Self-learning Virtual Sensor Networks using low-cost electronics in Urban Geosensor Networks

Daniel Henzen
Technische Universität Dresden
Helmholtzstraße 10
01069 Dresden, Germany
Daniel.Henzen@tu-dresden.de

Pierre Karrasch
Technische Universität Dresden
Helmholtzstraße 10
01069 Dresden, Germany
Pierre.Karrasch@tu-dresden.de

Lars Bernard
Technische Universität Dresden
Helmholtzstraße 10
01069 Dresden, Germany
Lars.Bernard@tu-dresden.de

Abstract

Several studies indicate a close correlation of environmental pressure (e.g. air pollution) and human health impacts. An effective development of political and academic measures and programs relies on the availability of sufficient and suitable data for the characterization of the considered environmental system. Currently, this is realised primarily by using data of administrative observation stations that are provided by public authorities. These administrative observation stations provide highly accurate measurements. However, spatial coverage and resolution are limited. We address this issue by applying an innovative approach based on Crowdsourcing and Citizen Science methods. By equipping citizens with a new kind of low-cost environmental sensor system an additional environmental data source is established. The observations of low-cost sensors in the urban area are used to densify data from administrative observation networks. These established approaches are based on the assumption that a real sensor is available at a particular location in the observed area. To overcome this mandatory requirement, we introduce the concept of Virtual Sensor Networks consisting of Virtual Sensor nodes. Different statistical models on particular locations characterise these Virtual Sensor nodes and the self-learning character of this approach enables a permanent monitoring and improvement. Finally, the use of crowdsourcing strategies increases the number of available Virtual Sensor nodes, arises the Virtual Sensor Network and leads to an improvement of spatial modelling of environmental parameters.

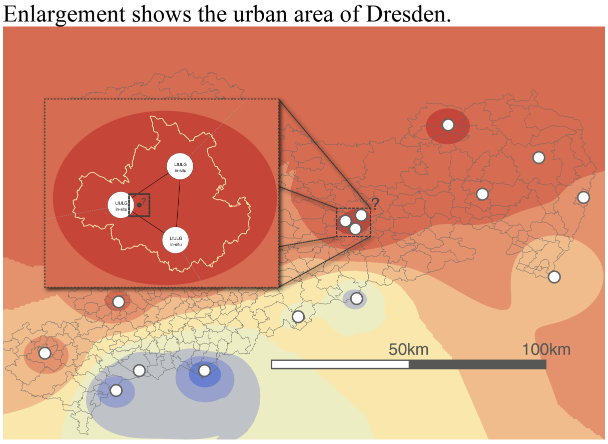
Keywords: Virtual Sensor Networks, low-cost sensors, Self-learning statistical models, Citizen Science

1 Urban Geosensor Networks

Availability of environmental data for research purposes has increased significantly over the last years. The data is either used as source for model building or as a proxy to deduce further environmental information (cf. [5, 10, 21]). This includes meteorological parameters and complex air pollution parameters that are typically available in different spatial and temporal resolutions. However, spatial and temporal resolution of officially available data are often insufficient e.g. for the consideration of local effects on human health and welfare and for the subsequent assessment of related measures. Related legal obligations as the European Air Quality Directive [6] require that "one sampling point shall be installed every 100,000 km^2". Although administrative observations provide highly accurate data, their spatial coverage and resolution are often insufficient for large-scale analyses. Furthermore, general distribution density of administrative observation stations often depends on the respective national stage of development.

Figure 1 shows the locations of available administrative observation stations within the Free State of Saxony, Germany (approximately an area of 18,240 km^2). A schematic representation of the interpolation of a measured air quality parameter is shown for a certain timestamp. Working with this interpolation in the context of large-scale analysis is prone to misinterpretations, particularly in urban areas where the high heterogeneity of urban environment (traffic, degree of sealing, vegetation cover, etc.) itself has a major impact on the spread of environmental parameters, e.g. on the occurrence of heat islands within sealed surfaces in urban areas. We enable large-scale analysis by placing low-cost sensors in urban areas to gather additional data supplementing administrative observations, and thereby facilitate the application of methods for data fusion from these different observation types.

Figure 1: Schematic representation of the interpolation of an air quality parameter for Saxony based on the administrative observation network (medium-scale official data). Enlargement shows the urban area of Dresden.



