Time Series Forecasting For Energy Consumption

Computer Science Dept. COMPGI15: Information Retrieval and Data Mining

Rupert Chaplin – Artemis Dampa – Mégane Martinez

Context

The last few years, preserving the environment has become one of the priorities of most countries. Therefore, being able to monitor and, even more, predict energy consumption of households and, at a larger scale, entire areas is a powerful asset.

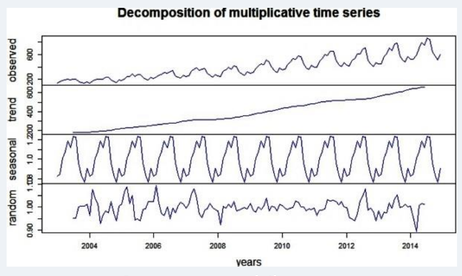
A time series is a sequence of values that represents the evolution of a quantity over time. By analysing the past behaviour and evolution of past energy consumption values, we aim to predict the future consumption at both a household and larger scale.

Problem Setting and Data

For this project we will use two data sets. The first contains the past electrical loads for 20 zones covered by a US electricity provider, as considered under the 2012 Global Energy Forecasting Competition (GEFCOM) 2012. The second contains the measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years.

Regression & ARIMA

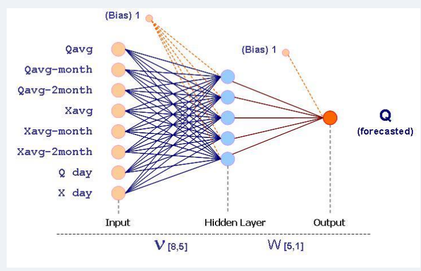
In order to complete or project we used two different methods. ARIMA is a regression based solution for time series forecasting that separates the different layers of the time series such as the trend, the seasonality and the residuals and then compute the predicted values by processing one layer at a time.



Source: ucananalytics.com

Neural Network Methods

We will also explore developing Neural Network approaches to forecasting. Such techniques allow for the consideration of a large number of parameters, and can be trained with gradient descent techniques to seek optimal fit to training data. Regulation or random drop out elements can be introduced to mitigate risk of over-fitting. Neural Nets have the advantage of easily introducing non-linear relationships in the model; as seen in our data exploration, it is likely that such non-linear effects are considerable in this data.

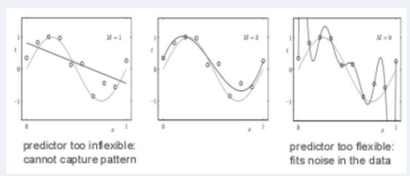


Source: Ahmad Shatnawi, University of Tabuk

Assessing model performance

Care needs to be taken to avoid overfitting to training data, which would risk designing models that will not generalise well to new data. We will use hold-out data for validation.

ARIMA and similar time series techniques also allow for the natural calculation of confidence bounds around forecasts.



