Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

A Region-Based Model for Estimating Urban Air Pollution

Arnaud Jutzeler[1], **Jason Jingshi Li**[2], **Boi Faltings**[1]
[1]Ecole Polytechnique Federale de Lausanne
[2]The Australian National University

Abstract

Air pollution has a direct impact to human health, and data-driven air quality models are useful for evaluating population exposure to air pollutants. In this paper, we propose a novel region-based Gaussian process model for estimating urban air pollution dispersion, and applied it to a large dataset of ultrafine particle (UFP) measurements collected from a network of sensors located on several trams in the city of Zurich. We show that compared to existing grid-based models, the region-based model produces better predictions across aggregates of all time scales. The new model is appropriate for many useful user applications such as exposure assessment and anomaly detection.

Introduction

The prevalence of urban air pollution is a major concern in both the developed and developing world. High levels of surface-level air pollutants are responsible for a range of respiratory and cardiovascular diseases, and an estimated 3.2 million people died prematurely from air pollution in 2010 according the Global Burden of Disease Study (Lim et al. 2013). Certain air pollutants, such as ultrafine particles (UFP), have high spatial variability in the urban environment, and thus it is in the public interest to develop detailed mapping of pollutants in order for scientists and governments to evaluate public exposure and develop new policies to minimise harm. Currently, major air pollutants are typically monitored by networks of static stations funded and operated by government authorities, collecting highly reliable and accurate measurements on a continuous basis. However, these stations are also costly to acquire and maintain, which results in limited information being collected about the spatial distribution of air pollutants. Recently, smaller and more affordable, albeit less accurate air quality sensors are becoming increasingly available to the market. This, coupled with an increased awareness of urban air pollution, creates the need for models to interpret measurement data and provide reliable estimations in a community-sensing setting (Krause et al. 2008; Aberer et al. 2010).

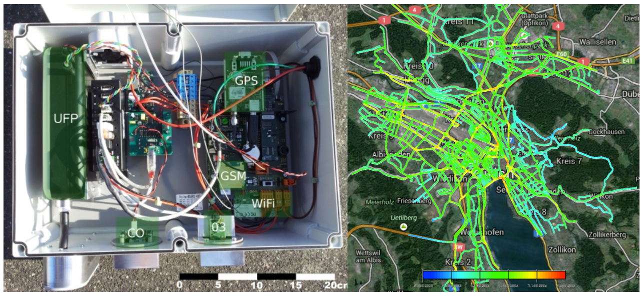

Figure 1: Left: A typical OpenSense sensing node from the trams; Right: Annual estimates of UFP levels from our model

Tradition air-quality modelling uses first principles to simulate the actual physical dispersions of air pollutants. The model typically uses emission statistics and meteorological variables as inputs to a system of physics equations, and deterministically derives the expected pollutant concentration for different grid-cells in the map. The current state-of-the-art mesoscale models have resolution with grid cells of around 1-2 km^3. The purpose of these models is typically to allow policy makers to simulate the outcomes of changing emission scenarios as a result of change in public policy. More recently, statistical, data-driven models have been developed for the purpose of analysing population exposure. A popular variant of such models, known as land-use-regression (LUR), uses land-use characteristics of the grid cells, such as average building density or proximity to a major road, as features for which they train a parameterized model to predict the pollutant levels.

In this paper, we propose a novel region-based approach for developing data-driven urban air pollution models. Instead of building model for predicting grid-cells, we partition the urban environment into regions of supposedly homogeneous emission: road segments with consistent traffic volume within the region. We then construct a Gaussian process using the spatial positions and land-use characteristics to estimate the average pollution level within these regions. Our approach is similar to existing land-use regression, but it is a non-parametric method that also considers the spatial nature of the phenomenon. We implemented and applied our model to a dataset of UFP measurements of an entire year across the city of Zurich (Fig. 1). We show that this region-based approach produces a more accurate estimation maps across all temporal scales, and it allows users to avoid exposure to abnormally high pollution sites.

