GI_Forum 2018, Issue 1
Page: 65 - 81
Full Paper
Corresponding Author:
julian.bruns@uni-heidelberg.de
DOI: 10.1553/giscience2018_01_s65

Automated Quality Assessment of (Citizen) Weather Stations

Julian Bruns[1,3], Johannes Riesterer[2], Bowen Wang[2], Till Riedel[2] and Michael Beigl[2]

[1]Forschungszentrum für Informatik, Karlsruhe, Germany
[2]Karlsruher Institute for Technology, Germany
[3]Heidelberg University, Germany

Abstract

Today, we have access to a vast amount of weather, air quality, noise or radioactivity data collected by individuals around the globe. This volunteered geographic information often contains data of uncertain and of heterogeneous quality, in particular when compared to official in-situ measurements. This limits their application, as rigorous, work-intensive data-cleaning has to be performed, which reduces the amount of data and cannot be performed in real-time. In this paper, we propose a method to evaluate dynamically learning the quality of individual sensors by optimizing a weighted Gaussian process regression using an evolutionary algorithm. The evaluation was carried out in south-west Germany in August 2016 for temperature data from the Wunderground network and the Deutsche Wetter Dienst (DWD), in total 1,561 stations. Using a 10-fold cross-validation scheme based on the DWD ground truth, we show significant improvements for the predicted sensor readings: we obtained a 12.5% improvement on the mean absolute error.

Keywords:

crowdsourcing air temperature; data quality assessment; Evolutionary Learning; Gaussian process regression; volunteered geographic information.

1 Introduction

Today, we are living in an era where sensors are cheap, can be easily obtained, and can be put into use with little effort – they are becoming ubiquitous. In the field of geo-science in particular, this leads to many new data sources and opportunities. In addition to classical data sources such as government organizations, individuals are now providing data voluntarily, so called volunteered geographic information (VGI). These information sources range from smartphones and GPS-equipped mobile devices to privately owned weather stations on such

sites as Wunderground[1] or OpenSenseMap[2]. Projects such as OpenStreetMap (OSM)[3] empower individuals, encouraging them to provide data and participate in the creation of an open map. All these possibilities could lead to 'collective [geo] sensing' (Blaschke et al., 2011). The increased availability of data sources leads to a greatly increased resolution in both the spatial and the temporal dimensions.

Measurements can be made in-situ, at any given area of interest, and can be re-located if the need arises.

But these new data sources come with new challenges regarding their use. To 'produce results that can be trusted' (Stewart, 2011), the quality and location of measurements have to be known. Traditional data sources are often standardized measurements provided by government agencies. Most of the time, their quality and the exact location to which they refer are well-known; they are calibrated regularly and can be collected more or less 24/7. VGI does not have this advantage. VGI is provided by different organizations and acquired differently. The resulting diversity in credibility, data structure etc. can add additional uncertainty to the results, which prevents the use of VGI without appropriate pre-processing. A good example is the recent study of Meier et al. (2017), in which they discuss the use of crowdsourced weather data for the city of Berlin in 2015. During their quality assessment, they had to filter out over 50% of the available data and stated that 'rigorous data quality assessment is the key challenge' (Meier et al., 2017). And while this quality assessment can be done by experts and on historical data, the associated workload is high. This is not feasible for 'big data' or in real time.

The goal of this paper is to assess the quality of citizen science weather data from these new data sources to improve predictions and meteorological models. To solve the problem, we propose an automated quality assessment based on an evolutionary algorithm. Based on benchmark measurements, the algorithm learns the quality of each sensor. We then apply the calibrated data in a Gaussian Process Regression (GPR) to predict the measurement of interest. Our approach allows us to incorporate expert knowledge as a-priori information in the evolutionary algorithm, as well as iterative improvement of the quality assessment with each new measurement. It is derived from the field of ubiquitous computing as well as well-known approaches from spatial statistics. We evaluate the proposed approach with a temperature prediction for the area of south-west Germany using data from the Deutsche Wetterdienst and the Wunderground network. We use the equivalent to ordinary kriging as our basic GPR to show the improvement even without additional background information.

[1] https://www.wunderground.com/
[2] https://opensensemap.org/
[3] https://www.openstreetmap.org/

2 Related Work

Crowdsourced Sensing

In crowdsourced sensing, a group of private and/or professional users collect and contribute sensor information collaboratively to form a body of knowledge. The rise of smart phones in particular and 'the increasing ability to capture, [classify], and transmit a wide variety of data (image, audio, and location) [have] enabled a new sensing paradigm' [Reddy et al., 2007]. Civic agencies of several countries across the world are already harnessing the swarm intelligence of the public by accelerating and scaling the use of such open innovation methods to help address a wide range of urban and societal problems, ranging from wildlife observations to air quality sensing (obamawhitehouse, 2014)[4]. Participatory geosensing has been applied successfully as an alternative to traditional environmental monitoring to study physical phenomena, particularly in city contexts, such as urban noise levels (D'Hondt et al., 2013).

In an Internet of Things, anything can be measured using „a set of observations that reduce uncertainty where the result is expressed as a quantity' (Douglas, 2007). This statistically motivated view on measurement partially contradicts the classical view on measurement processes that use the DIN 1319 standard, which shaped much of the last century. Considering the poor spatial and temporal resolution of many measurements available today, anything (in addition to existing knowledge) that is better than guessing can potentially contribute to a measurement, even if by strict definitions it is not itself even considered a measurement. However, this (as is also addressed in our work) requires algorithms to cope with 'the problem of interrelationship between reliability of information sources, their number, and the reliability of fusion results' (Rogova et al., 2004).

