High Resolution Air Pollution Maps in Urban Environments Using Mobile Sensor Networks

Ali Marjovi, Adrian Arfire and Alcherio Martinoli

***Abstract*— We propose three modeling methods using a mobile sensor network to generate high spatio-temporal resolution air pollution maps for urban environments. In our deployment in Lausanne (Switzerland), dedicated sensing nodes are anchored to the public buses and measure multiple air quality parameters including the Lung Deposited Surface Area (LDSA), a state of the art metric for quantifying human exposure to ultrafine particles. In this paper, our focus is on generating LDSA maps. In particular, since the sensor network coverage is spatially and temporally dynamic, we leverage models to estimate the values for the locations and times where the data are not available. We first discretize the area topologically based on the street segments in the city and we then propose the following three prediction models: i) a log-linear regression model based on nine meteorological (e.g., temperature and precipitations) and gaseous (e.g., NO_2 and CO) explanatory variables measured at two static stations in the city, ii) a novel network-based log-linear regression model that takes into account the LDSA values of the most correlated streets and also the nine explanatory variables mentioned above, iii) a novel Probabilistic Graphical Model (PGM) in which each street segment is considered as one node of the graph, and inference on conditional joint probability distributions of the nodes results in estimating the values in the nodes of interest. More than 44 millions of geo- and time- stamped LDSA measurements (i.e., more than 14 months of real data) are used in this paper to evaluate the proposed modeling approaches in various time resolutions (hourly, daily, weekly and monthly). The results show that the three approaches bring significant improvements in R^2, RMSE and FAC metrics compared to a baseline K-Nearest Neighbor method.**

I. Introduction

More than 7 millions of premature deaths are annually linked to air pollution from which 2.6 millions are particularly caused by urban outdoor air pollution [1]. Many studies on human health have concluded that environmental stress is a major factor for morbidity and has a negative impact on the quality of life especially in urban areas (e.g., [2]). One of the major challenges in these studies is to obtain or estimate high resolution (spatial and temporal) air quality data to be able to analyze the correlation between health and the exact air to which people are exposed.

Among all the airborne pollutants (SO_x, NO_x, CO, NH_3, O_3, etc.), recently there has been a growing attention to study particulate matters due to their significant adverse impact on human health. In urban environments, this measure is closely linked to urban traffic conditions [3]. Most of the recent studies (e.g., [4] and [5]) have focused on PM10 or PM2.5 which describe the amount (mass/number) of particles smaller than 10 μm or 2.5 μm in a given volume. However, the mass or number of particles do not necessarily represent the best measures for all risks to human health. The size and the surface area of the particles also matters. It is well-known that finer particles are potentially more toxic than coarse particles [6]. Studies have shown that measuring the surface of nanoparticles, rather than their mass or number, is more meaningful for quantifying their health impact [7], [8], [9], [10]. In fact, ultrafine particles (UFPs) are able to travel deeper into the lungs and, due to their large surface-to-volume ratio, have higher reactivity which can result in higher toxicity. Therefore we are interested in measuring and estimating the Lung-Deposited Surface Area (LDSA) which is a measure that describes the deposited surface of particles per volume of air inhaled.

The established method for monitoring air pollution, in most countries, is through the use of static air pollution monitoring stations. These reference stations provide highly accurate measurements from a limited number of specially selected sites, which should be representative of different types of locations (e.g., the National Air Pollution Monitoring Network - NABEL - in Switzerland, consists of 16 stations in total over the whole country). The stations are expensive, large, and power hungry, and so this type of monitoring networks can only provide spatial resolutions in the order of several hundred kilometers which must be interpolated with dedicated, state-of-the-art physico-chemical modeling techniques in order to reach a resolution of about 1 km^2.

A. Mobile Sensing

As opposed to traditional air quality monitoring stations, the use of networks of low-cost sensors is quickly emerging, aiming at providing air quality data with unprecedented temporal and spatial resolution. In this application field as well as others (e.g., surveillance [11], crowdsensing through smart-phones [12] and dynamic coverage [13]) there is a growing trend towards mobile sensing platforms. For air pollution monitoring in particular, innovative sensing strategies such as wearable air quality sensing nodes [14] and smart-phones used as mobile air quality sensors [15] are proposed. This will open exciting new opportunities for the study of urban air quality and its impact on health. An important issue for obtaining accurate and spatially highly resolved air pollution data is the trade-off between high cost of accurate air pollution monitoring sensors and the number of such devices required for succinctly monitoring a given geographical area. Fig. 1 depicts this trade-off and classifies the various techniques for

The authors are with the Distributed Intelligent Systems and Algorithms Laboratory, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. E-mail addresses: firstname.lastname@epfl.ch

978-1-4799-8856-3/15 $31.00 © 2015 IEEE
DOI 10.1109/DCOSS.2015.32

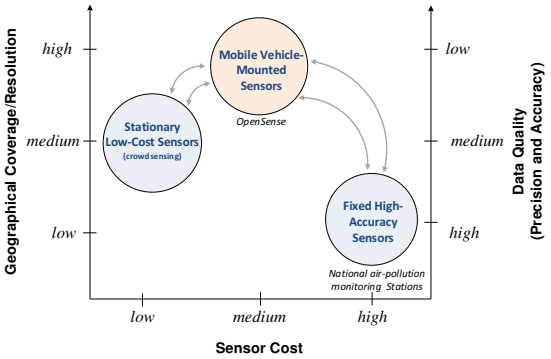

Fig. 1. Sensing trade-off for a given budget.

gathering data. In the context of the OpenSense II[1] project funded by the Swiss national research initiative Nano-Tera.ch and aiming at investigating mobile sensing technologies to monitor air pollution, we consider data gathered via all types of devices and stations shown in Fig. 1, i.e., from [high quality / low spatial resolution] to [low quality / high spatial resolution].

We have anchored our sensing platforms on top of ten public buses in Lausanne. This innovative deployment which adds mobility to monitoring platforms brings significant benefits in comparison to canonical static Wireless Sensor Networks (WSNs): finer spatial resolution, coverage of wider area with fewer nodes, cheaper maintenance, etc. However, not much literature exists on field estimation using non-stationary sensor networks. The movements of the nodes are not under our control and not even predictable since the buses are assigned to different lines every few hours depending often on real-time needs of the public transportation company. The coverage of the network dynamically changes over time and generating consistent maps with high spatio-temporal resolution is a tough challenge.