Source: School of Informatics, University of Edinburgh

Data Exploration

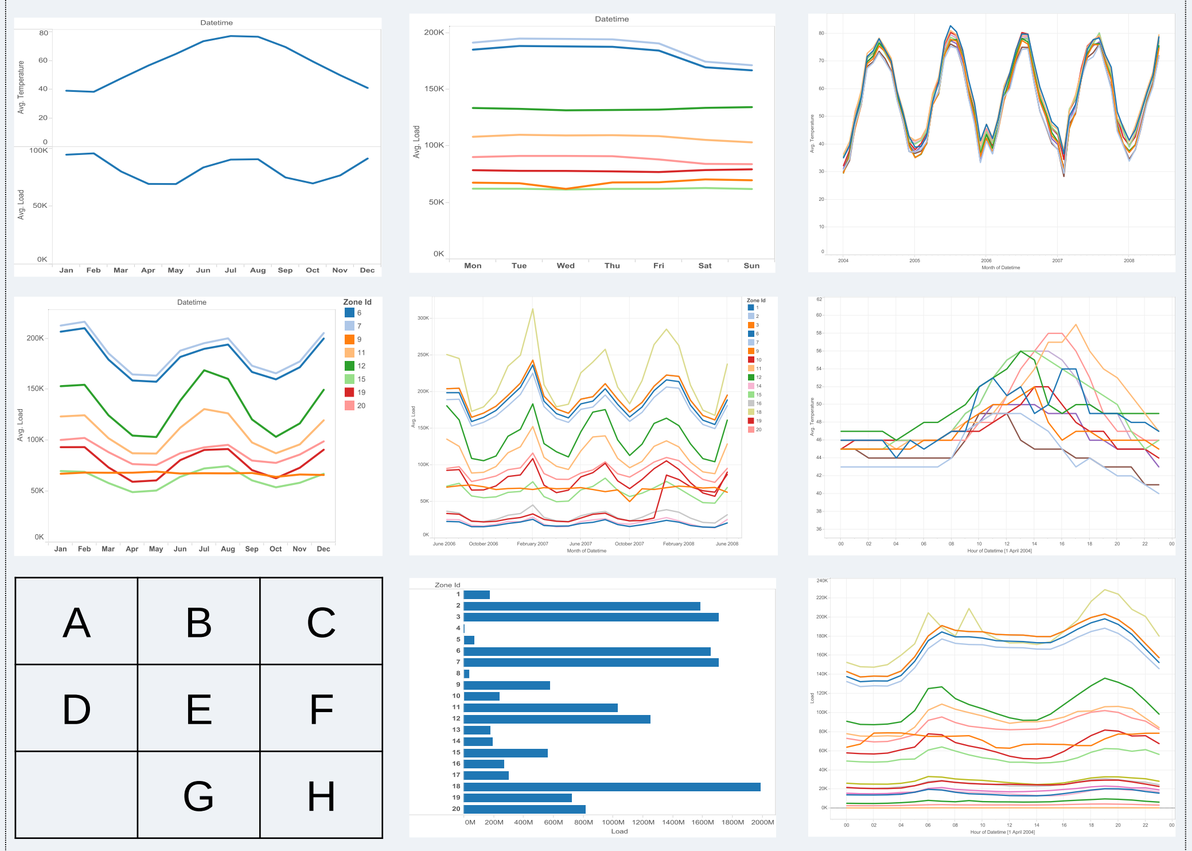
We conducted initial data exploration using Tableau, a product designed for quick visualisation of large datasets. We found this a useful tool to efficiently identify patterns and anomalies within the data, providing useful insights for the design of features to consider for future model development.

Household data

The household dataset contains measurements of the electric power consumption of one household gathered per minute over a period of four years from December 2006 until November 2010. The data contain three different sub-metering measurements; one of which corresponds to the kitchen containing mainly a dishwasher, an oven and a microwave, the other measures the laundry room consumption which contains a washing-machine, a tumble-drier, a refrigerator and a light, while the last one represents the water-heater and air-conditioner consumption. Our observations have shown that the water-heater and air-conditioner consumption is always higher than others'. Generally the consumption increases a lot in 2007 to drop a bit after some fluctuations in 2010. The general consumption of the house seems to decrease a lot every August. Depending on the day of the week each sub-metering measurement fluctuates differently. However, during the weekends the general consumption shows a small decrease. It is also observed that the consumption rises in the morning and late in the evening.

US Regional Energy Load

At the regional / zonal level, aggregate energy usage incorporates demand from a wide range of sources. A portion will be accounted for by the many individual households in a region, perhaps usage patterns of the type observed in the micro dataset. Additionally there will be industrial, commercial and large public sector users, with different patterns of usage. The allocation of energy across the different zones represented in the dataset is also an unknown: load patterns may be affected by infrastructural changes to the energy network itself, or by operational decisions made by the energy supplier. In this setting it is not practical to model / forecast energy usage from the micro level of individual usage patterns. Instead we will seek to identify means of predicting from within the aggregate time series, or from other exogenous variables where we can identify some predictive power on overall energy usage levels.



A – Clear seasonal pattern to temperature and energy usage. We anticipate a predictive relationship between temperature and energy, but this clearly it will not be linear. We will need to introduce non-linear features such as polynomial terms based on temperature.

B – Shows how average use varies across day of week. Different patterns evident for different zones. Model will need to allow day of week to have differential impact across zones. This observation also prompts us to consider if national holidays will have an impact.

C – Clear high correlation of weekly average temperature readings at different weather stations; but [F] considerable variation on a daily basis [chart shows variation over a single day]. H – variation in energy usage over the same day for different zones. Location of zones and weather stations are unknown; we will try to learn if certain weather stations are better predictors for certain energy zones.

D – Seasonal patterns evident in most zones, but different levels of impact. Perhaps due to different mix of energy consumers served (e.g. industrial vs residential).

E – Additionally, some clear anomolies / step changes for certain zones (e.g. change in pattern for zone 19). Model will need to account for these changes over time.

G – Considerable variation in overall load volumes by zone.

UCI Machine Learning Repository · kaggle · tableau · GEFCom2012 Load Forecasting