Analysing electricity consumption in the municipality of Lleida

CIMNE - BEE Group

2025

Contact for Doubts:
Gerard Mor (gmor@cimne.upc.edu)

Download datasets:
https://nextcloud.tech.beegroup-cimne.com/s/ZT4XYZEpxNpqEBc

1 Introduction

This document outlines the open assignment for a prospective role within the Data Analytics Department at CIMNE. The assignment centers on analyzing and understanding hourly electricity consumption data for the municipality of Lleida (Spain). It aims to assess the candidate's ability to apply data analytics techniques to real-world datasets, with a focus on identifying key factors influencing electricity usage. This practical exercise highlights the intersection of data-driven decision-making and urban sustainability, reflecting the department's commitment to leveraging data analytics for impactful solutions.

2 Objective of the Assignment

The primary objectives of this assignment are:

- **To demonstrate proficiency in data analytics and Machine Learning (ML) techniques in real-world applications:** The assignment is designed to evaluate the candidate's ability to process, analyze, and model complex datasets. By developing and evaluating predictive models, candidates will showcase their technical expertise in managing real-world challenges such as handling missing data, integrating diverse datasets, and selecting optimal ML techniques to deliver actionable insights.

- **To emphasize the practical utility of open data sources and their role in analytics-driven decision-making:** This task involves the use of publicly available datasets, including energy consumption, meteorological, and socio-economic data, to build a comprehensive analytical framework. Candidates will illustrate their capability to preprocess and combine these datasets, demonstrating the potential of open data to address pressing urban challenges.

- **To identify key features and methodologies for predicting urban electricity consumption and to benchmark various predictive approaches:** A core component of this assignment is feature selection and the exploration of predictive methodologies. Candidates will identify variables such as weather patterns, socio-economic indicators, and building characteristics that most significantly influence electricity consumption. Through experimentation with predictive models —ranging from regression to time-series forecasting— candidates will demonstrate their ability to critically evaluate model performance and interpret results, skills that are essential for data analytics in a professional setting.

3 Datasets

3.1 Electricity Consumption Dataset

Filename: electricity_consumption.parquet

The primary dataset for this assignment is sourced from Datadis, the data platform of the main Distribution System Operators in Spain. This open API provides hourly electricity consumption data at the postal code level across Spain.

The dataset is stored in Apache Parquet format and contains the following columns:

- **postalcode**: The postal code associated with each data point.
- **time**: The start time of each record in UTC, formatted as a datetime.
- **contracts**: The number of contracts included in the reported consumption data.
- **consumption**: The total electricity consumption during the 1-hour timestep, measured in kWh.

3.2 Weather Conditions Dataset

Filename: weather.parquet

Weather conditions play a crucial role in understanding variations in electricity consumption, as temperature, humidity, solar radiation, and other environmental factors often influence how much energy is used for heating, cooling, and lighting. For instance, during colder periods, higher heating demands can increase energy use, while warmer weather may result in more air conditioning usage.

In this assignment, we use data from the ERA5Land dataset[1], a popular source of meteorological data due to its global, high-resolution coverage (approximately 9 km) and free availability. The dataset has been pre-processed from its original GRIB format to a tabular form for easier analysis. It is provided in Apache Parquet format and contains the following columns:

- **postalcode**: The postal code corresponding to each data point, allowing spatial aggregation.
- **time**: Timestamp in UTC, enabling time-series analysis.
- **airtemperature**: Average dry-bulb outdoor temperature during each hour (1h), measured in °C. Temperature variations can affect heating and cooling needs, impacting electricity usage.
- **relativehumidity**: Average relative humidity during each hour (1h), in %. Higher humidity may influence cooling needs and comfort levels, potentially increasing electricity demand.
- **ghi**: Global horizontal irradiance accumulated over each hour (1h), measured in kWh/m^2. Solar irradiance affects natural lighting and heating, which may reduce or increase electricity consumption.
- **sunelevation**: Average solar elevation angle during each hour (1h), in degrees. Sun elevation provides insights into available natural light, which can reduce artificial lighting needs.
- **sunazimuth**: Average solar azimuth angle during each hour (1h), in degrees. This data can help model solar exposure on buildings, influencing passive heating and lighting requirements.
- **highvegetationratio**: Average proportion of high vegetation (trees) during each hour (1h), in %. Vegetation can provide natural shading, potentially lowering cooling demands.
- **lowvegetationratio**: Average proportion of low vegetation (plants) during each hour (1h), in %.
- **winddirection**: Average wind direction during each hour (1h), in degrees, potentially relevant for understanding localized weather effects on building heating and cooling.
- **windspeed**: Average wind speed during each hour (1h), in m/s. Wind can impact heating requirements by influencing building heat loss.
- **totalprecipitation**: Accumulated precipitation over each hour (1h), measured in mm/m^2. Weather conditions, including rain, can affect energy usage patterns, such as increased heating or lighting during overcast and rainy days.

[1] `https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=overview`

3.3 Postal Codes Boundaries

Filename: postal_codes_lleida.gpkg

This dataset defines the administrative boundaries of postal codes within the municipality of Lleida. It is useful for visualizing results on a map and for estimating electricity consumption at various geographical levels. The data format is Geopackage, with key features including **geometry** and **CODPOS** (postal code).

3.4 Cadastral Data for Buildings

Filenames: cadaster_{lleida,alcarras,alpicat}.gml

Cadastral data provides detailed information about the buildings within an urban area, which is essential for understanding patterns of electricity consumption. Building characteristics, such as age, size, and type, can significantly influence energy demand. For instance, older buildings might have poorer insulation, leading to higher heating or cooling needs, while larger buildings generally consume more energy due to their greater space requirements. Additionally, residential and non-residential buildings exhibit different electricity usage patterns due to differences in occupancy and activity.