To state the problem, consider a heterogeneous sensor network which consists of mobile and stationary nodes. Stationary nodes provide meteorological and gaseous information from fixed locations of a city. The mobile nodes measure LDSA while dynamically navigating various streets of the city on trajectories which are not systematically predictable. The question is how to generate spatio-temporal high resolution LDSA maps in this city. We are aiming at hourly maps with spatial resolution of small street segments. Since the coverage area of the limited number of mobile nodes changes from one hour to the other, there are always many street segments that do not have any measurements. We address this problem by building statistical models for the street segments of the city.

B. Air Pollution Modeling

Most of the works on air pollution modeling fall into two categories [16]: deterministic and statistical. Deterministic dispersion models simulate the physico-chemical processes of airborne gas dispersion, using the sources of emissions as input. GRAL [17] is an advanced example of this category which mathematically models the motion of pollution plume particles in the atmosphere using a Lagrangian dispersion model. A drawback of this category of models is that they need accurate information about emission inventories, structural and geographical details of the environment, and meteorological data, which are not always available in high temporal resolutions.

Alternatively, statistical models do not describe the actual physical processes, but they treat the input measurements as random variables to derive a statistical description of the target distribution. These methods can be divided into two subcategories. The first subcategory is represented by the purely field-driven models which aim at finding all the dependencies and variables from the measurement data. Spatial interpolation methodologies (e.g., inverse distance weighting interpolation [18], and K-Nearest Neighbor (KNN) [19]) are the most common approaches in this subcategory. The performance of such methods drops drastically if the field is dynamic and multi-variant (which is usually the case for urban environments under short term observational conditions). The second subcategory are the statistical models which work not only based on the field measurements but also take one or more explanatory variable(s) into account. The explanatory variables are usually other related modalities to the target variable. These methods usually show higher performance compared to the purely field-driven models. In the next paragraph, we will provide additional details about latter subcategory.

Hussein et al. [20] fitted a linear regression model on the data of a monitoring station to predict aerosol particles in Helsinki. Mølgaard et al. [21] used a Bayesian regression model to predict ultrafine particles concentrations of an urban monitoring station using meteorological and traffic data as inputs. To obtain better prediction performance, Clifford et al. [22] proposed a generalized additive model using meteorological data, time, solar radiation and rainfall as explanatory variables. Reggente et al. [23] employed a Gaussian process regression to estimate UFPs in an urban air pollution monitoring network based on local and remote concentrations of NO_x, O_3 , CO, and UFPs. None of the mentioned works have considered mobile sensor networks.

One stream of research has focused on modeling the air pollution based on land-use data. Land-use features (in the context of urban environmental modeling) are measures of average traffic volume, population density, building heights, heating type, terrain elevation, terrain slope, types of roads, etc. Li et al. [24] proposed a Gaussian process regression (AKA Kriging) model using land-use characteristics to estimate urban UFP levels from measurements collected from the trams in Zurich (Switzerland) within different grid-cells. The main problem with land-use data is that usually they are not available in high temporal resolutions. For instance, the most recent traffic counts data available for Lausanne (our targeted city) that was available to us, was gathered in 2010. This data obviously does not represent the dynamics of the traffic

[1] http://opensense.epfl.ch

from one day to the other in 2015. Therefore, this kind of data is usually considered as long-term representative for trends and can therefore produce only long-term predictions (i.e., low temporal resolution) of air pollution. One way to overcome this issue is to generate different models for every target time period. Hasenfratz et al. [25] and Li et al. [24] (in two separate works) built up two sets of a thousand models, each targeting one time period (e.g., one model per day) for one city. These models cannot be used for time periods other than the training ones. These two contributions used the mobile sensor network dataset gathered from the Zurich deployment of the former phase of the OpenSense project. In the method proposed by Hasenfratz et al. [25], measurements gathered in a previous period are also used in the model to increase the accuracy of high resolution maps. In particular, they annotate the UFP measurements obtained during one year with the corresponding meteorological (e.g., temperature) and time data (e.g., weekday). Then based on the current meteorological conditions and time, they fetch the most relevant historic UFP measurements and use them to augment the current dataset represented by the real-time UFP measurements. This method significantly increases the accuracy of the maps, although the real-time meteorological data are not directly used in the model itself. On the other hand, Li et al. [24] did not consider meteorological parameters at all.

C. Our Contribution

To address the stated problem, we propose three statistical modeling methods using data from our mobile sensor network. The details of our sensor network are presented in Section II. Differently from many previous works which partition the space into square grids, we discretize the area topologically based on the street segments in the city (explained in section II-E). Then we propose the following three models to predict LDSA values in each street segment:

1) A log-linear regression model based on nine meteorological (e.g., temperature and precipitations) and gaseous (e.g., NO_2 and CO) explanatory variables obtained from the two static stations (explained in section III-A). Although log-linear regression modeling has been vastly used in the literature, the number of explanatory variables, the scale of the data, and the time resolutions which we consider in this paper are beyond the framework of many previous works in this area.
2) A novel network-based log-linear regression model that takes into account the measurement (LDSA) values of the most correlated streets and also the nine explanatory variables from two static stations. The proposed virtual network captures the dependencies between the street segments and also takes into account the explanatory variables in the model of each street. Moreover, it automatically handles the issue of dynamic coverage of the mobile sensor network. To the best of our knowledge, no previous work has ever proposed such a network-based model for extending the spatiotemporal mapping capabilities of a mobile sensor network. Section III-B provides details of this contribution.
3) A novel Probabilistic Graphical Model (PGM) in which each street segment is considered as one node of the graph, and inference on conditional joint probability distributions of the nodes results in estimating the targeted modality (in our case, LDSA) in the nodes of interest. None of the previous works in this area have designed a PGM to capture automatically all the cross-correlations between the explanatory variables and LDSA values in different streets and also deal with the dynamic coverage of mobile sensor networks. This powerful tool is explained in section III-C.

Finally section IV presents the evaluation of the proposed methods by comparing them with each other and with a baseline KNN model.

II. The System

A. Sensing nodes

In our Lausanne deployment, dedicated sensing nodes are anchored to ten public buses and measure multiple air quality parameters including LDSA. The localization of the mobile nodes is achieved through fusion of GNSS and the vehicle dead-reckoning. Accurate time is also obtained from the GNSS module. All the measurements are geo- and time-stamped locally by the sampling node and sent through GPRS to a database server. Along with these, there are several meta-data information that are sent to the server to indicate the health state of the measurements. The final deployment of our mobile sensor network started in October 2013. The LDSA sensors are Naneos Partector [26] devices and have been added to the nodes starting from December 2013. The sampling rate for LDSA is 1 Hz. Fig. 2 shows one of the sensing nodes used in this project.

