Project title:

Development of sensor-based Citizens' Observatory Community for improving quality of life in cities

Acronym: **CITI-SENSE** Grant Agreement No: **308524**

EU FP7- ENV-2012

Collaborative project

Deliverable D 6.6

Data fusion of crowdsourced air quality observations and dispersion model data for urban-scale air quality mapping

Work Package 6

Date: 27.09.2016

Version: 1.0

Leading Beneficiary:	Norwegian Institute for Air Research (NILU)
Author(s):	P. Schneider (NILU), N. Castell (NILU), W. Lahoz (NILU), I. Vallejo (NILU)
Dissemination level:	RE (Restricted)

Versioning and contribution history

Version	Date issued	Description	Contributors
0.1	15. 09. 2016	Initial draft	P. Schneider (NILU), W. Lahoz (NILU)
0.2	22. 09. 2016	Comments from W. Lahoz incorporated	P. Schneider (NILU), W. Lahoz (NILU)
0.3	26. 09. 2016	Comments from Milos Davidovic incorporated	P. Schneider (NILU), W. Lahoz (NILU), Milos Davidovic (Vinca)
0.4	26. 09. 2016	Comments from David Kocman incorporated	P. Schneider (NILU), W. Lahoz (NILU), David Kocman (IJS)
1.0	27. 09. 2016	Submitted version	P. Schneider (NILU), W. Lahoz (NILU)

Peer review summary

Internal review 1			
Reviewer	Milos Davidovic (Vinca Institute of Nuclear Science, University of Belgrade, Serbia)		
Received for review	23. 09. 2016	Date of review	26. 09. 2016

Internal review 2			
Reviewer	David Kocman (Jožef Stefan Institute, Ljubljana, Slovenia)		
Received for review	23. 09. 2016	Date of review	26. 09. 2016

Executive Summary

We developed a methodology to use crowdsourced observations of air quality as collected in the CITI-SENSE project for deriving high-resolution urban-scale air quality maps. In this document we report about these efforts and how we used data fusion/data assimilation techniques to add value to the observations of pollutants (e.g., NO_2, PM_{10}, $PM_{2.5}$) from low-cost microsensors. We discuss the general concept of data assimilation and, as its subset, data fusion, and its application to citizen science observations. We then present details of the data fusion system used in CITI-SENSE, and illustrate its potential by applying it to the city of Oslo. We discuss the results and provide a summary of the lessons learned as part of this activity.

The method used for the mapping is based on geostatistical techniques and combines near real-time information from a network of low-cost air quality sensor platforms with information from an urban-scale air pollution dispersion model. The results indicate that the method is capable of providing realistic up-to-date maps of urban quality under moderate to strong pollution episodes when the sensors exhibit a good signal-to-noise ratio. The maps inherit the spatial patterns provided by the model and are thus able to contain information even in areas where no observations are available, while at the same time the maps inherit the overall magnitude of the sensor observations which are used to locally modify the underlying spatial patterns and concentration values.

The accuracy of the maps resulting from the data fusion methodology is dependent on the quality of both the observations and, to a somewhat lesser extent, on the accuracy of the model used. For the latter, it is primarily the spatial patterns rather than the actual modelled concentration values that are important. High systematic biases in particular can negatively affect the quality of the data fusion maps, but also unrealistic random outliers in the data can be problematic. To some extent the systematic biases of the sensors can be corrected for using appropriate co-location with reference stations before deployment and applying the resulting calibration equations, however some of the sensors exhibit biases that shift over time due to the environmental conditions and such errors will negatively affect the overall mapping accuracy.

The methodology is limited by the overall number of sensor platforms deployed throughout a city. Based on theoretical considerations as well as tests using simulated observations we currently consider a number of approximately 50 sensors within a location to be the minimum number that allows relatively robust mapping results. This number primarily stems from being able to derive an appropriate function for spatial autocorrelation and is not necessarily related to the area of the mapping domain. However, while this was not tested specifically, based on theoretical consideration the mapping accuracy in general should increase with a decreasing area of the mapping domain.

Overall, we think that using data assimilation and data fusion methods has significant potential for generating realistic and accurate maps of urban air quality in an up-to-date fashion, particularly given the likely future evolution of the sensor devices. Over the next several years, the sensor technology will mature significantly, resulting in devices with much higher accuracies (particularly lower bias and lower inter-sensor differences) as well as a smaller size, and they will be available at significantly reduced cost. As such, air quality sensors are likely to be ubiquitous in the future and robust methods for making sense of all the observations will be very important.

Table of contents

EXECUTIVE SUMMARY **3**
TABLE OF CONTENTS **4**
1 INTRODUCTION **5**
2 BACKGROUND **6**
2.1 DATA ASSIMILATION/DATA FUSION 6
2.2 APPLICATION FOR CROWDSCOURCED DATA 7
3 METHODOLOGY **9**
4 RESULTS **12**
4.1 SIMULATED OBSERVATIONS 12
4.2 REAL-WORLD CROWDSOURCED OBSERVATIONS 14
5 CONCLUSIONS **18**
ACKNOWLEDGEMENTS **20**
6 REFERENCES **21**
ANNEX **23**

1 Introduction

Small and low-cost sensor platforms measuring air quality have considerable potential for detailed mapping of urban air quality. Due to their relatively low price and small size, they can be deployed throughout the urban environment in much larger numbers than would be feasible with traditional air quality monitoring stations based on standard reference equipment. This significantly increased deployment density allows for a much more comprehensive set of observations of parameters relevant for air quality and for creating highly detailed urban air quality maps.

We present here a summary of the work carried out along these lines within Work Package 6 of the CITI-SENSE project. We describe a data fusion technique for combining near real-time crowdsourced observations of urban air quality with output from an urban-scale air pollution dispersion model that allows for providing highly detailed, up-to-date maps of urban air quality. Data fusion is conceptually similar to data assimilation [*Kalnay*, 2003; *Lahoz and Schneider*, 2014]. Data fusion describes a set of techniques for merging two or more datasets and thus generating a product of higher overall quality. Data fusion techniques, as a subset of data assimilation [*Lahoz and Schneider*, 2014], allow for combining observations with model data in a mathematically objective way (through the best linear unbiased estimate) and therefore provide a means of adding value to both the observations and the model. The method fills in the gaps in the observations and constrains the model with the observations. The model further provides detailed spatial patterns in areas where no observations are available. As such, data fusion of observations from high-density low-cost sensor networks together with models can contribute to significantly improving urban-scale air quality mapping.

This document is structured as follows: First we give some background on data assimilation and data fusion techniques and discuss how these methods apply to crowdsourced [*Howe*, 2006] information and similar data coming from citizen science techniques [*Hand*, 2010]

2 Background

Observations obtained using crowdsourcing and citizen science methods often contain significant data gaps and have issues regarding accuracy, which can complicate their use for detailed urban air quality mapping. One method to overcome such issues is to use data assimilation or data fusion methods that combine the observations with a spatially exhaustive information acquired from a model. In the CITI-SENSE project, we used data fusion to combine information from crowdsourced observations derived from AQMesh sensor pods [*Environmental Instruments*, 2016] that measure NO, NO_2, O_3, CO, PM, temperature, humidity, noise and air pressure, with high-resolution data from a model. The model information was derived from the urban air pollution dispersion model EPISODE [*Slørdal et al.*, 2003] in the case of Oslo and from statistical land-use regression models (for all other CITI-SENSE locations). The result of the data fusion process are up-to-date spatially exhaustive maps of urban air quality that inherit the spatial patterns from the model while at the same time being adjusted with regard to the actual observations of air pollutant concentrations at the various AQMesh pods. In the following, we give some general background on data assimilation and data fusion as a subset, and discuss the application of such methods for use with observations coming out of crowdsourcing and citizen science projects.

2.1 Data assimilation/Data Fusion

A brief overview of the general background of data assimilation and data fusion is presented here. For more detail, please see an already published article resulting from the work that has been carried out as part of the CITI-SENSE project [*Lahoz and Schneider*, 2014].

Information. We have two broad sources of information of the Earth System: measurements, i.e., "observations"; and understanding of the spatio-temporal evolution, typically embodied in "models," e.g., representing equations describing relationships between variables and/or parameters. Model information typically embodies our understanding of the system of interest, e.g., the Earth System. The observational and model information have uncertainty, and a key task is to understand and quantitatively estimate this uncertainty.