Early research focused mostly on managing distributed sensors on sensor webs, like Intel's IrisNet (Gibbons et al., 2003) or Microsoft's SenseWeb (Grosky et al., 2007). Such sensors and networks have long since become a reality with the broad availability of devices like NetAtmo, and have attracted the attention of researchers who are particularly interested in higher-resolution data (Chapman et al., 2017; Meier et al., 2017). In our study, the data is used to interpolate fine-grained temperature distributions. However, little objective research exists on the quality of this VGI, which uses a large number of measurements.

Prediction of environmental factors

The main advantage of VGI is that it provides more data and information about the environment which can be used to formulate and evaluate hypotheses and to gain valuable insights into the environment. Data gained is used to train models to predict environmental factors such as temperature and pollution. The basis for all spatial prediction models is

[4] https://obamawhitehouse.archives.gov/blog/2014/12/02/designing-citizen-science-and-crowdsourcing-toolkit-federal-government

Tobler's First Law (Tobler, 1970), which states that 'everything is related to everything else, but near things are more related than distant things'. One of the most frequently used approaches to incorporate this law is kriging (Krige, 1951), which was developed to estimate ore deposits, but has since been used for predictions in numerous spatial applications and has been modified to be more powerful and general.

Hengl et al. (2012) used a kriging approach to predict temperatures. They included a temporal component to predict (with an accuracy of 2.4°C) the daily mean temperature in Croatia for a resolution of $1km^2$, combining Modis satellite images with 57,282 ground measurements of daily temperatures in 2008. In a follow-up study, Kilibarda et al. (2014) introduced an automated mapping framework for predictions of daily mean, minimum and maximum air temperatures using regression-kriging for a resolution of 1km, with a root-mean-square error between 2 and 4°C.

Gräler et al. (2016) developed an R-package called gstat, which uses copulas to enable spatio-temporal kriging. They show the application and benefit of their approach with a prediction of daily mean PM10 concentration, a chemical responsible for air pollution, in 2005 in Germany.

Another modification of the kriging approach can be found in Bhattacharjee et al. (2016). They propose a semantic kriging approach, where a high-resolution satellite snapshot is used to learn the systematic temperature differences between various locations based on the underlying land-use and the semantic information of those locations. The different land-use classes are learned in a semantic hierarchical network.

Hjort et al (2011) presented another approach to predict local temperatures in the city of Turku, Finland. They used generalized linear models combined with regression trees and data from 36 stationary weather stations over a period of six years.

An overview and the theoretical background, as well as applications of spatio-temporal statistics can be found in Cressie and Wikle (2015).

3 Method

Our approach combines a novel evolutionary learning algorithm to automatically assess and determine the quality of each sensor, and models this information as an uncertainty kernel. This is then combined with a typical ordinary kriging kernel as a GPR to predict temperature.

Gaussian Process Regression

We wanted to modify a regression model so that it could take into account the individual quality of an observation. For this purpose, classical kriging with noise is not suitable, since the noise factor can only model a constant additional quantity and not the non-constant quality of the data points. It turned out that the more general GPR meets our requirements, since it is determined by defining a covariance function. We were therefore able to model the quality of measurements by constructing the appropriate covariance function. In particular, we combined a Matern covariance function with a covariance function that maps a quality

parameter of an observation to an uncertainty of its correctness. We use a Matern covariance function for the following reasons. Since every physical process is of local nature, we may assume that the measurement of temperature on earth follows Tobler's first law of geography. Furthermore, we may assume that local fluctuations can still occur due to meteorological and topographical effects. The limit of a Matern covariance function yields an exponential covariance function, and thus realizes Tobler's first law of geography. However, appropriate choices of the parameters result in less smooth functions, which are more suitable to fit the local fluctuations but are still smooth enough to be robust against statistical noise.

For the following paragraphs about Gaussian Processes, regressions and modelling, compare Edward et al (2006), in particular Chapter 4 for definitions and properties of covariance functions.

Let y be the quantity we want to predict at a point p, and $D = \{(p_i, y_i, q_i) | \, i \in [1 \dots n]\}$ be a set of data points, where p_i denotes a specific point in geo-coordinates, y_i an observation of y at point p_i, and q_i the quality parameter of the measurement. We assume that the observations are measurements of a physical process; they can thus be assumed to follow Tobler's first law of geography. If we furthermore assume that the errors of the measurements follow a normal distribution, it is reasonable by definition to model the quantity y as a Gaussian process.

We define the function

$$\kappa_Q(q_i, q_j) := \begin{Bmatrix} \frac{\lambda}{q_i^2} & \text{if } i = j \\ 0 & \text{else} \end{Bmatrix}$$

for the quality parameters of two observations, where $\lambda > 0$ is a fixed scaling parameter. It is a covariance function, since it is positive everywhere and only non-zero on the diagonal. Furthermore let $\kappa_M(d(p_i, p_j))$ be a Matern covariance function with respect to the distance $d(p_i, p_j) = |\, p_i - p_j \,|$. Since the sum of two covariance functions is itself a covariance function,

$$\kappa\Big((p_i, y_i, q_i), (p_j, y_j, q_j)\Big) := \kappa_M(d(p_i, p_j)) + \kappa_Q(q_i, q_j)$$

also defines a covariance function.