For this paper we only focus on the LDSA mapping due to following reasons:

- The LDSA sensors are calibrated by the manufacturer and therefore ready to use without further calibration efforts.
- The LDSA sensors are the fastest sensors in our deployment. The response time for this actively sniffing sensor is in the order of fractions of a second. Considering the fact that our sensors are mobile, the fast response of the sensors implies that the measurements are already spatially and temporally associated with the local field.
- The aging and time drift are negligible for the LDSA sensors, although minimal maintenance effort is required (approximately once per year).

In addition to the LDSA measurements collected by the buses, we consider two static monitoring stations in our system. One is the NABEL station located near the city center (on the César-Roux street) which monitors many air quality parameters (e.g., CO, NO, NO_2). The other is the meteorological monitoring station operated by the national weather service of Switzerland (MeteoSwiss) in Pully

Fig. 2. One sensor node anchored on top of a public bus (left). One of the LDSA sensing modules (right).

which provides meteorological parameters (e.g., precipitations, radiation and humidity). These two stations report their measured values every ten minutes.

B. LDSA data

About 44.5 millions of geo- and time- stamped real LDSA measurements gathered during more than 14 months are used in this paper. This amount of data is available after data cleaning (i.e. applying several simple filters based on the device health meta-information).

C. Explanatory variables

The data of the meteorological and pollution monitoring stations are about 70,000 rows each in total, due to their lower sampling rate. From these two stations, the data of the following 9 parameters are used in this paper:

- CO, NO and NO_2: These gases are mainly produced by combustion of fossil fuels and so they can be a good measure for traffic conditions in the city. Since ultrafine particles are produced from the same sources in the urban environment, these are good candidate parameters for our models.
- Ground level O_3: Ozone is the primary oxidant of pollutant gases present in the atmosphere and since it plays an important role in the balance between NO and NO_2 in the atmosphere, we include it as a parameter in our LDSA models.
- Radiation, precipitations, temperature and wind speeds: The stability of the atmosphere is highly dependent on these parameters. Solar radiation and temperature change the size of eddies which eventually affect the concentration of particles through dispersion. When a precipitation event (e.g., rain) starts, the concentration of the particles drastically drops. Also wind is generally expected to disperse locally the aerosols from one place to another. It is therefore important to take these parameters into account.
- Relative Humidity: The growth pattern of ultrafine particles is related to adsorption of water vapor, so humidity is a parameter to model the aerosols particles.

Statistical analysis of the correlation between the aforementioned parameters and the ultrafine particle data has been already studied in the literature (e.g., [27]). Throughout this paper we refer to these 9 parameters as "explanatory variables".

D. Time discretization

This paper considers "hour", "day", "week", and "month" time resolutions. In less than one hour, there are not enough LDSA measurements for most parts of the city, making it impossible to fit models. Depending on the time resolutions, the LDSA measurements and the explanatory variables are partitioned and aggregated in time and in space.

E. Space discretization

Most of the previous works (e.g., [25], [28]) partition the area to uniform grid cells and assume that the measurements inside a cell have the same conditions (e.g., in terms of weather, wind and traffic). Depending on the cell-size, one cell can cover several streets which have different environments and traffic conditions. To overcome this issue, Jutzeler et al. [29] proposed to use regions of homologous emissions to divide the city into partitions with similar daily traffic estimations. They associated every measurement to the closest road segment based on Euclidean distance. They showed that compared to grid-based, the region-based partitioning produces better predictions across aggregates of yearly to daily time scales. We follow this concept while using a more advanced street matching algorithm.

In this paper, the data of the street segments of the city are acquired from the online OpenStreetMap [30] database. Then we split the very long streets into multiple smaller streets in order to not lose high spatial resolution. The use of this space discretization will naturally result into higher resolutions in the downtown areas where street segments are shorter and the heterogeneity of the measured field is expected to also be higher than in suburban areas. Fig. 3 shows the length histogram of the street segments. Using the localization data of the measurements and estimating the azimuth of the bus, the LDSA values are assigned to their corresponding road segments based on the algorithm explained in [31]. The general idea of this algorithm is to continuously track the buses based on their location and to keep a list of route candidates for them during their movements in the streets. Each route has a score that defines how well the traced trajectory of the bus matches this route. Fig. 3 shows one snapshot of our street matching software and shows an example of how well the measurements are assigned to one street segment.

All the LDSA measurements are projected on 1377 street segments covering the region of interest depicted in Fig. 4. As the figure shows the measurements are unevenly distributed in various segments, representing a dynamic non-uniform coverage. An important metric of the goodness of a particular spatial discretization is how homogeneous are the measurements inside a given partition element (i.e., segment or grid cell), with better discretizations having lower deviation from the mean of the partition element. We have compared the standard deviation of LDSA values in our street segments with grid-cell partitioning considering six different cell numbers in Fig. 5. This figure shows that the standard deviation we obtain from street segmentation with 1377 segments is better then when we use 4900 (and for some cases 10000) grid

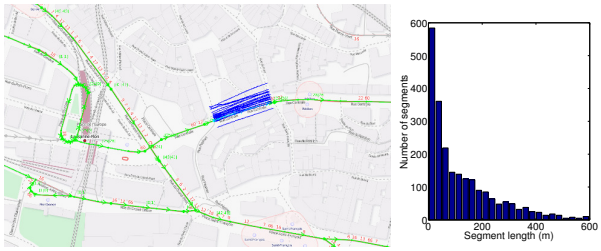

Fig. 3. Left: One snapshot of the developed street matching software. In this example the blue lines show the reported trajectory traces of the buses during one week in one segment. Due to localization errors, the blue traces are deviated form the actual position of the street segment. Our street matching algorithm has matched all of them to one segment. The green lines show the other segments that buses have passed during this period. Right: The histogram of length of the street segments. Most of our segments are shorter than 25 meters.

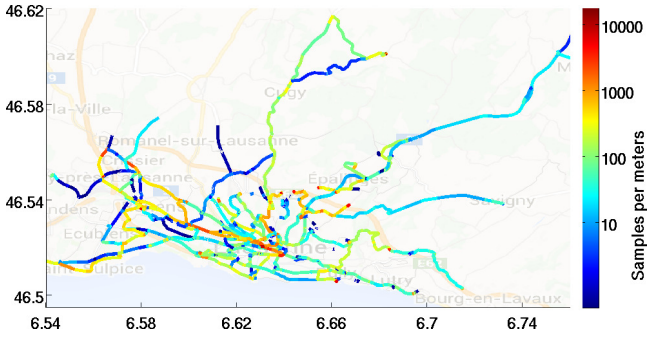

Fig. 4. The coverage area of the sensors from Dec. 1st. 2013 to Feb. 1st 2015. The number of samples is normalized considering the length of each link. The segments with less than 1 sample per meter are not considered in the coverage area.

cells. These results show the benefit of "street segmentation" over "grid partitioning" in urban environments.

