Bayesian Optimisation for Active Monitoring of Air Pollution

Sigrid Passano Hellan, Christopher G. Lucas and Nigel H. Goddard

School of Informatics, University of Edinburgh, UK
{s.p.hellan, c.lucas, nigel.goddard}@ed.ac.uk

Abstract

Air pollution is one of the leading causes of mortality globally, resulting in millions of deaths each year. Efficient monitoring is important to measure exposure and enforce legal limits. New low-cost sensors can be deployed in greater numbers and in more varied locations, motivating the problem of efficient automated placement. Previous work suggests Bayesian optimisation is an appropriate method, but only considered a satellite data set, with data aggregated over all altitudes. It is ground-level pollution, that humans breathe, which matters most. We improve on those results using hierarchical models and evaluate our models on urban pollution data in London to show that Bayesian optimisation can be successfully applied to the problem.

Introduction

Ambient air pollution is one of the leading causes of death globally, with particulate matter with diameter less than $2.5\,\mu\text{m}$, $PM_{2.5}$, causing 3-4 million deaths each year (Cohen et al. 2017; Lelieveld et al. 2015). One commonly monitored pollutant is nitrogen dioxide, NO_2, as it is used to indicate the presence of traffic-generated air pollution (Katsouyanni 2003). NO_2 is also part of an ozone-generating process, and short-term increases in ozone concentration are followed by short-term increases in mortality (Katsouyanni 2003). Due to these adverse effects the WHO have produced guidelines limiting pollution concentrations, and many countries have adopted air pollution regulations (World Health Organization 2006, p. 174-175).

Traditional air pollution monitoring uses large and expensive sensors that are typically managed by national or municipal authorities deciding where to locate sensors based on domain knowledge and constraints posed by the bulky nature of the sensors (Carminati, Ferrari, and Sampietro 2017). More recently, low-cost air pollution sensors have become available, e.g. Liu et al. (2020) and Kelly et al. (2017), which open up air pollution monitoring to more locations, and allow groups with limited budgets and domain expertise to create sensor networks, including local governments in developing nations as well as community groups around the world. By automating the decision-making process for placing air pollution sensors, we can help these groups make the most of their limited resources.

This goal, to simplify the process for setting up new monitoring networks in a low-cost way, motivates our decision to focus on computationally relatively inexpensive models and simple feature sets. Although models are generally improved by using data on road traffic and other pollution sources, these data can be difficult or expensive to obtain for developing nations or citizen scientists. We demonstrate that even a minimal feature set containing only sensor locations and readings can be useful. Additionally, all the computations are done on CPUs, so access to GPUs is not needed.

Simplicity and explainability also guide the model design. Experimental exploration showed the need for informative priors, as simpler approaches without hyperparameter priors or with only basic priors did not perform well. Instead, we found a pragmatic solution using related data to construct a prior through a hierarchical model. In that way data from other cities can be used for the prior, and the relevance of each other city inferred. Gaussian processes are used to model the data, which have interpretable hyperparameters, aiding explainability. And the number of hyperparameters is limited by keeping the covariance functions simple.

Bayesian optimisation (BO) has been used for pollution monitoring previously, but has tended either to estimate hyperparameters or their distributions in ways that do not generalise to new sites in a sample-efficient way (Ainslie et al. 2009) or to focus on robot trajectories (Morere, Marchant, and Ramos 2017; Singh et al. 2010; Marchant and Ramos 2012), a distinct problem from the one we consider here. Work has also been done on monitoring gas leaks (Reggente and Lilienthal 2009; Asenov et al. 2019), but again by using a single moving agent. An exception is Hellan, Lucas, and Goddard (2020), but it is limited to satellite data as a proxy for ground measurements. We extend on it with a more principled methodology and better results, as well as an evaluation on a more relevant data set taken from the London Air Quality Network (LAQN) (Imperial College London 1993). Our motivations are similar to Smith et al. (2019), who use Gaussian processes for calibrating low-cost pollution sensors, but do not address the problem of sensor placement.

The main contribution of this paper is showing Bayesian optimisation to be useful for automating planning of pollution sensor networks. Specifically, we consider the prob-

lem of iteratively placing stationary sensors and locating the maximum average pollution in the given area. Knowledge of the maximum lets us know whether regulations are being adhered to. We also discuss the general usefulness of the methodology adopted.