For a subset $S \subset D$, we denote by $GPR_\kappa(p \mid S)$ the corresponding GPR for the quantity y at a point p under the observation S, which is implemented by our new (combined) Kernel function: $\kappa\Big((p_i, y_i, q_i), (p_j, y_j, q_j)\Big)$.

Evolutionary Algorithm

We use an evolutionary algorithm to train the quality parameter. The algorithm iteratively generates new variants of the set of data points with modified qualities. We evaluate the fitness of each variant by considering the prediction error obtained using the GPR.

As defined above, let $D = \{(p_i, y_i, q_i) | \, i \in [1 \dots n]\}$ be a set of data points, $S \subset D$ a subset, and $GPR_\kappa(p \,|S)$ the corresponding GPR for the quantity y at point p under the observation S. For another subset $S' \subset D$, we define the fitness function $fit(S' \,|\, S) := \sum_{s' \in S'} (s' - GPR_\kappa(s' \,|S))^2$, which measures the error between observations in S' and their prediction by the GPR under the observations S.

Let DWD denote the dataset of the Deutsche Wetterdienst and WG the dataset of Wunderground. They contain tuples of the form (p_i, y_i), where $p_i = (lat_i, long_i)$ are geo-coordinates and y_i is the measured temperature at this point. To evaluate our model using a 10-fold cross validation scheme, we apply a test / train split to DWD, which yields the decomposition into DWD_{valid} and DWD_{train}.

In the training process (Figure 1, 1–5), we build and use an evolutionary algorithm without crossover. For each generation, the D_{cur} is divided into D_{pred}, $D_{unchanged}$ and D_{mut} in the proportion 0.3 : 0.5 : 0.2. D_{pred} is chosen to contain 20% of the data points with the highest quality in D_{cur}. The remaining points in D_{cur} are assigned randomly.

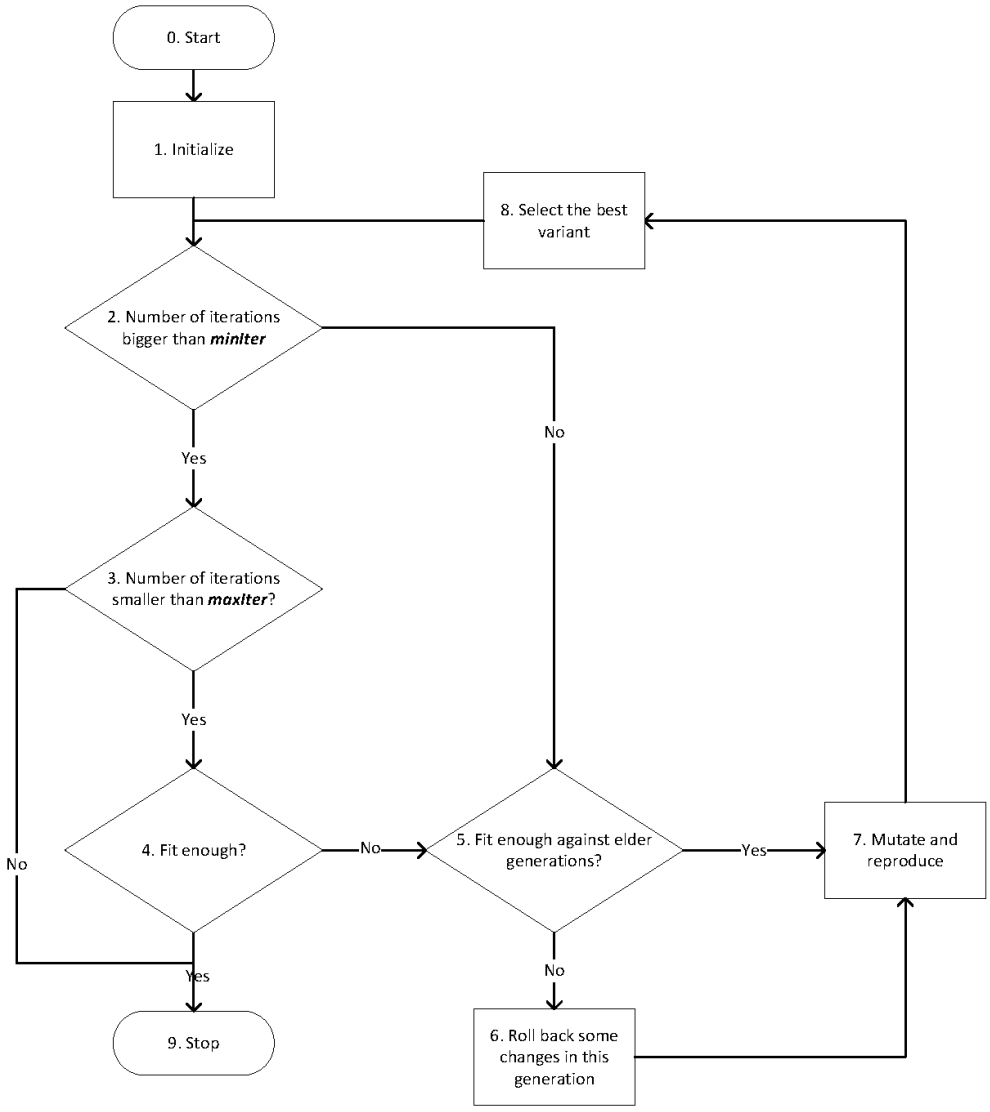

Figure 1: Graphic representation of the evolutionary algorithm. At each iteration, the termination criteria are checked, and if negative, the mutation and learning process is performed.