III. Modeling Approaches

A. Log linear explanatory variables

It has been experimentally shown that the mathematical links between gaseous parameters in air are logarithmic [32], [21]. In the first attempt, similar to many previous works (e.g., [25]) we use a log-linear regression model to estimate LDSA values in every street using the data of explanatory variables (defined in Section II-B) as inputs. The mathematical formulation of this model is defined by the following equation:

$$\log(L_m) = \alpha_0 + \sum_{i=1}^{9} \alpha_i \cdot \log(v_i) \quad (1)$$

where L_m denotes the LDSA estimated value in segment m, α_0 the intercept, v_i the explanatory variables i, and α_i the coefficient of each variable.

We divide the available data into two subsets, the "training set" and the "validation set", using 10-fold cross validation. On the training set, we use the QR decomposition algorithm [33] to solve the linear least squares problem in order to find the coefficients of the model for each street segment. Working on four time resolutions and 1377 street segments,

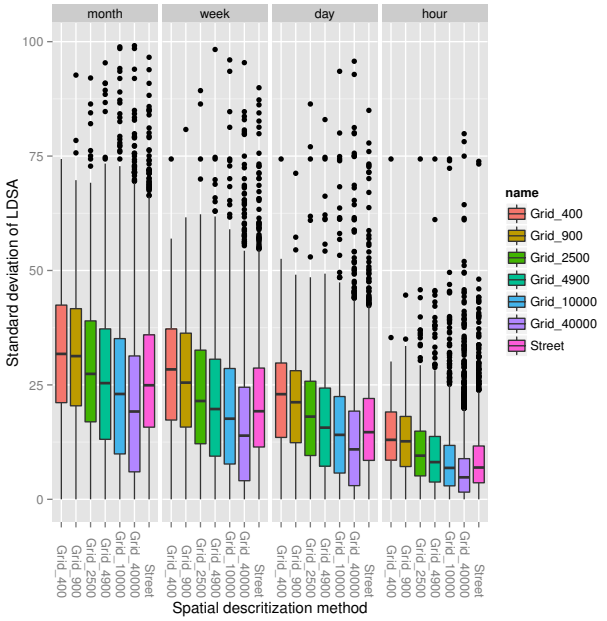

Fig. 5. Standard deviation of LDSA values in one spatial partition (cell or street segment). Lower standard deviation indicates more homogeneity in the measurements in one cell. The measurements are integrated based on the four time resolutions and then the standard deviation in each cell/segment is computed for every space-time tessellation. Street segmentation shows good results considering that the number of segments is 1377.

we developed 5508 models. The results are reported in Section IV.

B. Network-based log-linear regression

The goal is to estimate the LDSA values for the locations-times that the mobile sensor network has not covered (but still covered previously at least once). The previous model computes the LDSA values based on the values of the explanatory variables (which are always available through the static stations). Here, the idea is to take also into account the measured LDSA values of other segments for predicting the LDSA values in a given segment in a given time window. However, this is not trivial considering the fact that the sensors are mobile and the coverage is dynamic, an especially important factor when high temporal resolutions (e.g., hour) are considered. For instance, if the model of the segment S_m is dependent on the LDSA value of the segment S_n, then the model cannot work when there is no bus covering the segment S_n. To address this challenge we propose to build a virtual dependency network on the segments. In this network, each segment is one node and a directed edge is drown between node S_m and S_n if node S_n is considered as a variable in the model of node S_m. As we will see this network is able to address the problem of dynamic coverage of the mobile sensors.

Now the question is how to build the network and define how the models work on the network. We propose to connect node S_i to S_j if the following two conditions hold:

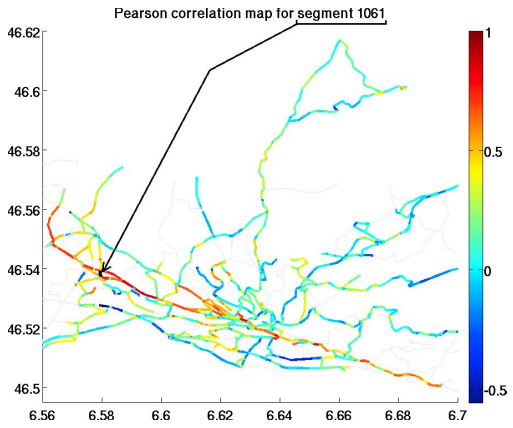

Fig. 6. The correlation map for segment 1061 (shown by a small black dot and the arrow). The segments along one long street show high correlation while some nearby streets in other directions do not show high correlations.

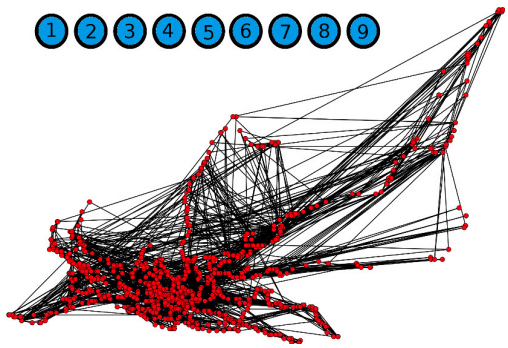

Fig. 7. The generated network for Lausanne's street segments for LDSA estimation model. Each red node represents the center of one segment while the black lines between two nodes represent the correlation between the LDSA values of two segments. The big 9 blue nodes represent the nine explanatory variables which are in fact connected to all other nodes. For increasing the readability of this figure we have removed all the edges between the blue nodes and the red nodes.

1) If the Pearson correlation between the LDSA values of node S_i and S_j is high. This is due to the fact that correlation is a basic need for every variable in a model. We have noticed that the segments which are geographically close do not necessary show high correlations, especially if they are not in the same direction. Fig. 6 shows a correlation map for one exemplar segment.
2) If node S_j has reported enough "complementary data" relative to the available data of node S_i. We define "complementary data" as the number of time-slots when there are LDSA values reported for the segment S_j but no value was reported for S_i. Based on the first condition, some of the segments which are along one street have the highest correlations and since they are most probably covered by the same bus they do not have any "complementary data" meaning that either all segments have data or none of them (making the models inefficient). With this second condition we make sure that the edges in the network are going to be efficient for the models.