A second aspect of observational information is that it has spatio-temporal gaps [*Lahoz and Schneider*, 2014]. To fill in the gaps we need a model, which can be as simple as linear interpolation or as complex as representing the Navier-Stokes equations of the atmosphere. We can understand the model as an intelligent interpolator of the observational information. We would like to fill in the gaps in an objective manner, e.g., by minimizing a penalty function calculated from observational information and prior information of the system (e.g., from a model forecast). A methodology that allows this intelligent interpolation is data assimilation [*Lahoz and Schneider*, 2014]. It has strong links to several mathematical disciplines, including control theory and Bayesian estimation.

Data assimilation. Data assimilation adds value to the observations by filling in the observational gaps, and adds value to the model by constraining it with observations – see Fig. 2 in [*Lahoz and Schneider*, 2014]. This allows self-consistent and realistic representation of the Earth System on a regular grid. In this way, data assimilation allows one to "make sense" of Earth Observation. In particular, data assimilation provides methods for combining in an objective way observations and models with different spatio-temporal characteristics and errors: local footprint vs. quasi-global footprint; local coverage vs. global coverage; differences in sampling frequency; and errors arising from matching different spatio-temporal scales. When we combine the observational and model information and their errors in data assimilation, we term the result the "analysis." We will never know precisely the errors

in the observations, models and the analyses, so we need to estimate them. This means we have to state the data assimilation problem in statistical terms. The weather forecasting agencies provide an example of how data assimilation combines heterogeneous observational and model information [*Kalnay*, 2003].

Bayesian estimation defines a systematic and rigorous approach to data assimilation [*Rodgers*, 2000]. However, its full-scale implementation in many areas, including weather forecasting, is impossible, chiefly due to the size of the problem. The typical dimension of current weather forecasting models is $\sim 10^7$ elements, while the number of observations available over 24h is $\sim 10^6$–10^7 [*Lahoz and Errera*, 2010; *Lahoz and Schneider*, 2014]. As a result, error covariance matrices for the model and observational information have $\sim 10^{14}$ elements. However, the Bayesian approach is still useful in that it provides general guidelines for developing a data assimilation system and evaluating its results. Nevertheless, in many practical applications we need to make simplifying assumptions to the data assimilation methodology. There are two main lines of work: (i) statistical linear estimation and (ii) ensemble assimilation.

In the statistical linear approach, there exist two broad classes of numerical algorithms for data assimilation: variational and sequential [*Bouttier and Courtier*, 1999]. These algorithms take respectively the form of the 4-D variational method (4D-Var), or the Kalman filter (KF). These are two different algorithms for determining the best linear unbiased estimate (BLUE) and they are equivalent only under the condition of linearity. Statistical linear estimation achieves Bayesian estimation when the system is linear and the errors are Gaussian.

The ensemble assimilation approach is a form of Monte-Carlo approximation, which attempts to estimate the error covariance matrices using a finite number of ensemble members. In the ensemble Kalman filter, EnKF [*Evensen*, 2003], we use a Monte-Carlo ensemble of short-range forecasts to estimate the short-term forecast error.

The major drawback of the algorithms introduced above (variational methods; sequential methods; ensemble methods such as the EnKF) is the underlying assumption that the model states have a Gaussian distribution. A modification to the KF (the extended Kalman Filter, EKF) can handle some departure from Gaussian distributions of model errors and non-linearity of the model operator (which evolves the model forward in time). However, if the model becomes too non-linear or the errors become highly skewed or non-Gaussian, the trajectories computed by the EKF will become inaccurate. A development in data assimilation using ensemble methods that addresses non-linear and non-Gaussian aspects is the particle filter, PF [*van Leeuwen*, 2009].

2.2 Application for crowdscourced data

Citizen Science. Activities from citizens involved in science ("Citizen Science") provide a novel and recent development in platforms for observing the Earth System, potentially complementing the traditional ways of observing the Earth System, viz., satellite and ground-based and in situ platforms. Citizen Science activities have been described as people accumulating knowledge in order to learn about and respond to environmental threats [*Irwin*, 1995], and as public participation in scientific research [*Rosner*, 2013]. In the EU, the Seventh Framework Programme for Research has funded several new Citizen Science initiatives – these include the CITI-SENSE project (http://www.citi-sense.eu). These projects explore the potential of Citizen Science to provide information on the environment (e.g., air quality, meteorological conditions), and inform environmental policymaking, and complement established environmental data and information systems.

While they were not used for the purposes of the mapping techniques described in this document, portable sensors are a very important tool in the toolbox of Citizen Science. They often rely on the fact that smartphones are increasingly ubiquitous, given growth in mobile use, changes in mobile usage, and the increasing range of features provided to mobile phone users. Through a smartphone, the citizen can provide and receive information on their immediate environment, e.g., at the most basic level using only the phone's internal sensors on temperature, noise, movement, location, or on a wide variety of other parameters using external sensor packs. This includes air quality parameters such as NO_x ($NO+NO_2$), CO, ozone, and aerosols (particulate matter, PM), measured by deploying small, low-cost external microsensors, and using the smartphone as the main communications device. In addition, smartphones allow users to provide geo-located observations on nearly any generic parameter using specific apps. There are several applications of the concept of Citizen Science. Examples include (i) the WOW (Weather Observations Website; http://wow.metoffice.gov.uk) project at the Met Office, UK – a way to obtain information on various meteorological parameters (temperature, rainfall rate, and snowfall) in the UK, and (ii) temperatures in an urban environment using solely the internal battery temperature sensors of smartphones [*Overeem et al.*, 2013]. Portable instruments measuring air quality were tested within the framework of the CITI-SENSE project, however due to the high spatio-temporal variability of air pollution, the non-systematic fashion in which such sensors are generally handled by the public, and the difficulty of assigning an area of representativeness to such measurements, they were not applied for the air quality mapping technique reported on here. Only non-moving (static) sensors deployed at fixed locations were used for such purposes.

Citizen Science and data assimilation. A natural path to follow with Citizen Science information is the use of data assimilation to add value to it, in the same way that it does for traditional observation platforms (see above). The use of Citizen Science for data assimilation brings its own challenges. These include the following. (i) Significantly different spatial scales compared to those at which data assimilation is traditionally performed (10–100 km vs. street level, i.e., 10–100s of metres)—see Fig. 9 in [*Lahoz and Schneider*, 2014]. (ii) Model development, e.g., the need to simulate smaller spatial scales. (iii) Noisy information from users and from microsensors [*Shanley et al.*, 2013]. (iv) Representation of uncertainty in a way that is user-friendly and informative [*Spiegelhalter et al.*, 2011]. A further challenge is the merging of data from traditional sources such as satellite and ground-based and in situ platforms, and data provided by Citizen Science.

In the CITI-SENSE project we evaluate several approaches of increasing complexity for providing users with gridded fields and for investigating the feasibility of using data assimilation (and its subset data fusion) techniques with observations acquired by Citizen Science. The efforts in the CITI-SENSE project focus on data fusion, where we replace the short model forecast information in data assimilation by a climatology.

3 Methodology

The data fusion methodology applied here is based on geostatistical principles [*Isaaks and Srivastava*, 1989; *Cressie*, 1993; *Goovaerts*, 1997; *Kitanidis*, 1997; *Wackernagel*, 2003; *Webster and Oliver*, 2007; *Sarma*, 2009; *Chilès and Delfiner*, 2012]. It uses universal kriging to combine observations with model data by predicting the concentrations at unknown locations by simultaneously interpolating the observations and using the model data to provide information about the spatial patterns.

In contrast to ordinary kriging, universal kriging allows the overall mean to be non-constant throughout the domain and to be a function of one or more explanatory variables. Universal kriging is similar to kriging with external drift and mathematically equivalent to regression kriging [*Hengl et al.*, 2007] or residual kriging [*Denby et al.*, 2010; *Horálek et al.*, 2013] but can perform the linear regression against auxiliary variables and the spatial interpolation of the corresponding residuals in a single step. Universal kriging assumes a non-stationary mean and in addition the presence of local spatial variation. As such, the parameter in question is modelled by a deterministic regression component that provides the large-scale spatial variation and provides spatial patterns in areas where no observations are available, and a kriging component that provides the small-scale random variation.