The paper is arranged as follows. We first introduce the background of existing models for estimating the spatio-temporal phenomena in the urban environment, the Zurich UFP dataset that we used for our experiments, and the Gaussian process regression (GPR) framework on which our models are based. We then introduce the region-based model that we developed for estimating the UFP levels in the greater Zurich metropolitan area, and evaluate it against state-of-the-art grid-based land-use regression approaches using generalized additive models (GAM) and GPR.

Background

Dispersion of air pollution is a well studied topic in environmental science. The traditional dominant approach is to reconstruct a complete picture of air pollution for a given urban area base on a set of physical and chemical equations, which describes the behaviour of the target pollutants within different grid cells (Godish 2003). The goal of these models is to determine the relation between the source of emission and ground-level pollutant concentrations. These physical models typically do not directly use any air quality measurements, though measurements may be used for tuning or validating the models. As the size of the grid cells are typically 1km^2 or larger, the measurement stations are required to be located at sites where the measured pollution level would typically reflect the average pollution level across the grid cell. Hence measurements are typically located at building roof tops or large parks, where no emission sources may introduce bias to the measurements.

By contrast, statistical approaches typically construct a model based on a dataset of measurements from a few static measurements stations. Instead of determining the relation between source and measurement, it simply uses measurements to build estimations of the average pollution that people are exposed to in the grid cells. However, given that the measurements are typically located away from the streets, cars and other emission sources, they are also away from where the people are. Hence their output may underestimate the actual exposure of people to air pollution.

The Zurich UFP Dataset

Since 2011, trams in Zurich, Switzerland have been fitted with sensing nodes measuring ambient ozone and UFP levels. By 2014, ten trams are equipped with the sensing nodes and operate inside the greater Zurich metropolitan area. Their installation and coverage is shown in Fig. 2. In contrast with traditional measurement equipment in static stations, these deployments use sensors that are much more affordable, energy-efficient and more mobile, but consequently produce less accurate and reliable measurements. The location of the sensor is also very different to the traditional setting in literature, as they are placed on top of trams that may run very close to emission sources that would introduce measurement bias and high small-scale variability. Therefore, the data-driven statistical models from literature may not be applicable, and more robust methods may be needed to produce reliable estimations. A more complete description of the dataset can be found in (Li et al. 2012; Hasenfratz et al. 2014).

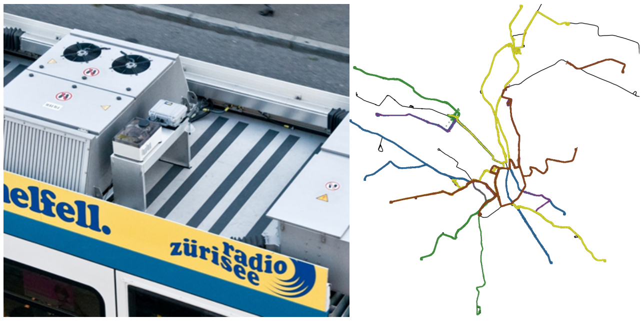

Figure 2: Left: A deployed sensing node on top of a tram in Zurich; Right: the map of sensor measurements collected in one week, the colors denote different sensor nodes.

Land-Use Regression

Recently, a popular data-driven approach, known as land-use regression (LUR), has been applied to assess the spatial distribution of airborne particulate matter in the urban environment (Hoek et al. 2008). In a typical measurement campaign over 1-2 weeks, 20-100 air pollution monitoring sites are set up across the study area. A model is developed using explanatory variables of various grid cells obtained from public geographic information systems. The explanatory variables consist of land-use information of the grid cells, such as traffic density, population density, proximity to highways, altitude and slope.

To study the spatial and temporal variability, an equation is then used to model the relationship between the pollution level (p) and the set of explanatory variables $\{A_1, \ldots, A_n\}$:

$$ln(p) = a + s_1(A_1) + s_2(A_2) + \cdots + s_n(A_n) + \epsilon$$

This allows predictions be made on grid cells where measurements are absent. The model is typically validated through standard random 10-fold cross-validation. A validated model can then be used to assess the ambient exposure of people located in the grid cells.