Each generation is evaluated using a fitness function based on the value of the MSE for predicting D_{pred} of the generation. The fitness value determines whether the observations in D_{pred} can be better reproduced by the parent generation $D_{cur} \setminus D_{pred}$, or by the new generation.

Our algorithm performs the following steps (the numbering is analogous to that in Figure 1):

1: The training process is initialized with the population $D_{cur} := DWD_{train} \cup WG$ as the union of DWD_{train} with WG, where the qualities are set to 1 for datapoints of DWD_{train} and to some fixed value $\mu \in (0,1]$ for datapoints of WG.

2–4: After a minimum number of iterations given by the hyper-parameter ***minIter***, the training process can be terminated if the improvement of the last few iterations is below a certain threshold and seems to have been converged. If the training process exceeds the maximum number of iterations ***maxIter***, the training process will be forcibly terminated. Here, ***minIter*** is set to 20 and ***maxIter*** to 100.

5–6: The fitness value of the current population D_{cur} is evaluated against that of its previous generations. If D_{cur} results in a worse fitness score, the current generation will roll back towards the last generation.

7: From D_{mut} two mutations are reproduced:

$D^1_{mut} := \{(p_i, y_i, (0.9 * q_i + 0.1)) | (p_i, y_i, q_i) \in D_{mut}\}$ and $D^2_{mut} := \{(p_i, y_i, (0.9 * q_i)) | (p_i, y_i, q_i) \in D_{mut}\}$ are created by randomly raising or lowering the quality of the elements in D_{mut}.

8: The variant from $\{D_{mut}, D^1_{mut}, D^2_{mut}\}$ that results in the highest fitness score will be selected. We replace D_{mut} with the selected variant $D_{selected}$ to create the next generation.

9: The result of the algorithm is a quality value for each sensor, which is then used in the GPR with our new combined covariance function.

4 Evaluation

Dataset

Our dataset is based on temperature measurements, taken each day at 12:00 MET within the latitude / longitude range of [47’ 5”; 49’ 5”; 7’5”; 9’5”] for all weather stations of the DWD and Wunderground station networks. The models were trained on the data from 01.08.2016 to 04.08.2016 and evaluated from 05.08.2016 to 08.08.2016. 42,966 observations were used from 1,561 weather stations (48 DWD stations; 1,513 Wunderground network stations).

We used a ten-fold cross-validation for the learning approach and predicted temperatures at randomly chosen DWD weather stations, which were removed from the training data set.