Currently, the usage of low-cost sensor systems for research purposes is steadily increasing. There are also various established environmental sensing systems such as the *Copenhagen Wheel* [16]. It is developed by Massachusetts Institute of Technology's Senseable City Lab as a modified bicycle with sensors to detect air and noise pollution, road conditions and traffic density. The *AirQualityEgg* (http://airqualityegg.com/) is an egg-shaped sensor box primarily designed for indoor use. Control of how and where the data is stored is limited. To address this issue and to improve usability, the research group GI@School extended the AirQualityEgg concept and designed it to be more autonomous and location-aware [3]. The result is known as *SenseBox* [1]. It can be operated out-of-the-box and allows an easy and customized configuration of sensors and communication.

Taking the SenseBox as an exemplary hardware basis for our approach, we propose to extend administrative Urban Geosensor Networks by wireless sensor networks. They consist of miniaturised computers called the sensor nodes [23] to monitor environmental phenomena [15]. Our sensor platform is based on microcontrollers and a set of varying shields and sensors that can be combined arbitrary to suit a specific analysis. An analysis may require a simple input like meteorological data but may also require more complex environmental parameters of air pollution.

2 Virtual Sensor Network

The observations of low-cost sensors in the urban area are used to complement administrative observation networks. As mentioned, crowdsourcing strategies applying low-cost sensors can be used to improve the spatial modelling of environmental parameters. Nevertheless, this approach is based on the assumption that a real sensor is available at a particular location in the observed area. To link administrative and low-cost sensors we introduce the concept of Virtual Sensor Networks. A Virtual Sensor Network is a combination of multiple Virtual Sensor nodes and offers capabilities to interact with these nodes as with physical sensors.

Unlike the concept proposed above (and detailed below), the notion of Virtual Sensors has been applied before, but with varying meanings: Havlik et al. [7] introduced the term Virtual Sensor in terms of an application of a model of raw sensor data to produce a higher-level data product [8]. Extended by Hill et al. [9] the concept includes not only derived data products but also the applied workflow. Both approaches lack in the definition of standardised workflows and output products. Watson and Watson [20] discussed how open standards can be used to facilitate the workflow steps and to improve workflow flexibility to provide statistical models. Already Schade and Craglia [17] described the necessity to integrate the domains of environmental models and the Sensor Web Enablement (SWE; http://www.ogcnetwork.net/swe). Nevertheless, Watson and van der Schaaf [19] the first time considered this as *model as a sensor*.

Contrary to this, Ngai et al. [14] suggest a different approach: (virtual) sensor nodes do not collect the data themselves, but gather the data from mobile sensors passing by. Thus, in this paper Virtual Sensors are understood as sensor data gatherer.

2.1 Virtual Sensor Network concept

We define a Virtual Sensor node as a combination of the approaches of Watson and van de Schaaf [19] and Ngai et al. [14]: Virtual Sensor nodes provide observations from varying low-cost sensors taken within a certain area and transferred via various communication paths to our (centralised) sensor platform. The collected data can be heterogeneous – low-cost sensors can have different levels of accuracy and resolution – and are the basis for statistical models (see section 2.2). These statistical models involve a geographic location and scope including the area of application.

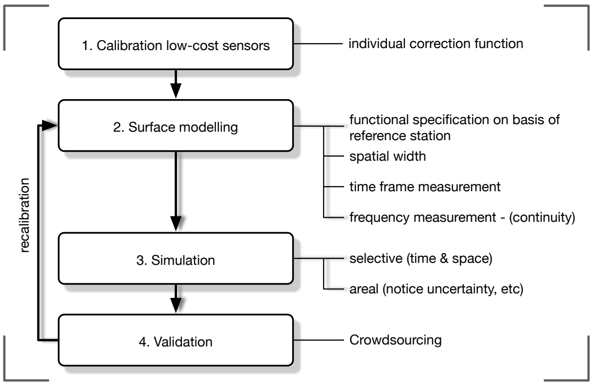
The determination of statistical models improves with increasing amount of measured data. Thus, we are able to deduce the value of a certain parameter in space and time based on the measured parameters from an administrative observation station. In turn, a low-cost sensor has to be calibrated via a suitable administrative observation station to correct and validate the low-cost sensor's individual measurement. However, even for a non-calibrated low-cost sensor the delta values (i.e. changes of parameter values over time) are still available for further analyses.