Gaussian Process

Our model uses Gaussian progress regression (GPR) for learning about the spatial phenomenon and making predictions. GPR is a non-parametric approach that has been successfully applied in the last decades to various fields. Originally known as kriging, it has especially been used in the geostatistics community to model phenomena such as soil concentrations, weather-related or even pollutant concentration at lower scale (Cressie and Cassie 1993; Rasmussen and Williams 2006). The GPR framework is still nowadays a very active research topic as evident in recent works such as (Bonilla, Guo, and Sanner 2010), (Cao et al. 2013) and

(Nguyen and Bonilla 2014). Advantages of this approach are numerous:

- It does not require any particular prior structural knowledge on the modelled function.
- It provides a value of certainty on the predictions.
- It incorporates native mechanisms to handle noisy data.
- It allows the automatic determination of the input variables relevance.

Technically, a Gaussian process (GP) is the generalization of a multivariate normal distribution to an infinity of random variables. It is defined by a mean function $m(\mathbf{x})$ and a covariance function or kernel $k(\mathbf{x}, \mathbf{x}')$. It can be viewed as a distribution over whole functions where every random variable represents a value of the function of interest f at specific point. We note:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

The idea behind GPR is to start with a prior Gaussian process that encodes the prior knowledge that we might have on the covariance structure of the function $f(\mathbf{x})$. We know from GP's marginalization property that the values $f(\mathbf{x}_i)$ of any set of points $\mathbf{x}_i \in X$ are sampled from a multivariate normal distribution that has its the mean vector drawn directly from the mean function m, and the covariance matrix directly drawn from the covariance function k. Now let us say that we have two sets of points X and X_* and that we make noisy observations $\mathbf{y}$ of the function f at the points X. We write:

$$p(\mathbf{y}|X) = \mathcal{N}(m(X), k(X, X) + \sigma_n^2 I)$$
$$p(\mathbf{f}(X_*)) = \mathcal{N}(m(X_*), k(X_*, X_*))$$

where σ_n^2 is the additional noise on the observations. What we seek is the distribution of $\mathbf{f}(X_*)$ given the observation. Such distribution $p(\mathbf{f}(X_*)|X_*, X, y)$ known as the predictive distribution is given by conditioning the normal joint distributions. For a single points $\mathbf{x}_* \in X_*$ it gives us the following equations:

$$\begin{aligned}\bar{f}(\mathbf{x}_*) &= m(\mathbf{x}_*) \\ &\quad + k(\mathbf{x}_*, X)(k(X, X) + \sigma_n^2 I)^{-1}(\mathbf{y} - m(\mathbf{x})) \\ \mathbb{V}[f(\mathbf{x}_*)] &= k(\mathbf{x}_*, \mathbf{x}_*) \\ &\quad - k(\mathbf{x}_*, X)(k(X, X) + \sigma_n^2 I)^{-1} k(X, \mathbf{x}_*)\end{aligned}$$

The main challenge with this approach is learn adequate covariance functions. This can be achieved by learning (hyper-)parameters contained in some families of covariance functions. We describe in the following section what kinds of covariance functions were used and how their parameters were learned by maximizing the marginal likelihood (ML). For additional information on the subject of GPR we suggest the reader to refer to (Rasmussen and Williams 2006).

A Region-Based Model

The Region Partition

A drawback of previous models for evaluating population exposure to air pollutants is that people are not usually found at sites with the average ambient concentration such as building roof tops, but at sites close to emission sources such as footpaths adjacent to traffic. This means that even though the ambient concentrations predicted by the model can closely match the measurements traditionally made in static stations, they can still systematically underestimate the actual exposure. Furthermore, as the models are based in grid cells of uniform size, there is an issue of reconciling the scale of the phenomenon: If the grid cells are too large, then there can be considerable variance within the grid cell to render the average concentration useless for practical purposes; if the cells are too small, then they may be overtly influenced by small-scale events, and there may be insufficient data in each grid cell to make any useful estimates.