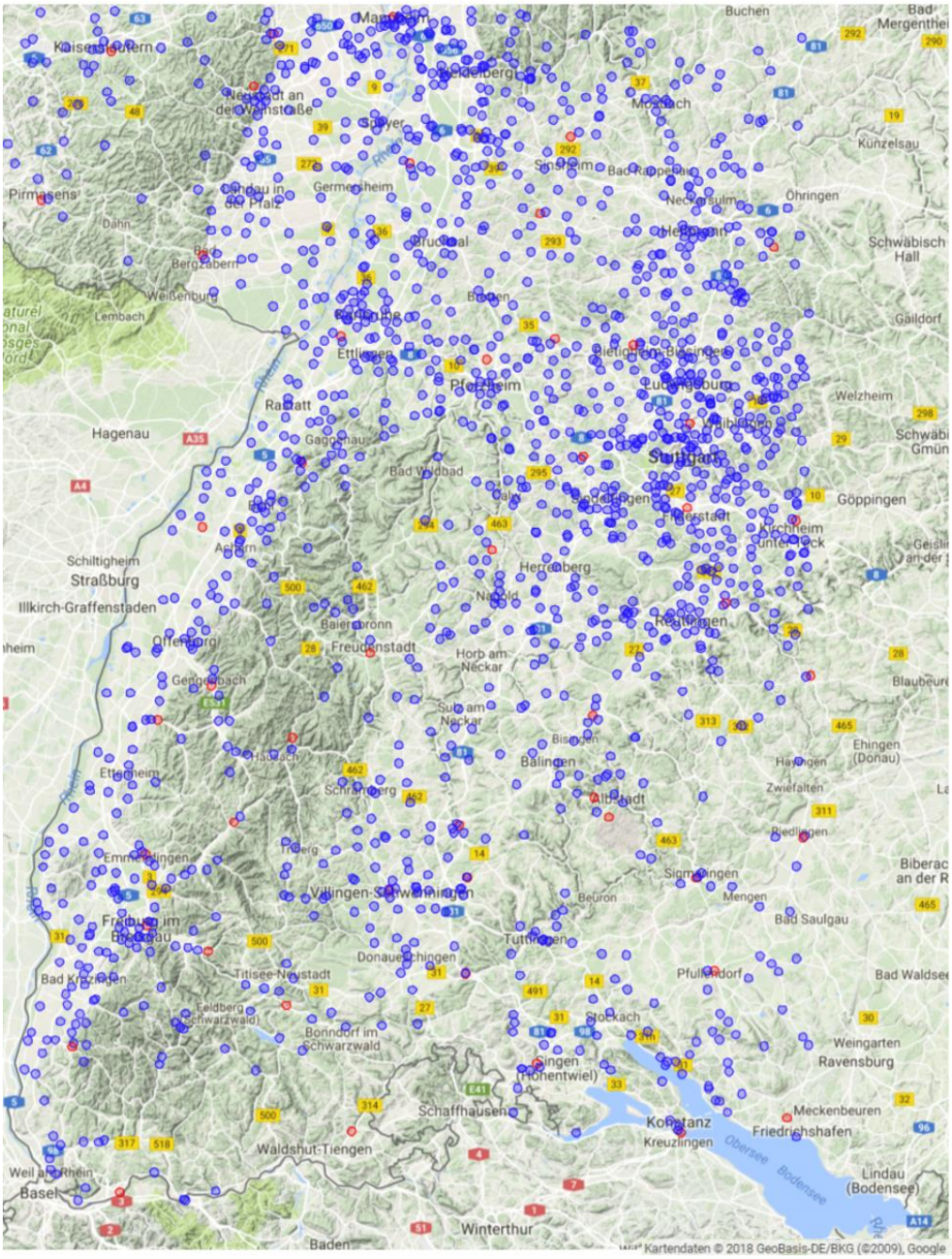

Figure 2: Spatial distribution of the stations in Google Maps. The DWD stations are shown in red.

Parametrization

To evaluate the impact of our proposed approach, we compared four different parametrizations for the predictions:

Table 1: Overview of Parametrizations

Model	Prediction Method
Baseline (Benchmark)	Ordinary kriging using only DWD stations.
Naïve Fusion	Ordinary kriging using all weather stations.
A-Priori Information	Adapted GPR with a-priori values for uncertainty for all stations.
Learned Model	Proposed new model.

The Baseline model represents the state-of-the art prediction without the benefit of VGI data. The Naïve Fusion represents the blind use of the additional data without regard to the data quality. To our knowledge, this blind use has not been done before. It shows the potential risks of VGI but also provides a second benchmark for using in quality assessment. A-Priori Information represents the knowledge of experts regarding the quality of measurements, e.g. experience of prolonged use or specifications of sensors. In this study, we determined the quality value for each station class by a simple grid-search. We assumed in this parametrization that the quality of each sensor class was the same; we did not differentiate between sensors. The DWD stations had a quality value of 0.98, the Wunderground sensors a quality value of 0.81. The Learned Model represents the proposed new model. Based on the A-Priori Information parameter, for each sensor a unique uncertainty value is learned iteratively via the combined model presented.

5 Results and Discussion

The results for the temperature prediction can be seen in Table 2, and graphically in Figure 3.

Table 2: Summary of prediction results in degrees Celsius. In brackets, the percentage improvement compared to the Benchmark.

MODEL	MEAN ABSOLUTE ERROR	STANDARD DEVIATION
Baseline (Benchmark)	1.12°C	0.83°C
Naïve Fusion	1.26°C (-12.5%)	1.03°C
A-Priori Information	1.21°C (-8.0%)	0.99°C
Learned Model	0.98°C (12.5%)	0.76°C

We used the Mean Absolute Error (MAE) as error metric, as this shows the quality of the prediction in a single value and is well established. The standard deviation (SD) is used to show the volatility of the quality of the results.

We see that the ranking of the MAE and SD is the same for every model. Not surprisingly, the Naïve Fusion model performs the worst. Without any quality assessment, the influence of false measurements and of the high variance in placement decreases the quality of the prediction compared to the traditional approach, the Baseline model. The increased availability of information inherent in VGI is overshadowed by the poor and heterogeneous quality of the measurements. The inclusion of a very simple quality assessment in the A-Priori information model already shows an increase in prediction quality, even though there is no differentiation between the stations within each class. However, it still performs less well than the Baseline model. The Learned model performs the best overall, performing more than 20% better than the Naïve Fusion model. This is the result of the learning process and the covariance function used. Sensors which perform badly overall give less credence to the prediction result over time. Low quality sensors are automatically filtered out based on their data, e.g. when they are inside buildings, are defective, or produce constant values.

But while the accuracy of the prediction is important, the increased spatial resolution of the prediction is one of the main advantages VGI presents. Figure 3 shows the resulting predictions of the different models; the A-Priori information model was omitted as it is almost identical to the Naïve Fusion model.

