Street Art & Urban Creativity | Vol. 12, No. 5, 2026 | ISSN 2183-9956
Street Art & Urban Creativity. Scientific Journal
https://doi.org/10.62161/sauc.v12.6400

MACHINE LEARNING FOR URBAN HEATING VECTOR ESTIMATION TO SUPPORT CITY DECARBONIZATION

Relevant features, transferability, and the role of spatial patterns

ELENA USOBIAGA FERRER (ELENA.USOBIAGA@TECNALIA.COM)[1]

[1]Tecnalia, Spain.

KEYWORDS

Artificial Intelligence (AI)
Heating vector
Decarbonization
Machine learning
Spatial Analysis
Climate Zone
Transferability

ABSTRACT

To support urban decarbonization, XGBoost classifiers were developed to predict residential heating energy vectors in Spain using open cadastral, census, and Energy Performance Certificate data. Training across climate zones C1, D1, and D3 evaluated the impact of non-spatial, coordinate-derived, and neighbour-corpus features. Spatially aware models achieved up to 73.6% accuracy, with neighbour energy probabilities providing the most significant gain and reducing misclassification clustering. However, transfer to cities without ground truth revealed instability (54.6% building-level agreement), underscoring the necessity of neighbourhood-level ground-truth collection to ensure reliable city-scale predictions.

RECEIVED: 12 / 06 / 2026
ACCEPTED: 01 / 09 / 2026

1. Introduction

The decarbonization of cities constitutes a fundamental challenge in the fight against climate change. Currently, urban areas are responsible for approximately 70% of global CO2 emissions related to energy consumption (Luqman et al., 2023). This situation is expected to worsen due to demographic trends, as it is estimated that by 2050, more than 70% of the world's population will reside in urban areas («Cities - United Nations Sustainable Development Action 2015», s. f.), significantly increasing the demand for energy, water, and resources. In this scenario, the European Union's Mission for Climate-Neutral and Smart Cities (*EU Mission*, 2026) underscores the decisive role of urban centers in achieving climate goals.

Within the urban environment, heating and cooling account for approximately half of the final energy consumption in Europe, where fossil fuels remain the primary energy source for a large portion of the building stock. Consequently, the decarbonization of the residential sector has become a strategic priority, driven primarily by the Energy Efficiency Directive - EED (Directive 2023/1791 of the European Parliament and of the Council on Energy Efficiency and Amending Regulation 2023/955, 2023) and the Energy Performance of Buildings Directive - EPBD (Directive 2024/1275 of the European Parliament and of the Council on the Energy Performance of Buildings, 2024). This regulatory framework promotes integrated energy planning, deep building renovation, and the deployment of low-emission technologies.

In Spain, the latest transposition of the regional decarbonization aspirations is the Zero Emission Architecture framework (ARCE 2050)—developed for the draft National Building Renovation Plan (Ministerio de Transportes, Movilidad y Agenda Urbana. Gobierno de España., 2025) presented in December 2025—, which builds upon the former Long-Term Strategy for Building Renovation (Secretaría de estado de transportes, movilidad y agenda urbana. Gobierno de España., 2020), establishing bold target reductions in residential primary energy consumption of 25% by 2030 and 33% by 2035 (Ministerio de Transportes, Movilidad y Agenda Urbana. Gobierno de España., 2025).

Among the key operational instruments to materialize these objectives are Heating and Cooling Plans. These plans allow for the analysis of thermal energy production, distribution, and consumption within a territory to identify opportunities for demand reduction, the utilization of renewable energies, and the recovery of waste heat. However, the effectiveness of these plans depends strictly on a precise characterization of building stocks. Such characterization must not be limited to physical aspects (such as age or typology) but must integrate detailed knowledge of the energy vectors used in each dwelling (e.g., natural gas, electricity, biomass, etc.).

The importance of identifying these vectors lies in the fact that they determine the associated emissions and the available alternatives for substitution; renovation strategies will differ drastically depending on whether a dwelling relies on natural gas or is connected to a renewable heat network. In the Spanish context, studies such as SPAHOUSEC III (IDAE, 2026) already demonstrate that, although heating is the primary energy use, there is significant heterogeneity based on location and building typology, reinforcing the need for disaggregated data.

Despite this necessity, one of the primary obstacles to implementing these strategies is the lack of data and the inequality of available information. One of the primary sources for obtaining building information is cadastral records. These records provide data on key building characteristics, such as construction year, number of floors, and building use; however, they do not include information regarding heating systems or the energy carriers employed. To obtain such details, energy efficiency certificates can be consulted, as they contain the necessary data to determine a building's energy rating. Nevertheless, despite being collected for these certificates, this information is not always publicly disseminated. A review of these sources in Spain reveals that only five of the seventeen autonomous communities—the Basque Country, Navarre, Madrid, Castilla-La Mancha, and the Canary Islands—provide open access to data on heating systems. The remaining communities generally publish only the final certification results, occasionally including additional data, but omit specific system details. The following figure (Figure 1) illustrates the location of the regions that provide this data. Given the distribution of the country's climatic zones, zones A3 and A4 (characteristic of southern Spain), as well as zones B3 and B4 (found in the south and the Mediterranean arc), are not represented. Furthermore, climatic zones C2, C3, C4, and E1 are underrepresented in the municipalities of these territories, accounting for less than 10% of the housing in these areas.

Figure 1. Location of regions that provide heating vector data in the EPCs

Source: image generated using Microsoft Copilot (Microsoft Corporation, 2026).

A marked disparity in digitalization and transparency is observed across different geographical regions and competent administrations, resulting in 'information gaps' in areas where the fuel types or heating systems remain unknown. This data asymmetry precludes an accurate assessment of decarbonization potential and hinders evidence-based decision-making at urban and regional scales. Furthermore, it renders it impossible to monitor the effectiveness of incentive policies for energy rehabilitation or system replacement.

Within this context, the REGEN Project, funded by the European Commission under the Horizon Europe program, aims to neighbourhood decarbonization by employing digital technologies and sustainability assessments within the framework of urban regeneration. Specifically, it focuses on the decarbonization of the building stock through active and passive renovation measures, addressing precisely the technical information gap required to transform buildings into climate-neutral assets, and framing the development of the present work.

Given these circumstances of policy aspirations, operational instruments, and digital barriers, the REGEN Project has unveiled a clear need to develop tools capable of estimating energy carriers when direct data are unavailable. This study proposes the development and validation of an Artificial Intelligence approach designed to predict the heating system vector used in residential buildings, based on related features available in public records. To ensure the robustness and generalizability of the model, training was conducted using datasets from various climatic zones across Spain (those largely included in the Basque Country, Navarra and Madrid). This way, we enable a relevant benchmark in the Spanish context, in which climate zone classification functions as a discrete design criterion for building conditioning systems, defining boundary environmental conditions that dictate thermal demand limits, envelope transmittance thresholds, and HVAC system capacity sizing (Ministerio de Transportes, Movilidad y Agenda Urbana. Gobierno de España., 2019).

2. State of the art

2.1. Decarbonization of the building stock and the role of energy vector characterization

The building sector represents one of the most critical challenges for the EU's climate neutrality goals, accounting for approximately 40% of final energy consumption and 36% of energy-related greenhouse gas emissions (Commission for the Environment Climate Change and Energy et al., 2022). In response, the regulatory landscape has shifted from aspirational targets to binding obligations through the recast EPBD (Directive 2024/1275 of the European Parliament and of the Council on the Energy Performance of Buildings, 2024) and the EED (Directive 2023/1791 of the European Parliament and of the Council on Energy Efficiency and Amending

Regulation 2023/955, 2023). These frameworks introduce mandatory instruments such as Minimum Energy Performance Standards (MEPS), National Building Renovation Plans (NBRP), and the pursuit of Zero-Emission Buildings (ZEB) (Directive 2024/1275 of the European Parliament and of the Council on the Energy Performance of Buildings, 2024).

Despite this heightened ambition, the annual renovation rate in the EU has remained stagnant at approximately 1% for a decade. Literature suggests that this "implementation gap" is not due to a lack of policy ambition, but rather to systemic structural misalignments and territorial capacity constraints (Economidou et al., 2020; OECD, 2020). A primary obstacle is the fragmentation of building-stock data, where information resides in disconnected cadastral records, certification databases, and permitting systems, often lacking a universal identifiers (Gómez-Gil et al., 2023). This fragmentation results in a significant "performance gap," where policy decisions are based on assumptions rather than operational consumption data, hindering the ability of local and regional authorities to identify and prioritize the worst-performing buildings (BPIE (Buildings Performance Institute Europe), 2025; Directive 2024/1275 of the European Parliament and of the Council on the Energy Performance of Buildings, 2024).