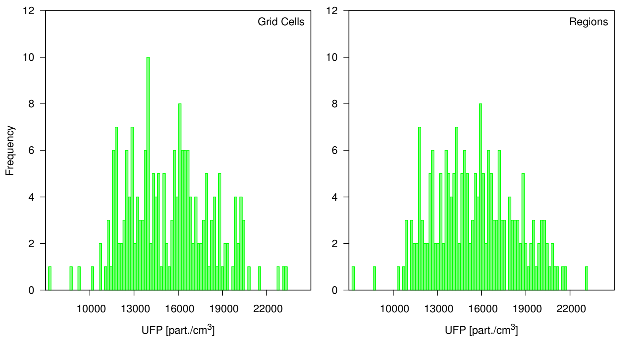

Figure 3: Distribution of the annual average UFP concentrations of (left) the top 200 grid cells with most measurements; and (right) the top 200 regions with most measurements

To counter these issues, we propose an approach that instead of using uniform grid cells as the fundamental spatial unit, we use regions of homologous emissions. Given that most of the emission come from traffic, we divided the space with road segments with supposedly homologous traffic density. This is done by parsing road segments from the daily traffic data from the office of Canton of Zurich. We first merged all the segments with the same traffic density value that touch each other. Then we cut them at every major intersection. Finally we cut the resulting segments in equal parts as to obtain no segment longer than 100 meters. The measurements are then aggregated by associating every measurement to the closest segment with a tolerance of 20 meters thus creating a valid space partitioning. Our regions are in fact the areas around segments that represent all the points that are closer to a given segment than any other with a maximum distance of 20 meters. The distribution of the annual average UFP concentration is shown in Fig. 3.

The LU Prior

Motivated by replicating the land-use regression from (Hoek et al. 2008; Hasenfratz et al. 2014) we first developed a GP model that reasons exclusively in terms of land-use variables. We note $\mathbf{x}_{LU}$ the vectors taken as input by this model that contain such variables. As our goal was to specify into the model as little a priori structure as possible only two very simple prior mean functions were considered: the trivial 0-mean function, and a constant function c. Concerning the covariance function, we only addressed stationary kernels that is to say kernels for which the stationary assumption $k(\mathbf{x},\mathbf{x}') = k(\mathbf{x}-\mathbf{x}')$ holds. We tried several well-known families of stationary kernels such as the squared exponential, Matérn, and piecewise polynomial functions. Each of these functions carries different assumptions on the smoothness of the modelled function. Preliminary tests showed that the prior GP that gives the best overall performance was the combination of the constant mean function and the squared exponential covariance function. Then our pure land-use prior is defined by:

$$m(\mathbf{x}_{LU}) = c$$
$$k(\mathbf{x}_{LU},\mathbf{x}'_{LU}) = \sigma^2_{f_{LU}} \exp\Big(-\frac{1}{2}D^\top MD\Big)$$

where $D = \mathbf{x}_{LU} - \mathbf{x}'_{LU}$, and $M = \text{diag}(\boldsymbol{\ell}_{LU})^{-2}$. $\sigma_{f_{LU}}$ is the magnitude hyper-parameter and $\boldsymbol{\ell}_{LU}$ contains one length-scale hyper-parameter per input feature. We note that among all the covariance functions that were tested, the squared exponential function is the one that carries the strongest assumption on the smoothness of the process.

