Developing a Probabilistic Graphical Model for fine particles estimations
Prof. A. Martinoli

This project is part of the OpenSensoll project between EPFL, EHZ and Naneos Particle Solutions GmbH. The objective is the estimation of \( \text{PM}_{10} \) concentrations using mobile sensing platforms placed on top of buses. Models have already been implemented to estimate concentrations in locations that already have measurements scattered over time.

**Data**

**Land use parameters**
- Population density
- Building density
- Energy sources used for heat and hot water
- Slope
- Altitude
- Industries

**Traffic**
- Counting points from a TRANSITEC survey in 2010

**Meteorological data**
- One value for all Lausanne with a time resolution of 10 mins
- NABEL station: CO, NO, NO2 and O3 concentrations
- Pully station: rain, temperature, radiations, wind velocity and humidity

**OpenSense data**
- Daily averages from buses measurements over the year 2014

**Log-linear regression**

A Log-linear Regression was performed in order to estimate \( \mu \) and \( \sigma \) and thus build a probability distribution for each street segment without measurements.

\[
\log \mu = \alpha_0 + a \log P_1 + b \log P_2 + \cdots + n \log P_n \\
\log \sigma = \alpha_0 + a \log P_1 + b \log P_2 + \cdots + n \log P_n
\]

The parameters \( P_i \) correspond to the space-dependent variables (traffic and land-use), thus it is a static model.

One regression per class was performed and thus a set of coefficients was computed for each class.

- \( \hat{\sigma} \) from \( \mu \) estimation
- \( \hat{\sigma} \) from \( \sigma \) estimation

The probability distributions could then be used as inputs in the Probabilistic Graphical Model.

**Problem statement**

The objectives for this project are:

1. Extend predictions outside the buses network (blue Network in the figure)
2. Use a Probabilistic Graphical Model (PGM)

In order to estimate the \( \text{PM}_{10} \) in the blue network, we first had to build probability distributions.

To do so, the streets segments with \( \text{PM}_{10} \) measurements were used to build a Log-linear Regression and determine the coefficients.

To facilitate the process, the area of Lausanne was clustered into 10 categories, the Urban Ecosystems. These categories were built according to the environmental characteristics of the streets (land use parameters and traffic). A regression was performed on each category.

Once the probability distributions were available for all street segments, the Probabilistic Graphical Model was used to predict the values to complete the time-space continuum.

**Probabilistic Graphical Model**

The objective of the PGM is to fill the missing \( \text{PM}_{10} \) locations-times. The nodes represents the streets segments. The edges link \( \text{PM}_{10} \)-correlated streets. Meteo stations are also included in the network.

A MATLAB code was used to run the Loopy Belief Propagation, an algorithm that predict values for missing locations-time. The inputs are the \( \text{PM}_{10} \) marginal probability distributions (2D blue curves) for each street and the joint probability distribution (3D red curves) for \( \text{PM}_{10} \)-correlated streets and meteo stations.

The R2 were calculated for each street segment. 75% of the distribution are negative outliers, which mean an overestimation of the predicted values (left boxplot) Apart from these outliers, the results are quite satisfying, with a median of 0.75.