Background

Bayesian Optimisation

Bayesian optimisation (Shahriari et al. 2015) is an optimisation method based on maintaining a probabilistic model $m(x)$ of the underlying problem $f(x)$. At each iteration the model is fitted to the collected data, the next sampling location chosen according to the acquisition function $a(m(x))$ and a sample collected. The acquisition function balances the exploration/exploitation trade-off. At each iteration, the problem $\max_x a(m(x))$ is solved instead of $\max_x f(x)$. While evaluating $f(x)$ requires taking a measurement, $a(m(x))$ can be calculated from the model. One acquisition function is expected improvement (Jones, Schonlau, and Welch 1998), $a(x) = \mathbb{E}[\max(0, f(x) - f')|\, \mathcal{D}]$ where $\mathcal{D}$ is the data observed and f' is the highest sample from previous iterations. The model is usually a Gaussian process (GP) (Rasmussen and Williams 2006).

Hierarchical Bayesian Modelling

Hierarchical Bayesian modelling is the principle of having layers of random variables built on top of each other (Shiffrin et al. 2008). The advantage of using a hierarchical model for BO is that it can be used to transfer information about the distribution of GP hyperparameters across different contexts – such as cities – without assuming that these contexts are identical. This is advantageous in applications like pollution monitoring where there are plentiful observations for some contexts but not others, e.g. cities with and without extensive pollution monitoring programs, and sample efficiency is paramount as each sample is expensive to collect. Hierarchical Bayesian modelling has been used for air pollution monitoring and modelling, e.g. Dawkins et al. (2020), Cocchi, Greco, and Trivisano (2007) and Sahu (2012), but not, to our knowledge, in combination with Bayesian optimisation.

Approximate Inference

Although the standard approach for BO is to use a single point estimate for hyperparameters, a more Bayesian approach has several advantages, including better sample efficiency and more accurate estimates of posterior uncertainty (De Ath, Everson, and Fieldsend 2021; Snoek, Larochelle, and Adams 2012). Instead of a single sample, a set of samples (Monte Carlo approximation) or a parameterised distribution (variational inference) can be used to represent the hyperparameters. However, it makes the inference more computationally challenging. The posterior can be calculated in closed form for the standard GP model with a point estimate of the hyperparameters (Rasmussen and Williams 2006), but this is not possible for the hierarchical models. Following previous work, we use Monte Carlo approximation rather than variational inference, as De Ath, Everson, and Fieldsend (2021) found the latter to be less accurate. The Bayesian treatment of hyperparameters has been considered more widely for non-BO Gaussian processes, e.g. in Lalchand and Rasmussen (2020), Murray and Adams (2010) and Williams and Rasmussen (1996).

In general when using approximate inference for model hyperparameters, one tries to solve Eq. (1), where X^* and Y^* are the test features and values, θ_j the model hyperparameters and $\mathcal{D}_j$ the observed data. In contrast, we are solving Eq. (2) where $\mathcal{D}$ is observed data from related problems, i.e. our prior information. By factoring out the dependence on $\mathcal{D}$, we can split the inference problem into two separate ones, which we can solve separately. The first one takes into account $\mathcal{D}$ and only needs to be solved once using Markov chain Monte Carlo. The second one takes into account $\mathcal{D}_j$ and is solved each time a new observation is added through the BO iterations, using importance weighting.

$$p(Y^*|X^*, \mathcal{D}_j) = \int p(Y^*|X^*, \theta_j) p(\theta_j|\mathcal{D}_j) d\theta_j \quad (1)$$

$$p(Y^*|X^*, \mathcal{D}, \mathcal{D}_j) = \int p(Y^*|X^*, \theta_j) p(\theta_j|\mathcal{D}_j, \mathcal{D}) d\theta_j \quad (2)$$

We found a simple Metropolis-Hastings based approach (Chib and Greenberg 1995) worked well, but gradient-aware methods, e.g. Hoffman and Gelman (2014), may be more appropriate for more complex kernels and higher-dimensional parameter spaces.

Methodology

While the heart of our method is a standard Bayesian optimisation loop, its effectiveness relies on problem-specific design decisions, data pre-processing, and pre-training in order to obtain an appropriate joint prior over GP parameters.

Pollutant concentrations tend to follow a log-normal distribution (Cats and Holtslag 1980), so we log-transformed our concentration data to match the distributional assumptions behind Gaussian processes. We also standardised our data using the mean and standard deviation from the tuning set. For evaluation, the observed data is considered to be the ground truth, so the variance of the noise in the GP likelihood is clamped to a small value (1e-06). The acquisition function used is expected improvement (Jones, Schonlau, and Welch 1998) and importance weighting is used to adapt the hyperparameter samples to the problem at hand. Each problem is initiated with some randomly selected samples; 10 for the satellite data and 5 for the London data. The computing resources used are described in the appendix.