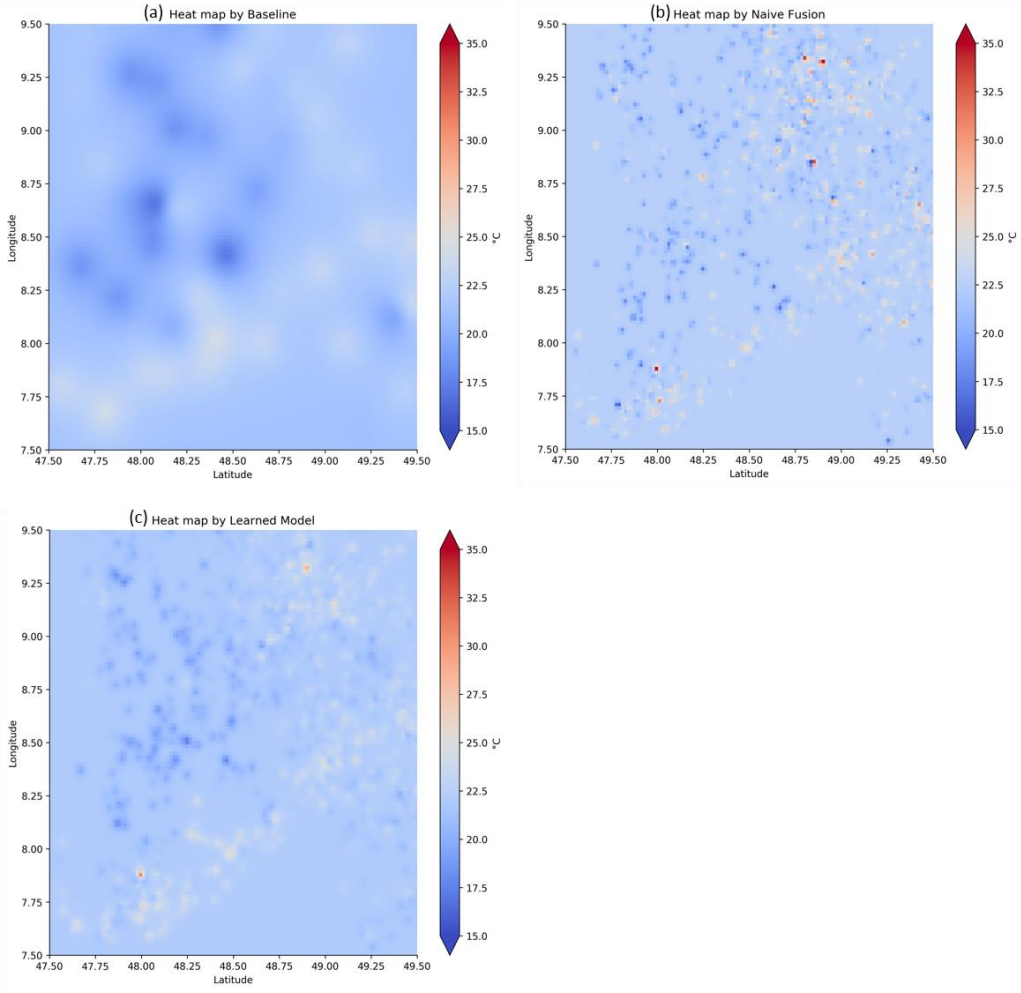

Figure 3: Resolution of predictions

We can see the differences in resolution and prediction for each model. The Naïve Fusion presents a highly detailed map of temperature as the number of available sensors is quite high. There is a strong but fluid transition between the different areas. When compared to Figure 1, the number of weather stations is directly connected to the sharpness of the map, which can be seen in the south and the west. Outliers are seen on this map. The temperature range is from 14.6°C to 35.23°C. The Baseline model shows a low spatial resolution and a low overall temperature, as well as a small temperature range, from 17.27°C to 24.12°C. The small number of DWD weather stations can be seen by the rough transition between the different prediction areas. In the south-west of the map, there is a higher number of DWD stations and the map is much smoother. The Learned model strikes a balance between the other models. The temperature is between 17.77°C and 27.52°C and the transitions are smooth overall. A clear distinction between warmer and colder areas can be seen, allowing a

detailed temperature map. The overall trend of the temperature distribution over our study area stays the same for all predictions. The graphic analysis shows clearly the advantages of VGI. Whereas in the MAE prediction results, the Baseline outperforms the Naïve Fusion quite strongly, in the practical use to create informative maps, in our opinion the Naïve Fusion outperforms the Baseline.

Error Distribution

To further evaluate and understand the different models, we examine their error distribution. Figure 4 shows the histogram of the prediction errors for each model. One can see that the Naïve Fusion and A-Priori models are almost identical. This is not surprising, as the two are quite similar in their parametrization and underlying modelling logic. Both resemble a broad normal distribution with, overall, a large range, of around 10°C. There is a slight negative bias in the predictions overall, indicating that these models overestimate the temperature. This is most likely the result of the difference in placement of the reference stations and the citizen weather stations. As mentioned earlier, the standardized placement of DWD stations leads to the exclusion of several climatic conditions, e.g. urban heat islands. These are captured by the citizen weather stations and lead to an overestimation of the temperature, as these effects are not filtered out. The baseline model on the other hand underestimates the temperature, as the majority of its errors lie between 0 and 2, which leads to a skewed distribution. We assume this is the result of a small number of outlier stations, which decrease the mean temperature of the overall distribution, in particular as stations in the warmest cities of Germany are in this area (Freiburg and Karlsruhe), as well as of the presence of different climatic regions, such as the Black Forest and the Upper Rhine valley. But we also see the effect of the standardized placement, as the standard deviation of the errors is lower than for the other two models. This presents a more coherent prediction, which can also be seen if the Baseline and Naïve Fusion models are compared (see Figure 3). Finally, the Learned model shows a similar distribution to the Baseline model, but the spread of the distribution is even smaller and the centre of the errors lies between 0 and 1. Similar to the graphic prediction in Figure 3, the shape is a combination of the Baseline and Naïve Fusion models. The histogram supports the hypothesis that our Learned model manages to leverage the advantages of VGI successfully.