The LU+Spatial Prior

The idea behind our second model was to make use of the geographical locations in addition to the land-use features. Indeed, our intuition was that there still might be some local variations in the process that cannot be explained with the land-use variables. Unlike GAM approach, GPR allows us to integrate geographical coordinates in the input vectors without adding unrealistic assumptions. In fact, GPR was historically used only for pure spatial regression. This extension of the previous covariance function was here again selected based on preliminary tests results. The prior of our second model is defined by:

$$m(\begin{bmatrix}\mathbf{x}_{LU}\\ \mathbf{x}_S\end{bmatrix}) = c$$
$$k(\begin{bmatrix}\mathbf{x}_{LU}\\ \mathbf{x}_S\end{bmatrix},\begin{bmatrix}\mathbf{x}'_{LU}\\ \mathbf{x}'_S\end{bmatrix}) = k(\mathbf{x}_{LU},\mathbf{x}'_{LU}) + \sigma^2_{f_S}\exp\Big(-\frac{\|\mathbf{x}_S-\mathbf{x}'_S\|}{\ell_S}\Big)$$

The covariance function is actually the sum of the covariance function of the LU prior and the Matérn function of degree $\frac{1}{2}$ (also known as the exponential covariance function) on the Euclidean geographical distance between the two points. By contrast with the squared exponential function, the exponential function was the tested one that assumed the process to be the less smooth.

Conversely our mixed model can be seen as an extension of purely spatial GP models. Although it is not the focus of the present study, it still represents an interesting interpretation of our mixed model. Indeed, spatial GP regression is a tricky task when it comes to air pollutant at urban scale as the stationary assumption has been shown to be unrealistic (due to street canyons mechanisms for example). To address this issue several techniques that aim at deriving complex non-stationary covariance functions have been developed. However, these methods tend to be computationally very demanding and require a lot of data. In our case, our intuition is that the process viewed as a function of both land-use variables and spatial coordinates will already be more stationary than the process viewed as a function of the spatial coordinates solely. A simple stationary function might already give us good performance at reasonable computational cost.

Prior Fitting

We showed previously that our GP prior models contain several hyper-parameters $\theta = (c, \sigma^2_{f_{LU}}, \sigma^2_{f_S}, \ell_S, \boldsymbol{\ell_{LU}}, \sigma^2_n)$ whose values are not known a priori. Those hyper-parameters were learned during the regression using the same training data by maximizing the marginal likelihood (ML) given by:

$$p(\mathbf{y}|X,\theta) = -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}log|K| - \frac{n}{2}log2\pi$$

where $K = K(X,X) + \sigma^2_n I$ is the covariance matrix for the points X. A non-linear conjugate gradient optimizer was used to carry out that task. Every iteration of the optimizer has a time complexity of $O(n^3)$ with n being the number of training points in X. Details of the derivations of ML's partial derivatives with regards to the hyper-parameters can be found in (Rasmussen and Williams 2006).

Implementation

We implemented our own java-based platform to carry out GPR. We coded it nearly from scratch using only EJML[1] as linear algebra library. The conjugate gradient optimizer was taken from the Matlab toolbox GPML v.2 (Rasmussen and Nickisch 2010) and translated in Java. We showed previously that the regression task and particularly the fitting of the hyper-parameters can be very costly as each iteration takes $O(n^3)$. Therefore, we designed our platform to exploit modern multi-core architecture. Experiments are entered under the form of XML files and broken into tasks that are automatically distributed among the cores in an efficient way to keep busy as many cores as possible at any time. It took roughly 60 hours on a 64-cores AMD Opteron 6272 @2.1Ghz to run all the evaluation tests of next sections which gathered more than 14,000 fitting tasks with a maximum of 500 iterations each on 180-points training sets.

Evaluation

In this section, we empirically evaluate our region-based model with the aforementioned Zurich dataset. The primary purpose of most data-driven statistical models is to assess the degree of exposure to air pollution of population in certain areas, such as questions like "If I take a run along certain streets everyday at 6pm, how much bad air am I breathing in?" To do this, the model uses sensor and land-use data to determine the average concentration of the target pollutant of a specific temporal window. Models are validated through standard random 10-fold cross-validation.