To overcome the lack of building-level data, urban energy planning has traditionally relied on deterministic models and aggregate simulations. However, there has been a recent paradigm shift toward bottom-up methodologies that seek higher resolution while managing computational costs. A prominent approach is Archetype-Based Energy Modeling (ABEM), which groups buildings into categories based on shared physical and operational characteristics (Cerezo Davila, 2017; Reinhart & Cerezo Davila, 2016). Archetypes can be engineering-based (relying on detailed simulations of representative buildings), regulatory-based (based on historical building codes), or hybrid (Shen & Wang, 2024). For instance, the ERESEE strategy in Spain and the EU-funded TABULA and EPISCOPE projects have utilized ABEM to estimate residential energy demand across diverse European contexts by using construction periods and typologies as proxies for energy performance (Eguiarte et al., 2020, 2022).

While ABEM provides a scalable framework, traditional archetypes often rely on expert judgment and simplified assumptions, which may overlook behavioral variations and micro-scale disparities in energy demand. Recent advancements have integrated Machine Learning (ML) to automate archetype segmentation, allowing for data-driven clustering based on real-world building stock data and occupancy patterns (Ali et al., 2019).

However, a critical gap remains in the characterization of the "energy vector"—the specific energy source (e.g., natural gas, electricity, biomass, or district heating) used for heating and domestic hot water (DHW). While physical characteristics (age, area, typology) are frequently included in building stock models, the energy vector is often treated as a secondary variable. Yet, the vector is the primary determinant of the carbon footprint and the technological pathway for decarbonization; a building's transition strategy differs fundamentally depending on whether it relies on an individual gas boiler or is connected to a renewable district heating network.

Accurately identifying the energy vector is essential for the development of Heating and Cooling Plans (HCPs), as it allows planners to detect fossil fuel dependencies and evaluate the potential for electrification or the integration of local renewable resources. In Spain, efforts such as the SPAHOUSEC III study (IDAE, 2026) highlight the importance of these vectors in understanding consumption patterns and socio-economic disparities. Currently, the most robust methodologies for characterizing these vectors employ hybrid approaches, combining top-down official statistics with bottom-up data from Energy Performance Certificates (EPCs) and measured consumption data (Badec & Salam, 2024; Li et al., 2020).

Integrating the prediction of the energy vector through AI models—moving beyond static archetypes to dynamic, predictive characterizations—represents a necessary step toward achieving the goals of the EPBD and EED. By filling this data gap, it becomes possible to design evidence-based retrofit strategies that prioritize not only passive efficiency measures but also the systemic change of energy vectors, accelerating the transition toward climate-neutral cities.

2.2. Machine Learning in Building Heating Vector Characterization

The application of machine learning (ML) to the characterisation of building energy systems has grown significantly over the past decade. Chen et al. (Chen et al., 2025), in a systematic review of the field, identify three dominant trends: the transition from physics-based (deterministic physical assumptions to heating and cooling demand derived from climatic and material conditions of buildings) to data-driven models (in which complex relationships among driving factors for energy vectors are handled by ML models), with the consolidation of ensemble algorithms such as XGBoost and LightGBM as the architectural standard for tabular building energy data, and the increasing integration of spatial and contextual descriptors that go beyond the physical characteristics of the individual building.

In the Spanish context, Beltrán-Velamazán et al. (Beltrán-Velamazán et al., 2025) developed a national-scale building energy characterisation model (nUBEM – Urban Building Energy Model) using XGBoost over an integrated dataset of Energy Performance Certificates, cadastral records and Digital Terrain Model data, applied to approximately 1.2 million certificates from Spain's national registry. The methodological overlap of the latter with the present work is substantial: both use EPCs as ground truth, adopt XGBoost as the predictive architecture. However, the key distinction lies in the output variable: nUBEM predicts continuous energy consumption and CO2 intensities, whereas the present study addresses multiclass classification of the heating energy vector itself, which we consider critical for a finer characterization of consumption itself, but acknowledge that the information to achieve reasonable modeling is scarcer.

In the broader literature, Sisman et al. (Sisman et al., 2025) address an analogous problem in the Turkish context, embedding EPCs within Turkey's National Spatial Data Infrastructure and training Random Forest, Gradient Boosting and XGBoost models for building energy performance prediction. Their consistent finding that XGBoost outperforms Random Forest and Gradient Boosting when features are heterogeneous provides independent validation for the modelling architecture adopted here.

Jiang et al. (Jiang et al., 2024) extend the problem to neighbourhood scale, combining Urban Building Energy Modelling (UBEM) simulations with ensemble algorithms to predict heating and cooling loads at block level, and demonstrate that predictive accuracy increases substantially when spatial context and urban morphology variables supplement individual building attributes (a conclusion directly applicable to the Spanish case).

Other approaches, such as data-driven or synthetic archetype building using ML, have also been applied in energy modelling, increasingly aware of socioeconomic, energy market, and occupancy conditions (Harun et al., 2026).

Despite this active line of research, a review of the literature reveals a specific gap worth flagging for discussion: prediction of the heating energy vector itself, as a categorical outcome rather than a continuous consumption metric, remains comparatively unexplored. Most cited works predict consumption, performance scores or emissions; classifying the underlying technology from indirect signals is a related but distinct problem.

Finally, the marked class imbalance typical of building stock datasets is a recurring methodological challenge across this literature. The Synthetic Minority Over-Sampling Technique (SMOTE), introduced by Chawla et al. (Chawla et al., 2002), addresses this problem by generating synthetic minority-class instances through interpolation, and constitutes the foundation of the class rebalancing strategy adopted in the present study, and a constraint noted in the reviewed works.

2.3. Hypothesis: Spatial patterns in Energy-Related characterization

In the absence of general ground truth data for energy characterization, in cases such as the Spanish, assesment of the value of different data sources and, potentially, their indirect modeling, supports the advances in the construction of this policy-relevant information.

Within this particular idea of indirect modeling, a rather unexplored question in the literature is the actual characterization of the available energy infrastructure as a factor in modeling, being commonly a propietary and sensible information, not open for general use.

A shy trend for the indirect consideration of infrastructural constraints is the in-model use of spatial patterns, treating geographic proximity not as a generic smoothing variable (a rather extended practice), but as a proxy for the physical topology of energy supply infrastructure. Fang et al. (Fang et al., 2025) propose a data-driven urban area delineation framework using DBSCAN to detect coherent urban nuclei separated by sparse, disconnected areas, underpinning the spatial clustering approach adopted, while Long et al. (Long et al., 2025) use DBSCAN clustering with a Graph Attention Network to significantly improve air conditioning load prediction.

This last trend is particularly relevant to the present study, and its main hypothetical pillar. Essentially, adding spatial features in energy characterization model follows either a philosophy of accounting for the marked uneven spatial patterns that all urban sets exhibit, or introducing a notion of general urban form which acts as a proxy for the representation of physical logics followed by the energy infrastructure. For instance, transport infrastructure of Natural Gas will most probably follow urban density for efficiency and service maximization reasons and, in the same way, will have strong feasibility constraints in sparsely populated urban fabric.

Our working hypothesis is that, in energy related predictive models, adding spatial features that capture urban structure which can constrain energy infrastructure adds significant gains in predictive power and generalization.

3. Materials and methods

This section describes the end-to-end workflow used to predict residential heating energy vectors from open geospatial and certificate data in Spain (see Figure 2). We first assemble and harmonise multi-source inputs—Energy Performance Certificates, cadastral records, census indicators, and terrain descriptors—and filter the training corpora to three climate zones (C1, D1, and D3) to support a factorial comparison of spatial information. Building on exploratory analysis of class imbalance and spatial structure, we engineer three feature tiers (NOTSPATIAL, BASICSPATIAL, and TAGGEDSPATIAL), train and tune an XGBoost classifier with SMOTE applied inside cross-validation folds, and evaluate nine zone-specific models using stratified out-of-fold metrics together with spatial diagnostics of misclassification (Global and Local Moran's I). Finally, we outline the label-free transfer protocol applied to Laredo (Cantabria), where only corpus-free spatial tiers can be deployed, using marginal technology mix and tier agreement as stability indicators in the absence of local ground truth.

Figure 2. General approach

Source: Authors

3.1. Data collection and preprocessing

For our study intent, multiple general classes of data have been considered, regarding the potential availability of heating vector data in a rather limited number of regions, and the need for indirect modelling in the remaining territory.

First, EPC sources form the closest source of ground truth. Even with the inherent issues of a crowdsourced data source as such (technicians fill and upload the data inputs for each certificate), when EPC databases are sufficiently informed and open, fine-grain information about building characteristics and calculation methods constitute the greatest level of detail.