Of further interest is that the highest errors for all models are negative. Interestingly, the outlier is most noticeable for the Baseline model. We would have expected that such a high error would only be present in VGI measurements. This indicates that there is at least one station among the reference stations used for the evaluation which has a relatively low temperature compared to all other stations nearby. Therefore, even when using official data sets we would urge caution and the need for a rigorous understanding of the data before analysing the data sets for the generation of insights.

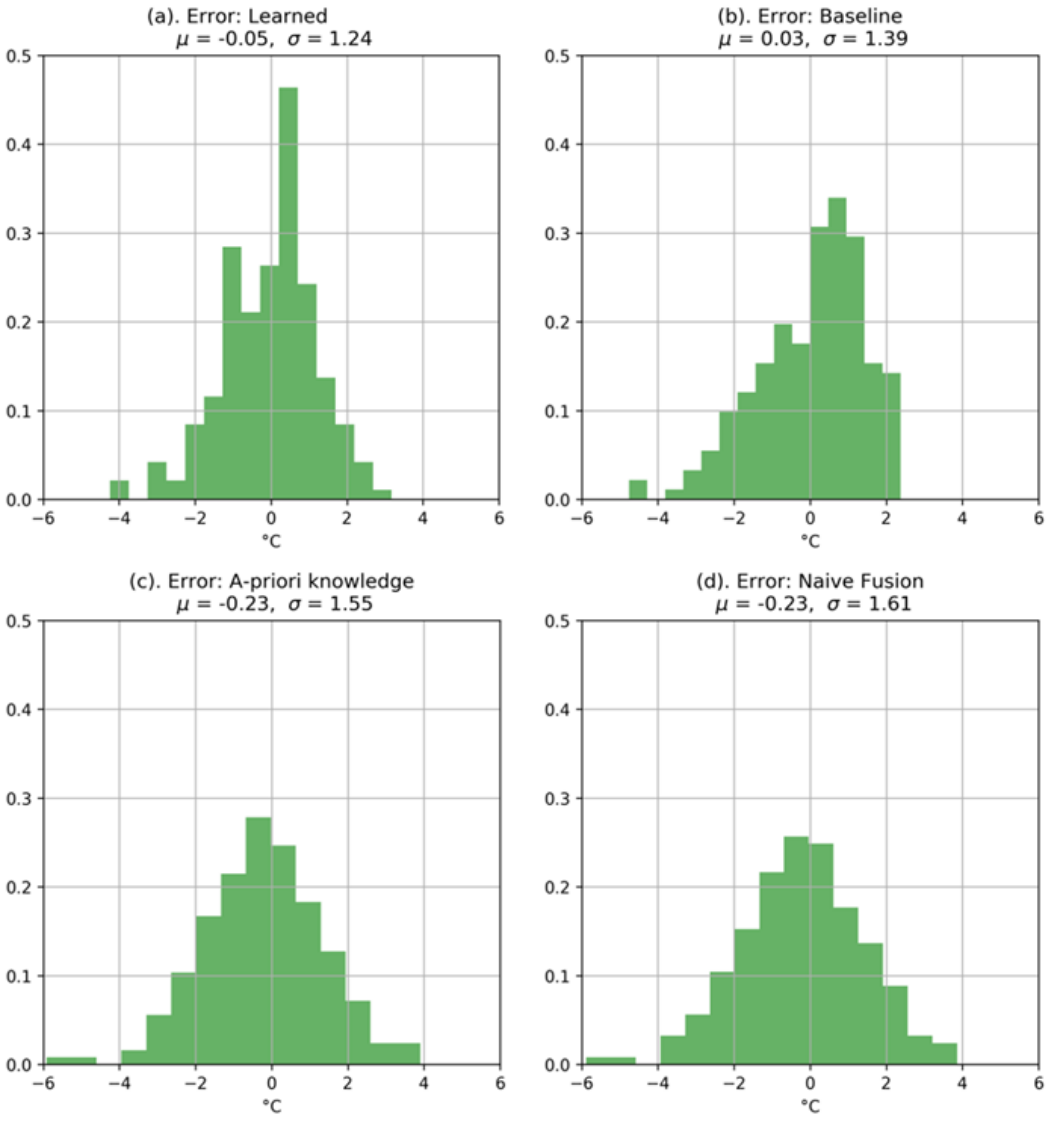

Figure 4: Histograms of the Prediction Error Out Of Sample.

6 Conclusions and Future Work

In this study, we proposed an automated quality assessment of VGI sensors, with weather stations as the concrete use-case. The proposed approach combined a new evolutionary learning algorithm with GPR to learn and model the quality of sensors in order to produce reliable and accurate predictions without the need to clean the data beforehand. We evaluated the approach on weather data, as this is the most accessible type of data and is therefore suitable for use by researchers and practitioners alike. Our results showed an

improvement in the prediction quality of 12.5% over the established benchmark of DWD weather stations, simply by including in addition the quality of the measurement. Furthermore, we showed that the naïve use of citizen weather stations improves the spatial resolution of the temperature prediction immensely. The proposed approach preserved this improvement of the spatial resolution while providing the full benefit of VGI, as discussed in e.g. Blaschke et al. (2011) and Meier et al. (2017). For the smart cities of the future and urban climate, this approach allows for more in-depth analyses, as to date the existing measurement networks are rather sparse (e.g. for temperature and air pollutants). New (crowdsourced) measurement approaches, as in the SMARTAQNET project[5], or involving cars equipped with sensors, are currently being developed. Our approach allows full advantage to be taken of these innovations.

