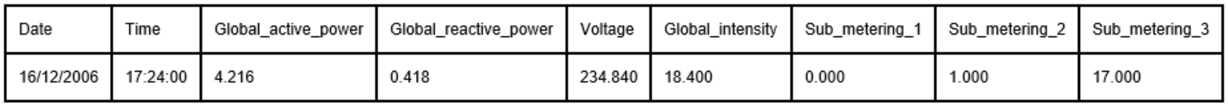
Team members: David Montes, Aaron Choi, Wil Carrasco

Team Name: The Ocho

Project Tile: Clustering Individual Household Electric Power Consumption and Future Consumption Regression Analysis

Description: Our group proposes to use the Individual household electric power consumption data set to look for power consumption trends over time. We plan on clustering the data using descriptive methods to discover patterns and trends. Applying predictive methods such as regression we plan to predict future power consumption.

https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption



Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
16/12/2006	17:24:00	4.216	0.418	234.840	18.400	0.000	1.000	17.000

Methodology: The dataset itself contains 2,075,259 power consumption measurements from a house in Sceaux, France between December 2006 and November 2010. The measurements were taken every minute and consist of the following:

- Date in dd/mm/yyyy
- Time in hh:mm:ss
- Global Active Power: household global minute-averaged power (kilowatt)
- Global Reactive Power: household minute-averaged reactive power (kilowatt)
- Voltage: minute-averaged voltage(volt)
- Global Intensity: household global minute-averaged current intensity (ampere)
- Sub Metering 1: Energy sub-metering which corresponds to the kitchen, containing a dishwasher, an oven, and a microwave. (watt-hour of active energy)
- Sub Metering 2: Energy sub-metering which corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator, and a light
- Sub Metering 3: Energy sub-metering which corresponds to an electric water heater and an air-conditioner

The first step in processing the data will be to convert the time-stamp into a single value. Since the data is provided as a .txt file, we split the data. The data and time will be combined and can be converted into a real value between 0 and 1 which represents the time of year. For example, the first minute of the year (Jan 1st 00:00) would be 0.00 and the end of march would be approximately 0.25… This way, the potential similarity between two data points from the same time of year will be reflected in the time attribute. During data processing, we will have to take into consideration the missing value in the measurements. There’s approximately 1.25% of the rows that contain missing data.

We would also normalize parameters Global Active Power, Global Reactive Power, Voltage, Global Intensity, and the Sub Metering 1 thru 3.

After completing pre-processing steps and normalization of data, we plan to use a partitional clustering approach implementing k-means and/or bi-kmeans.

The program proposal will be an application that can show cluster patterns appliance usage given a time of day. Additionally if there is time, we would like to apply regression to predict energy consumption usage. The input would be the time and the output would be the Active/Reactive Power.

By observing the patterns and trends from the clustered data, we can have a better understanding of the relationship between the time and power consumptions. If time permits, we could also observe trends and see future energy consumptions.