Beyond EPC databases, the geolocation and identification of individual buildings allow to connect with cadastral building databases, and other environmental features of the climatic, demographic, and energy-market type, either built at the census tract or the municipal level, which can potentially be built for regions without EPC data

For the construction of data features representing these realms, this study combines different public georeferenced data sources. We use three open EPC datasets in which a connection can be made between energy vector and geolocated individual building or property data, which we thoroughly clean and homogenize for training purposes.

The relation with property data is leveraged to enrich our datasets with cadastral information, out of which land use, area, year of construction, number of dwellings and floors, is used. Having informed coordinates, lets us also feed our datasets with environmental (area-related) information in varying degrees of granularity.

Sociodemographic information such as resident population, share of secondary dwellings, or household income, are built for census tract boundaries, in the case of Spain representing around 1000-2000 inhabitants in official statistical operations.

At a broader level, the municipal boundary, we also can access information about the predominant heating systems and degree of centralization present in dwellings, represented as percentages, and available for the whole country in the census decennial operations.

Finally, other contextual information such as the climatic zone (at the municipality level) and the median height above sea level (using official DTMs) has been used for characterization and, as detailed below, for the final filtering of training datasets.

All these sources were preprocessed using Geographic Information Systems (GIS) and assigned to the lowest available level, generating a point layer with all the associated information, which served as the basis for the subsequent analyses. Table 1 shows an overview of data sources along with key attributes and granularity.

Table 1. Overview of data sources used in the study

Source	Category	Type of Data	Key Attributes	Granularity employed
Energy performance certificates	Dwelling data	Dwelling information	Heating system, type of energy, coordinates	Dwelling
Cadastral data	Building data	Building information	Building use, surface area, year of construction, number of dwellings, number of underground floors	Building
National Institute of Statistics	Urban contextual data	Population census (2021)	Demographics, population density, secondary dwelling	Census tract
Household Income Atlas	Urban contextual data	Socioeconomic indicators	Household income	Census tract
Digital Terrain Model (DTM)	Urban contextual data	Orography	Average elevation	Census tract
Gas network availability	Municipal contextual data	Access to gas natural network	Percentage of dwellings with natural gas	Municipality
Heating systems centralization	Municipal contextual data	Population census (2011)	Percentage of heating centralised systems	Municipality
Climatic zone	Municipal contextual data	Climate context	Climate zone	Municipality

Municipal boundaries	Urban contextual data	Administrative boundaries	Names, codes, boundaries of municipalities	Municipality

Source: Authors.

3.2. Exploratory analysis and Feature Engineering

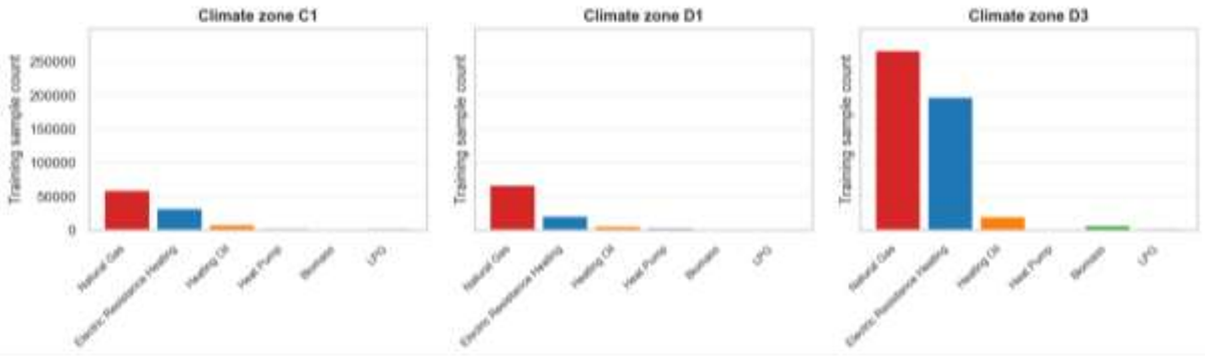
So as to filter out small portions of our target regions with minoritary climatic conditions, the datasets were filtered and merged to represent the three largest climate zones (namely C1, located mainly in coastal Basque Country; D1, located mainly inland Basque Country and Navarra; and D3, totally located within the Madrid Region). Ideally, this filters out energy systems design which might be tied to climatic bindings. Final datasets were composed of 101,083 rows in region C1, 96,298 in region D1, and 489,228 in D3.

For the following data extraction, raw records were consolidated through a deterministic preprocessing pipeline. The target variable originally contained approximately 35 distinct text values, administrative variants of the same underlying technology across different certification software versions; these were mapped to six master classes (Natural Gas, Electric Resistance Heating, Heating Oil, Heat Pump, LPG and Biomass) to obtain categories with clear commercial meaning for energy planning. Records corresponding to obsolete or non-energetic labels (coal heating, absent from the residential stock since the early 1990s, and "district heating" categories) were removed, together with a small number of records presenting physically inconsistent combinations of declared energy vector and heating equipment. Duplicate certificates for the same building were resolved by retaining the most recent record, and coordinate quality filtering discarded entries without valid geolocation.

A specific analytical decision taken at this stage is the unification of two originally separate labels, to obtain meaningful categories. Records administratively coded as "Electricidad" were consolidated into a single Electric Resistance Heating class, as both described the same physical phenomenon (direct resistive conversion of electrical energy into heat) and their original separation stemmed from inconsistent nomenclature across certification software versions rather than a genuine technological distinction. Similarly, aerothermal systems, administratively coded as a separate category, were initially merged into Heat Pump, since aerothermal technology is a subcategory of heat pump systems rather than an operationally distinct one.

The target variable holds a very imbalanced state in the final six-class classification in all three climatic zones selected (
Figure 3). Minority classes (Biomass, LPG and Heat Pump) are under-represented by one to two orders of magnitude relative to Natural Gas. A classifier optimising overall accuracy would therefore tend to ignore rare technologies unless explicitly corrected. This distribution carries a direct modelling implication: a classifier predicting Natural Gas unconditionally would already achieve a substantial baseline accuracy without learning any genuine pattern, and the minority classes require explicit rebalancing to avoid being treated as statistical noise during training. As a first methodological decision, we opted to apply SMOTE (Chawla et al., 2002) inside each cross-validation step fold to avoid leakage.

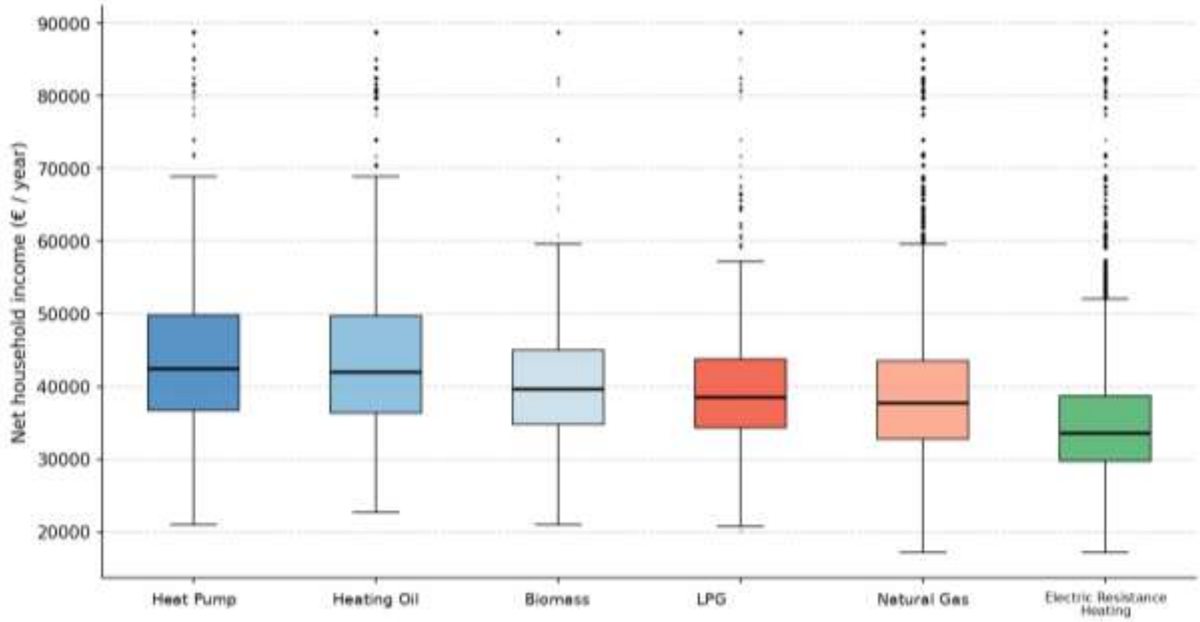
Figure 3. Training Class Distribution by Climate Zone



Source: Authors

Regarding leakage, a dedicated audit identified and removed variables that directly or derivatively encoded the target (equipment fields, DHW twin columns and downstream computed outputs such as energy rating or annual emissions), preventing the model from learning to reproduce the certification form rather than the underlying physical pattern.
A temporal analysis showed that Natural Gas dominates across all post-1960 construction cohorts, peaking among buildings built between 1970 and 1990, coinciding with the period of most rapid expansion of gas distribution networks in the selected regions. A structurally significant shift appears in the most recent cohort, where Heat Pump adoption grows markedly, driven by the increasing cost competitiveness of aerothermal systems and by tightening energy performance requirements; Natural Gas shows a corresponding decline in the same period.
Income data at census section level revealed a clear economic stratification across classes. Heat Pump and Natural Gas buildings cluster in higher income sections, while Electric Resistance Heating and LPG are over represented at lower income levels, consistent with the substantially higher upfront cost of heat pump installation relative to resistive heating (Figure 4).