This research has several restrictions, however, which should be taken into account. First, we evaluated our approach only on temperature data in South-West Germany for a short period at the beginning of August 2016. While the data set for training as well as the set for evaluation are quite big, they are based on just a small fraction of the total data available. In particular, seasonal and daily cycles have not been examined. Second, we only fully implemented and compared one prediction method (ordinary kriging) and one kernel approach to incorporate the uncertainty. While the reasoning for this is discussed in our section on method, a more in-depth comparison could lead to different results. Third, we did not compare our results to those of a manually cleaned data set (as in Meier et al., (2017)). We assume this could lead to an improvement of both the Naïve Fusion and the A-Priori methods, but this is beyond the scope of the current study.

In the future, an evaluation using different data sets would be of great interest, especially for air pollutants and in different climatic regions. Another interesting question would be the inclusion of different kernels as well as of background information. The work of Bhattacharjee et al. (2016) presents an example using semantic kriging which includes land use information and could be used as an alternative kernel to ordinary kriging. Another approach is found in regression-kriging, discussed in Hengl et al. (2007). Arnfield (2003) presents an overview of causal factors for the influence on temperature. The use of spatio-temporal prediction instead of only spatial prediction could lead to further insights. Kilibarda et al. (2014) show the application of such spatio-temporal kriging and the benefits it provides. The challenges here lie in the selection and modelling of the suitable kernel as well as in the computational complexity. Finally, the results of our error analysis for the Baseline model show a strong skewness. Further investigations into this error could lead to interesting insights.

[5] http://www.smartaq.net/

Acknowledgements

This work is part of the research project BigGIS (reference number: 01IS14012) funded by the Federal Ministry of Education and Research (BMBF) within the frame of the program 'Management and Analysis of Big Data' in 'ICT 2020 – Research for Innovations'. We thank the Smart Data Innovation Lab for providing us with a high-performance computing infrastructure on which all computations presented in this paper were performed. This work has been partially funded by the German Federal Ministry for Traffic and Digital Infrastructure (BMVI) as part of project SmartAQnet (grant no. 19F2003B), and by the German Federal Ministry of Education and Research (BMBF) as part of project SDI-X (grant no. 01IS15035).

References

Arnfield, A. J. (2003). Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *International Journal of Climatology*, 23(1), 1-26.

Bhattacharjee, S., Das, M., Ghosh, S. K., & Shekhar, S. (2016, October). Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (p. 38). ACM.

Blaschke, T., Hay, G. J., Weng, Q., & Resch, B. (2011). Collective sensing: Integrating geospatial technologies to understand urban systems—An overview. *Remote Sensing*, 3(8), 1743-1776.

Chapman, L., Bell, C., & Bell, S. (2017). Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. *International Journal of Climatology*, *37*(9), 3597-3605.

Cressie, N., & Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

D'Hondt, Ellie, Matthias Stevens, and An Jacobs. (2013), Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. *Pervasive and Mobile Computing* 9.5: 681-694

Douglas, Hubbard. How to measure anything: finding the value of intangibles in business. (2007): 46.

Edward, C., Rasmussen, Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*. USA: The MIT Press.

Gibbons, Phillip B., et al. (2003). Irisnet: An architecture for a worldwide sensor web. *IEEE pervasive computing* 2.4: 22-33

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *R Journal*, 8(1), 204-218.

Grosky, W. I., Kansal, A., Nath, S., Liu, J., & Zhao, F. (2007). Senseweb: An infrastructure for shared sensing. *IEEE multimedia*, 14(4).

Hengl, T., Heuvelink, G. B., & Rossiter, D. G. (2007). About regression-kriging: from equations to case studies. *Computers & geosciences*, 33(10), 1301-1315.

Hengl, T., Heuvelink, G. B., Tadić, M. P., & Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. *Theoretical and applied climatology*, 107(1-2), 265-277.

Hjort, J., Suomi, J., & Käyhkö, J. (2011). Spatial prediction of urban–rural temperatures using statistical methods. *Theoretical and applied climatology*, 106(1-2), 139-152.

Kilibarda, M., Hengl, T., Heuvelink, G., Gräler, B., Pebesma, E., Perčec Tadić, M., & Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, 119(5), 2294-2313.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52 (6), 119-139.

Meier, F., Fenner, D., Grassmann, T., Otto, M., & Scherer, D. (2017). Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, *19*, 170-191.

Reddy, S., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2007, November). A framework for data quality and feedback in participatory sensing. In *Proceedings of the 5th international conference on Embedded networked sensor systems* (pp. 417-418). ACM.

Rogova, Galina L., and Vincent Nimier. (2004). Reliability in information fusion: literature survey. *Proceedings of the seventh international conference on information fusion*. Vol. 2.

Stewart, I. D. (2011), A systematic review and scientific critique of methodology in modern urban heat island literature. *International Journal of Climatology*., 31: 200–217. doi:10.1002/joc.2141

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1), 234-240.