Analysis and visualization of the Heating Degree Days in the Benelux states

László Kiss
Senior Data Scientist,
ETH, Ericsson Hungary

March 2019

Abstract

Heating Degree Days is a technical measurement for calculating the energy demand to heat a buildings. Space heating is responsible for a large fraction of energy use. Heating Degree Days is a proxy for energy demand to heat a home or a business. This technical number derived from measurement of outside air temperature. However, it is also depend on a large number of other factors, in particular building design, energy prices, income levels and behavioural aspects. Space heating is responsible for a large component of energy use, so a decrease in the use of space heating has the potential to lead to a significant decrease in overall energy use.

Analyzing the changes of the Heating Degree Days (HDD) can lead to an efficient energy use and it can support a better energy market strategy.

With the methodology of the multivariate time series analysis I will show the trend and the seasonality of the HDD time series data. Based on the historical data sets of the Eurostat, I will make a forecast and data visualization about this.

Introduction

Heating Degree Days is not just a measurement for an energy demand to heat a building, it is a measure of severity and duration of cold weather. The colder the weather in a given month, the larger the degree day value for that month. In essence this a difference between a reference or 'base' and the outside temperature. The define of a base temperature based on outside temperature which is not required to operate a building heating systems. The exact value varies from building to building depending on the characteristics of the building and on its use.

The analysis of the Heating Degree Days and reveal the new connections in this dataset can support a better strategy in the energy market and more efficient performance in the use of energy to heat buildings.

Data sources

This article and the related analysis based on dataset from Eurostat, which is the statistical office of the European Union. The start of the dataset is January of 1975 and it's frequency is monthly. By geographically the monthly value of Heating Degree Days is ordered to NUTS2 regions in European Union.

The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU for the statistical and other purposes.

Definition of Heating Degree Days

The calculation of HDD relies on the base temperature, defined as the lowest daily mean air temperature not leading to indoor heating. The value of the base temperature depends in principle on several factors associated with the building and the surrounding environment. By using a general climatological approach, the base temperature is set to a constant value of 15 °C in the HDD calculation.

$$If\ T_m \leq 15°C \text{ THEN HDD} = \sum_i (18°C\ -\ T_i^m) \text{ ELSE HDD} = 0$$

where T_i^m is the mean air temperature of day i.

The methodology of the analysis of the HDD in Benelux States - Multivariate Time Series

The Benelux states - Belgium, Belgium, Netherlands and Luxembourg are divided to 24 NUTS2 regions. The dataset from the Eurostat contains the HDD values related to all regions so we can handle this as multivariate time series. The geographical and weather conditions are same in this countries, we can consider these states as a geographical unit.

During the analysis I will examine the inner connections and I try to build a connection between the values of the HDD and the energy consumption in this area.

Against the univariate the multivariate time series has more than one time dependent variable. The variables depends not just on their past values but also has some dependency on other variables.

Vector Auto Regression model is the one of the most commonly used methodology for the multivariate time series analysis and forecasting. In a VAR model the each variables has a linear function of the past values of itself and the past values of all other variables. For instance

HDD in Benelux states

Time	Belgium	Luxembourg	Netherlands
1990-01-01	2662.83	2951.31	2581.58
1991-01-01	3045.18	3310.82	3017.65
1992-01-01	2780.03	3019.43	2720.82

If we consider the Belgium time series values as a y_1 variable and the Luxembourg is y_2

y1	y2
$y1_{t-n}$	$y2_{t-n}$
$y1_{t-2}$	$y2_{t-2}$
$y1_{t-1}$	$y2_{t-1}$
$y1_t$	$y2_t$

we can calculate with this mathematical formula.