In general, we compute the estimated concentration $\hat{Y}(s_0)$ at point s_0 as

$$\hat{Y}(s_0) = c + a_1 \cdot x_1(s_0) + a_2 \cdot x_2(s_0) + \cdots + a_p \cdot x_p(s_0) + \varepsilon(s_0) \quad (1)$$

Where c is a constant, a_1, a_2, etc. are regression coefficients, X_1, X_2, …, X_p are the values of the p predictor variables of the regression component, and ε is a stationary random process with a given semivariogram. In matrix notation we have

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1(s_0) & \cdots & x_p(s_0) \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_1(s_n) & \cdots & x_p(s_n) \end{bmatrix} \begin{bmatrix} c \\ a_1 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon} \quad (2)$$

Where $\boldsymbol{Y}$ indicates the estimated values at all prediction locations, $\boldsymbol{X}$ represents the values of the predictor variables at all locations, $\boldsymbol{a}$ is the vector of regression coefficients, ε indicates the vector of residual errors that is estimated using kriging with the known semivariogram model, n is the number of prediction locations, and p is the number of predictor variables.

In practice, the spatial trend or drift ε of the mean is estimated here using a single predictor variable, which is the long-term average concentration map provided by a model (for example, a dispersion model as was used in the case of Oslo, or a statistical land-use regression model as was used in the case of all other CITI-SENSE locations). In this case, there is only a single predictor variable, so Equation 1 simplifies to

$$\hat{Y}(s_0) = c + a_1 \cdot x_1(s_0) + \varepsilon(s_0). \quad (3)$$

The observations are provided by the AQMesh air quality sensors deployed throughout the urban environment. As such, the system takes the overall spatial patterns of the concentration field from the annual average map, which acts as a climatology (essentially a long-term mean), and adjusts this field based on the observations. The theoretical semivariogram required for calculating the covariances in the kriging process is fitted automatically to the empirical semivariogram for each new set of observations jointly with the respective basemap at specified time intervals.

Before the actual data fusion takes place, we first transform both the modelled and observed concentrations into log-space using the natural logarithm. This approach follows previous work such

as that carried out by [*Denby et al.*, 2008], [*De Smet et al.*, 2010], and [*Horálek et al.*, 2014] and is done because the frequency distribution of observed and modelled concentrations most often resembles the lognormal distribution. A log-transformation therefore is able to convert these distributions into an approximately Gaussian distribution, which we assume for universal kriging. Taking the lognormal distribution of the concentrations into account has further been shown to provide superior mapping accuracy [*Denby et al.*, 2008; *Horálek et al.*, 2013].

The theoretical semivariogram required for calculating the covariances in the kriging process was fitted automatically to the empirical semivariogram for each new set of observations (generally at hourly intervals). We kept the variogram model types the same as those derived for the model-derived basemaps, while the respective range of the models was allowed to vary by up to 30% around the values derived for the basemaps. We allowed the nugget and sill parameters to vary freely.

After universal kriging is carried out in log-space, the resulting concentration field and the corresponding mapping uncertainty have to be back-transformed from log-space. [*Denby et al.*, 2008] showed that the theoretical back-transformed expectation value of a concentration C is given as

$$E[C] = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{3}$$

where μ and σ represent the mean standard deviation of the log-normal transformed data, respectively. In practice, the concentration values resulting from the data fusion process are thus back-transformed by exponentiation with the kriging error as

$$\hat{Z}(s_0) = \exp\left[\hat{Y}(s_0) + \frac{\sigma^2(s_0)}{2}\right] \tag{4}$$

where $\hat{Z}(s_0)$ is the estimated back-transformed concentration value at point s_0, $\hat{Y}(s_0)$ is the concentration at point s_0 resulting from the data fusion process, and $\sigma(s_0)$ is the kriging standard deviation at point s_0 [*De Smet et al.*, 2010].

The theoretical back-transformed variance of the log-normal distribution is computed as

$$\mathrm{var}[C] = [\exp(\sigma^2) - 1] \cdot \exp[2\mu + \sigma^2] \tag{5}$$

Where μ and σ represent the mean and standard deviation of the log-normal transformed data, respectively [*Denby et al.*, 2008]. Thus, we can calculate the back-transformed standard deviation (uncertainty) $\delta(s_0)$ at point s_0 of the fused map in practice as

$$\delta(s_0) = \sqrt{\exp[(\sigma^2(s_0) - 1] \cdot \exp[2 \cdot \hat{Y}(s_0) + \sigma^2(s_0)]} \tag{6}$$

Where $\sigma(s_0)$ is the kriging standard deviation at point s_0 and $\hat{Y}(s_0)$ represents the concentration at point s_0 resulting from the data fusion process [*Denby et al.*, 2008; *De Smet et al.*, 2010].

In practice, the methodology was implemented in the R programming language. The R version used was R 3.2.4 Revised (2016-03-16 r70336). In addition, a variety of R packages was used, most notably the `sp` package (version 1.2-3) for providing spatial foundation classes [*Bivand et al.*, 2013], the `gstat` package version 1.1-3 [*Pebesma*, 2004] for performing the universal kriging and the `automap` package version 1.0-14 [*Hiemstra et al.*, 2009] for automated fitting of the semivariogram.

For mapping in Oslo we used the EPISODE dispersion model. EPISODE is a 3-D Eulerian/Lagrangian dispersion model that provides urban- and regional-scale air quality forecasts of atmospheric pollutants. The model, which is described in detail in [*Slørdal et al.*, 2003], is a Eulerian grid model with embedded subgrid models for computing the various pollutant concentrations that result from area-, point-, and line-based emission sources. Applying finite difference numerical methods, EPISODE integrates forward in time and solves the time-dependent advection and diffusion equation on a three-dimensional grid. EPISODE provides schemes for advection, turbulence, deposition, and chemistry. EPISODE contains a sub-grid line source model based on a standard integrated Gaussian model [*Peterson*, 1980], which computes the concentration levels of non-reactive pollutants from road traffic over distances up to hundreds of meters downwind. Most commonly, EPISODE is used for modeling airborne species such as NO_2, NO_x, PM_{10}, $PM_{2.5}$, CO, and SO_2. Validation studies have shown good correspondence between modelled and measured concentrations of NO_2, PM_{10}, and $PM_{2.5}$ [*Oftedal et al.*, 2009].

For more detail on the specific methodology and in particular for example on how to derive the basemaps from dispersion models or land-use regression models, please see deliverable D6.3.

4 Results

Here we present examples of the data fusion mapping approach. We first present examples using simulated observations and then show examples using real-world observations taken using the AQMesh sensors.

4.1 Simulated observations

Simulating artificial observations from a known ”true” concentration field has the advantage that the applied mapping methodology can be tested thoroughly and validated against the assumed true reference field.

Figure 1 shows an example of this process. The top left panel shows the ”true” concentration field, which we assume is the actual state of the atmosphere at a given moment in time. In practice, this field is of course unknown, and we wish to reproduce this field as closely as possible given the actual point-based observations that we have. In the top center panel we can see the point-based observations that were derived from the truth field using random errors and biases, as well as in the background the long-term average concentration map derived from the chemical transport model EPISODE (also called a *basemap*). It is these two datasets that we like to fuse or combine in order to get as close as possible to the original truth field. Note that the observations are significantly higher than the corresponding concentrations found in the basemap.

