Forecasting - Homework 3 Final project

Penet, Sébastien b00639637 Jacomet, Hippolyte b00713284
Leblanc, Emilie b00529713

April 4, 2017

Abstract

This project aims to produce multivariate forecasts for the monthly U.S. gas consumption for the next 12 months (from July 2016 to June 2017). The data sources are the National Center for Environmental Information and Google Trends. We study 3 datasets:

- **gas.csv** containing the gas consumption and the log gas consumption from January 1973 to July 2016.
- **temp.csv** containing the different temperature metrics from January 1895 to February 2017.
- **gas.csv** containing the gas consumption and the log gas consumption from January 2004 to March 2017.

In this document, we quickly present our process and how to navigate between the different documents.

1 Process

In this study, we mainly predicted the log gas consumption in order to avoid the trusting effect of extreme values. Our predictions are often log predictions: when this is the case, we also provided the exponential in order to give the resulting gas consumption predictions.
The methodology followed in our study is the following:

1. Graphical analysis to understand our data
2. Do univariate predictions of gas consumption alone
3. Do multivariate predictions of gas consumption also with temp and google data
4. Select the most relevant temp and google features for predicting gas consumption and do univariate predictions of these selected features (aim: multiple linear regression)
5. Do a multiple linear regression to predict the gas consumption according to these selected features from temp and google

2 Where to find the documents

All the data used and our R output files (mainly R.markdown) can be found in the provided zip file and in the Git repository at the following link : https://github.com/leblacomenet/HW3.
The files are :

- The **data** file contains the initial datasets (**gas.csv**, **temp.csv** and **google.csv**) and the predictions of the necessary temperature (resp. google) features from March 2017 (resp. April 2017) to June 2017.
- All **graphical_analysis** documents (R markdown and pdf file) are the result of our first graphical analysis of our datasets and allows us to present the data.
- All **univariate_forecasts** documents (R markdown and pdf file) correspond to our univariate analysis and predictions of the gas consumption.

- All **multivariate_forecasts** documents (R markdown and pdf file) are the result of our multivariate analysis and predictions of gas consumption based on temp and google datasets.
- All **temperature_selection** documents (R markdown and pdf file) explain the process of selecting the most interesting temp and google features for our multiple linear regression predictions.
- All **univariate_forecasts_for_google** documents (R markdown and pdf file) correspond to our univariate analysis and predictions of the useful google features for our multiple linear regression analysis.
- All **univariate_forecasts_for_temp** documents (R markdown and pdf file) correspond to our univariate analysis and predictions of the useful temp features for our multiple linear regression analysis.
- All **multiple_linear_regression** documents (R markdown and pdf file) are the result of a multiple linear regression predicting the gas consumption according to either the actual temp and google data or their predictions when needed.
- The **final_predictions** file contains the **univariate_predictions.csv**, the **univariate_decomposition_predictions.csv** and the **multiple_regression_predictions.xlsx**, that is to say the results of our final models for the univariate processes (predicting consumption only based on consumption) and for the multiple linear regression process (predicting consumption based on the three datasets available).

3 Final predictions

Since the multivariate process does not yield good forecasts, our final predictions that we wish to present are:

- **univariate_predictions.csv**: predictions with a SARIMA(4,0,0)(0,1,1)[12] on log gas consumption.
- **univariate_decomposition_predictions.csv**: our predictions based on a decomposition via STL directly on the time series (not the log gas consumption). Then, we do an exponential smoothing on the seasonally-adjusted series and finally add the seasonality.
- **multiple_regression_predictions.xlsx**: our predictions of a multiple linear regression based on min_temp_val, heat_days_val and extreme_heat (computed dummy variable based on the google features). In the file, we provide both the predictions based on the entire historical data and based on two years of historical data.