[1] http://code.google.com/efficient-java-matric-library/

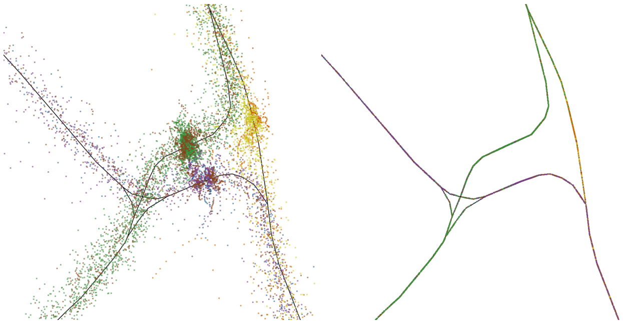

Figure 4: Before (left) and after (right) map-matching of measurement data. The colors denote different sensor nodes

In this paper, we compare the following approaches:

- **GAM** A grid-based generalized additive model that uses land-use features of hectare grid cells from (Hasenfratz et al. 2014).
- **Grid_LU** A grid-based GP model that uses land-use features of hectare grid cells.
- **Grid_LU+Spatial** A grid-based GP model that uses both land-use and spatial features of hectare grid cells.
- **Regions_LU** A region-based GP model that uses land-use features of road-based regions.
- **Regions_LU+Spatial** A region-based GP model that uses both land-use and spatial features of road-based regions.

Preprocessing

In addition to the initial sensor calibration, we took the following steps of further pre-processing the raw data. Similar to (Hasenfratz et al. 2014), we filtered abnormally high concentrations that can be attributed to sensor error, with threshold set at 100,000 $part./cm^3$. We then built a spatial filter with the help of OpenStreetMaps[2]. It removes measurements that were taken at the indoor depots and those that are too far away (more than 20 meters) from the tram lines due to GPS error. Finally, it map-matches the remaining measurements to the closest tram line (see Fig. 4). For a fair comparison to the approach from (Hasenfratz et al. 2014), we also aggregated the data to yearly, seasonally, monthly, weekly and daily windows, and took only the top 200 grid cells / regions with the most measurements for the cross-validation.

Land-Use Variables

We took the following land-use data from the respective government offices to use as input features to the models.

- From the Swiss Federal Statistical Office
  - population density
  - industry density
  - building heights
  - heating type

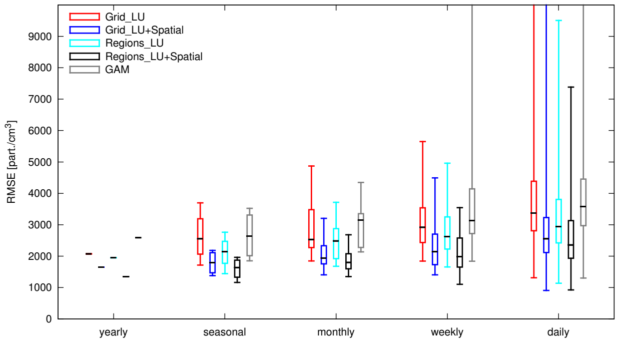

Figure 5: RMSE of model predictions in random 10-fold cross-validation (the lower the better)

  - terrain elevation
  - terrain slope
  - terrain orientation
- From the Canton of Zurich government
  - average daily traffic volume
  - PM_{10} emission estimates
  - NO_x emission estimates

The traffic density values are associated with line-strings that correspond to street segments, and all other land-use values are associated with hectare-sized grid cells. In our region-based model, we first compute a land-use buffer region of a street segment by adding a 20 meters buffer. The land-use values of a region are then derived from the weighted average of the land-use values of the grid-cells or line-strings that overlap the land-use buffer, with the weights computed by the proportions of the overlap.

Results

RMSE First we compare the root-mean-square error (RMSE) of the predictions from different models under our validation setting (Fig. 5). It is a well-known standard metric for comparing the quality of predictions. On the scale of the annual, seasonal, monthly, weekly and daily data, we see a general trend where the length of the temporal window inversely correlates with the RMSE of the predictions. In comparing the models, we see that the grid-based GP models produce equal or less error to the generalized additive model from standard land-use regression, the spatial component in the GP reduces the amount of error, and that the region-based models perform better than their grid-based counterparts.