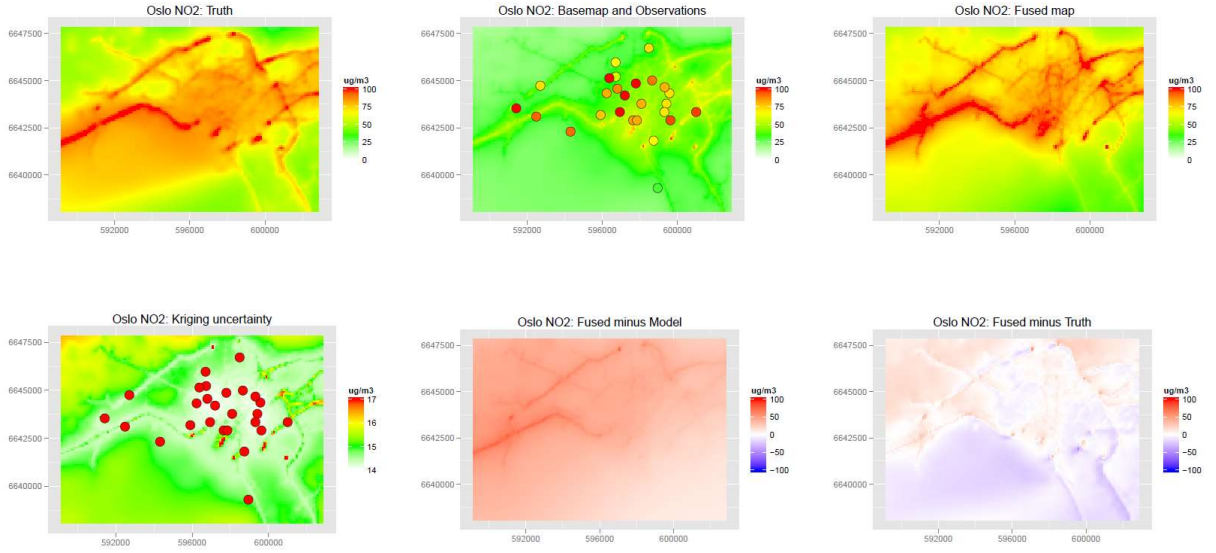

Figure 1 - Example of the data fusion methodology demonstrated using simulated observations. See text for explanation of the individual panels.

The top right panel shows the results of the data fusion process. We can see that the overall concentration field is now much close to the truth field in terms of overall concentration levels but also in terms of spatial patterns. While the fused concentration field cannot exactly replicate the truth field, it manages to do so at least in those areas where observations are abundant. The lower left panel shows the uncertainty associated with the mapping process as well as the locations of the observations. We can see that, as would be expected, the uncertainty of the fused concentration field is lowest in areas where there are many observations, whereas in those areas where there are only few or no observations at all, the uncertainty of the fused map increases. The bottom centre panel

shows the difference between the data fusion map and the modelled basemap, i.e., it indicates to what extent the basemap was modified by the algorithm in order to obtain the fusion map. Here we see that the basemap was increased throughout most of the domain but not everywhere equally. Stronger modifications were carried out in the centre and western part of the domain whereas the southeastern part was only weakly modified.

Finally, the bottom right panel shows the difference image between data fusion map and the original ”truth” map, i.e., this panel shows us how well the method performed (and where it performed better or worse). What we can see here is that in all areas that appear white the concentrations estimated by the data fusion process replicate the original true values very well. In areas where this difference map is red or blue the data fusion map over- or under-estimates the original true concentration field. The areas with the largest errors are located primarily where none or only a few observations were available.

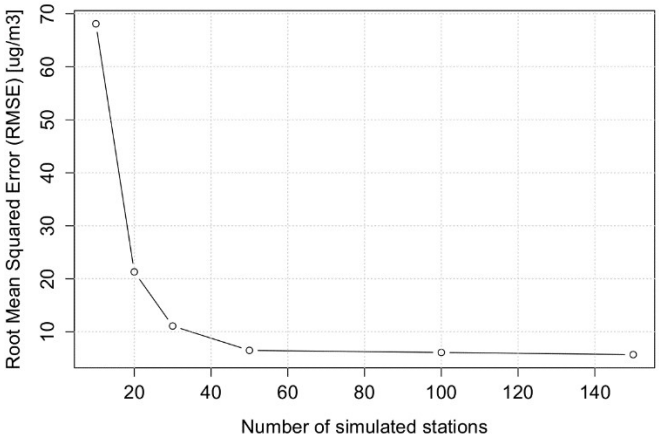

Figure 2 - The impact of the number of observations in a location on the mapping accuracy given as the RMSE derived from leave-one-out cross-validation. This is an example derived for mapping of NO_2 in Oslo.

Using simulated observations we were able to carry out a simple analysis for studying the minimum number of observations that are required for obtaining an acceptable mapping accuracy. We used continuously larger subsets randomly sampled from a set of 150 simulated observations with the data fusion algorithm to map NO_2 in Oslo. Each map was then validated against the observations using leave-one-out cross-validation. The results are illustrated in Figure 2. It shows how the mapping error rapidly decreases with an increasing number of observations. At a number of approximately 50 observations it reaches a value of around 5 µg/m^3, which does not improve significantly anymore with increasing numbers of observations as it is close to the original simulated uncertainty of the observations. This empirically found treshold of at least 50 observations matches the general rule of thumb in the geostatistical literature of the minimum number of samples required for providing a good estimate of the semivariogram [*Edwards and Fortin*, 2001]. It should be noted that this test was carried out using leave-one-out cross-validation against the actual sensor observations and not against independent referent observations and as such the RMSE cannot be interpreted as an overall uncertainty but rather represents the uncertainty related to the mapping procedure. In addition, Figure 2 shows only one example of this relationship – using a different set of simulated observations will results in a different outcome.

4.2 Real-world crowdsourced observations

Within the CITI-SENSE project, AQMesh sensors were deployed at volunteers' premises throughout several locations, including Oslo, Norway and Barcelona, Spain on which we focus here. The deployment strategy for most locations followed guidelines regarding achieving a suitable spatial distribution for land-use regression modelling. Among other purposes, the data were then used to calculate maps of urban air quality.

Figure 3 shows an example of how we applied the data fusion methodology for real-world observations carried out within the CITI-SENSE AQMesh monitoring network deployed in Oslo, Norway. In the top left panel we can see the model-derived basemap from the EPISODE model overlaid by the locations and values of the observed concentrations, in this case shown for NO_2 for 6 January 2016 at 09:00 UTC.

The top left panel of the figure illustrates the two input datasets that are required by the data fusion algorithm: The background map shows the long-term average concentration of NO_2 as modelled by the EPISODE chemical dispersion model, whereas the point markers indicate both the location of the crowdsourced measurement devices as well as the magnitude of the NO_2 concentration observed by each device. It can be observed that in this instance the observations are overall significantly higher than the long-term average concentrations. When the data fusion algorithm is applied to these databases the resulting concentration field (top right panel) is much more consistent with the observations. Each fused concentration field is associated with a map of uncertainty (bottom left panel) which illustrates qualitatively how the reliability of the mapping result varied in space and gives quantitative information about each grid cell's mapping uncertainty. Finally, the bottom right panel of Figure 3 shows the basemap correction, i.e., the amount by which each grid cell of the basemap (top left panel) had to be adjusted in order to achieve the data fusion result (top right panel). In this case all correction values are positive as all of the crowdsourced observations were significantly higher than concentrations given by the basemap, however the correction can also take negative values.