Figure 4. Distribution of net household income by heating energy class



Source: Authors

Although the split between Electric Resistance Heating and Heat Pump was motivated by nomenclature rather than economic criteria, this income gradient confirms that the two classes are also economically distinct, lending additional support to their treatment as separate categories in the final target variable.

Finally, the geographic analysis exposed a sharp territorial pattern that maps closely onto altitude and provincial boundaries, further motivating our adoption of a stronger spatial approach. Low altitude, high density municipalities concentrate Natural Gas and Electric Resistance Heating, while high altitude and dispersed rural settlements show markedly higher proportions of Heating Oil and LPG. The transition between the two regimes proved threshold like rather than gradual, consistent with the economics of gas network extension becoming prohibitive beyond a certain altitude and directly motivated the spatial and altitude standardised features included in the final models. This specific spatial feature generation stage addressed the project's central hypothesis: that heating technology adoption is not determined by the physical attributes of a building in isolation, but by its position within or outside the economic reach of energy supply infrastructure.

Two families of spatial descriptors were engineered to operationalise this hypothesis in the absence of official energy supply network cartography. First, a K-Nearest Neighbours density estimator (log_density_knn) approximates local urban density from inter-building distances, while a DBSCAN clustering layer distinguishes continuous urban nuclei from geographically isolated buildings, assigning the latter to a noise class that functions as an explicit marker of exclusion from network-served territory. A less sophisticated residential density on the census tract was built as a benchmark. Second, a neighbourhood energy fingerprint (NeighborProb{*} with * being the energy vector) was additionally computed for each building as the class-proportion vector of its five nearest neighbours, calculated out-of-fold to prevent information leakage into subsequent validation. A further set of features standardises altitude relative to each municipality's local profile, rather than in absolute terms, and combines building volume with density to distinguish rural from urban properties.

In Appendices A1 to A7, we present a complete description of the dataset used.

3.3. Model Selection

We adopted gradient-boosted decision trees (XGBoost) as the sole classifier family for all models reported in this study. The choice follows prior work on heterogeneous tabular building-energy data, where boosted ensembles typically outperform bagged tree methods on imbalanced multiclass targets. Three characteristics of our problem motivated this decision: sequential error correction concentrates capacity on minority and ambiguous classes; the training pipeline applies SMOTE within each cross-validation fold to mitigate class imbalance; and the classifier outputs class probabilities (via multiclass log-loss optimisation), which align with the intended use case of probabilistic triage rather than hard labelling alone.

All nine experimental models share an identical preprocessing and learning configuration, held fixed across climate zones and spatial tiers so that reported differences isolate feature content rather than classifier tuning. We did not perform nested hyperparameter search per tier; absolute accuracies in models may therefore be conservative, but within-zone tier comparisons remain valid. Each model is implemented as an imbalanced-learning pipeline: median imputation and ordinal encoding of categorical inputs, SMOTE resampling, and an XGBoost classifier with fixed hyperparameters (max_depth = 15, learning_rate = 0.0699, subsample = 0.733, colsample_bytree = 0.722, min_child_weight = 8, reg_alpha = 0.211, reg_lambda $\approx 8.6 \times 10^{-8}$, objective = multi:softmax, eval_metric = mlogloss).

Performance is assessed with stratified 5-fold out-of-fold cross-validation on each zone-specific training corpus (C1, D1, D3). In every fold, the model is trained on 80% of labelled buildings and generates predictions for the held-out 20%; the five out-of-fold prediction sets are concatenated so that each building is scored exactly once. This protocol was applied uniformly to all spatial tiers and climate zones. We preferred cross-validation to a single held-out test split for two reasons: rare classes contribute few observations, so permanently withholding a partition would reduce effective training support; and averaging over five independent train–validation partitions yields more stable accuracy and log-loss estimates than one arbitrary split.

3.4. Contribution Analysis

A Gain-based feature importance inspection confirmed that the spatial neighbourhood fingerprint variables (NeighborProb{*}) occupy the highest positions in the ranking, ahead of the density estimator and the altitude and construction year features, with structural cadastral attributes contributing predictive signal as contextual modifiers rather than primary discriminators. This sign is consistent with our initial hypothesis of spatial distribution being a strong predictive factor in our predictive efforts.

We acknowledge that, given the nature of our available data, most cases in Spain will need predictions in which no ground truth data exists and, therefore, no NeighborProb will be available. Thus, keeping in mind generalization signs (no significant changes in accuracy and confusion metrics after train-test splits) and accuracy, we perform a contribution analysis by controledly removing spatial features, first those which imply having ground truth data, then those only using geolocation, and then the rest of the features.

Three versions of each of the three climate zone-defined datasets are re-trained, including: 1) only un-engineered data such as climatic data, available indicators of heating systems predominance in municipal levels, sociodemographic characteristics at the census tract level, and building-level features such as size, typology, and age, henceforth noted NOTSPATIAL; 2) the former features, and some basic spatial derivatives, including KNN, DB scan and area-based density, henceforth noted BASICSPATIAL; and 3) using the original datasets, the NeighborProb features, as described above, and noted TAGGEDSPATIAL.

A strong hypothesis in our models is the assumption that strong spatial clustering will happen in our predictive efforts, as identified in the literature, on account of infrastructure and market constraints. As broadly noted in the literature, buildings violate the random spatial distribution. Therefore, as an extention to the spatial features contribution, we evaluate whether misclassifications cluster in space using exploratory spatial data analysis. This last analysis is carried out in the form of Global Moran's I (measures overall spatial autocorrelation of a binary out-of-fold error indicator ; 1 = wrong class, 0 = correct, where values near 0 imply randomly distributed errors; and positive I indicates geographically clustered failures or similar error rates among neighbours) (Anselin, 1995; Rey & Anselin, 2007).

3.5. Case study and model agreement

One of the primary focal points of the REGEN project is to establish a comprehensive diagnosis of the current building stock of a neighbourhood prone to regeneration, and its rehabilitation needs, as well as to identify the rehabilitation solutions and system upgrades that best contribute to decarbonization, based on the baseline of each building. To do so, generating a complete prediction of the heating energy vector will help establishing this baseline in a sound statistical manner, prior to other -more expensive- capturing or surveying efforts.

The selected neighbourhood, La Puebla Vieja, is the original population center of the municipality of Laredo, located in Cantabria (Spain) (Figure 5). Declared a Historic-Artistic Site in 1970, this enclave follows the model of a fishing village strategically positioned on a slope: the port is located at the base, sheltered from northeasterly winds, while the upper area provides a defensive position with an elevation gain of over 300 meters. This configuration results in a medieval layout characterized by six octagonal streets. Currently, the area exhibits a significant degree of physical and functional degradation, plagued by depopulation, loss of commercial vitality, and abandonment, underscoring the urgent need for rehabilitation.

Figure 5. Laredo and La Puebla Vieja

Source: Authors

Regarding data availability, this case study exemplifies the majority of the spanish building stock in terms of open data description of energy related characteristics. The region of Cantabria does not display additional data to their EPC datasets beyond rating and very basic (and already present in cadastral information) characteristics. The available information is, however, consistent with the data sources considered, so the contemplated models without the neighbour probability (namely BASICSPATIAL and NOTSPATIAL models), can be built for Laredo.

The lack of ground truth does not allow us, at present, to evaluate model transfer against actual information in the 1,140 buildings identified in Laredo. The only connection point with the proposed models is that the city lays on the C1 climate zone. Therefore, we propose to evaluate model agreement and stability of predictions using the C1 models, inspecting label-free stability diagnostics: First, we inspect marginal technology mix — predicted class histogram (% per technology) for each deployable model. Large shifts between tiers or regions indicate city-scale narrative instability. Then, we inspect within-region tier agreement — share of buildings with identical predictions under NOTSPATIAL vs BASICSPATIAL for the same C1 model. Low agreement means coordinate-derived features materially change deployment outcomes (Pan & Yang, 2010; Roberts et al., 2016).