GP Models

The GP models are constructed from two base covariance functions, the RBF kernel k_{R} in Eq. (3) and a directed version k_{W} from Hellan, Lucas, and Goddard (2020), given in Eq. (4). The latter only considers the spatial distance orthogonal to a reference direction γ, conceptualised as the wind direction. τ is the difference between feature vectors. In Eq. (4), $A = \begin{bmatrix} \sin(\gamma)^2 & -\sin(\gamma)\cos(\gamma) \\ -\sin(\gamma)\cos(\gamma) & \cos(\gamma)^2 \end{bmatrix}$.

$$k_{\mathrm{R},i}(\tau) = \sigma^2_{\mathrm{r},i} \exp\left(\frac{-\tau^T\tau}{l^2_{\mathrm{r},i}}\right) \quad (3)$$

$$k_{\mathrm{W},i}(\tau) = \sigma^2_{\mathrm{w},i} \exp\left(\frac{-\tau^T A\tau}{l^2_{\mathrm{w},i}}\right) \quad (4)$$

From k_{R} and k_{W} three composite covariance functions are constructed, which are the ones evaluated in this paper. They are given in Eqs. (5) to (7) and consist of a sum of two terms, which are meant to capture the slowly-varying and faster-varying parts of the observed signal. Including the noise hyperparameter, the RBF-RBF model has 5 hyperparameters, the RBF-Product model 7 and the Sum model 6.

$$k_{\mathrm{RBF-RBF}}(\tau) = k_{\mathrm{R},1}(\tau) + k_{\mathrm{R},2}(\tau) \quad (5)$$

$$k_{\mathrm{RBF-Product}}(\tau) = k_{\mathrm{R},1}(\tau) + k_{\mathrm{R},2}(\tau)k_{\mathrm{W},3}(\tau) \quad (6)$$

$$k_{\mathrm{Sum}}(\tau) = k_{\mathrm{R},1}(\tau) + k_{\mathrm{W},2}(\tau) \quad (7)$$

Hierarchical Structure

Problem-appropriate hyperparameter settings are essential for Bayesian optimisation to be efficient. Too large a lengthscale $l_{\mathrm{r},i}$ and the process is assumed to vary too slowly and relevant locations are not explored. Too small a lengthscale and resources are wasted checking locations that are unlikely to be high and provide little new information. These pitfalls can be avoided using a prior that is informed by relevant data from other contexts. To that end, our model is structured hierarchically as described in Fig. 1, and the prior inferred from a related tuning set $(\mathcal{D}_1, \cdots, \mathcal{D}_N)$. The lowest level of the hierarchical model is the data snapshots, which are the only variables that are observed directly. GP models are fitted to the data snapshots, and the hyperparameters θ of the GP models constitute the middle layer of the model. The top level is the parameters defining the distribution of the GP hyperparameters. This can be expressed mathematically as $p(\eta, \theta, \mathcal{D}) = p(\eta)p(\theta|\eta)p(\mathcal{D}|\theta) = p(\eta)\prod_i p(\theta_i|\eta)p(\mathcal{D}_i|\theta_i)$.

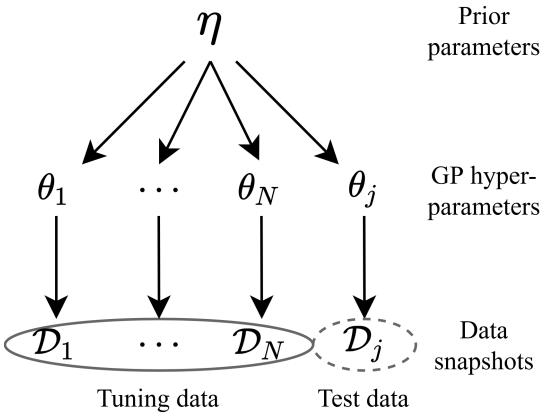

Figure 1: Visualisation of the hierarchical structure adopted. The bottom level is the observed data $\mathcal{D}$. It is modelled using GPs, which are defined by their hyperparameters θ. The distribution of the GP hyperparameters is captured by η. The hyperparameters θ are independent given η.

The structure reflects the assumption that the data snapshots are generated by distinct but similar underlying processes. Therefore, the tuning hyperparameters $(\theta_1, \cdots, \theta_N)$ cannot be used directly, but their distribution η can be.