When multiple low-cost sensors are involved in a data collection process for a Virtual Sensor node, it is necessary to calibrate all statistical models for the low-cost sensors according to at least one administrative observation station to get valid observations. If calibration cannot be performed, multiple statistical models have to be stored for a Virtual Sensor node. When a non-calibrated sensor passes an administrative observation station, a recalibration is triggered, i.e. the correction of the low-cost sensors individual error in measurement and, hence, a recalibration of the non-calibrated statistical models, is carried out.

There are different levels of quality for the calibration of a Virtual Sensor. The involved (low-cost) sensors can be a) fully calibrated; b) partially calibrated; or c) not calibrated regarding to an administrative observation station. Further, it is assumed that a non-moving (stationary) sensor has a higher quality of measurement than the equivalent mobile version, because of the settling time (the time required for the response curve to reach and stay within a range of certain percentage of the final value [18]) of a sensor has a less impact on the observation. Furthermore, the Virtual Sensor node qualities for a given location, and therefore the statistical models are strongly related to the amount and the temporal resolution of available data. Data modelling is more precise if all necessary time intervals are covered by the measured data.

Figure 2 shows the workflow for creating Virtual Sensor nodes in a network. The first step (calibration of the low-cost sensors by means of an administrative observation stations to get an individual calibration function) is optional, however benefits all following steps. As a second step, low-cost observations are used to determine the nexus to administrative observations and hence determine further spatial interpolations and surface rendering. As a third step, the statistical model is validated by simulation with either a point or an areal representation as discontinuities of urban areas need to be considered, e.g. different pollution levels along a street. Eventually, the statistical models are validated and recalibrated by the power of crowdsourcing [2, 7] whenever required (step 4).

Figure 2: Workflow for creating a self-learning Virtual Sensor node.


1. Calibration low-cost sensors
individual correction function
2. Surface modelling
functional specification on basis of reference station
spatial width
time frame measurement
frequency measurement - (continuity)
recalibration
3. Simulation
selective (time & space)
areal (notice uncertainty, etc)
4. Validation
Crowdsourcing

In conclusion, the above deliberations lead to an operational definition of Virtual Sensor nodes. For each Sensor Node the following elements are essential:

i. sensor identification,
ii. observed environmental parameter,
iii. location of the Virtual Sensor node (including the specific coordinate and range),
iv. raw data and
v. statistical model, including the algorithm, the temporal coverage, the quality and uncertainty, the related administrative observation station, the statistical models provenance and the used data.

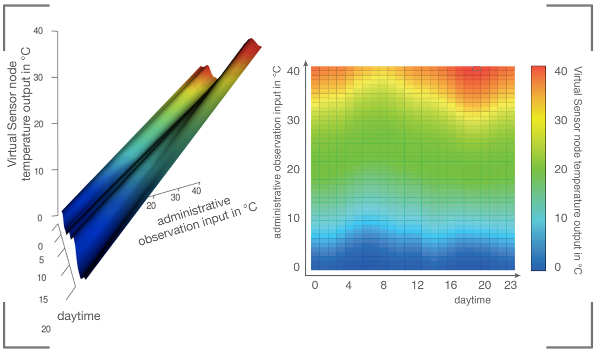
The following subsection introduces the generation of a single Virtual Sensor node based on surface modelling, simulation and validation (cf. Figure 2). Subsequently, the perspective is broadened from an individual Virtual Sensor node to a Virtual Sensor Network.

2.2 Generation of a Virtual Sensor node

The idea of the developed spatio-temporal statistical model is based on the assumption that every considered location has a specific behaviour concerning an environmental parameter in time. Therefore, the observations of a particular location can be written as a function of its characteristic environment, the time or time interval of measurement as well as an additional random component. It is further assumed that an administrative observation network exists and that it is possible to extract a functional relationship between the measurements of any pair of data of an administrative observation station and a low cost sensor for a particular location and time or time interval.

The implementation of the statistical model consists of two main steps. The first phase is the initialization. The design of the model requires comprehensive measurements for a particular location at different times. Based on these observations a first statistical model is deduced and describes the relation between the low-cost sensor observations as dependent variable and the observations of an administrative observation station as independent variable. Detailed statistical analysis during this phase of model building allows a further extension of the model. This implies the complexity of the statistical model itself (e.g. polynomial order) as well as a temporal clustering. Figure 3 demonstrates for example the variability of an ordinary linear regression model of temperature measurements over a day.