4. Results

4.1. Model Results

The best-performing model (D1-TAGGEDSPATIAL) achieved an out-of-fold accuracy of 73.6%, a log loss of 0.666, and a macro-averaged F1 of 0.47 on the full D1 training corpus (n = 96,298 labelled buildings from País Vasco and Navarra with ZC_Municipio = "D1"). Log loss is the most operationally relevant of the three headline metrics, since the system is intended for probabilistic triage rather than deterministic labelling: the model assigns calibrated class probabilities suitable for ranking buildings for inspection or survey, whereas a naive majority-class rule (always predicting Natural Gas) reaches 68.8% accuracy yet assigns zero probability to five technologies—making it unusable for probability-based workflows. Relative to the non-spatial baseline on the same corpus (D1-NOTSPATIAL, log loss 0.833, accuracy 66.4%), the full spatial tier reduces cross-entropy by 20% while gaining 7.2 percentage points in accuracy, confirming that neighbour-corpus probabilities and coordinate-derived indices jointly improve both discrimination and probabilistic calibration.

Per-class performance follows a gradient tied to class prevalence and spatial coherence, detailed in Table 2. **Natural Gas** is identified most reliably (precision 0.84, recall 0.83), benefiting from sheer volume (68.8% of the corpus) and a spatially contiguous urban–suburban footprint reinforced by NeighborProbNaturalGas. **Electric Resistance Heating** achieves balanced intermediate performance (F1 0.57), with moderate confusion against gas-served blocks where morphologically similar buildings coexist. **Heating Oil** is recovered with reasonable recall (0.56) despite low prevalence (5.2%), exploiting a distinctive high-altitude, low-density signature combined with neighbour-oil probabilities. **Heat Pump** remains harder (F1 0.38), reflecting structural overlap with gas-served, higher-income construction and limited support (*n* = 2,531; 2.6%). **LPG** and **Biomass** are the hardest to discriminate (F1 0.28 and 0.21), sharing geography and building typology with Heating Oil and Gas Natural beyond what SMOTE can fully resolve.

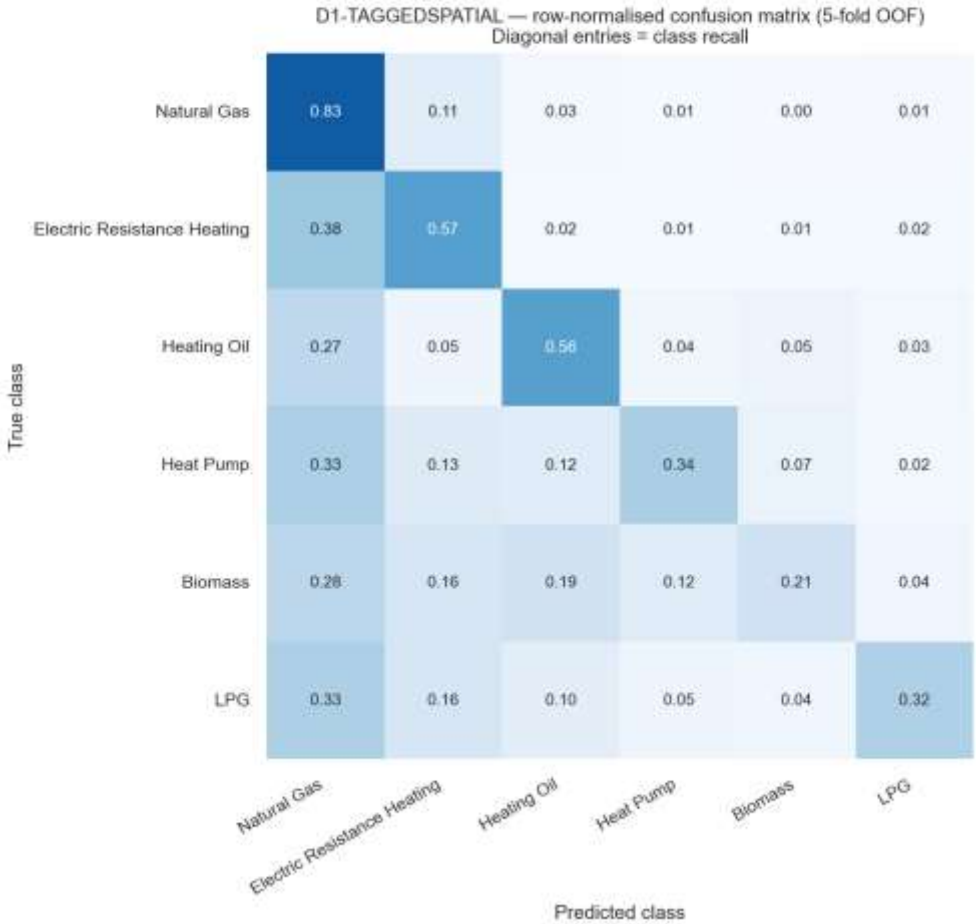
Table 2. Out-of-fold per-class metrics — D1-TAGGEDSPATIAL (stratified 5-fold CV).

Class	Precision	Recall	F1-score
Natural Gas	0.84	0.83	0.84
Electric Resistance Heating	0.57	0.57	0.57
Heating Oil	0.49	0.56	0.52
Heat Pump	0.43	0.34	0.38
LPG	0.25	0.32	0.28
Biomass	0.20	0.21	0.21
Macro average	**0.47**	**0.47**	**0.47**
Weighted average	**0.74**	**0.74**	**0.74**

Source: Authors

In a row-normalised out-of-fold confusion matrix (Figure 6) we can visualize the fraction of buildings in a true class (row) assigned to each predicted class (column); while diagonal values are class recall. The confusion matrix reveals a structural limitation that further tabular feature engineering alone cannot resolve: Natural Gas and Electric Resistance Heating are consistently confused within dense Atlantic urban blocks, where both technologies coexist in morphologically identical buildings that are indistinguishable from the exterior without internal energy-billing data. A similar pattern affects Heat Pump versus Natural Gas in recently built, gas-networked areas.

Figure 6. Row-Normalised confusion matrix (OOF) in D1-TAGGEDSPATIAL model



Source: Authors

4.2. Model Comparison and Contribution of Spatial Features

Table 3 shows a comparison of incremental accuracy and diminishing clustering of errors for all three sets of three models. **Table** 4 summarises tier increments across climate zones. Coordinate-derived features (BASICSPATIAL) add a modest but consistent gain over NOTSPATIAL (+0.7 to +0.8 pp in C1 and D1; +0.2 pp in D3). Neighbour-corpus probabilities (TAGGEDSPATIAL) dominate the remaining improvement (+5.6 to +6.4 pp in northern zones; +10.6 pp in D3). The pattern supports the hypothesis that local technology composition—where a labelled spatial corpus exists—is the primary spatial signal, while geometric descriptors capture urban-form proxies (density, cluster membership, distance to urban cores) without requiring labelled neighbours, yet adding a more modest improvement.

Table 3. Model Metrics Summary

Model	N rows	N features	CV accuracy (%)	CV log loss	Moran's I	LISA significant (%)
C1-BASICSPATIAL	101083	23	64,2	0,825	0,141	28,8
C1-NOTSPATIAL	101083	18	63,5	0,846	0,139	29,7
C1-TAGGEDSPATIAL	101083	29	69,8	0,705	0,118	16,9
D1-BASICSPATIAL	96298	23	67,2	0,81	0,194	34,6
D1-NOTSPATIAL	96298	18	66,4	0,833	0,201	36,7

D1-TAGGEDSPATIAL	96298	29	73,6	0,666	0,179	20,1
D3-BASICSPATIAL	489228	23	52,4	0,976	0,093	20,7
D3-NOTSPATIAL	489228	18	52,1	0,979	0,093	34,6
D3-TAGGEDSPATIAL	489228	29	63	0,762	0,043	22,4

Source: Authors

Table 4. Tier gains by Climate Zone

Climate zone	Δ BASICSPATIAL OVER NOTSPATIAL	Δ TAGGEDSPATIAL OVER BASICSPATIAL	Δ TAGGEDSPATIAL OVER NOTSPATIAL
C1	0,7	5,6	6,3
D1	0,8	6,4	7,3
D3	0,2	10,7	10,9

Source: Authors

D3 models exhibit low NOTSPATIAL/BASICSPATIAL accuracy (~52%) but a large TAGGEDSPATIAL uplift (~11 pp), reflecting Madrid's electric-heating-dominated corpus and the decisive role of neighbour probabilities once class collapse (no Heat Pump labels) is handled in preprocessing. Model identifiers use the D3 climate-zone label for consistency with the factorial design.