For the BO loop we want the expectation of the acquisition function on snapshot j at point x_* given the distribution of GP hyperparameters, It can be approximated as

$$a_j(x_*) = \mathop{\mathbb{E}}_{p(\theta_i|\mathcal{D},\mathcal{D}_j)}[a_{j,i}(x_*)] \quad (8)$$

$$= \mathop{\mathbb{E}}_{p(\theta_i|\eta)p(\mathcal{D}_j|\theta_i)/Z}[a_{j,i}(x_*)] \quad (9)$$

$$\approx \frac{1}{M}\sum_{i=1}^{M} a_{j,i}(x_*), \;\; \theta_i \sim \frac{1}{Z}p(\theta_i|\eta)p(\mathcal{D}_j|\theta_i) \quad (10)$$

$$\approx \frac{1}{W}\sum_{i=1}^{M} a_{j,i}(x_*)p(\mathcal{D}_j|\theta_i), \;\; \theta_i \sim p(\theta_i|\eta) \quad (11)$$

where $W = \sum_{i=1}^{M} p(\mathcal{D}_j|\theta_i)$ is the sum of the importance weights and Z a normalising constant. $a_{j,i}(x_*)$ is the acquisition function evaluated on snapshot j using hyperparameter sample θ_i on the point x_*. In the second line η is used to summarise $\mathcal{D}$. In the third the expectation is approximated using Monte Carlo. In the fourth importance weighting is used so samples can be generated from the prior η, and weighted by their fit to the specific problem $\mathcal{D}_j$. By taking MCMC samples from $p(\theta_i|\eta)$ and not $p(\theta_i|\eta)p(\mathcal{D}_j|\theta_i)$ the computations are sped up greatly because we can approximate the distribution once and then reuse it at each iteration of the BO loop and for each test data snapshot.

MCMC Sampling

The hyperparameter samples used for the prior are inferred from the tuning set using MCMC. It is done separately for each of the model kernels (RBF-Product, Sum and RBF-RBF). Joint samples are collected of the top two levels of the hierarchical structure from Fig. 1; η and $\theta = \bigcup_{n\in N}\theta_n$. Note that θ_j is not included. θ_n is the hyperparameters for the GP model of $\mathcal{D}_n$, e.g. $(\sigma_{\mathrm{r},1,n}, l_{\mathrm{r},1,n}, \sigma_{\mathrm{r},2,n}, l_{\mathrm{r},2,n})$ for the RBF-RBF kernel. Except for γ, each hyperparameter $\theta_{n,k}$, e.g. $l_{\mathrm{r},1}$, is assumed to be from a gamma distribution, $\theta_{n,k} \sim \Gamma(\psi_k, \phi_k)$, and η is the set of parameters of those gamma distributions, $\eta = \bigcup_{k\in K}\{\psi_k, \phi_k\}$. Thus $\theta_n = \bigcup_{k\in K}\{\theta_{n,k}\}$.

The sampling is done by alternatingly sampling η and θ. The tuning set $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ is used to calculate the likelihood of the samples. The details of this sampling process are given in the appendix. The result is H samples of $(\eta^{(h)}, \theta^{(h)})$. The M samples of θ_i for the importance weighting are generated by doing the following M times: select h at random, sample $\theta_{i,k} \sim \Gamma(\psi_k^{(h)}, \phi_k^{(h)})$ for each k. We use M=100 and H=1200 (satellite) or H=2000 (London), with a burn-in of 200 samples.

Data

Two data sets are used for evaluation. The first comes from the TROPOspheric Measuring Instrument (TROPOMI) aboard the Sentinel-5P satellite from the EU's Copernicus

programme. The data set consists of 1083 images of 28x28 pixels, each giving the NO_2 concentration in $\mathrm{mol/m^2}$ within an area of about 7x7 km from the ground to the upper troposphere. The images are from October and November 2018, and have been selected for higher pollution concentrations. Images with more than 10 % missing data were excluded. The NO_2 concentrations vary greatly between images, so different subsets were considered. We use the same division as in Hellan, Lucas, and Goddard (2020). The 'Strong' subset consists of the 50 images with the highest maxima, the 'Median' subset of the 50 images with the median maxima and the 'Weak' subset of the 50 images with the lowest maxima. Tuning sets consist of 10 images adjacent to these. The 'Selection' subset consists of 100 images selected representatively from the whole set, without overlapping the above sets. Fig. 2 shows example images. The median tuning set is used for the selection subset. For this data set, a data snapshot $\mathcal{D}_n$ refers to one image.