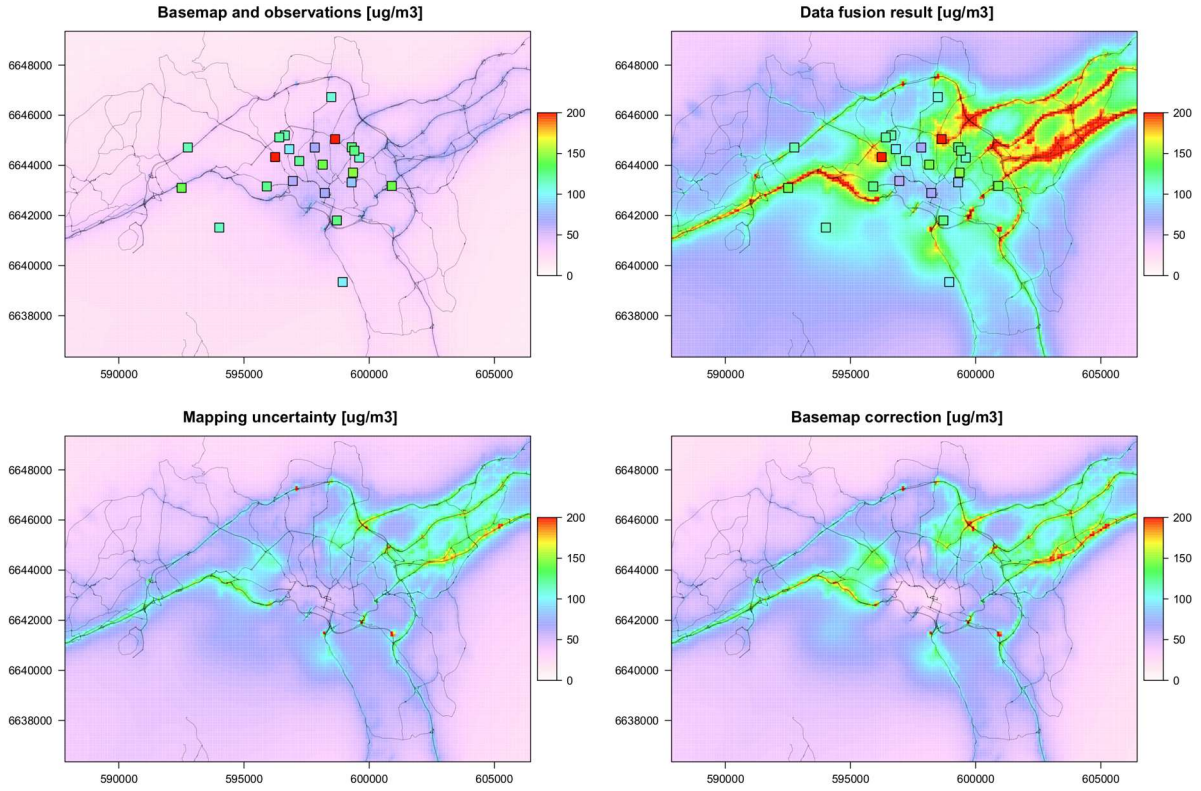

Figure 3 - Data fusion example with real-world observations in Oslo. The top left panel shows the model-derived annual average concentration basemap and the point-based observations from the AQMesh pods on 6 January 2016 at 09:00 UTC. The top right panel shows the result of the data fusion in the background with the observations shown as squares for reference. The bottom left panel shows the uncertainty associated with the geostatistical mapping, while the bottom right panel shows the amount of modification to the basemap to obtain the fusion map in the top right panel. For more explanation, please see the text.

The final fused maps can be easily displayed using a web interface. Both Figure 4 as well as Figure 5 show examples of this. Figure 4 shows how the actual concentration field coming out of the data fusion process can be displayed in a simple zoom-able and pan-able web mapping interface (in this case based on the Javascript library Leaflet). However, while such maps could be used internally within the CITI-Sense project, it was decided not to show such maps of actual concentrations of the individual species to the public. This was to some extent related to the relatively high uncertainty associated with these maps but also because the public is usually not aware of what certain concentration values as well as the individual species (NO_2, PM_{10}, $PM_{2.5}$) mean in general and for their personal health in particular. As such, a map showing an air quality index with coarse classes is generally a better choice for displaying this type of information to the public. Figure 5 shows an example of how we handled this within the CITI-Sense project. It illustrates the concept using a screenshot of the web portal developed by Dunavnet, which shows not only the observations from the static AQMesh sensors but also from mobile sensors units and user perception observations acquired by means of the CityAir app. In addition, the screenshot shows a data fusion-based map here converted from the actual concentrations of the various species to and Air Pollution INdex (APIN), which is based on the Common Air Quality Index (CAQI) [*van den Elshout et al.*, 2008; *Van Den Elshout et al.*, 2014]. These types of map still shows the overall spatial patterns of air quality in Barcelona, however it does so while remaining

easily interpretable without having to understand the difference between the different pollutants or the various pollutant-dependent thresholds for health effects.

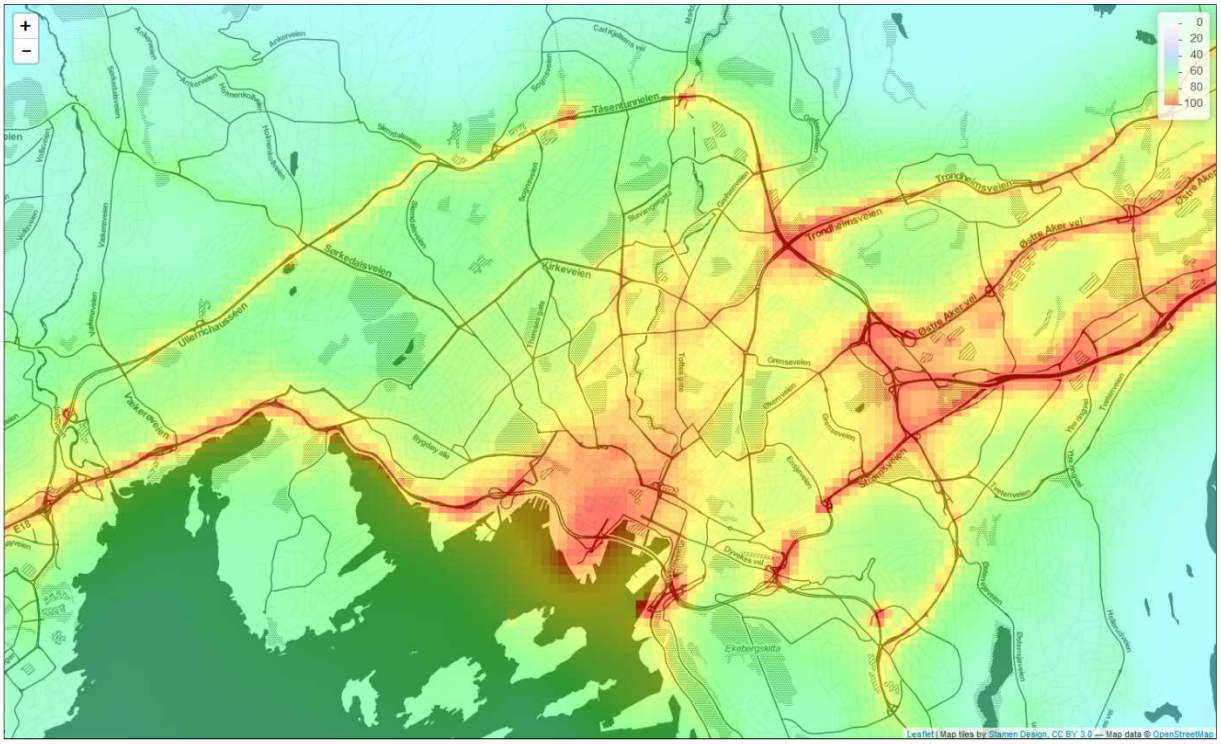

Figure 4 - Example of a possible online visualization of the data fusion maps including zooming and panning functionality. This type of interface is very useful for exploring the air quality in certain neighbourhoods of the city. Within CITI-SENSE the actual concentrations were only used for internal purposes and not shown to the general public. Instead the maps were converted into a more easily understandable air quality index called APIN.

