Exploratory Analysis of Energy Performance Certificates in the Community of Madrid

Laia Garcia-Aguilera and Daniel Garcia-Gonzalez

Universidade da Coruña, Centro de Investigación CITIC, A Coruña, Spain
laia.garcia@udc.es, d.garcia2@udc.es

Abstract

This project is part of a broader study on household energy consumption and emissions using artificial intelligence applied to Energy Performance Certificates from the Community of Madrid. It is currently in its first phase, focused on exploratory data analysis and data cleaning: detecting and removing errors and duplicates, standardizing formats, handling fictitious and outlier values, and reviewing fields with missing data. The goal is to establish a solid foundation for developing predictive models that can identify the factors with the greatest influence on energy efficiency and sustainability in buildings.

1 Introduction

Optimizing building energy efficiency is a key strategy to reduce energy consumption and CO_2 emissions, contributing to sustainability and climate goals. As residential buildings account for a significant proportion of total energy demand, studying their energy performance is particularly important.

Energy Performance Certificates (EPCs) provide detailed information on building characteristics, energy systems, and consumption, offering a valuable resource for data-driven analysis. This study aims to prepare and analyze EPC records from the Community of Madrid [1] as a first step towards building predictive models of household energy consumption and CO_2 emissions, and to identify the most relevant features for these models.

2 Study

EPC records from 2021 to 2025 were combined into a single dataset. Inconsistencies between years, such as differences in encoding, delimiters, and column formats, were standardized. The dataset was then filtered to include only residential buildings in the province of Madrid, which accounted for 92.25% of the original data.

The target variable is non-renewable primary energy consumption intensity (kWh/m^2·year), while CO_2 emission intensity (kg CO_2e/m^2·year) is considered as an alternative target. Input variables include building characteristics (typology, construction year, regulation, useful area, envelope compactness and glazing ratios) and system-related features (heating, cooling and domestic hot water parameters).

Several preprocessing steps were applied to ensure data quality and consistency. Categorical variables were normalized by merging redundant or inconsistent labels. The construction year field was cleaned and converted into four-digit numeric values within the 1800–2025 range. Glazing ratios were split by façade orientation.

Some variables were excluded due to redundancy, low variability or poor data quality, including envelope compactness, climatic zone, renewable energy features, and equipment nominal powers. Equipment seasonal efficiencies, which contained many placeholder values, were retained and discretized into three levels: low, medium, and high.

Missing values were substantially reduced by interpreting heating and cooling coverage equal to zero as the absence of the corresponding system, which resolved most null entries in related variables. Remaining missing values were removed from the dataset.

Outliers in numeric columns with impossible or extremely high values —specifically useful area, domestic hot water daily demand, and the target variables— were detected using a modified z-score and removed.

Exploratory analysis revealed several noteworthy patterns. Non-renewable primary energy consumption tends to decrease in newer constructions, reflecting improvements in building regulations and energy efficiency over time. A strong linear relationship was observed between energy consumption and CO_2 emissions, confirming their consistency as performance indicators. Other features, such as useful area or glazing ratios, exhibited weak or highly scattered relationships, suggesting that their influence may depend on building typology or system characteristics. Among the input variables, useful area and domestic hot water demand showed the highest correlation ($r \approx 0.88$), indicating that larger dwellings require proportionally more hot water. These findings validate the internal consistency of the dataset and confirm its suitability for subsequent predictive modeling.

3 Conclusions

This first phase focused on cleaning, standardizing, and exploring EPC data for residential buildings in the Community of Madrid. The results confirmed expected patterns, such as lower energy consumption in newer buildings and the strong link between energy use and CO_2 emissions, while other features exhibited weaker or context-dependent relationships.

Despite the overall quality of the dataset, several limitations remain. Some relevant features, such as renewable energy contributions or detailed system capacities, were excluded due to poor data quality. The data also lack certain geometric or environmental descriptors, such as building element thermal properties and shading, that could improve model precision. Additionally, some residual noise and reporting inconsistencies may persist due to the administrative nature of EPC data.

Future work will involve developing predictive models of energy consumption and emissions, evaluating different algorithms and hyperparameters, and studying feature importance. The dataset may be expanded with additional building, socioeconomic, or environmental data to improve model accuracy and interpretability.

Acknowledgments

This work was partially funded by Ministry for Digital Transformation and Civil Service and "NextGenerationEU"/PRTR under grant TSI-100925-2023-1.

References

[1] Comunidad de Madrid - Dirección General de Transición Energética y Economía Circular. Registro de certificados de eficiencia energética de edificios. https://datos.comunidad.madrid/dataset/registro_certificados_eficiencia_energetica, 2024–2025. Dataset licensed under Creative Commons Attribution. Last viewed October 2025.