nano-tera.ch
swiss scientific initiative in health / security / environment systems

inUse Strategic action

FNSNF

Towards Real-Time Estimation of Urban Air Pollution with Region-Based Models

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Arnaud Jutzeler, Jason J. Li, Boi Faltings

Artificial Intelligence Laboratory, EPFL

1. The Purpose

- Estimating Exposure
"Sensing the air we breathe"
- Estimating Emission
"Where do the bad air come from"

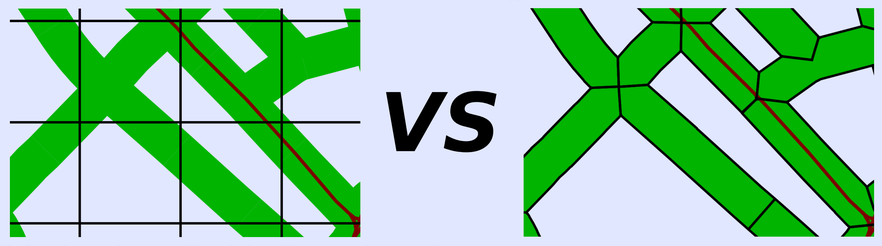
To do that we need to derive prediction maps with sufficiently high temporal and spatial resolutions.

2. The Data

Mobile station mounted on trams in Zürich measuring every 4 seconds the ultra-fine particles (UFP) concentration.

Zürich tram network is shown in red, the roads for which we have traffic data in green and the Nabel reference station in blue.

3. The Spatial Partitioning

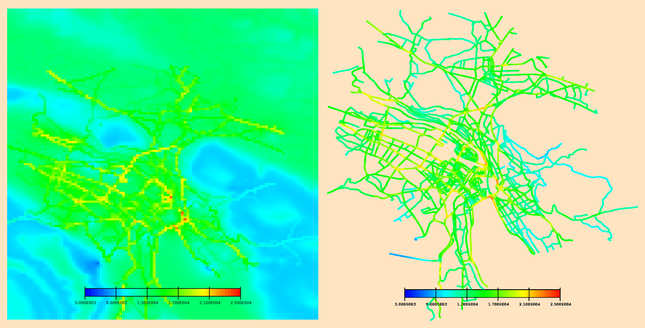


Motivations behind road regions:
- People are on the streets!
- Road traffic = main UFP sources
- Ambient concentration supposedly more homogeneous
- Produces more robust short-term aggregates as trams follow the roads

4. Spatial Regression on Long-term Aggregates

We first derived spatial prediction maps for average concentration for long time window such as years, seasons, months, weeks and days.

After some pre-processing, all the measurements taken within a specific time window were aggregated temporally and spatially using the hectares grids or regions. Then standard Gaussian process regression over those aggregates was performed using the following spatial features:
- Position
- Traffic density
- Building density
- Population density
- Heating types
- Topography (altitude, slope, exposition)
- Emissions inventory

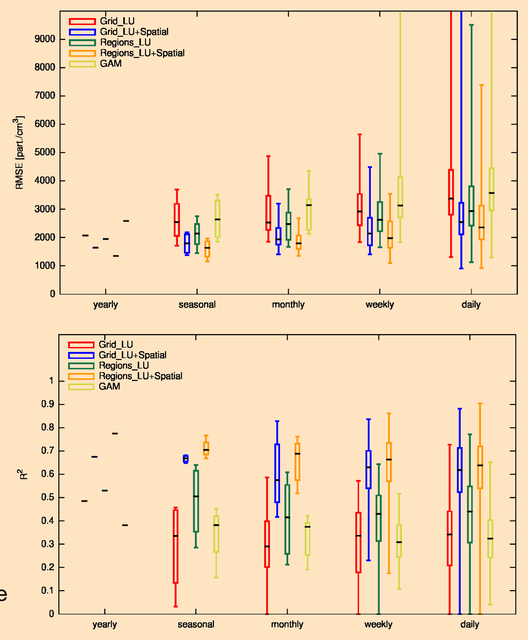


Example of yearly prediction maps for both hectare grid cells and road regions produced by Gaussian process regression on spatial features

Cross-validation using a random 10-fold testing scheme over 200 aggregates

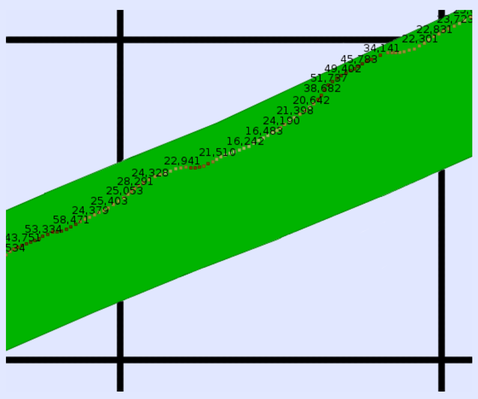
- Performance gain by adding spatial distance between the regions in the covariance structure [1]
- Region-based models perform better than hectare-based ones [2]
- The shorter the time window, the higher variation on performances

GAM denotes the Generalized Additive Model given in [3].



