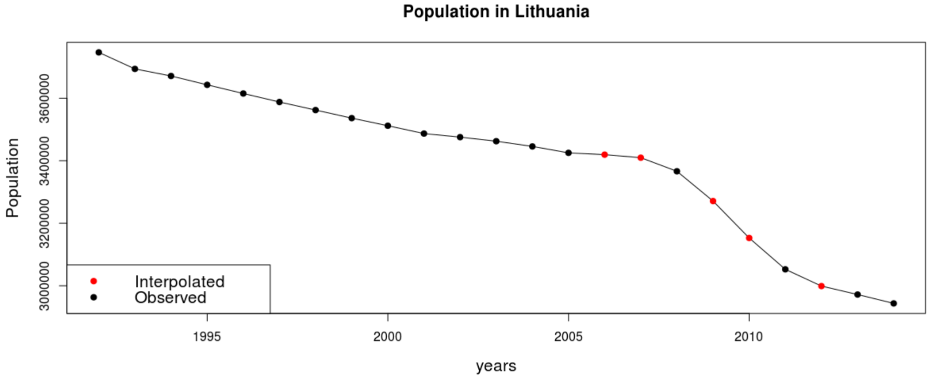
Milestone IV - Application

Computing Project 2016/17
14D003/14D004

Carlos Isaac Rodriguez Prado
Hans-Peter Höllwirth
Veronika Kyuchukova

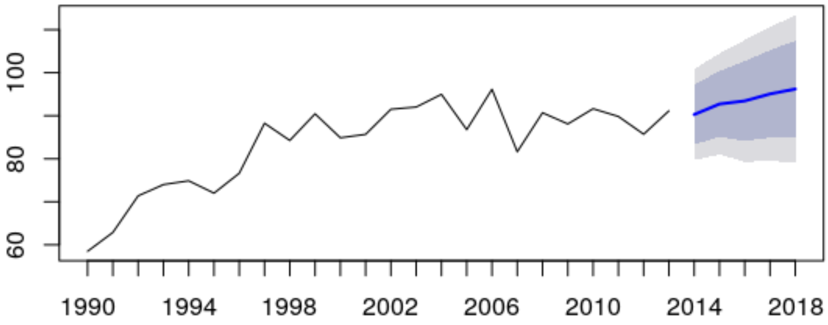
The analytics part involves 3 steps: **[1]** data cleaning, **[2]** descriptive analysis of potential correlations between pollutant emissions, population size, and measured pollutant concentrations, and **[3]** predictive analysis which focuses on the prediction of future pollutant concentrations.

[1] In a first step, we interpolate missing data points from the population dataset (both city and national level) in order to make effective use of the data in the regression and forecasting analysis. We use ***spline interpolation*** (rather than polynomial interpolation) to avoid oscillation. The following graphs gives an interpolation example:



[2] Having annual concentration data for around 10-20 years (depending on the pollutant type and country) and corresponding annual emission quantities, broken down into roughly 150 different sectors for most European countries (though most countries report the numbers for only 50-70 different sectors), we want to find potential correlations between the emissions in different sectors and the reported concentrations. In a first step, we apply country-wise ***LASSO regression*** in combination with cross-validation to identify the most relevant variables (population or emission sectors). In most countries, less than 5 variables are selected (lowest MSE criteria). In a second step, the selected variables are used for **OLS regression**. However, in hardly any country variables prove to be significant and the R-squared values are generally low (<0.10), implying that neither population nor reported emission numbers explain the pollutant concentrations well.

[3] The result in the correlation analysis suggests that we can not predict future pollutant concentrations based on emission or population data. Hence, we concluded to apply time-series analysis on directly on the annual pollutant concentration dataset. We use **ARIMA models** for this purpose (selected based on the Akaike criterion). The integrating order is chosen with the KPSS test. An example for a 5-year forecast with error bars is shown below:


60
80
100
1990
1994
1998
2002
2006
2010
2014
2018