Figure 3: Example for an hourly-based array of linear regression models for temperature between low-cost sensor and administrative observation station (same statistical model, two different forms of visualization).



An in-depth residual analyses of all available (sub)models is the basis for an evaluation process. Based on these results the necessity for further or other temporal clustering can be estimated. The statistical results also provide valuable information about the model's accuracy. Thus, depending on the use case requirements for accuracy, these parameters can also lead to an exclusion of a specific statistical model.

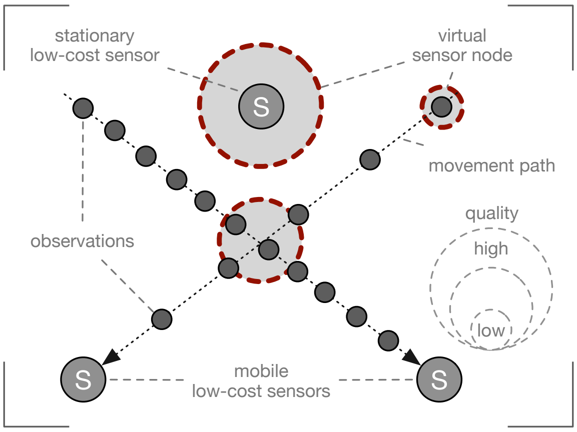
After the initial phase follows the self-learning-phase (cf. Figure 2; 2.-4.). In this phase the statistical models resulting from the initial phase are used to predict the observations only by means of the administrative observation stations for all locations for which particular statistical models are available. Regarding the crowdsourcing idea, new low-cost observations will be available from time to time. These observations can be used to recalibrate the existing statistical models or even to change their inherent structure.

Over time, the introduced approach also allows statements about the range of the appropriate statistical models. Depending on spatial heterogeneity, some statistical models are applicable in larger areas than others. A characterisation of the environments heterogeneity, derived from further data sources (topography etc.), can serve as an additional input variable.

2.3 Generation of a Virtual Sensor Network

Figure 4 shows the creation of Virtual Sensors in a system with three low-cost sensors. Based on the differing characteristics of the low-cost sensors – various measurement accuracies, ranges, etc. – the upper Virtual Sensors location (dashed circle) measures the highest data quality (largest diameter) since the statistical model bases on data from just one stationary low-cost sensor. The other low-cost sensors are mobile sensor nodes with different sampling rates (under the assumption of same movement speed or different movement speeds but same recording time) are moving from the top-left to the bottom-right and from the top-right to the bottom-left. The Virtual Sensor node on the right measures the lowest possible data quality (smallest circle) for the statistical model. Consisting of just one sample point in space and time, the results of the statistical model according to an administrative observation station are inaccurate. The Virtual Sensor node in the centre is based on multiple sampling points in space and time, and has thus a higher accuracy.

Figure 4: Concept of a Virtual Sensor Network data collection - different qualities in terms of spatial (mobile/stationary) and temporal (amount of data over time) characteristics of low-cost sensors.



Due to the high variability of the model and the permanent necessity for adjustment, it is mandatory to provide the statistical model in a flexible but still easily accessible way. This can be achieved by employing standardized Web Services (cf. [19]) like the Web Processing Service (WPS) or the Sensor Planning Service (SPS). However, the decision for either one does not guarantee the straightforward readjustment of a statistical model. For this purpose the concept and variability of the Moving Code paradigm can be applied [11, 13].

3 Conclusion and Future Work

In this paper, we establish our idea for Urban Geosensor Networks with low-cost electronics and a concept for a Virtual Sensor Network to fill the gap between medium-scale and large-scale observations with the aid of crowdsourcing strategies. For future work we focus on (amongst others):

i. automatic adaption of the currently static statistical models and how this can be integrated into an automatic process workflow,

ii. find similar locations where an extrapolation of a statistical model is feasible (cf. [22]),

iii. a detailed concept for the underlying technologies and the offered web service interfaces, as e.g. the recently published standard WPS 2.0 [12] – the current use case is restricted to a small number of sensor nodes integrated into a centralized system,

iv. an integration into an decentralized system as described by Duckham [4].