The second data set was extracted from the London Air Quality Network (Imperial College London 1993). The full data set consists of years of data across multiple pollutants. For this paper, the NO_2 readings were used, and a tuning set constructed from the 2015 data and a test set from the 2016 data. Examples from the tuning set are shown in Fig. 3. The days with less than 40 readings were discarded, and only the readings from 'Roadside' sensors used. The former was done as days with more data are more suitable to show the benefits of BO. The latter was done as sensors with different classifications behaved differently, so this simplified the modelling. Each day was used as a data snapshot $\mathcal{D}_n$. The tuning set consists of 214 such days and the test set 365 days.

Results

Two metrics are used to evaluate the model performance. Maximum ratio is the ratio between the true value at the estimated maximiser, $\hat{y}$, and the true maximum y^*. Since there are multiple data snapshots, the average is used, $R = \frac{1}{P}\sum_{p=1}^{P} \hat{y}_p/y_p^*$ where P is the number of snapshots evaluated. The ratio metric is ideally one, with a higher score being better (it cannot exceed one). The second metric is the Euclidean distance between the estimated maximiser $\hat{x}$ and the true maximiser x^*. Again, the average is used, $D = \frac{1}{P}\sum_{p=1}^{P} |\hat{x}_p - x_p^*|$. The distance metric is ideally zero, with a higher score being worse. As in Hellan, Lucas, and Goddard (2020), we compare to a random baseline. At each iteration a location is chosen at random from those available. The result is given as the average over 100 runs on each of the data snapshots in each subset. For BO one run is used. The metrics are calculated on the pre-processed data, i.e. after log transform and standardisation.

Satellite Data

The results on the satellite data are given in Figs. 4 and 5 and Tables 1 and 2, which also give the results from Hellan, Lucas, and Goddard (2020) for comparison. The performance has improved on both the strong and selection subsets. As can be seen, the previous work did not do better than the random guessing baseline on the selection subset.

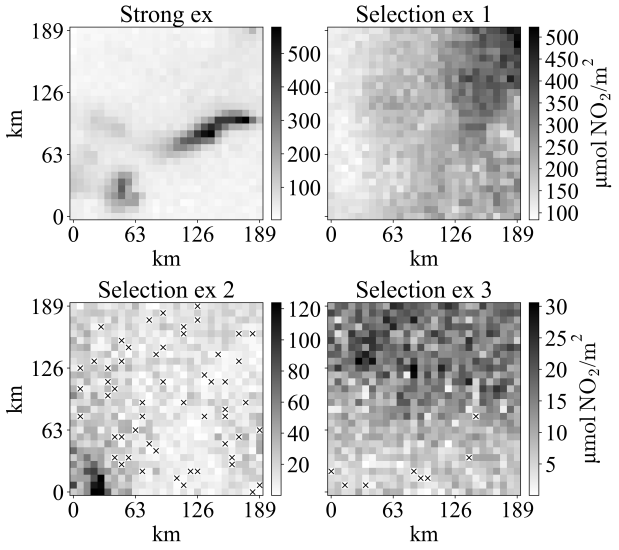

Figure 2: Examples of data snapshots from the satellite data set. Selection examples 1, 2 and 3 are the strongest, median and weakest images from the selection subset, respectively. The strong example is adapted from Hellan, Lucas, and Goddard (2020). Crosses indicate missing data.

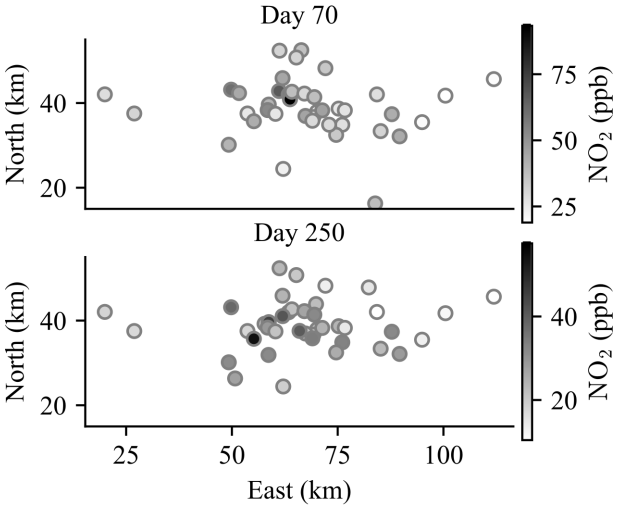

Figure 3: Examples of data from the LAQN data set. The distances are from the most south-westerly monitoring station, in Beech outside Alton in Hampshire. Note that not all sensors are available each day, that the location of the maximum varies and the strong clustering in central London.

