Second Assignment: Analyzing Hourly Electricity Consumption in Lleida

Artificial Intelligence
Department of Computer Engineering and Digital Design
University of Lleida

November 2024 - January 2025

Contact for Doubts:
Please direct any questions through Virtual Campus.
Download datasets:
https://nextcloud.tech.beegroup-cimne.com/s/ZT4XYZEpxNpqEBc

1 Introduction

This document presents the second assignment for the subject of Artificial Intelligence, focusing on the analysis and understanding of hourly electricity consumption data within Lleida. This assignment builds upon foundational Machine Learning concepts and applies them to real-world data, providing insights into how environmental and social factors influence electricity usage.

2 Objective of the Assignment

The main objectives of this assignment are as follows:

- **To gain practical experience with Machine Learning (ML) techniques applied to real-world data environments:** This assignment provides hands-on experience with ML models, allowing students to apply theoretical knowledge to real datasets. Through the development and testing of predictive models, students will learn to navigate the complexities of real-world data, such as managing diverse datasets, handling missing values, and selecting appropriate ML algorithms for specific tasks. This experience is essential for developing the analytical and technical skills needed in modern data-driven industries.

- **To illustrate the utility of open datasets, emphasizing their potential value in both public and private sector roles:** By working with publicly available datasets, such as energy consumption, weather, cadastral, and socio-economic data, students will understand the value and potential impact of open data. Open datasets are powerful tools in research, policy-making, and business, as they enable transparency, innovation, and the generation of new insights. In this assignment, students will learn how to source, preprocess, and integrate these datasets to draw meaningful conclusions that could apply to both public sector projects and private industry applications.

- **To identify relevant features and methodologies for predicting urban-level electricity consumption and to experiment with and compare various predictive techniques:** A key goal of the assignment is to determine which features (e.g., temperature, household income, building age) are most relevant for accurate electricity consumption forecasting. Students will explore different ML methodologies, such as regression, time-series forecasting, and classification models, to identify the most effective approaches for urban-level predictions. By experimenting with various models and comparing their results, students will learn critical evaluation skills, enabling them to select the best-suited methods for similar tasks in future projects.

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.

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

- **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.

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.

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

- **numberOfDwellings**: The number of residential units within a building, which helps estimate occupancy and usage intensity, both of which impact electricity demand.

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.

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

- **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.
- **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 (Max Score: 2 Points)

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

Methodology: Apply clustering techniques, such as K-means or hierarchical clustering, to identify daily load patterns. Use dimensionality reduction (e.g., PCA) for better visualization and interpretation.

Impact: Understanding typical load patterns at a postal code level can help in planning energy resources, identifying unique demands, and managing peak load periods more effectively.

4.2 Predict Day-Ahead Load Curve Probability (Max Score: 3 Points)

This task involves building a supervised classification model to predict the probability of specific load curve types for the next day by postal code.

Methodology: Use classification models (e.g. decision trees) trained on historical data and external variables, such as weather and socio-economic data, to forecast probabilities.

Impact: Predicting daily load patterns enables better resource planning, allowing for proactive adjustments to energy supply (energy flexibility services) and improving grid reliability.

4.3 Electricity consumption short-term forecast (Max Score: 4 Points)

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.

Methodology: Employ time-series forecasting models with exogenous inputs to capture trends and seasonality in electricity consumption. Use the time-series cross validation methodology to avoid model overfitting. Remember that you will need to aggregate cadaster data to postal code geographical level.

Impact: A 96-hour forecast for the municipality supports better energy resource allocation, anticipates peak demand, and aids in implementing demand-response strategies.

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].

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

5.1 Documentation

The documentation must include the full names of all group members. Assignments can be completed individually or in pairs.

Your report will be evaluated based on the clarity of writing, quality of analysis, effectiveness of plots, and depth of conclusions. Rather than including large code snippets in the documentation, refer to specific lines in your source files (e.g., `clustering.py#L22-202`).

5.1.1 Recommended Table of Contents

While optional, the following structure is recommended:

1. An initial section detailing all data wrangling, merging, alignment, and aggregation steps across datasets, along with any transformations or filters applied to prepare data for ML algorithms. Include relevant visualizations and explanatory analysis.
2. For each task in Section 4, provide a section with:
   - A summary of data wrangling, transformations, and filtering specific to the task, describing the dataset provided to the ML algorithms.
   - A comparison of implemented ML algorithms, noting the best-performing method.
   - Analysis of the top-performing model, with evaluation metrics over time and visualizations of the results, including textual interpretation. Plots can be static or interactive.

5.1.2 Documentation Format

Submit your documentation as a PDF or HTML file, with a maximum length of approximately 8 DIN A4 pages.

- **PDF Format:** Use a single or double-column scientific paper style (e.g., with an online LaTeX editor[5]).
- **HTML Format:** Only use HTML if including interactive plots. Markdown and grip[6] can be used to export and link HTML files.

[5] `https://es.overleaf.com/`

[6] `https://github.com/joeyespo/grip`