R^2 Another standard metric for evaluating model performance is the R^2 score, also known as the coefficient of determination (Fig. 6). The score gives an indication of how well the model predictions replicate the observed outcome, as the proportion of total variation of outcomes explained by the model. Similar to the RMSE results, we see that the GP models with spatial features have higher R^2 scores than the pure land-use models, and the region-based models perform better than the grid-based models across all time scales.

[2] http://www.openstreetmaps.org

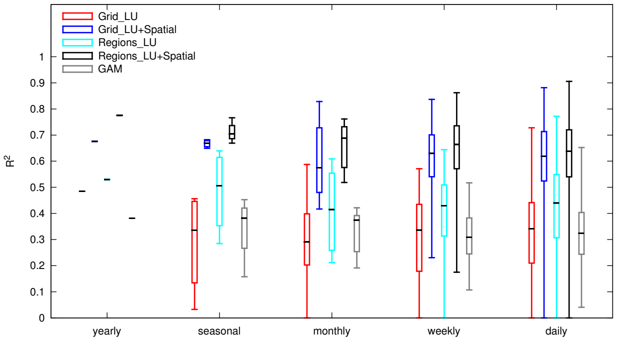

Figure 6: R^2 score of model predictions in random 10-fold cross-validation (the higher the better)

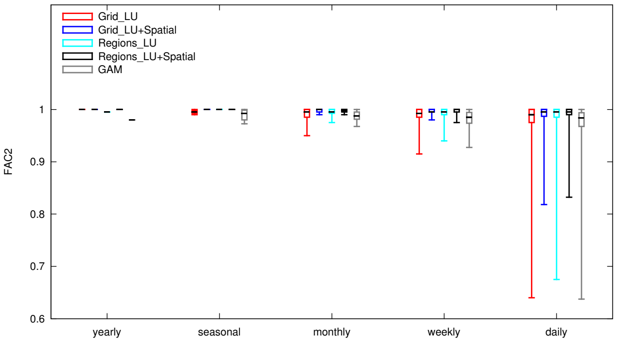

Figure 7: FAC2 score of model predictions in random 10-fold cross-validation (the higher the better)

FAC2 Score We also test the $FAC2$ score of the model predictions (Fig. 7). It measures the fraction of data points that lie inside the factor of two area. As opposed to the two previous ones, this metric is not easily influenced by small number of high and low outliers. We see that this score is always close to one for except for the daily models, and the region-based GP model with spatial features is still the best-performing model.

Discussion

Our results show that a region-based approach for modelling long-term average UFP levels in urban environment produces more accurate predictions across all temporal scales than traditional grid-based approaches under standard random 10-fold cross-validation. However, we must also keep in mind that it is difficult to compare between the grid-based and region-based models, as the data are aggregated by different type of spatial regions. We note that for the purpose of exposure assessment, how the data is spatially aggregated is only secondary to the spatial dispersion of the target population. If we are looking at how much bad air we breathe for jogging along a particular route everyday, it makes more sense to look at the average pollution concentrations along the route, rather than the average concentrations of the grid cells that happen to overlap the route.

The evaluation also has the drawback that all the data came from tram deployments, and hence even though the approach is general for all regions, in this paper we have only tested how well we predict the average concentrations of tram routes. However as the air quality sensors such as the ones used for the Zurich deployment are becoming ever smaller and more affordable, completing the picture is only a matter of further deployments at different sites such as on top of electric buses, lamp posts, balconies, or even backpacks. This would lead to an ever more spatially-detailed estimations of air pollution dispersion in the urban environment.

Conclusion and Future Work

We proposed and implemented a novel region-based approach for estimating UFP concentrations in the urban environment. We showed that when applying the approach to a dataset of one year of continuous measurements from Zurich trams, it produces quality predictions that are comparably better than the current state of the art. It indicates that the estimations from the model is appropriate for evaluating population exposure to UFP pollution.