To create the network based on the two mentioned edge conditions, using the available LDSA data of the segments, we compute the cross correlation of LDSA values of all combinations of the segments and then for every segment S_i we find the M (= 10 in this paper) most correlated segments. Among the most correlated segments we find the ones that have more complementary data relative to S_i. We establish an edge for every node S_i and then keep adding edges to the network (while considering the two conditions) until the network is minimally connected.

To take the nine explanatory variables into account, we insert them as nodes to this network and connect them to every other node in the network. In our deployment, this process generated a network with 1386 nodes (1377 street segments + 9 explanatory variables) and found 15040 edges. Fig. 7 shows this network.

Since we have ten buses, during one time slice (e.g., one hour) a fraction of nodes of the network will have actual measurements, while the rest should be estimated using the models. It is obvious that if the network is (at least) minimally connected, then theoretically all the nodes can be predicted one by one even if only one segment has actual measurement. Denoting E as the edge list of this network, and L_{S_m} as the LDSA value of segment S_m, here is the proposed model for node S_m:

$$\log(L_{S_m}) = \alpha_0 + \sum_{i=1}^{9} \alpha_i \,.\, \log(v_i) + \sum_{[m-n]\in E} \alpha_n \,.\, \log(L_{S_n}) \quad (2)$$

The optimal values for the coefficients of this model are found similar to the first method using the training set data.

This novel graph-based approach iteratively estimates the LDSA value of the segments based on the LDSA value of the correlated nodes and also based on the values of the explanatory (meteorological and gaseous) variables. This is an approach well-suited for mobile sensor networks where the coverage area of the network dynamically changes over time. Basically it does not matter which nodes have measurements and which nodes are to be estimated, as long as there is at least one node with a measurement, the LDSA value of all the other segments can be estimated iteratively. Section IV evaluates this approach in detail.

C. Probabilistic Graphical Model

In this section, we propose a probabilistic graphical model to infer the LDSA values from the observed values, their dependencies to other segments and to the explanatory variables. Our probabilistic model for street correlations is based on the assumption that values of LDSA in two correlated streets are more likely to be similar. To design the model, we use the framework of Markov networks or Markov random fields [34], very common in statistical physics, economy, and image processing. Assuming $X = \{X_1, X_2, ..., X_N\}$ a set of discrete random variables, a binary Markov network over X defines a joint distribution $P(X)$. The network is defined via a graph whose nodes correspond to variables in X and its edges E represent direct probabilistic

dependencies between those variables. Each variable X_i is associated with a potential $\varphi(X_i)$ and each edge $[X_i - X_j]$ is associated with a non-negative compatibility potential $\varphi(X_i, X_j)$. The joint distribution is then defined as:

$$P(X_1, ..., X_N) = \frac{1}{Z} \prod_{i=1}^{N} \varphi(X_i) \prod_{[X_i - X_j] \in E} \varphi(X_i, X_j) \quad (3)$$

where Z is a normalizing constant. Intuitively, $\varphi(X_i)$ encodes how likely the different values of X_i are, ignoring dependencies between the variables. For assigning a particular value x_i to variable X_i and value x_j to X_j, the potential $\varphi(X_i, X_j)$ specifies how "compatible" this assignment is; the higher the value, the more likely this pair of values is to appear together.

In our problem setting, the variables are the union of the LDSA values in the segments $S = \{S_1, S_2, ..., S_n\}$ and the explanatory variables $V = \{V_1, V_2, ..., V_9\}$, thus $X = S \cup V$, and the edges are defined by relationships between them. The network defined in Section III-B represents the interaction between our variables. Intuitively, an edge e_{ij} between S_i and S_j captures the basic intuition that, if S_i and S_j interact, they are more likely to have similar values (i.e., high Pearson correlation). Now we need to determine:

1) the marginal probability distribution of LDSA in all segments, $\varphi(S_i), i \in [1..n]$,
2) the marginal probability distribution of explanatory variables, $\varphi(V_j), j \in [1..9]$,
3) the pairwise probability distribution of LDSA in segment i and k, $\varphi(S_i, S_k)$,
4) the pairwise probability distribution of LDSA in segment i and explanatory variable j, $\varphi(S_i, V_j)$.

All the random variables in this problem are continuous which makes the computation of marginals intractable. Hence we discretize the values into equal width intervals. For each modality (the nine explanatory variables and the LDSA) we divide the range into 20 sections and discretize their values. Then we compute the normalized frequency tables and joint frequency tables of all the variables (nodes and edges) as marginals and pairwise probability distributions. Fig. 8 (left) shows the marginal probability distribution of LDSA in three segments. Fig. 8 (right) shows the pairwise probability distribution of LDSA in two correlated segments. These distributions are obtained from the LDSA data of the corresponding segments.

At a given time some of the segments have observations forming an "observed set" called S_o, while some are not observed forming an "unobserved set" called S_u, such that $S = S_o \cup S_u$. Our approach aims to find the value of LDSA for missing streets S_u based on the observed values S_o (other streets and the explanatory variables V) that can be formalized as:

$$P\big(S_u | (S_o, V)\big) = \frac{P(X1, ..., X_N)}{P(S_o, V)} \quad (4)$$

We use the inferred full joint probability (Eq. (3)) in the above equation to answer any query in which some of the streets in the graph are clamped to observed values. Fig. 9 presents this concept. Marginalization of $P(S_o, V)$ scales

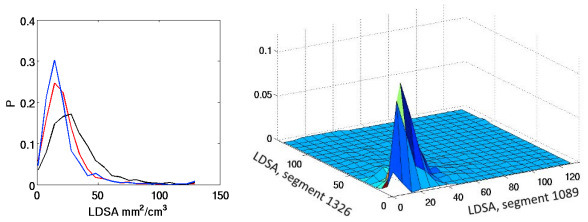

Fig. 8. An example that shows the marginal probability distribution of LDSA in three different segments (left). The blue and red (corresponding to segments 1089 and 1326) show very high correlations. The pairwise probability distribution of LDSA in two correlated segments is represented on the right plot.

