation and applications* (2nd ed.). Wiley.

Santos, A. M., Portugal, A., Marcelo, C., Magalhães, J., Cruz, P., Campos, P., … Mina, S. (2024). *The future of sampling frames*. Presentation by Statistics Portugal – DMSI – Department of Methodology and Information Systems at European Conference on Quality in Official Statistics 2024, Estoril - Portugal.

Sullivan, J., & Genadek, K. (2024). *Master address file–administrative records frame: Methodological overview* (Technical Report). U.S. Census Bureau. Retrieved from https://www.census.gov/...

Tillé, Y. (2019). *Sampling and variance decomposition*. Springer.

VLISS Virtual Laboratory. (2020). *Eurostat survey sampling guidelines*. (Accessed via Eurostat documentation)

Wallgren, A., & Wallgren, B. (2014). *Register-based statistics: Administrative data for statistical purposes*. John Wiley & Sons.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.

Appendix A

Data Dictionary

Table A.1: Full Data Dictionary

Column	Data Type	Definition
edif_cod	int64	Building code
aloj_cod	int64	Dwelling code (individual unit)
coord_x	float64	Longitude coordinates
coord_y	float64	Latitude coordinates
aloj_sit_cod_[year]	object	The registered situation/status of a dwelling of a specific year
aloj_tipologia_cod	int64	The typology of the individual dwelling unit
disp_cod	float64	Availability status of the dwelling to participate in surveys
data_ext	int64	Data extraction date
bpr_[year]	object	Indicator on whether the selected dwelling was included or participated in the BPR of a specific year
fr_cod	int64	Parish code
fr_dsg	object	Parish name
tipologia_cod	object	The typology code of the Building
tipologia_dsg	object	The typology description of the building
energia_[year]	float64	Energy consumption values for a specific year for the individual unit

Appendix B

Additional Household Status Heatmaps

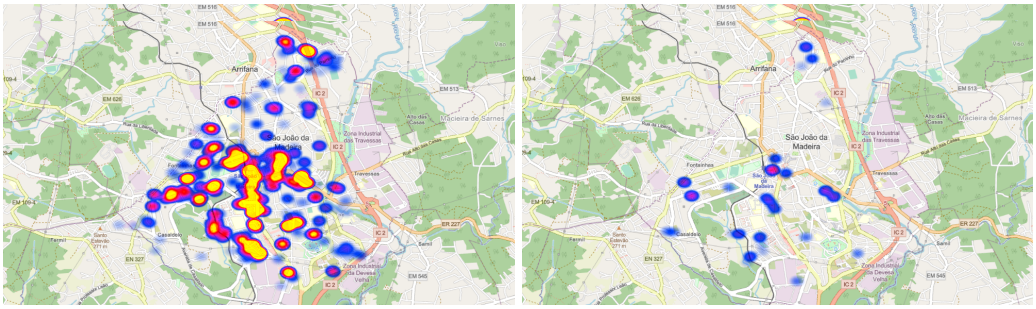


Figure B.1: Additional household status heatmaps generated in ArcGIS Pro. These maps complement the primary and vacant household visualizations in the main text. The first figure representing Households with Status Secondary, and the second figure representing a combination of filters for houselods that are Eliminated, Inaccessable, and Non-existent due to Census error

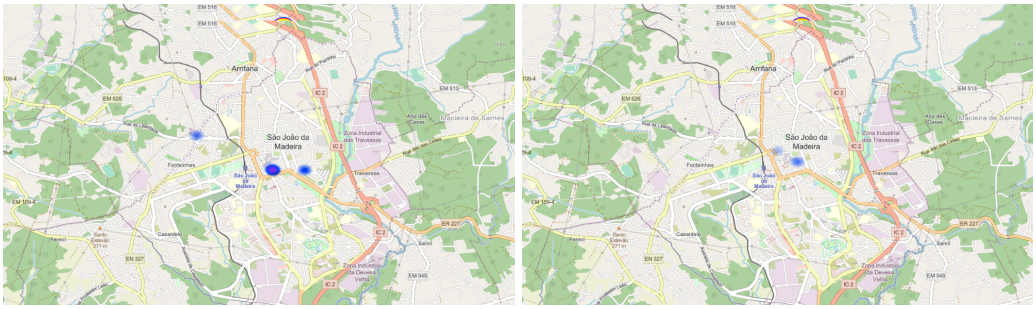


Figure B.2: Further household status heatmaps illustrating spatial distribution for other statuses.The first being Non-existent family accommodation because it was associated with another accommodation, and the second being Family accommodation occupied for other purposes