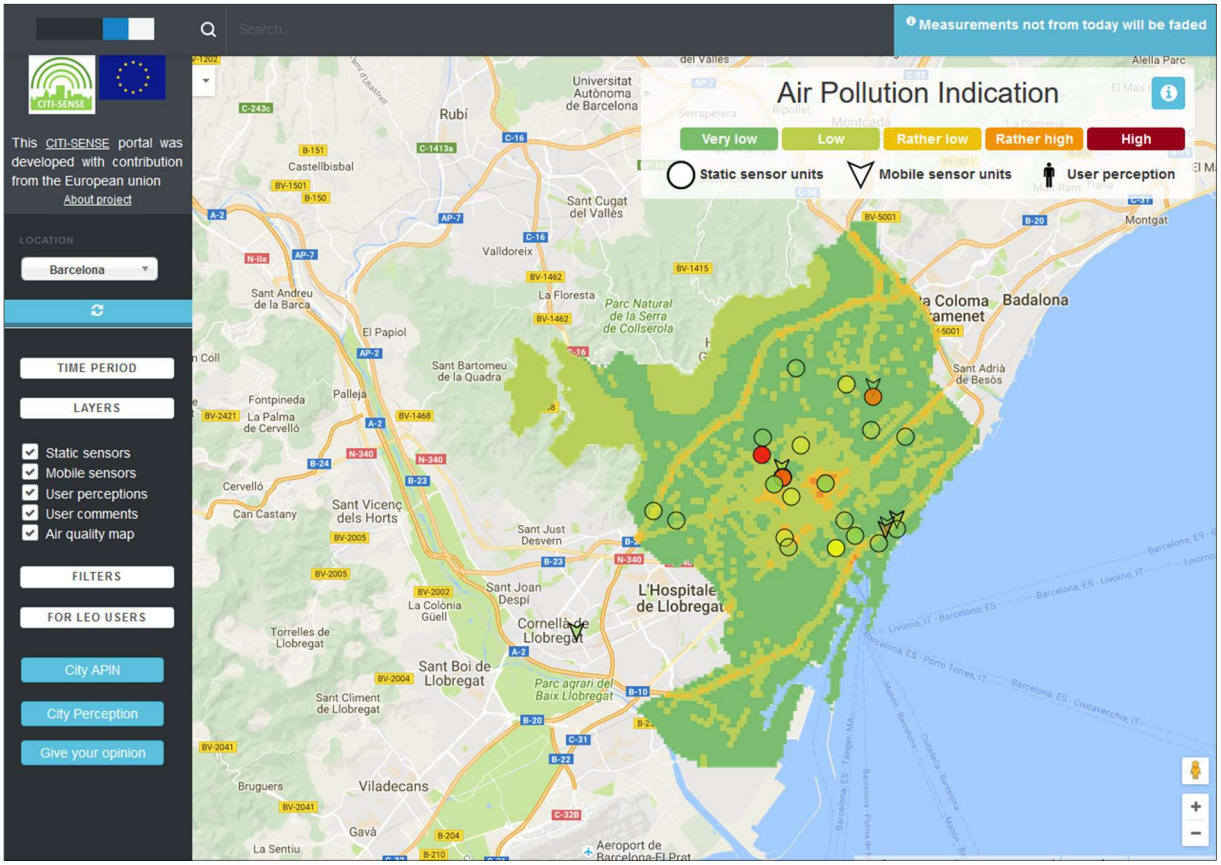

Figure 5 - Screenshot of the CITI-SENSE outdoor visualization portal, here showing an Air Pollution Indicator (APIN) map for Barcelona. The map is based on the methodology outlined above where the underlying concentration data was subsequently converted to the index classes of the APIN.

5 Conclusions

We developed a methodology to use crowdsourced observations of air quality as used in the CITI-SENSE project for deriving high-resolution urban-scale air quality maps. The method is based on geostatistical techniques and combines near real-time information form network of low-cost air quality sensor platforms with information from an urban-scale air pollution dispersion model. The results indicate that the method under moderate to high pollution conditions (resulting in a good signal-to-noise ratio of the sensors) is capable of providing realistic-looking up-to-date maps or urban quality. These maps inherit the spatial patterns provided by the model and are thus able to contain information even in areas where no observations are available, while at the same time the maps inherit the overall magnitude of the sensor observations which are used to locally modify the underlying spatial patterns and concentration values.

The accuracy of the maps resulting from the data fusion methodology is dependent on the quality of both the observations and, to a somewhat lesser extent, on the accuracy of the model used. For the latter it is primarily the spatial patterns rather than the actual modelled concentration values that are important. Strong biases of the sensor platforms in particular can negatively affect that quality of the data fusion maps, but also unrealistic random outliers in the data can be problematic. To some extent the latter were removed from the dataset using automated quality control methods, however these do not always work reliably. In addition, such methods are only capable of dealing with random errors, whereas systematic biases are significantly more difficult to correct for.

The methodology is clearly limited by the overall number of sensor platforms deployed throughout a city. Based on theoretical considerations as well as some real-world testing we currently consider a number of approximately 50 sensors within a city to be the minimum number that allows relatively robust mapping results. Within the CITI-SENSE project we were operating at numbers considerable below this threshold (on average around 20 sensor pods per city). The number of observations affects both the linear regression component of the method (erroneous outlying observations can affect the regression with a larger weight and sometimes cause negative/invalid slopes) and the determination of the empirical semivariograms, which in our tests due to the low number of samples did not always follow the traditional semivariogram shapes and were thus difficult to fit using a theoretical semivariogram, particularly in an automated fashion. A larger number of sensors deployed throughout a city is very likely to alleviate these issues.

Furthermore, with a large enough set of sensor platforms deployed throughout a city it is conceivable that the individual pods can be used to inter-calibrate each other in an automated fashion and to thus correct biases in calibration and high random errors. Such network calibration techniques have already been attempted in some settings, however for air quality applications in the urban environment they require relatively short maximum distances between individual pods as the spatial gradients of air quality are extremely steep and thus concentrations can vary considerably over distance of even only a few metres.

Independent of the implementation of such network-based calibration techniques, we believe that improved automated quality control of the data is absolutely crucial. Some rudimentary automated quality control mechanisms were implemented as part of the mapping activities in CITI-SENSE to pre-process the observational data before the actual data fusion, but more robust methods capable of dealing with calibration drift of the sensors due to environmental conditions are necessary.

Overall, we think that using data assimilation and data fusion methods has significant potential for generating realistic and accurate maps of urban air quality in an up-to-date fashion, particularly given the likely future evolution of the sensor devices. Over the next several years, the sensor technology will mature significantly, resulting in devices with much higher accuracies (particularly lower bias and lower inter-sensor differences) as well as a smaller size, and they will be available at significantly reduced cost. As such, air quality sensors are likely to be ubiquituous in the future and robust methods for making sense of all the observations will be very important.

Acknowledgements

Funding for this work has been provided by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 308524. The authors would like to thank the various CITI-SENSE locations officer for providing basemaps for testing the data fusion in the various cities.

6 References

Bivand, R. S., E. Pebesma, and V. Gómez-Rubio (2013), *Applied Spatial Data Analysis with R*, Second edi., Springer.

Bouttier, F., and P. Courtier (1999), *Data assimilation concepts and methods March 1999*.

Chilès, J.-P., and P. Delfiner (2012), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons.

Cressie, N. A. C. (1993), *Statistics for spatial data*, Wiley-Interscience, New York.

Denby, B., M. Schaap, A. Segers, P. Builtjes, and J. Horálek (2008), Comparison of two data assimilation methods for assessing PM10 exceedances on the European scale, *Atmos. Environ.*, *42*(30), 7122–7134, doi:10.1016/j.atmosenv.2008.05.058.

Denby, B., I. Sundvor, M. Cassiani, P. de Smet, F. de Leeuw, and J. Horálek (2010), Spatial mapping of ozone and SO2 trends in Europe., *Sci. Total Environ.*, *408*(20), 4795–806, doi:10.1016/j.scitotenv.2010.06.021.

Edwards, G., and M.-J. Fortin (2001), A Cognitive View of Spatial Uncertainty, in *Spatial Uncertainty in Ecology*, pp. 133–157, Springer New York, New York, NY.

van den Elshout, S., K. Léger, and F. Nussio (2008), Comparing urban air quality in Europe in real time. A review of existing air quality indices and the proposal of a common alternative, *Environ. Int.*, *34*(5), 720–726, doi:10.1016/j.envint.2007.12.011.

Van Den Elshout, S., K. Léger, and H. Heich (2014), CAQI common air quality index - update with PM2.5 and sensitivity analysis, *Sci. Total Environ.*, *488–489*, 461–468, doi:10.1016/j.scitotenv.2013.10.060.

Environmental Instruments (2016), Technical specification, Available from: http://www.aqmesh.com/

Evensen, G. (2003), The Ensemble Kalman Filter: theoretical formulation and practical implementation, *Ocean Dyn.*, *53*(4), 343–367, doi:10.1007/s10236-003-0036-9.

Goovaerts, P. (1997), *Geostatistics for natural resources evaluation*, Oxford University Press, New York.

Hand, E. (2010), Citizen Science - People power, *Nature*, *466*(August), 685–687, doi:10.1038/news.2010.106.

Hengl, T., G. B. M. Heuvelink, and D. G. Rossiter (2007), About regression-kriging: From equations to case studies, *Comput. Geosci.*, *33*(10), 1301–1315, doi:10.1016/j.cageo.2007.05.001.

Hiemstra, P. H., E. J. Pebesma, C. J. W. Twenhöfel, and G. B. M. Heuvelink (2009), Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network, *Comput. Geosci.*, *35*(8), 1711–1721, doi:10.1016/j.cageo.2008.10.011.