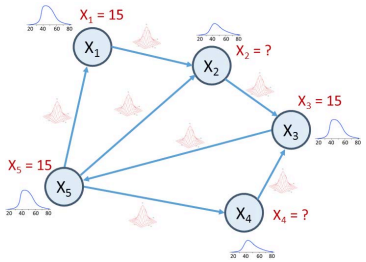

Fig. 9. A simple example showing a graphical model in this context. Each node (corresponding to one street segment) has one potential function (the small blue curves), the edges represent the interaction between the nodes corresponding to the joint potentials (the small pink 2D surfaces). LBP can solve this problem and find the best values for unobserved nodes (in this case X_2 and X_4) given the values of the other nodes.

exponentially with the length of $S_o \cup V$. As a more efficient approach we use the approximate algorithm of Loopy Belief Propagation (LBP) [35] to answer the conditional query of Eq. (4). LBP is a greedy strategy to sequentially update the value of each variable, keeping the value of the rest fixed. The algorithm performs the value assignment in random order for all variables. Each variable X_i is assigned to a value that maximizes the likelihood of joint probability. After all variables are assigned, they are randomly re-ordered, and the assignment process is repeated. This process continues until no value of any variable is changed between two successive iterations [36]. The output of this iterative algorithm is a probability distribution for all the unobserved variables (X_u). We consider the argument of the maximum of this distribution as the final output value of the model for the corresponding segments. The higher the maximum probability is, the lower the uncertainty on the value estimated by the model.

There are a few significant advantages of this proposed model compared to the other two (and to many other previous works e.g., [25]):

- There is only one single model built for every time resolution, i.e,, four models in total. This means that this model can capture all the dependencies between all the segments and also the dependencies between the LDSA values of each segment with the explanatory variables. This is a huge advantage considering the fact that most other methods are either location dependent (e.g., our network-based log linear regression model) or

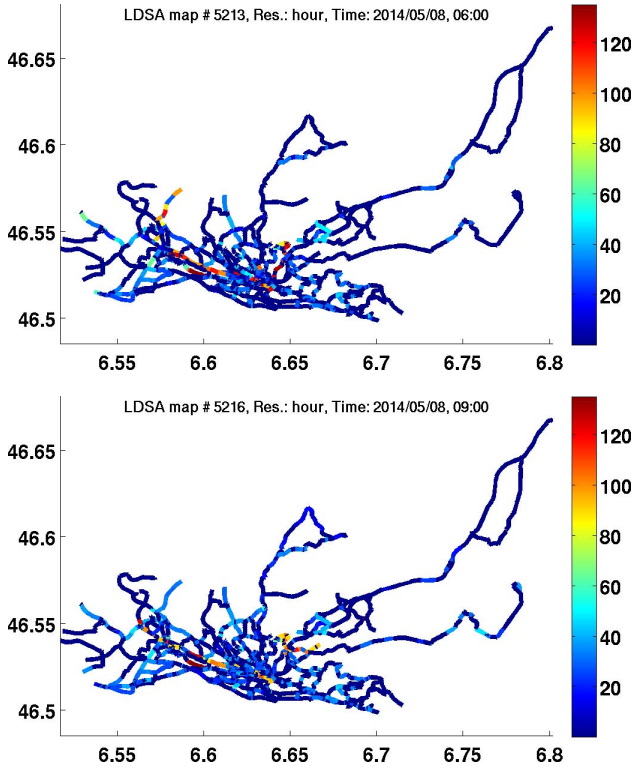

Fig. 10. Two examples of the hourly air pollution map of streets of Lausanne based on estimated (modeled) and measured LDSA values. These maps are generated based on the third proposed model.

time dependent (e.g., the one proposed in [25]).

- Since the output is probabilistic, the method always provides a metric of uncertainty on the possible output values. Many other modeling methods also provide a metric of uncertainty but this one gives the uncertainty not only for the output value but also on every other possible value.
- This method perfectly deals with the heterogeneity and dynamics of the system. It does not matter which segments are covered at a given time by the mobile sensor network, the model propagates the belief in the network and is able to predict the value of the other nodes. Of course the more segments report measurements, the higher is the accuracy of the predicted values for the other nodes.

IV. Evaluations and Results

The three proposed models are used to model LDSA values for locations/times of interest. Using the models we have generated complete hourly, daily, weekly and monthly air pollution maps of Lausanne. Fig. 10 presents some examples of the hourly maps resulting from the third proposed model.

In our evaluation sets, for every estimated value (model output), there is an observed value. Denoting M as the set of modeled values and O as the set of corresponding observations, we consider the following three metrics:

1) RMSE: The root mean square error is computed as the following:

$$RMSE = \sqrt{\frac{1}{L}\sum_i (O_i - M_i)^2} \tag{5}$$

where L is the number of estimations provided by the model. Obviously, the lower this metric is, the better the model works.

2) FAC2: The factor of two measure, is the percentage of ratios O_i/M_i that lay between 0.5 and 2. i.e.

$$0.5 < \frac{O_i}{M_i} < 2 \tag{6}$$

The more close to 1 this metric is, the better the model has estimated the values.

3) R^2: The coefficient of determination shows the linear dependence of observed and modeled values.

$$R^2 = 1 - \frac{\sum_i (O_i - M_i)^2}{\sum_i (O_i - mean(O))^2} \tag{7}$$

where $mean(O)$ denotes the mean of all observations which are considered in the validation sets. $R^2 = 1$ represents a perfect linear fit between the model and the observations.

In addition to the three proposed modeling approaches we have also implemented the conventional KNN regression model as a baseline to evaluate the results. We trained the KNN model (to find the optimal value of K) using a training set of LDSA values. No explanatory variable was used in this method (so the model is fully field-driven based on LDSA). Like the other 3 models, we have aggregated the data into street segments and used the center of the street segment as the geographical location of the measurements in KNN. Euclidean distance is used as the metric of distance and the search method was exhaustive.

There are two advantages for KNN in our problem setting: i) Using KNN we can build a single model for the city per time resolution (4 models in total). Among the three proposed methods only the third one is similar to KNN in this regard. ii) KNN is able to predict even for the segments which do not have any observations. None of the proposed methods in this paper have this ability. However, the results show that the performance of KNN is very poor on our data (see Fig.11). The RMSE is large, FAC2 is not satisfying and particularly R^2 is mostly negative.

We have evaluated the three proposed methods using exactly the same procedure and data. The data is divided into training and evaluation sets using a 10-fold cross validation method. The results are shown in Fig. 12, Fig. 13 and Fig. 14. The first method (log-linear regression modeling) shows good results in comparison to KNN. In fact this method only uses the explanatory variables to predict LDSA while KNN uses only LDSA measurements of the other segments to predict LDSA in a given one. The fact that log-linear regression shows better results proves the impact of the explanatory variables in our system.

The second method (network-based log linear regression) shows much better results than the first method and KNN.