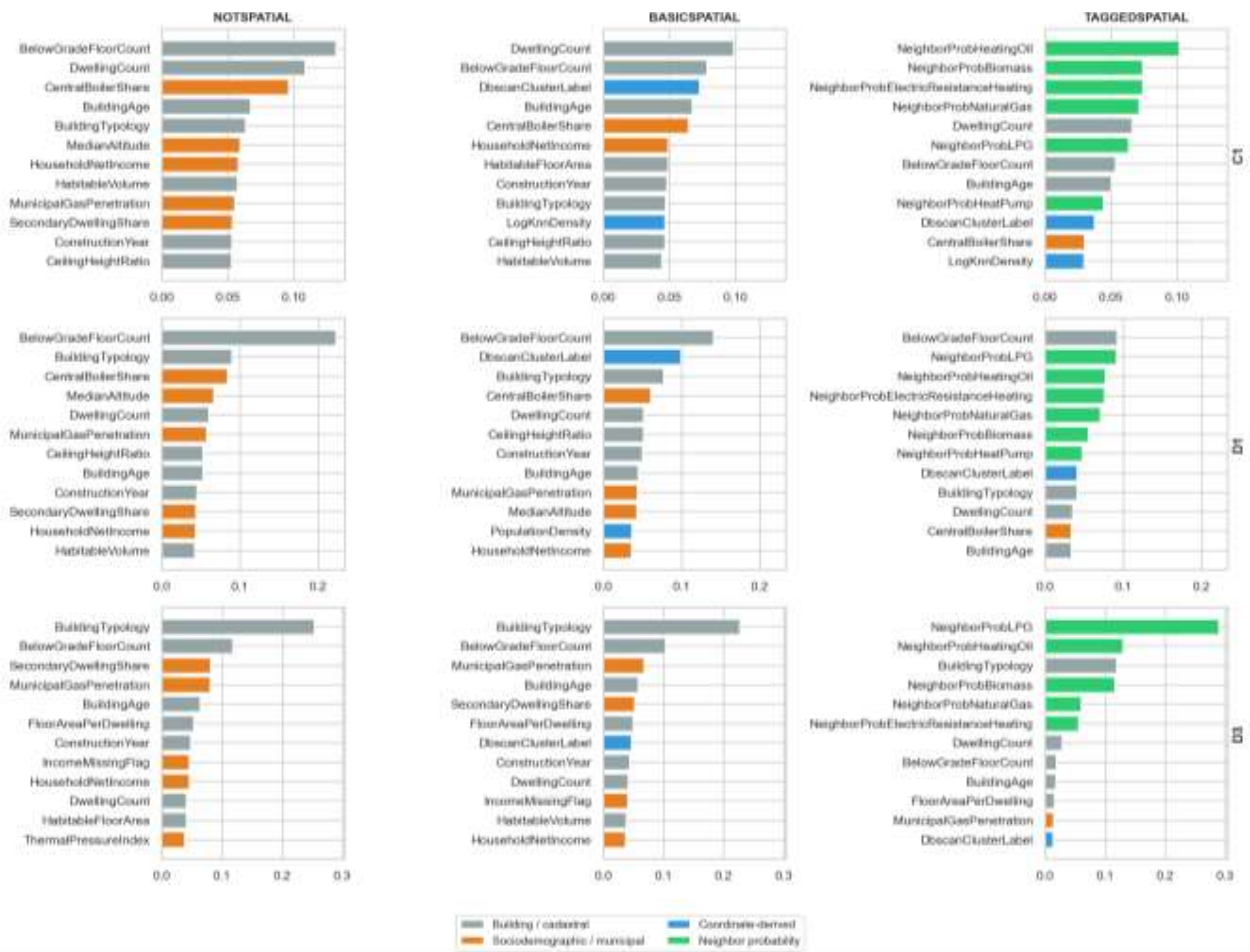
Global Moran's I on the out-of-fold binary error indicator confirms that misclassifications are spatially structured in all nine models ($p = 0.001$). In D1, NOTSPATIAL yields $I = 0.201$ versus 0.179 for TAGGEDSPATIAL, indicating that spatial inputs absorb part of the geographically clustered residual uncertainty. The effect is strongest where neighbour-corpus features are available (Δ Moran's I ≈ 0.02 in D1; ≈ 0.05 in D3 when comparing NOTSPATIAL to TAGGEDSPATIAL). LISA maps (Appendix) localise error hotspots along altitude and coastal–inland transitions, consistent with the infrastructure gradients described in Section 3.2.

Errors remain spatially structured (Global Moran's $I = 0.179$ on the binary misclassification indicator; $p = 0.001$), though reduced relative to the non-spatial variant ($I = 0.201$), indicating that spatial inputs absorb part—but not all—of the geographically clustered residual uncertainty. This pattern reflects the problem's Bayes error rate given the available observable characteristics: an honest limit of inference from public administrative and certificate data rather than a shortcoming of the modelling approach itself.

Gain-based feature importance (Figure 7) shows the growing non-trivial contribution from spatial features. Adding coordinate-derived inputs in BASICSPATIAL introduces a modest but recurrent contribution from urban-form features, notably DbscanClusterLabel and LogKnnDensity, which enter the top twelve alongside the same contextual variables. The step change appears with TAGGEDSPATIAL models: neighbour-corpus probabilities occupy the majority of top-ranked positions—17 of 36 top-twelve slots across C1, D1, and D3—and often dominate the upper ranks (e.g. all three leading features in C1-TAGGEDSPATIAL; NeighborProbLPG alone accounts for the largest single gain in D3-TAGGEDSPATIAL).

Structural and municipal attributes remain present as secondary discriminators rather than primary drivers. This pattern aligns with the tier-wise accuracy gains (Table 3): geometric spatial descriptors provide an incremental urban-form channel, while labelled neighbour composition—where available—is the dominant spatial mechanism explaining improved discrimination and calibration.

Figure 7. Top 12 features by gain importance

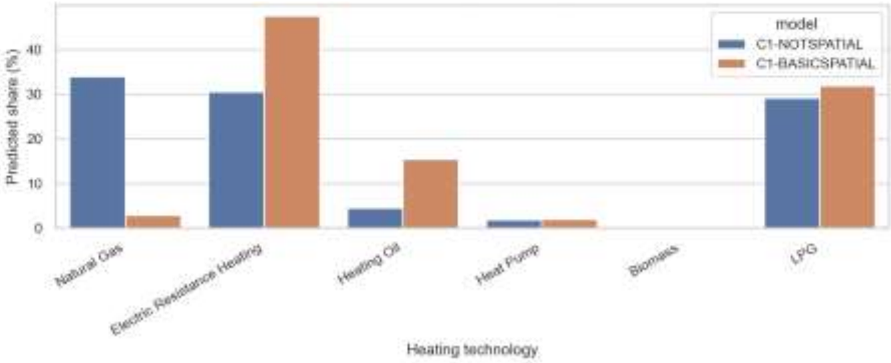


Source: Authors

4.3. Transferability without ground truth

Laredo (n = 1,140) lies in climate zone C1 but provides no heating labels or local EPC corpus. Only C1-NOTSPATIAL and C1-BASICSPATIAL models were deployed. Without reference labels, we assess marginal technology mix and within-tier agreement (Figure 8).