Horálek, J., P. De Smet, P. Kurfürst, F. De Leeuw, and N. Benešová (2013), *European air quality maps of PM and ozone for 2011 and their uncertainty*, Bilthoven, Netherlands.

Horálek, J., P. de Smet, P. Kurfürst, F. De Leeuw, and N. Benešová (2014), *European air quality maps of PM and ozone for 2010 and their uncertainty*, ETC/ACM Technical Paper 2014/4.

Howe, J. (2006), The Rise of Crowdsourcing, *Wired Mag.*, *14*(6), doi:10.1086/599595.

Irwin, A. (1995), *Citizen science: a study of people, expertise, and sustainable development*, Routledge.

Isaaks, E. H., and R. M. Srivastava (1989), *Applied geostatistics*, Oxford University Press, New York.

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK.

Kitanidis, P. K. (1997), *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press.

Lahoz, W., and Q. Errera (2010), Constituent Assimilation, in *Data Assimilation*, edited by W. A. Lahoz, B. Khattatov, and R. Ménard, pp. 449–490, Springer Berlin Heidelberg, Berlin, Heidelberg.

Lahoz, W. A., and P. Schneider (2014), Data assimilation: making sense of Earth Observation, *Front. Environ. Sci.*, *2*(16), 1–28, doi:10.3389/fenvs.2014.00016.

van Leeuwen, P. J. (2009), Particle Filtering in Geophysical Systems, *Mon. Weather Rev.*, *137*(12), 4089–4114, doi:10.1175/2009MWR2835.1.

Oftedal, B., S. E. Walker, F. Gram, H. McInnes, and P. Nafstad (2009), Modelling long-term averages of local ambient air pollution in Oslo, Norway: evaluation of nitrogen dioxide, PM10 and PM2.5, *Int. J. Environ. Pollut.*, *36*(2), 110–126, doi:10.1504/IJEP.2009.021820.

Overeem, a., J. C. R. Robinson, H. Leijnse, G. J. Steeneveld, B. K. P. Horn, and R. Uijlenhoet (2013), Crowdsourcing urban air temperatures from smartphone battery temperatures, *Geophys. Res. Lett.*, *40*(15), 4081–4085, doi:10.1002/grl.50786.

Pebesma, E. J. (2004), Multivariable geostatistics in S: The gstat package, *Comput. Geosci.*, *30*(7), 683–691, doi:10.1016/j.cageo.2004.03.012.

Peterson, W. B. (1980), Epa, *User's Guid. HIWAY-2. A Highw. Air Pollut. Model*.

Rodgers, C. D. (2000), *Inverse Methods for Atmospheric Sounding*, WORLD SCIENTIFIC.

Rosner, H. (2013), Data on Wings, *Sci. Am.*, *308*(2), 68–73, doi:10.1038/scientificamerican0213-68.

Sarma, D. D. (2009), *Geostatistics with Applications in Earth Sciences*, Springer Science & Business Media, Dordrecht, The Netherlands.

Shanley, L. A., R. Burns, Z. Bastian, and E. S. Robson (2013), Tweeting Up a Storm The Promise and Perils of Crisis mapping, *Photogramm. Eng. Remote Sens.*, (October 2013), 865–879.

Slørdal, L. H., S.-E. Walker, and S. Solberg (2003), *The Urban Air Dispersion Model EPISODE applied in AirQUIS 2003 - Technical Description*, Kjeller, Norway.

De Smet, P., J. Horálek, M. Conková, P. Kurfürst, F. De Leeuw, and B. Denby (2010), *European air quality maps of ozone and PM10 for 2008 and their uncertainty analysis*, Bilthoven, Netherlands.

Spiegelhalter, D., M. Pearson, and I. Short (2011), Visualizing Uncertainty About the Future, *Science (80-.).*, *333*(6048), 1393–1400, doi:10.1126/science.1191181.

Wackernagel, H. (2003), *Multivariate Geostatistics*, Springer Science & Business Media.

Webster, R., and M. A. Oliver (2007), *Geostatistics for Environmental Scientists*, John Wiley & Sons.

ANNEX

In the following we should additional figures that could not be included in the main text.

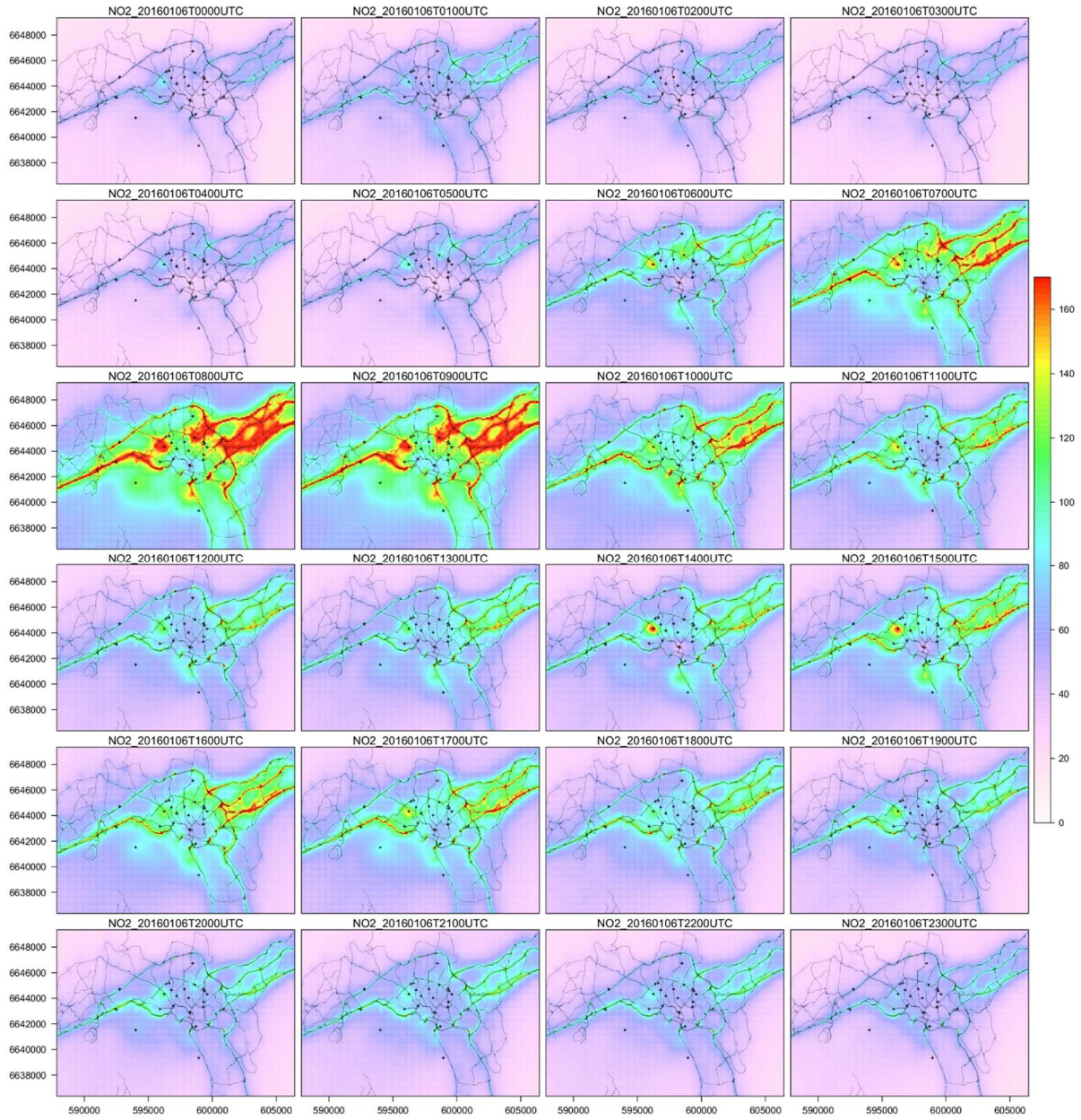

Figure 6 - Example of 24 hours of fused NO_2 maps for Oslo, Norway. The daily cycle of traffic emissions and the resulting higher concentration during morning and afternoon rush hour can be clearly seen.