This dataset, available from the Spanish National Cadaster[2], covers all municipalities in Spain, excluding Navarra and the Basque Country, and is stored in GML format. Relevant features for this assignment include:

- **conditionOfConstruction**: Indicates the structural condition of each building. Buildings in poor condition may have lower energy efficiency, increasing electricity use for heating or cooling.
- **beginning**: The year of construction, providing insights into building age. Older buildings might lack modern insulation, influencing energy requirements.
- **reference**: Cadastral reference number, allowing for unique identification and spatial analysis of buildings.
- **geometry**: The footprint geometry of each building, which is useful for calculating the area and understanding spatial relationships.
- **value**: Total gross floor area of the building, which correlates with potential energy consumption, as larger buildings generally require more energy.
- **numberOfDwellings**: The number of residential units within a building, which helps estimate occupancy and usage intensity, both of which impact electricity demand.

[2] `https://www.catastro.hacienda.gob.es/INSPIRE/buildings/ES.SDGC.BU.atom.xml`

3.5 Socio-economic Data

Filename: socioeconomic.parquet

Socio-economic factors are critical to understanding patterns of electricity consumption, as income levels, demographics, and household characteristics directly impact energy usage behaviors. For example, higher-income households may consume more electricity due to the use of additional appliances, while areas with higher elderly populations may exhibit increased heating needs. Additionally, the type of income sources (e.g., salary, pension) may correlate with the typical daily routines of the population, influencing peak electricity demand times.

This dataset is sourced from the National Statistics Institute's Rental Distribution Atlas[3], and provided for this assignment in Apache Parquet format. The dataset includes a range of socio-economic indicators that offer insights into the underlying factors driving electricity consumption, some of them are:

- **year**: The year related to the data point.
- **postalcode**: The postal code associated with each data point.
- **percentagepopulationover65**: The percentage of the population over 65 years. Older populations might have different energy needs, particularly for heating.
- **percentagepopulationunder18**: The percentage of the population under 18. Areas with more young people may have different usage patterns due to educational activities and family structures.
- **percentagesinglepersonhouseholds**: The percentage of single-person households, which may have distinct energy usage patterns compared to multi-person households.
- **population**: Total population count, offering a measure of density, which can impact overall electricity demand in an area.
- **incomesperhousehold**: Average income per household. Higher incomes can correlate with higher consumption due to greater appliance use and larger living spaces.
- **incomesperunitofconsumption**: Income per unit of consumption, which normalizes income by household size, giving a refined view of consumption capacity.
- **grossincomesperperson**: Gross income per person, affecting disposable income and potentially influencing energy usage.
- **incomessourceisotherbenefits**: Percentage of income from other benefits, giving insights into the socio-economic composition of an area.

[3] `https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177088&menu=ultiDatos&idp=1254735976608`

- **incomessourceisotherincomes**: Percentage of income from sources other than salaries and pensions.
- **incomessourceispension**: Percentage of income from pensions, indicative of the elderly population.
- **incomessourceisunemploymentbenefit**: Percentage of income from unemployment benefits, which can be associated with economic status.
- **incomessourceissalary**: Percentage of income from salaries, which may influence typical electricity consumption patterns tied to working hours.
- **giniindex**: Gini index for income inequality, as areas with high income disparity might show diverse consumption behaviors across income groups.
- **incomesratioq80byq20**: Income ratio (80th to 20th percentile), providing a measure of income distribution, which can influence overall energy demand.
- **averagepopulationage**: Average age of the population, which may relate to energy usage preferences and requirements.
- **peopleperhousehold**: Average number of people per household, which impacts total household electricity consumption.

4 Tasks

This section outlines the main tasks for analyzing and predicting electricity consumption data. Each task description includes the objective, suggested approach, and the potential applications of its results.

4.1 Identify Common Daily Load Curves

The objective is to detect the most common daily electricity load curve patterns by postal code, grouping similar behaviors to reveal typical daily usage trends.

Methodology: Apply unsupervised learning techniques to identify this patterns. You can use dimensionality reduction for better visualization and interpretation.

4.2 Predict Day-Ahead Load Curve Probability

This task entails developing a supervised classification model to predict the likelihood of the clustered daily load curves from the previous task for the following day, categorized by postal code.

Methodology: Use classification models trained on historical data and exogenous variables, such as weather, cadaster and socio-economic data, to forecast probabilities. The frequency should be daily.

4.3 Electricity consumption short-term forecast

Develop a supervised regression model to predict electricity consumption across the entire municipality for the next 96 hours, using a global forecasting approach that incorporates consumption data from all postal codes, cadastral, weather, and socio-economic data. The model frequency should be hourly.

Methodology: Employ time-series forecasting or regression models with exogenous inputs to capture trends and seasonality in electricity consumption. Use methods to optimise model hyperparameters and avoid overfitting.

5 Software implementation recommendations

We will consider the quality and efficiency of the implemented algorithms. It is required to follow these steps for your implementation:

1. The programming language is Python 3
2. The usage of the following libraries is recommended, not mandatory: Pandas, Geopandas or Polars for data wrangling; Scikit-learn or statsmodels for modelling; Matplotlib, Folium or Plotly for visualising.
3. Simplicity and readability of your code.
4. It is recommended to follow the style guide[4].

5.1 Documentation

You should prepare a 10' presentation showing the results of your work, include whatever plot or chunk of code you want to highlight. In addition, you must send us the source code of your solution in a ZIP file or in a Github (or alternatives) repository.

[4] https://www.python.org/dev/peps/pep-0008/