Gaussian Process Models for Low Cost Air Quality Monitoring

Michael T. Smith[1*], Joel Ssematimba[2], Mauricio A. Álvarez[1] & Engineer Bainomugisha[2]
University of Sheffield[1] & University of Makerere[2]

Air pollution contributes to over three million deaths [1] each year. Kampala has one of the highest concentrations of fine particulate matter (PM 2.5) of any African city [2]. Unfortunately, with the exception of the US Embassy, there is no programme for monitoring air pollution in the city due to the high cost of the equipment required. Hence we know little about its distribution or extent. Lower cost devices do exist, but these do not, on their own, provide the accuracy required for decision makers. We propose that using a coregionalised Gaussian process to combine the low cost sensors with the embassy's high quality results provides sufficiently accurate estimates of pollution across the city.

The air pollution data used in this study has been collected using a network of sensors developed at Makerere University, built around the Alphasense optical particle counter [3]. The US Embassy also monitors the air pollution using an EPA approved system which we assume provides an accurate baseline. The aim of this project is to extrapolate outwards from the embassy by using the correlation structure between the low-cost sensors and the embassy's sensor. The network is currently being commissioned and only contains a few weeks of data.

Coregionalisation with two sensors Both models in this abstract use Gaussian processes (GPs) to perform probabilistic regression. Besides providing uncertainty quantification, GPs allow us to define priors over the covariance between measurements. For both models we assume that the covariance can be described with an exponentiated quadratic kernel. In this first model we also describe covariance between two sensors using a coregionalisation kernel. The inputs to the GP are simply the date and time. Its outputs are the log of 3-hour averaged PM2.5 measurements. For comparison we repeat the GP model fitting but with the off-diagonals of the coregionalisation matrix set to zero, to disable coregionalisation.
To demonstrate the effectiveness of using coregionalisation we consider just the US embassy and one static sensor, ignoring the spatial aspect. We remove 300 hours from the US embassy's training set and try predicting these values using both models. We found the RMSE decreased from 60 to 36 μg/m^3 using the coregionalisation model. The results of which are shown in Figure 1. We have selected a particularly erratic time period to illustrate the effectiveness, but we found an improvement in the RMSE for all tested time periods. Intriguingly these sensors are sufficiently correlated even though they lie nearly 5 km apart, as illustrated in Figure 2.

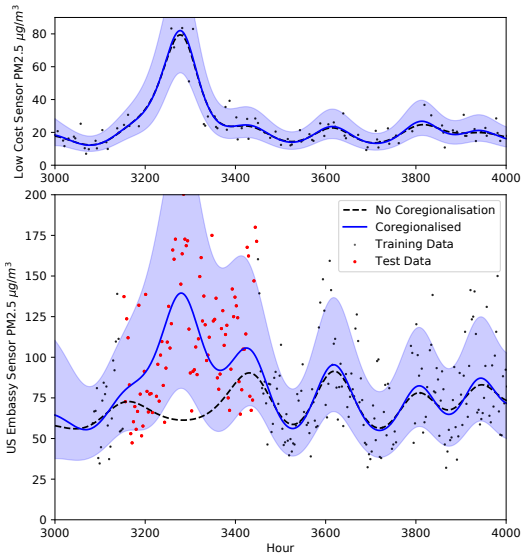

Figure 1: PM2.5 measurements from low-cost sensor at Makerere University (upper) and US Embassy (lower). The lines are the predictions from the two GPs (for 12 noon, to remove daily oscillations, inputs: date, time and sensor). The uptick in pollution during the test-period is predicted more accurately by the coregionalised model (in blue), than by the simple model (black-dashed). Note that although time-of-day is an input into the model, day-of-week is not, leading to a 168-hour oscillation. Confidence intervals are 1 std, and are non-symmetric due to the log transform.

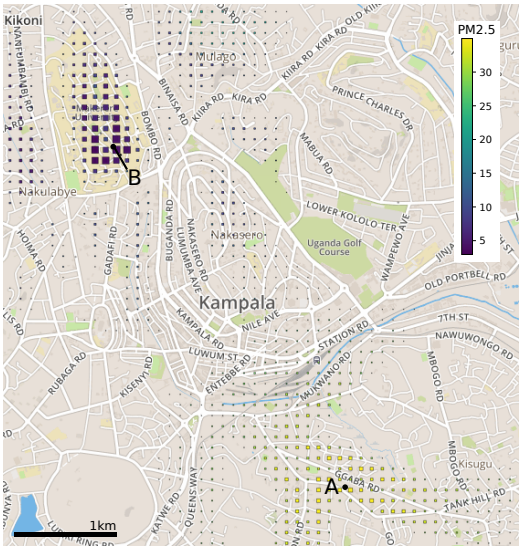

Figure 2: Predictions (from non-coregionalised) GP model combining all sensors. The confidence in the predictions is indicated by the size of the squares (larger=more confident). The static sensors include those at the embassy (A) and the university (B).

Spatial-temporal Model A second model is used which includes both mobile and static sensors, with inputs; latitude and longitude, date, time and distance from main roads. We aim to make coregionalised predictions across the whole city for the Embassy's output. Currently however, the lack of geographically distributed low-cost sensors means we are unable to determine whether differences in measurements are due to differences in sensor-type or differences in pollution. In lieu of the full network, therefore, we use a simple, non-coregionalised model, and assumed all the sensors provide unbiased PM2.5 results. Figure 2 demonstrates the system[+]. Though incomplete, useful features are already visible; the sensors near the university (B) are placed a considerable distance from the roads, while the sensor at the embassy (A) is nearly next to a road. This leads to more confident predictions nearer to roads in the region around the embassy.

Conclusions We have developed a system for collecting and analysing real-time data from a network of sensors. The use of coregionalisation allows us to benefit from both the spatial distribution of low-cost sensors and a high precision single sensor. Once calibrated we will be able to quantify the coregionalisation warranted between each sensor. In future work we will introduce active-learning to direct the mobile sensors to locations that will offer greatest information gain, for example, to allow the coregionalisation between sensors to be updated. We will also perform leave-one-sensor-out cross-validation to detect sensor failure. The system aims to provide a low-cost method for monitoring air quality across whole cities in resource-constrained regions.

Acknowledgemnts Project funded by USAID: Development Impact Lab and the UK EPSRC.

References

[1] J. Lelieveld *et al.*, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, p. 367, 2015.

[2] N. V. Mead, "Pant by numbers: the cities with the most dangerous air listed," *The Guardian*, Feb 2017.

[3] Alphasense, *Optical Particle Monitor*. OPC-N2, 2017.

[*]Corresponding author: m.t.smith@sheffield.ac.uk

[+]A real-time interactive version of this is available at https://lionfish0.github.io/air-quality-kampala