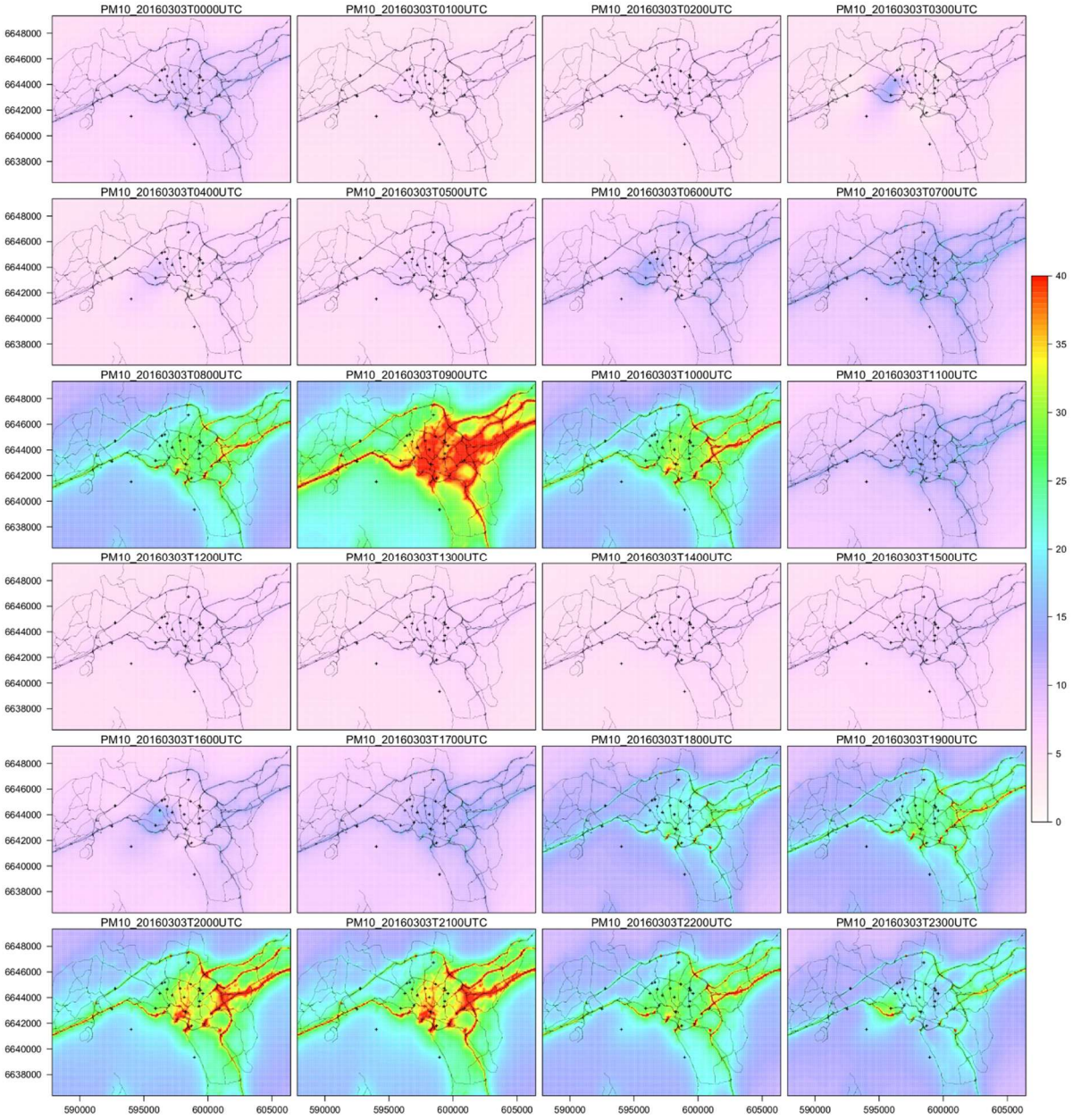

Figure 7 - Example of 24 hours of fused PM_{10} maps for Oslo, Norway. The morning rush hour pollution is clearly visible. In addition, we can observe the higher concentration in the evening due to heating from wood-fire ovens.

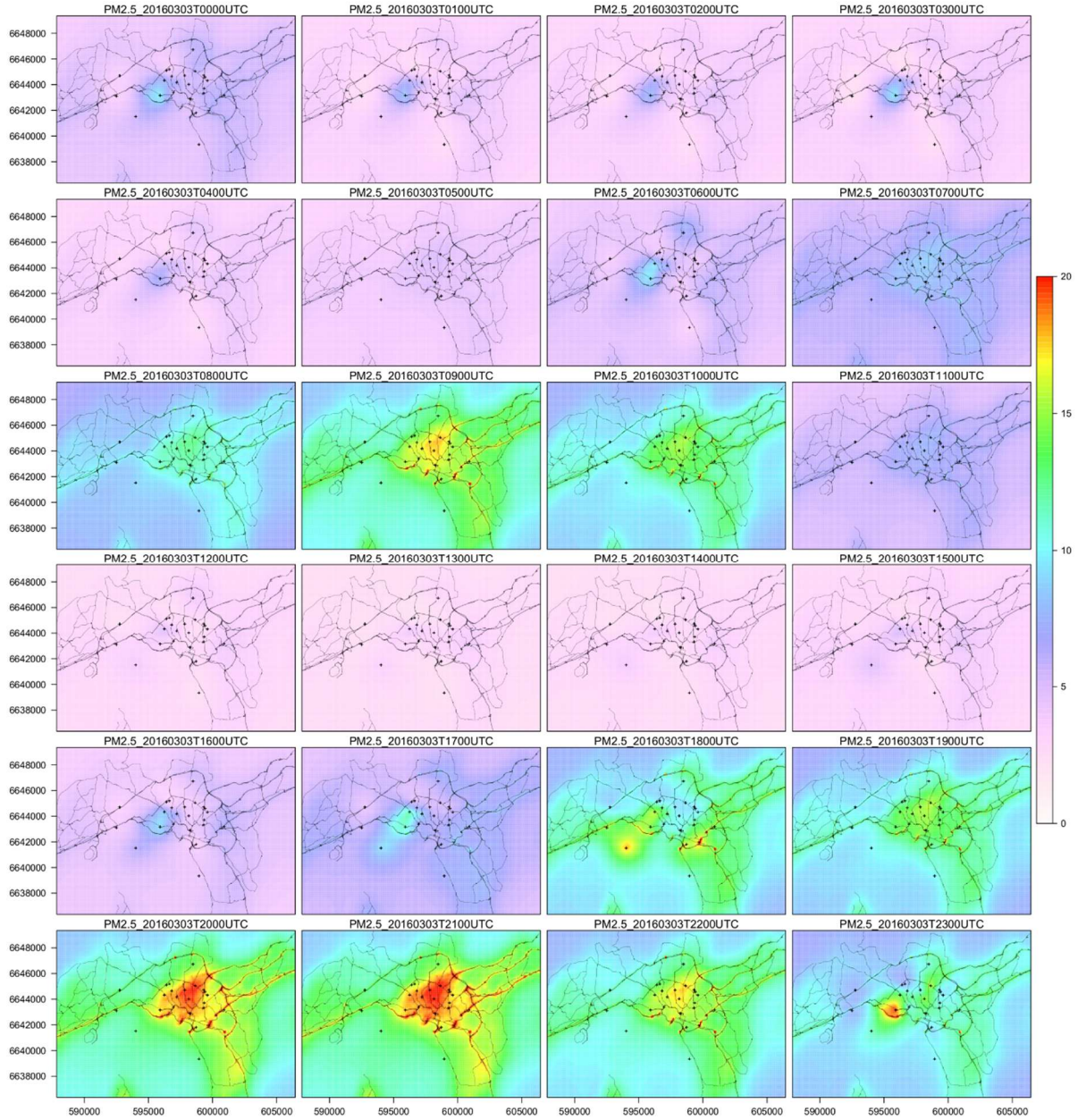

Figure 8 - Example of 24 hours of fused $PM_{2.5}$ maps for Oslo, Norway. The morning rush hour pollution is clearly visible. In addition, we can observe the higher concentration in the evening due to heating from wood-fire ovens.