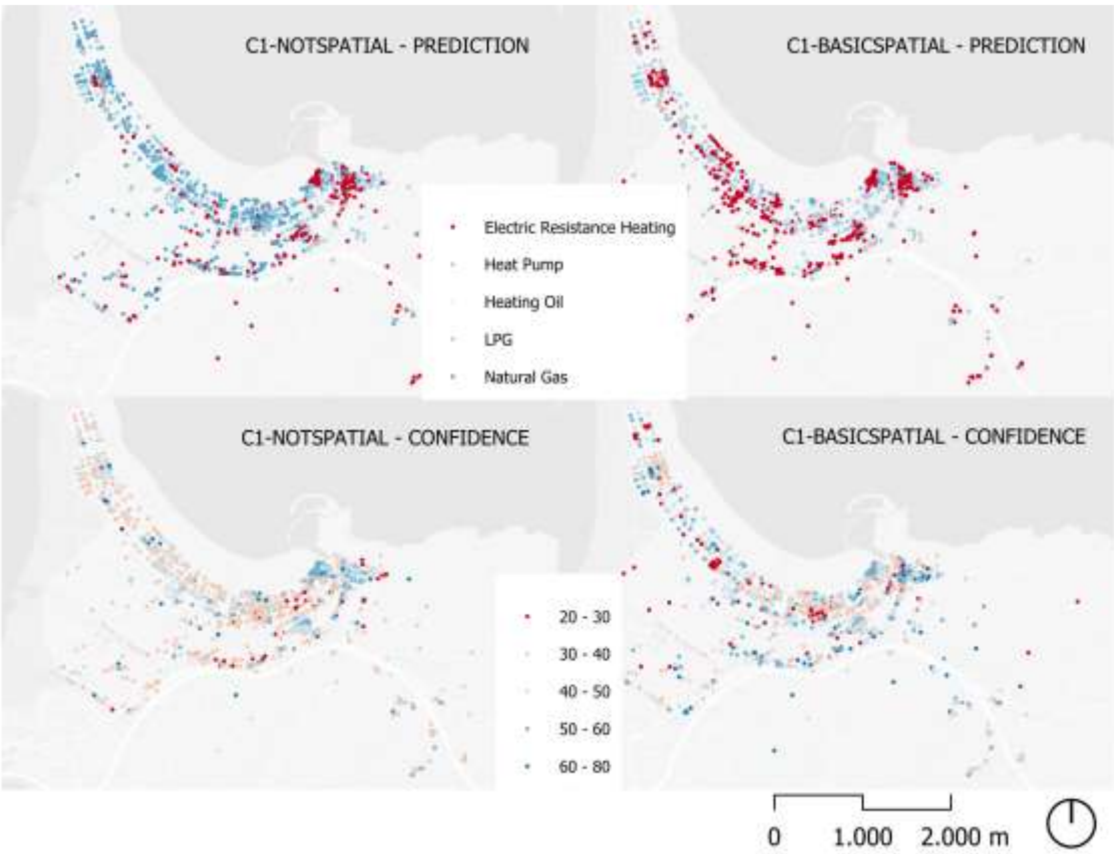
Figure 8. Predicted technology mix in C1 deployable models in Laredo



Source: Authors

Predicted city-scale landscapes diverge materially between tiers. C1-NOTSPATIAL assigns 33.9% Natural Gas and 30.5% Electric Resistance Heating, whereas C1-BASICSPATIAL reduces Natural Gas to 2.9% and raises Electric Resistance (47.5%), Heating Oil (15.5%), and LPG (31.9%). Building-level concordance between tiers is 54.6%, meaning coordinate-derived features change the predicted vector for nearly half of buildings. This instability does not invalidate in-domain spatial gains but shows that corpus-free spatial transfer cannot support a single deployment narrative without local validation.

Figure 9. Predicted technology maps for Laredo



Source: Authors

In Figure 9, we can visualize the distribution of predictions, both in tag and confidence. Although both can slightly be seen as holding a certain spatial pattern (differences in the Puebla

Vieja and the West beach expansion, the model with BASICSPATIAL features shows a more random distribution in space, which might suggest that the gain is not only good in accuracy terms, but also in the elimination of spatial effects not contemplated by the NOTSPATIAL model. Moving towards a TAGGEDSPATIAL situation would then benefit from a neighbourhood-stratified ground-truth sample (coastal core, historic centre, upper slope), guided by a further insights from this apparent clusterization.

5. Discussion

This study quantifies how spatial information improves heating-vector classification when EPC ground truth is available, and how that improvement translates—or fails to translate—to a label-free target city. Three findings stand out.

First, spatial tiers are not interchangeable: neighbour-corpus probabilities yield the largest accuracy gains where labelled spatial stock exists; coordinate-derived features add a smaller, independent increment. Second, error geography matters: models with lower Moran's I on misclassifications better respect the spatial structure of infrastructure-limited technology adoption. Third, transfer is not stable under spatial ablation: in Laredo, toggling BASICSPATIAL relative to NOTSPATIAL reshapes the predicted technology mix more than many policy audiences would expect from "adding a few spatial variables."

Relative to Beltrán-Velamazán et al. (2025), we predict the categorical energy vector rather than continuous consumption—a harder, planning-relevant target with scarcer labels. The REGEN use case (La Puebla Vieja, Laredo) shows how such models can support regeneration diagnostics when open EPC data are absent, provided uncertainty is communicated through probabilistic outputs and label-free stability checks rather than a single accuracy figure.

6. Limitations of the Study and Future Research Directions

This study was subject to several limitations. First, the availability of data was restricted, as the primary source for the analysis consisted of Energy Performance Certificates (EPC). This is not a homogeneous data source, neither in the rigor of their preparation (which affects their representation of reality) nor in the specific data published. This results in a disparity in the available data, subject to the discretion of regional administrations. Furthermore, it must be noted that certificates are issued for specific purposes (e.g., sale, rental, or renovation); consequently, not all the stock is certified, introducing bias into data and models.

Additionally, the inconsistencies in data collection necessitate territory-specific preprocessing to homogenize terminology, which often depends on the individual certifier's notes. This requires significant additional effort to achieve a comparable dataset across different territories. It is also important to highlight that the dataset contains both dwelling-level and building-level certificates. Since a single building may contain dwellings with different heating systems (particularly in those originally constructed without heating) this introduces further complexity in data processing, especially when encountering contradictory data for the same building. The decision regarding which data to assign at the building level, or how to validate the most accurate entry, is non-trivial and may introduce further error into the models' training sets.

Beyond the limitations inherent to the data, other factors not considered in this study may influence the models, such as the real estate market, socioeconomic aspects, territorial patterns, or specific climatic considerations. Further construction of relevant features, might as well improve the present modeling.

Furthermore, there is a pronounced class imbalance that affects the model's predictions. The predominance of natural gas or electricity in certain territories, compared to less frequent energy sources, considerably impacts the model's robustness when predicting minority classes. This presents a future challenge to refine the selection of features to improve the prediction accuracy for these classes. Beyond SMOTE, other strategies should be tested in future efforts.

Regarding spatial validation design, the random k-fold CV used here may optimisticise performance under strong spatial autocorrelation; spatial block CV is a natural extension and could be tested in future evolutions.

The transferability of the model is considerably limited by the lack of a ground truth for contrast in the case of Laredo, which would have allowed for the validation of the results. A potential approach to address this limitation would be the *ad hoc* development of a spatially controlled sample to validate the obtained findings, and minimize the need for ground truth samples (being as it is complicated to obtain).

Finally, regarding policy implications, a harmonised minimum open-data standard for heating systems—beyond final rating letters—would reduce the need for territory-specific label mapping and enable national-scale vector inventories, potentially complemented by register linkage (cadastre ↔ EPC ↔ census) as suggested by INE's sample-based heating statistics.

7. Conclusions

This study provides valuable insights across several domains. First, it highlights the critical need for more comprehensive data and further research in this field. Regarding data availability, the Spanish case clearly demonstrates a significant disparity in how Energy Performance Certificates (EPCs) are generated and published. The establishment of common guidelines to ensure homogeneity in EPC production, as well as standardized requirements for open data publication, would be highly beneficial. Such a framework would facilitate the creation of a unified baseline encompassing all climatic zones, thereby enabling more robust research into territorial differences and other influencing factors. Also, for this matter, further characterization of the problem, adding more climatic, sociodemographic, market and infrastructure relevant features can potentially improve model generalization.

Public certificate data, when treated spatially, can support multiclass prediction of residential heating vectors at building level, with TAGGEDSPATIAL models reaching 73.6% CV accuracy in zone D1 and calibrated class probabilities suitable for triage. Spatial ablation shows that neighbour-corpus features dominate gains, while coordinate-derived indices provide a corpus-free spatial channel relevant to cities without local labels. Application to Laredo demonstrates that this corpus-free channel is not deployment-stable without local evidence (54.6% NS/BS agreement). Advancing decarbonisation planning therefore requires both better open data harmonisation across Spanish regions and targeted neighbourhood validation in cities where models must substitute for missing inventories. This way, our conclusion is that incremental and, more importantly, spatially aware building of ground truth samples, could be extremely beneficial for complete stock modeling, without a huge and expensive effort from administrations.

Furthermore, access to detailed data on energy vectors allows for a better evaluation of building renovation strategies that extend beyond passive measures to include building heating system substitution.

This approach would maximize building energy efficiency and urban decarbonization, ultimately improving citizens' quality of life. Moreover, these data enable the development of global diagnostics and allow the design of tailored strategies to optimize regional decarbonization. Finally, such information contributes to the identification of common challenges that could be addressed collectively (leveraging economies of scale) or through targeted administrative measures and subsidies, particularly to combat energy poverty or to accelerate the transition away from fossil-based systems.

Since the decarbonization of the residential sector has become a strategic priority in the fight against climate change and given that heating and cooling in urban environments account for approximately half of Europe's final energy consumption while fossil fuels remain the dominant energy source for a large share of the building stock, there is an urgent need to accelerate the transition towards more sustainable energy systems. A comprehensive energy diagnosis is a fundamental first step in this process, as it enables a detailed understanding of current energy consumption patterns, system inefficiencies, building performance, and renewable energy integration opportunities.