In addition to exposure assessment, the results from our model can be used for other applications. Given that the measurements are obtained close to the source, the region-based model can also be used to create the estimate of emissions required by traditional, physics-based models. Furthermore, by comparing disparities between the expected pollution level from the model and sensor measurements, we can identify abnormal events that influences the air quality in the urban environment. In preliminary tests for measurement outliers, we easily detected ongoing construction work at Vulkanplatz, a quiet parking lot behind a railway station that should ordinarily not produce so much air pollution. However, further work is needed in acquiring data about abnormal events and performing a systematic analysis in order to evaluate the efficacy of this approach.

Another important line of future work is to include meteorological features that are important for real-time pollution levels. Handling the environmental factors in an intelligent way is crucial to the development of a reliable model for producing estimations and forecasts for real-time exposure assessments, and provides a basis for evaluating the quality of measurements collected in a community-sensing setting described in (Krause et al. 2008; Aberer et al. 2010). It would allow a centre to implement incentive mechanisms such as the ones described in (Papakonstantinou et al. 2011; Faltings, Li, and Jurca 2012; 2014) for eliciting truthful and helpful measurements from a community of sensors.

Acknowledgments

This work was supported by the OpenSense project and InUse project funded by the NanoTera.ch program from the Swiss National Science Foundation, and the ARC Discovery Project “Artificial Intelligence Meets Sensor Networks” (DP120103758) from the Australian Research Council.

References

Aberer, K.; Sathe, S.; Chakraborty, D.; Martinoli, A.; Barrenetxea, G.; Faltings, B.; and Thiele, L. 2010. Opensense: open community driven sensing of environment. In Ali, M. H.; Hoel, E. G.; and Shahabi, C., eds., *GIS-IWGS*, 39–42. ACM.

Bonilla, E. V.; Guo, S.; and Sanner, S. 2010. Gaussian process preference elicitation. In *NIPS*, 262–270.

Cao, Y.; Brubaker, M. A.; Fleet, D.; and Hertzmann, A. 2013. Efficient optimization for sparse gaussian process regression. In *NIPS*, 1097–1105.

Cressie, N. A., and Cassie, N. A. 1993. *Statistics for spatial data*, volume 900. Wiley New York.

Faltings, B.; Li, J. J.; and Jurca, R. 2012. Eliciting truthful measurements from a community of sensors. In *2012 3rd International Conference on the Internet of Things (IOT)*, 47–54. IEEE.

Faltings, B.; Li, J.; and Jurca, R. 2014. Incentive mechanisms for community sensing. *IEEE Transactions on Computers* 63(1):115–128.

Godish, T. 2003. *Air Quality*. CRC Press.

Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; and Thiele, L. 2014. Pushing the spatio-temporal resolution limit of urban air pollution maps. *Proceedings of the 12th International Conference on Pervasive Computing and Communications (PerCom14)*.

Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; and Briggs, D. 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* 42(33):7561 – 7578.

Krause, A.; Horvitz, E.; Kansal, A.; and Zhao, F. 2008. Toward community sensing. In *IPSN*, 481–492. IEEE Computer Society.

Li, J. J.; Faltings, B.; Saukh, O.; Hasenfratz, D.; and Beutel, J. 2012. Sensing the air we breathe - the opensense zurich dataset. In Hoffmann, J., and Selman, B., eds., *AAAI*. AAAI Press.

Lim, S. S.; Vos, T.; Flaxman, A. D.; et al. 2013. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 19902010: a systematic analysis for the global burden of disease study 2010. *The Lancet* 380(9859):2224 – 2260.

Nguyen, T. V., and Bonilla, E. V. 2014. Fast allocation of gaussian process experts. In *International Conference on Machine Learning*.

Papakonstantinou, A.; Rogers, A.; Gerding, E. H.; and Jennings, N. R. 2011. Mechanism design for the truthful elicitation of costly probabilistic estimates in distributed information systems. *Artificial Intelligence* 175(2):648–672.

Rasmussen, C. E., and Nickisch, H. 2010. Gaussian processes for machine learning (gpml) toolbox. *J. Mach. Learn. Res.* 11:3011–3015.

Rasmussen, C., and Williams, C. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press.