5. Spatial Regression for Real-time Data

Concentration also varies throughout the day so we would like to derive prediction maps for shorter periods. However the shorter the time window used to average, the higher the variation in the land-use regression performances (thus the worst-case). This become critical when we try to reason in real-time that is to say very short time intervals such as 30 minutes or 1 hour long. At this scale, the coverage is not as good but it is still sufficient (we can still easily gather up to 200 different regions). The main problem actually resides in the number of measurements available in each regions to compute the averages.

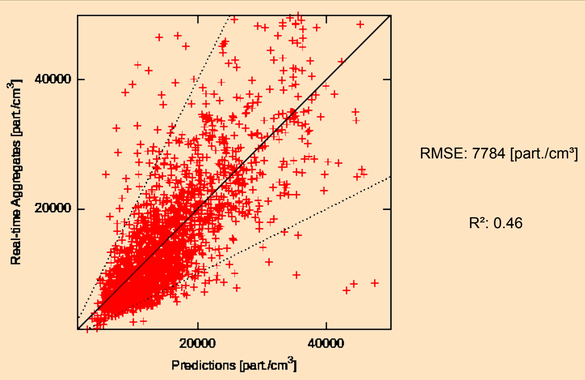


Example of successive UFP measurements taken on the same street with an interval of one second

Indeed, UFP measurements can have high variations at very small temporal and spatial scales. This is mainly due to the fact that **the sensors are very close to the sources**. In this setting, aggregates computed only from few measurements must be discarded or at least be weighted at models level. We achieve this in GP regression by adding an extra variance term for low count aggregates. We note that using a symmetric partition like hectare grid cells would result in less robust real-time aggregates. Also the length of the roads regions could be adjusted as to solve the trade-off between the strictness of the homogeneity assumption within regions and the robustness of the real-time aggregates.

6. Towards better Real-time Models

To improve the performance of real-time models, we want to counterbalance the lack of coverage and the bad quality of the real-time aggregates by using potentially beneficial informations contained in the historical datasets.
We want our models to exploit similarities between the UFP maps based on temporal features such as:
- Hour of the day
- Public holidays
- Meteorology (temp., humidity, wind, rain, ...)
- **Fixed reference stations**



2-fold chronologically ordered cross-validation of a temporal model on 30min aggregates **in a single region** gathered over 6 months with at least 30 measurements

In [3], short-term aggregates were augmented with historical measurements that shared similar temporal variables. Different combinations of temporal variables with different tolerance values were tried.
Our line of inquiry is learn more informative priors using GP regression on historical real-time aggregates. It has the advantage to be automatic in a sense that the relevance of every temporal variables is learned automatically as opposed to [3]. The first option is to consider the temporal and spatial processes as two separable GP processes, the other, more expensive, to consider a unique spatial-temporal GP process.

7. Conclusion

- Region-based models are appropriate to derive UFP prediction maps in urban environment.
- Modelling real-time UFP concentration is very challenging because of those small scale variations.
- The same approach applied to other pollutants that are more diffused could lead to better results.
- Some more work is still required to validate our different real-time models.

8. The Software Framework

We designed a modular and generic framework to carry out those experiments. It supports simple GP regression to state-of-the-art approximations. Applying it to other pollutants and other cities once the data are gathered would be straight-forward. As it is Java-based, it could also be simply packaged as a module of the GSN platform [4]. It also offers an XML API and it exploits multi-core architectures.

9. References

[1] J. J. Li, A. Jutzeler, B. Faltings
"Estimating Urban Ultrafine Particle Distributions with Gaussian Process Models" in Research@Locate 2014

[2] A. Jutzeler, J. J. Li, B. Faltings
"A Region-Based Model for Estimating Urban Air Pollution" in AAAI 2014

[3] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, L. Thiele
"Pushing the spatio-temporal resolution limit of urban air pollution maps" In PerCom 2014

[4] K. Aberer, M. Hauswirth, A. Salehi
"A middleware for fast and flexible sensor network deployment" In VLDB 2006