Project INF 442-1:
A feasibility study of predicting household power consumption based on meteorological data

Simon Bliudze
Simon.Bliudze@inria.fr

March 14, 2020

1 Introduction

The goal of this project is to analyse the data available for the power consumption of a household, in order to identify consumption patterns and establish their correlate (or absence thereof) with meterological data.

The project is inspired from [3], but uses a differet dataset.

As opposed to the exercise sessions, you are not limited to C++ but can use any programming language or tool you consider fit and feel comfortable with. For example, the very first step, data cleaning, might be easier to carry out using a scripting language. Similarly, it will be much easier to visualise your results by generating input data in a format accepted by an appropriate external tool.

For evaluation, priority will be given to qualitative results, how much information you can obtain from the data and how clearly you can present it. However, you will be expected to be capable of explaining whatever code or tool you will have used (within the limits of the required functionality).

2 Data

The data provided for the project consist of two datasets:

Household power consumption The first data set is obtained from a set of measurements of electric power consumption in one household with a *one-minute* sampling rate over a period of almost four years. The measurements [2] were gathered in a house located in Sceaux (France) between December 2006 and November 2010 (47 months). The original measurements are provided as a single file with data within rows separated by semicolons (;). For the purposes of this project this file was split into four: one for each of the years 2007–2010 (data for December 2006 were discarded).

The dataset contains some missing values in the measurements (nearly 1,25% of the rows). All calendar timestamps are present in the dataset but for some timestamps, the measurement values are missing: a missing value is represented by the absence of value between two consecutive semicolon attribute separators. For instance, the dataset shows missing values on April 28, 2007.

Attribute Information:

1. `date`: Date in format `dd/mm/yyyy`
2. `time`: time in format `hh:mm:ss`
3. `global_active_power`: household global minute-averaged active power (in kilowatt)
4. `global_reactive_power`: household global minute-averaged reactive power (in kilowatt)
5. `voltage`: minute-averaged voltage (in volt)
6. `global_intensity`: household global minute-averaged current intensity (in ampere)
7. `sub_metering_1`: energy sub-metering №1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
8. `sub_metering_2`: energy sub-metering №2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
9. `sub_metering_3`: energy sub-metering №3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

The quantity

$$\texttt{global\_active\_power} \cdot 1000/60 - \texttt{sub\_metering\_1} - \texttt{sub\_metering\_2} - \texttt{sub\_metering\_3}$$

represents the active energy consumed every minute (in watt-hour) in the household by electrical equipment not measured in sub-meterings №№1, 2 and 3.

Meteorological data The second dataset comprises meteorological data obtained from Météo France public database [1] for the same period of four years 2007–2010 at *three-hour* sampling rate. The data are provided in 48 separate files—one per month—with data within rows separated by semicolons (;). Measurements are provided for all meteorological stations in France. The station closest to Sceaux is located in Orly. Its identifier is 07149 (see `postesSynop.csv`). Table 1 shows the first ten measurement attributes (the full list, in French, is available in `doc_parametres_synop_168.pdf`).

Table 1: Attribute information for the meteorological data

Description	Mnemonics	Type	Unit/Format
WMO station identifier	`numer_sta`	int	
Date (UTC)	`date`	string	AAAAMMDDHHMISS
Pressure at sea level	`pmer`	int	Pa
Pressure variation in 3 hours	`tend`	int	Pa
Barometric tendancy type	`cod_tend`	int	code (0200)
Wind direction / 10 mn average	`dd`	int	degree
Wind speed / 10 mn average	`ff`	real	m/s
Temperature	`t`	real	K
Dew point	`td`	real	K
Humidity	`u`	int	%

The mnemonics are used as data-file row headers; WMO = World Meteorological Organisation; for codes, see `wmo_306-v1_1-2012_en.pdf`

3 Tasks

Data pre-processing Data is assumed to have errors and imperfections so cleaning and detecting outliers are needed. This is explicitly stated for the household power consumption measurements, but could also be the case for the meteorological data.

Detection of outliers can be performed using the Z-score for each point:

$$z = \frac{x - \mu}{\sigma},$$

where x is the data point, μ represents the mean of the whole dataset, σ is the standard deviation.

Cleaned data from housefold power consumption and meteorological data should be integrated by matching the data sets by time and location (only the Orly meteo-station is relevant). Missing data can be completed by interpolation; unnecessary data can be removed. Notice that the sampling frequencies of the two data sets are different.

Seasonality detection Intuitively, one would expect that there are periods in a year—more or less similar to natural seasons—when power consumption is very similar. Can definitive periods of the year be found, where same or very similar power consumption patterns are found? How many and why?

Use clustering, and by dividing data into days to represent each day as a vector and implement K-means algorithm to compare distances between vectors.

Distribution of hourly consumption Intuitively, one would expect that power consumption is distrubuted normally for any given hour of the day. An increase in consumption is expected during certain hours.

What is the shape of hourly power consumption distribution during the day? Does it, indeed, resemble a Gaussian distribution? Is there a difference in hourly distribution between "seasons" obtained by seasonality detection in the previous task?

Check normality with normality tests (the Kolmogorov-Smirnov test), normalize distributions if needed; compare across hours. Visualising the distributions would be a plus.[1]

Weather influence In winter, colder temperatures would lead to more power consumption due to greater use of heaters. During the summer, warmer weather would have the similar effect due to greater use of the air-conditioner. Similarly, different life and eating habits in summer and winter could affect the kitchen power consumption.

- Use the data available for years 2007–2009 to establish the correlation(s) between power consumption and weather conditions.
- Based on this correlation, estimate the power consumption for the year 2010 from the meteorological data.
- Compare the obtained estimates with the power consumption data for 2010.

References

[1] Météo France: *Données SYNOP essentielles OMM.* Available at `https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32`. Downloaded on 28/02/2019.

[2] Georges Hébrail & Alice Bérard (2012): *Individual household electric power consumption Data Set.* UC Irvine Machine Learning Repository. Available at `https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption`.

[3] Sava Ristić (2018): *Data science approaches for city-wide power consumption analysis.* Master's thesis, Politecnico di Milano, School of Industrial and Information Engineering, Italy. Available at `https://www.politesi.polimi.it/handle/10589/139040`.

[1] If you prefer to use C++ for visualisation, one suggestion is to use the `matplotlib` library available from GitHub.