$$Y_1(t) = a_1 + \omega_1 11 * y_1(t-1) + \omega_1 12 * y_2(t-1) + e_1 * (t-1)$$

$$Y_2(t) = a_2 + \omega_1 21 * y_1(t-1) + \omega_1 22 * y_2(t-1) + e_2 * (t-1)$$

Here,

- a_1 and a_2 are the constant terms,
- ω_{11}, ω_{12}, ω_{21}, and $\omega_2 2$ are the coefficients,
- e_1 and e_2 are the error terms

This equations is very similar to Auto Regression process, like AR(1):

$$y(t) = a + \omega * y(t-1) + e$$

In this case, we have only one variable – y, a constant term – a, an error term – e, and a coefficient – w. In order to accommodate the multiple variable terms in each equation for VAR, we will use vectors and we can write equation in a following form:

$$\begin{bmatrix} \text{y1(t)} \\ \text{y1(t)} \end{bmatrix} = \begin{bmatrix} \text{a1} \\ \text{a2} \end{bmatrix} + \begin{bmatrix} \omega 11 & \omega 12 \\ \omega 21 & \omega 22 \end{bmatrix}$$

$$* \begin{bmatrix} \text{y1(t-1)} \\ \text{y2(t-1)} \end{bmatrix} + \begin{bmatrix} \text{e1(t)} \\ \text{e2(t)} \end{bmatrix}$$

The two variables are y1 and y2, followed by a constant, a coefficient metric, lag value, and an error metric. This is the vector equation for a VAR(1) process. For a VAR(2) process, another vector term for time (t-2) will be added to the equation to generalize for p lags:

$$\begin{bmatrix} \text{y1} \\ \text{y2} \\ . \\ \text{yk} \end{bmatrix} = \begin{bmatrix} \text{a1} \\ \text{a2} \\ . \\ \text{ak} \end{bmatrix} + \begin{bmatrix} \text{w11} & . \\ \text{w21} & . \\ . & . \\ \text{wk1} & . \end{bmatrix}$$

$$* \begin{bmatrix} \text{y1(t-1)} \\ \text{y2(t-1)} \\ . \\ . \\ \text{yk(t-1)} \end{bmatrix} + ... \begin{bmatrix} \text{w'11} & . \\ \text{w'21} & . \\ . & . \\ \text{w'k1} & . \end{bmatrix} * \begin{bmatrix} \text{y1(t-p)} \\ \text{y2(t-p)} \\ . \\ . \\ \text{yk(t-p)} \end{bmatrix} + \begin{bmatrix} \text{e1(t)} \\ \text{e2(t)} \end{bmatrix}$$

The above equation represents a VAR(p) process with variables y1, y2 ... yk. The same can be written as:

$$\begin{bmatrix} \text{y} \end{bmatrix} = \begin{bmatrix} \text{a1} \end{bmatrix} + \begin{bmatrix} \text{w1} \end{bmatrix}$$

$$* \begin{bmatrix} \text{y1(t-1)} \end{bmatrix} + ... \begin{bmatrix} \text{wp} \end{bmatrix} * \begin{bmatrix} \text{y1(t-p)} \end{bmatrix} + \begin{bmatrix} \text{e} \end{bmatrix}$$

$$y(t) = a + \omega_1 * y(t-1) + ... + \omega_p * y(t-p) + \epsilon * t$$

The term ϵ_t in the equation represents multivariate vector white noise. For a multivariate time series, ϵ_t should be a continuous random vector that satisfies the following conditions:

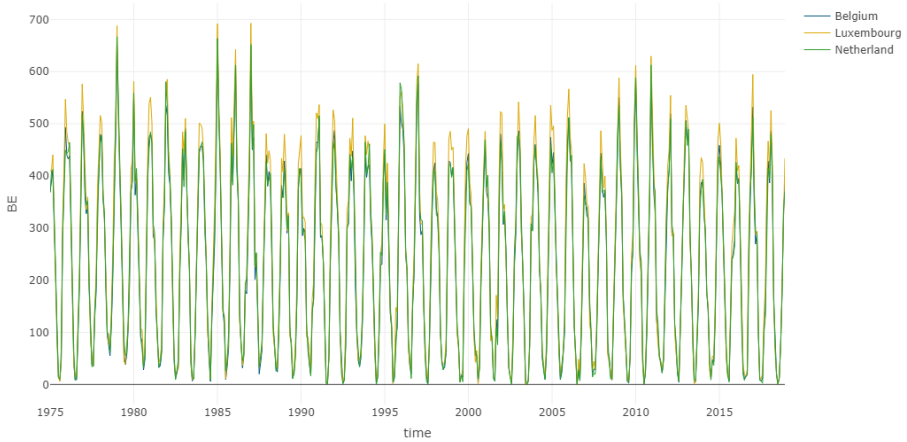
1. $E(\epsilon_t) = 0$
   Expected value for the error vector is 0

2. $E(\epsilon_{t1}, \epsilon_{t2}) = 0$
   Expected value of ϵ_t and ϵ_t' is the standard deviation of the series

From the equations 1 and 2 it is clear that each variable use it's previous values to make a predictions. Unlike AR, VAR is able to understand and use the relationship between several variables.

Time series data of Heating Degree Days in Benelux

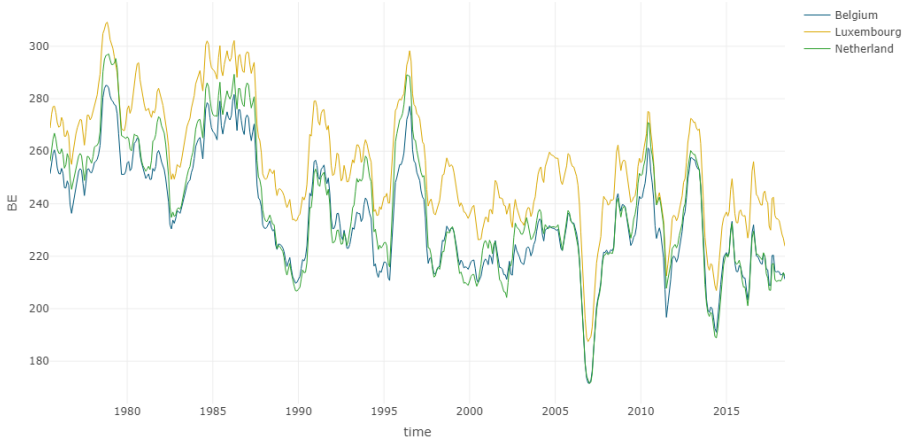
The data collection started in 1975 and it is available in yearly and monthly frequency.



There is an important technique for all types of time series data called decomposition. It seeks the construct of the data from an observations. Usually time series decomposed into:

- T_t, the trend component at time t which is refers to a long term progression.
- C_t the cyclical component at time t which reflects repeated but non-periodic fluctuations
- S_t the seasonal component at time t, reflecting seasonality (seasonal variation).
- $I_t$the irregular component (or "noise") at time t, which describes random, irregular influences.

For the forecasting I use the auto.arima function for the all column in the dataset, and after I select the common order for the After the decomposition we can see the trend in our data set.



Forecasting with VARMA model

Vector autoregressive moving-average (VARMA) processes are suitable models for producing linear forecasts of sets of time series variables. A vector sequence y(t) of n elements is said to follow an n-variate ARMA process of orders p and q if it satisfies the equation:

$$A_0 y(t) + A_1 y(t-1) + .. + A_p y(t-p) = M_0 \epsilon(t) + M_1 \epsilon(t-1) + .. + M_p \epsilon(t-p)$$

where A_0, A_1, ... ,A_p, $M_0\epsilon$(t), $M_1\epsilon$(t-1), $M_q\epsilon$(t-q) are matrices of order n $\times$ n and ϵ(t) is a disturbance vector of n elements determined by serially-uncorrelated white noise processes that may have some contemporaneous correlation.