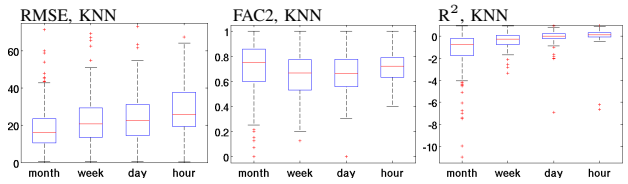

Fig. 11. The results of the KNN model: $RMSE$ (left), $FAC2$ (center), R^2 (right). Note: R^2 is mostly negative indicating that this statistical approach does not work well on this data.

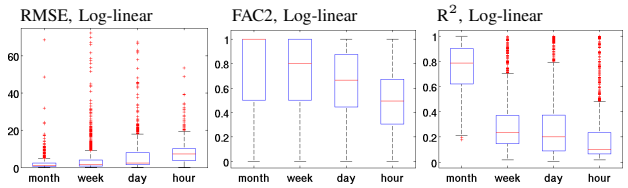

Fig. 12. The results of the log-linear regression model.

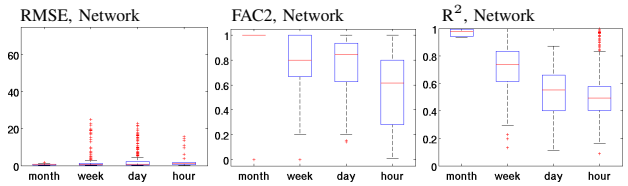

Fig. 13. The results of the Network-based log-linear regression model.

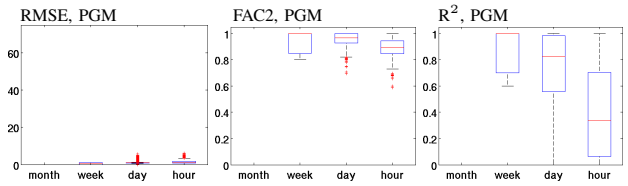

Fig. 14. The results of the probabilistic graphical model.

The RMSE, FAC2 and R^2 are all improved. This shows the impact of the proposed virtual network on the model.

The third method (probabilistic graphical model) outperforms the other three models. However, this model could not obtain results for the time resolution of "month" due to lack of enough data to build all the potentials correctly (we have less than 15 months of data so far). All the three metrics show good performance for this method, validating its effectiveness. An advantage about our PGM is that one model can capture all the dependencies between all the segments and also the dependencies between the LDSA values of each segment with the explanatory variables. Another good point about this method is that the results of the estimations are probabilistic and they show the certainty of an estimated value and the likelihood of another value.

V. Conclusion and Outlook

Three modeling methods using a real-world large scale mobile sensor network were proposed to generate high spatio-temporal resolution LDSA maps for an urban environment. The models can deal with dynamic coverage of the mobile sensor network. We topologically divided the city based on the street segments and showed that this way of space discretization is more efficient than its grid-based counterpart. The first method was a conventional log-linear regression model based on nine meteorological and gaseous explanatory variables. For the second and the third methods, we proposed creating a virtual network based on the dependencies of LDSA values in which each street segment is considered as one node of the graph and each edge represents correlations between two nodes. The second method is a novel network-based log-linear regression model that takes into account the LDSA values of the most correlated streets and also the nine explanatory variables from two static stations. The third model is a novel probabilistic graphical model which infers on the conditional joint probability distributions of the nodes and results in estimating the values in the nodes of interest.

More than 44 millions of geo- and time- stamped LDSA measurements (i.e., 14 months of real data) are used in this paper to evaluate the proposed modeling approaches in various time resolutions (hourly, daily, weekly and monthly).

Studying RMSE, FAC2 and R^2, we conclude that the proposed network-based models (the second and the third) show more promising results than the first method and KNN. In particular the third method (probabilistic graphical model) outperforms the other three models. One of the main advantages of the proposed probabilistic graphical model is that it builds one single model for the whole city and for the whole period. This model can capture all the dependencies between all the segments and also the dependencies between the LDSA values of each segment with the explanatory variables. The other good point about this method is that the results of the estimations are probabilistic and they show the certainty of an estimated value and likelihood of any other value. The only drawback is that when there is not enough data to accurately compute the joint distributions, the uncertainty grows in the output of this model.

In future, we will apply similar modeling methods for generating high resolution maps for other measured modalities (e.g., CO and NO_2). However this is a challenging task since the other sensors in our platform need to be carefully calibrated and their data need to be validated through non-trivial techniques considering their drift and aging.

Integrating land-use data into the models is another important future work which potentially can improve the quality of the maps. This is very useful specially for the third proposed method (PGM) since one single model was built for all streets. Differently from the temporal explanatory variables used so far in this paper, land-use data provides spatial characterization for the regions of the city. This complementary source of information would increase the performance of the models.

In the long-term, we also plan to crowd-source chemical sensors to citizens in order to increase the space- and time-resolution of the maps. However, going to this direction implies significant work on addressing privacy and data quality issues.

VI. Acknowledgment

The authors would like to acknowledge Thomas Coral, Jonathan Giezendanner, and Patrick Osterwalder for their student projects on analyzing the correlations between explanatory variables and the LDSA measurements and their log-linear regression models. We would like to thank Loïc Frund for his internship work on street matching algorithms. We also acknowledge Dr. Martin Fierz for helpful discussions and technical support on LDSA field measurements. Special thanks goes to Dr. Sarvenaz Choobdar for her insightful consults on probabilistic graphical modeling.

This work was funded by NanoTera.ch, a research initiative scientifically evaluated by the Swiss National Science Foundation and financed by the Swiss Confederation, in the framework of the OpenSense II project.

References

[1] World Health Organization (WHO). News release, 25 March 2014, Geneva. [Online]. Available: http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

[2] R. Beelen, O. Raaschou-Nielsen, M. Stafoggia, Z. J. Andersen, G. Weinmayr, B. Hoffmann, K. Wolf, E. Samoli, P. Fischer, and M. Nieuwenhuijsen, "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project," *The Lancet*, vol. 383, no. 9919, pp. 785–795, 2014.

[3] P. Avino, S. Casciardi, C. Fanizza, and M. Manigrasso, "Deep investigation of ultrafine particles in urban air," *Aerosol and Air Quality Resarch*, vol. 11, no. 6, pp. 654–663, 2011.

[4] J. Garcia, F. Teodoro, R. Cerdeira, L. Coelho, and M. Carvalho, "Developing a methodology to predict PM10 urban concentrations using GLM," *Air Pollution XXII*, vol. 183, p. 49, 2014.

[5] H.-J. Chu, B. Huang, and C.-Y. Lin, "Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship," *Atmospheric Environment*, vol. 102, pp. 176–182, 2015.