In particular, understanding the heating energy vector currently used by buildings, such as natural gas, heating oil, electricity, biomass, or district heating, is essential for identifying viable decarbonization pathways and evaluating the potential for fuel switching towards low-carbon alternatives.

By establishing a robust baseline, the diagnosis enables evidence-based decision making, supports the identification and prioritization of the most effective renovation and decarbonization measures, and helps maximize the environmental, economic, and social benefits of future investments. Moreover, it reduces uncertainty in planning processes and facilitates the design of tailored strategies that respond to the specific characteristics and needs of each neighbourhood or building stock. Ultimately, energy diagnosis serves as a key enabler for achieving a cost-effective, resilient, and sustainable energy transition.

7. Acknowledgements

This paper is published as part of the REGEN project, which received funding from the European Union’s Horizon research and innovation program under grant agreement N° 101123325.

References

Ali, U., Shamsi, M. H., Hoare, C., Mangina, E., & O'Donnell, J. (2019). A data-driven approach for multi-scale building archetypes development. *Energy and Buildings*, *202*, 109364. https://doi.org/10.1016/j.enbuild.2019.109364

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Badec, K. P., & Salam, P. A. (2024). Building Energy Performance Benchmarking: Current Methods, Applications, and Implications for the Philippines. *2024 International Conference on Sustainable Energy: Energy Transition and Net-Zero Climate Future (ICUE)*, 1-8. https://doi.org/10.1109/ICUE63019.2024.10795622

Beltrán-Velamazán, C., Monzón-Chavarrías, M., & López-Mesa, B. (2025). Predicting Energy and Emissions in Residential Building Stocks: National UBEM with Energy Performance Certificates and Artificial Intelligence. *Applied Sciences*, *15*(2), 514. https://doi.org/10.3390/app15020514

BPIE (Buildings Performance Institute Europe). (2025). *Delivering the EPBD: A guide towards better, affordable and more resilient buildings for all in Europe*. https://www.bpie.eu/delivering-the-epbd-a-guide-towards-better-affordableand-more-resilient-buildings-for-all-in-europe/

Cerezo Davila, C. (2017). *Building archetype calibration for effective urban building energy modeling* [Massachusetts Institute of Technology]. http://hdl.handle.net/1721.1/111487

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. https://doi.org/10.1613/jair.953

Chen, G., Lu, S., Zhou, S., Tian, Z., Kim, M. K., Liu, J., & Liu, X. (2025). A Systematic Review of Building Energy Consumption Prediction: From Perspectives of Load Classification, Data-Driven Frameworks, and Future Directions. *Applied Sciences*, *15*(6), 3086. https://doi.org/10.3390/app15063086

Cities—United Nations Sustainable Development Action 2015. (s. f.). *United Nations Sustainable Development*. Recuperado 29 de julio de 2026, de https://www.un.org/sustainabledevelopment/cities/

Directive 2023/1791 of the European Parliament and of the Council on Energy Efficiency and Amending Regulation 2023/955, EP, CONSIL, 231 OJ L (2023). http://data.europa.eu/eli/dir/2023/1791/oj

Directive 2024/1275 of the European Parliament and of the Council on the Energy Performance of Buildings, EP, CONSIL (2024). http://data.europa.eu/eli/dir/2024/1275/oj

Economidou, M., Todeschi, V., Bertoldi, P., D'Agostino, D., Zangheri, P., & Castellazzi, L. (2020). Review of 50 years of EU energy efficiency policies for buildings. *Energy and Buildings*, *225*, 110322. https://doi.org/10.1016/j.enbuild.2020.110322

Eguiarte, O., de Agustín-Camacho, P., & del Portillo-Valdés, L. (2022). Energy and economic analysis of domestic heating costs based on distributed energy resources: A case study in Spain. *Energy Reports, Selected papers from 2022 7th International Conference on Advances on Clean Energy Research*, *8*, 56-61. https://doi.org/10.1016/j.egyr.2022.10.214

Eguiarte, O., Garrido-Marijuán, A., de Agustín-Camacho, P., del Portillo, L., & Romero-Amorrortu, A. (2020). Energy, Environmental and Economic Analysis of Air-to-Air Heat Pumps as an Alternative to Heating Electrification in Europe. *Energies*, *13*(15), 3939. https://doi.org/10.3390/en13153939

EU Mission: Climate-Neutral and Smart Cities. (2026, julio 20). https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe/eu-missions-horizon-europe/climate-neutral-and-smart-cities_en

Fang, C., Zhou, L., Gu, X., Liu, X., & Werner, M. (2025). A data driven approach to urban area delineation using multi source geospatial data. *Scientific Reports*, *15*. https://doi.org/10.1038/s41598-025-93366-x

Harun, F., Pikas, E., Iliste, E., Hallik, J., & Kalamees, T. (2026). From static archetypes toward dynamic, condition-aware archetypes in UBEM: An integrative review with an illustrative case demonstration. *Frontiers in Energy Research*, *14*. https://doi.org/10.3389/fenrg.2026.1800126

IDAE. (2026). *SPAHOUSEC III. Estudio del consumo energético, equipamiento y hábitos de consumo de la energía del sector residencial en España.* https://www.idae.es/sites/default/files/documentos/publicaciones_idae/20260123_SPAHOUSEC_III.pdf

Jiang, Q., Huang, C., Wu, Z., Yao, J., Wang, J., Liu, X., & Qiao, R. (2024). Predicting building energy consumption in urban neighborhoods using machine learning algorithms. *Frontiers of Urban and Rural Planning*, *2*. https://doi.org/10.1007/s44243-024-00032-3

Li, K., Xue, W., Tan, G., & Denzer, A. S. (2020). A state of the art review on the prediction of building energy consumption using data-driven technique and evolutionary algorithms. *Building Services Engineering Research & Technology*, *41*(1), 108-127. https://doi.org/10.1177/0143624419843647

Long, C., Yang, X., Su, Y., Liu, F., Ma, R., Ma, T., Wu, Y., & Shen, X. (2025). Air Conditioning Load Forecasting for Geographical Grids Using Deep Reinforcement Learning and Density-Based Spatial Clustering of Applications with Noise and Graph Attention Networks. *Energies*, *18*(11), 2832. https://doi.org/10.3390/en18112832

Luqman, M., Rayner, P. J., & Gurney, K. R. (2023). On the impact of urbanisation on CO2 emissions. *Npj Urban Sustainability*, *3*(1), 6. https://doi.org/10.1038/s42949-023-00084-2

Ministerio de Transportes, Movilidad y Agenda Urbana. Gobierno de España. (2019). *Código Técnico de la Edificación: Documento Básico DB-HE Ahorro de Energía.* https://www.codigotecnico.org/pdf/Documentos/HE/DBHE.pdf

Ministerio de Transportes, Movilidad y Agenda Urbana. Gobierno de España. (2025). *Borrador del Plan Nacional de Renovación de Edificios (PNRE): Marco ARCE 2050.* https://www.mivau.gob.es/recursos_mfom/audienciainfopublica/recursos/borrador_pnre.pdf

OECD. (2020). Strengthening Governance of EU Funds under Cohesion Policy: Administrative Capacity Building Roadmaps. *OECD Multi-level Governance Studies*. https://doi.org/10.1787/9b71c8d8-en

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

Reinhart, C. F., & Cerezo Davila, C. (2016). Urban building energy modeling – A review of a nascent field. *Building and Environment*, *97*, 196-202. https://doi.org/10.1016/j.buildenv.2015.12.001

Rey, S. J., & Anselin, L. (2007). PySAL: A Python Library of Spatial Analytical Methods. *Review of Regional Studies*, *37*(1). https://doi.org/10.52324/001c.8285

Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J., Schröder, B., Thuiller, W., Warton, D., Wintle, B., Hartig, F., & Dormann, C. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*. https://doi.org/10.1111/ecog.02881

Secretaría de estado de transportes, movilidad y agenda urbana. Gobierno de España. (2020). *EREESE 2020. Actualización 2020 de la estrategia a largo plazo para la rehabilitación energética en el sector de la edificación en España.* https://www.mivau.gob.es/recursos_mfom/paginabasica/recursos/eresee_2020.pdf

Shen, P., & Wang, H. (2024). Archetype building energy modeling approaches and applications: A review. *Renewable and Sustainable Energy Reviews*, *199*, 114478. https://doi.org/10.1016/j.rser.2024.114478

Sisman, S., Kara, A., & Aydinoglu, A. C. (2025). Leveraging spatial data infrastructure for machine learning based building energy performance prediction. *PLOS ONE*, *20*(10), e0335531. https://doi.org/10.1371/journal.pone.0335531