Another future topic is the transfer of the concept to other places with smart sensor activities cities that deal with air quality (e.g. Barcelona, Amsterdam, Madrid). One major requirement for this process is the availability of administrative observation stations in the desired place.

An additional but not yet considered aspect is the availability of sufficient observations: Effective and successful crowd sourcing and citizen science strategies [7] will be the decisive factor for the usability of the developed statistical models and primarily an organizational challenge. Using already known strategies to involve citizens are feasible (e.g. gamification, Web 2.0 collaborative competitions, promotions). Apart from this, the work on an integrative web-based platform providing access to the results is in progress and will be part of future discussions.

Acknowledgements

This research has been funded by the Excellence Initiative by the German Federal and State Governments (Institutional Strategy, measure "support the best").

References

[1] A. Bröring et al.: SenseBox – A Generic Sensor Platform for the Web of Things. In: A. Puiatti and T. Gu, editors, *Mobile and Ubiquitous Systems: Computing, Networking, and Services*. pages 186–196 Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[2] J. Burke et al.: Participatory Sensing. Presented at the Workshop on World-Sensor-Web WSW Mobile Device Centric Sensor Networks and Applications, 2006.

[3] D. Demuth et al.: The AirQuality SenseBox. *EGU General Assembly 2013*. 15, 2013.

[4] M. Duckham: Decentralized Spatial Computing. Springer Science & Business Media, Berlin, Heidelberg, 2012.

[5] EEA: Air quality in Europe - 2014 report. European Environment Agency, 2014.

[6] European Parliament and Council of the European Union: Directive 2008/50/EC of the European parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe. 2008.

[7] M.F. Goodchild: Citizens as sensors: the world of volunteered geography. *GeoJournal*. 69, 4, pages 211–221, 2007.

[8] D. Havlik et al.: From sensor to observation web with environmental enablers in the future internet. *Sensors*. 2011.

[9] D.J. Hill et al.: A virtual sensor system for user-generated, real-time environmental data products. *Environmental Modelling & Software*. 26, 12, pages 1710–1724, 2011.

[10] N. Janssen and S. Mehta: Human exposure to air pollution. *Air quality guidelines Global update. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide*. pages 61–85 World Health Organization Regional Office for Europe, 2006.

[11] M. Müller and L. Bernard: Moving Code in Spatial Data Infrastructures – Web Service Based Deployment of Geoprocessing Algorithms. *Transactions in GIS*. 2010.

[12] M. Müller and B. Pross: OGC WPS 2.0 Interface Standard. 2015.

[13] M. Müller et al.: Moving code – Sharing geoprocessing logic on the Web. *ISPRS Journal of Photogrammetry and Remote Sensing*. 83, C, pages 193–203, 2013.

[14] E.C.H. Ngai and P. Gunningberg: Quality-of-information-aware data collection for mobile sensor networks. *Pervasive and Mobile Computing*. 11, pages 203–215, 2014.

[15] S. Nittel et al.: Report from the first workshop on geo sensor networks. *ACM SIGMOD Record*. 33, 1, pages 141–144, 2004.

[16] C. Outram et al.: The Copenhagen Wheel: An innovative electric bicycle system that harnesses the power of real-time information and crowd sourcing. *EVER International Conference on Ecological Vehicles and Renewable Energies*. 2010.

[17] S. Schade and M. Craglia: A Future Sensor Web for the Environment in Europe. Presented at the Proceedings of the 24th International Conference on Informatics for Environmental Protection, 2010.

[18] T.-T. Tay et al.: High Performance Control. Springer Science & Business Media, 1998.

[19] K. Watson and H. van der Schaaf: Describing models on the web using OGC standards. *International Congress on Modelling and Simulation*. pages 1–7, 2013.

[20] V. Watson and K. Watson: Design of a software framework based on geospatial standards to facilitate environmental modelling workflows. *International Environmental Modelling and Software Society iEMSs*. 2012.

[21] WHO: Global Health Risks - Mortality and burden of disease attributable to selected major risks. 2009.

[22] S. Wiemann et al.: Classification-driven air pollution mapping as for environment and health analysis. Presented at the International Environmental Modelling and Software Society, 2012.

[23] F. Zhao and L.J. Guibas: Wireless Sensor Networks. Morgan Kaufmann, 2004.