[6] J. Schwartz and L. M. Neas, "Fine particles are more strongly associated than coarse particles with acute respiratory health effects in schoolchildren," *Epidemiology*, vol. 11, no. 1, pp. 6–10, 2000.

[7] A. Nel, T. Xia, L. Mädler, and N. Li, "Toxic potential of materials at the nanolevel," *Science*, vol. 311, no. 5761, pp. 622–627, 2006.

[8] M. Auffan, J. Rose, J.-Y. Bottero, G. V. Lowry, J.-P. Jolivet, and M. R. Wiesner, "Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective," *Nature nanotechnology*, vol. 4, no. 10, pp. 634–641, 2009.

[9] T. M. Sager and V. Castranova, "Surface area of particle administered versus mass in determining the pulmonary toxicity of ultrafine and fine carbon black: comparison to ultrafine titanium dioxide," *Part Fibre Toxicol*, vol. 6, no. 15, pp. 1–11, 2009.

[10] V. J. Soppa, R. P. Schins, F. Hennig, B. Hellack, U. Quass, H. Kaminski, T. A. Kuhlbusch, B. Hoffmann, and G. Weinmayr, "Respiratory effects of fine and ultrafine particles from indoor sources - a randomized sham-controlled exposure study of healthy volunteers," *International Journal of Environmental Research and Public Health*, vol. 11, no. 7, pp. 6871–6889, 2014.

[11] J. Q. Cheng, M. Xie, R. Chen, and F. Roberts, "A latent source model to detect multiple spatial clusters with application in a mobile sensor network for surveillance of nuclear materials," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 902–913, 2013.

[12] Y. Chon, N. D. Lane, Y. Kim, F. Zhao, and H. Cha, "Understanding the coverage and scalability of place-centric crowdsensing," in *ACM Int. joint Conf. on Pervasive and Ubiquitous Computing*, 2013, pp. 3–12.

[13] B. Liu, O. Dousse, P. Nain, and D. Towsley, "Dynamic coverage of mobile sensor networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 2, pp. 301–311, 2013.

[14] K. Hu, T. Davison, A. Rahman, and V. Sivaraman, "Air pollution exposure estimation and finding association with human activity using wearable sensor network," in *Proceedings of the ACM Workshop on Machine Learning for Sensory Data Analysis*, 2014.

[15] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," *Mobile Sensing*, 2012.

[16] A. Jakeman, R. Simpson, and J. Taylor, "Modeling distributions of air pollutant concentrations-III. The hybrid deterministic-statistical distribution approach," *Atmospheric Environment*, vol. 22, no. 1, pp. 163–174, 1988.

[17] D. Oettl, "Documentation of the Lagrangian Particle Model GRAL (Graz Lagrangian Model Vs. 12.10," *Amt d. Stmk. Landesregierung, FA17C, Technische Umweltkontrolle, Bericht: Lu-03-12*, 2012.

[18] D. W. Wong, L. Yuan, and S. A. Perlin, "Comparison of spatial interpolation methods for the estimation of air quality data," *Journal of Exposure Science and Environmental Epidemiology*, vol. 14, no. 5, pp. 404–415, 2004.

[19] E. G. Dragomir, "Air quality index prediction using k-nearest neighbor technique," *Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII*, vol. 1, pp. 103–108, 2010.

[20] T. Hussein, A. Karppinen, J. Kukkonen, J. Härkönen, P. P. Aalto, K. Hämeri, V.-M. Kerminen, and M. Kulmala, "Meteorological dependence of size-fractionated number concentrations of urban aerosol particles," *Atmospheric Environment*, vol. 40, no. 8, pp. 1427–1440, 2006.

[21] B. Mølgaard, T. Hussein, J. Corander, and K. Hämeri, "Forecasting size-fractionated particle number concentrations in the urban atmosphere," *Atmospheric Environment*, vol. 46, pp. 155–163, 2012.

[22] S. Clifford, S. L. Choy, T. Hussein, K. Mengersen, and L. Morawska, "Using the generalised additive model to model the particle number count of ultrafine particles," *Atmospheric Environment*, vol. 45, no. 32, pp. 5934–5945, 2011.

[23] M. Reggente, J. Peters, J. Theunis, M. Van Poppel, M. Rademaker, P. Kumar, and B. De Baets, "Prediction of ultrafine particle number concentrations in urban environments by means of Gaussian process regression based on measurements of oxides of nitrogen," *Environmental Modelling & Software*, vol. 61, pp. 135–150, 2014.

[24] J. J. Li, A. Jutzeler, and B. Faltings, "Estimating Urban Ultrafine Particle Distributions with Gaussian Process Models," *In S. Winter and C. Rizos (Eds.): Research@ Locate14*, pp. 145–153, 2014.

[25] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268–285, 2015.

[26] Naneos Particle Solutions GmbH (2012). Partector nanoparticle dosimeter. Switzerland. [Online]. Available: www.naneos.ch/products.htm

[27] M. Zeri, J. F. Oliveira-Júnior, and G. B. Lyra, "Spatiotemporal analysis of particulate matter, sulfur dioxide and carbon monoxide concentrations over the city of Rio de Janeiro, Brazil," *Meteorology and Atmospheric Physics*, vol. 113, no. 3-4, pp. 139–152, 2011.

[28] X. Hu, L. A. Waller, M. Z. Al-Hamdan, W. L. Crosson, M. G. Estes, S. M. Estes, D. A. Quattrochi, J. A. Sarnat, and Y. Liu, "Estimating ground-level PM 2.5 concentrations in the southeastern US using geographically weighted regression," *Environmental Research*, vol. 121, pp. 1–10, 2013.

[29] A. Jutzeler, J. J. Li, and B. Faltings, "A region-based model for estimating urban air pollution," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 424–430.

[30] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 12–18, 2008.

[31] N. Schuessler and K. W. Axhausen, "Map-matching of GPS traces on high-resolution navigation networks using the Multiple Hypothesis Technique (MHT)," *Working paper 568, institute for transport planning and system (IVT)*, Oct. 2009.

[32] E. Limpert, W. A. Stahel, and M. Abbt, "Log-normal distributions across the sciences: Keys and clues," *BioScience*, vol. 51, no. 5, 2001.

[33] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 936.

[34] R. Kindermann and J. L. Snell, *Markov random fields and their applications*. American Mathematical Society Providence, RI, 1980, vol. 1.

[35] A. Chechetka and C. Guestrin, "Focused belief propagation for query-specific inference," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 89–96.

[36] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. of the 7th Conf. on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.