London Data

The results on the London data are given in Fig. 6 and Table 3. As there are no previously published results for comparison, an additional random baseline is provided – random selection without replacement. As all readings are assumed to be noise-free this leads to more efficient exploration. For the satellite data the difference between the baselines is small due to the larger number of available samples.

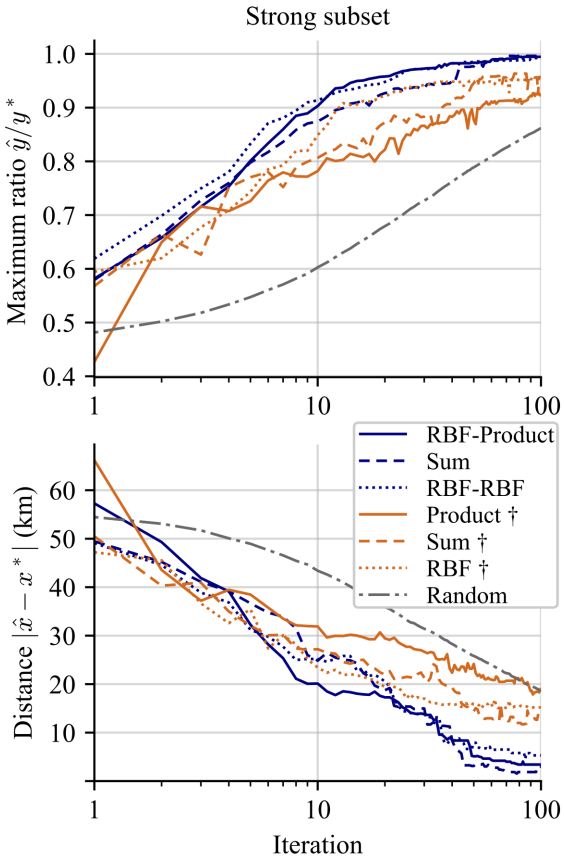

Figure 4: Results on strong subset of satellite data. Our values in navy (darkest) compared to values in Hellan, Lucas, and Goddard (2020) in orange (lightest) marked with †. The baseline 'Random' is shown in grey (medium, different pattern). The results obtained are better than those in existing work, and both out-compete the random baseline. $\hat{x}$ is the estimated maximiser and x^* the true maximiser. $\hat{y}$ and y^* are the true concentration values at $\hat{x}$ and x^*, respectively.

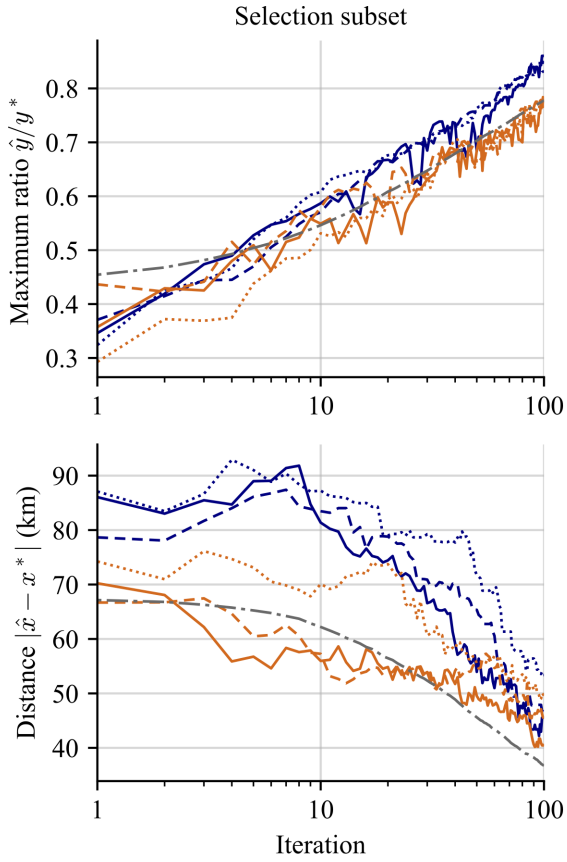

Figure 5: Results on selection subset of satellite data. The legend and variable definitions are given in Fig. 4. Values in orange marked with † are from Hellan, Lucas, and Goddard (2020). The new results are an improvement on those in existing work when considering the ratio metric; the three new results mostly lie on top of each other, and the three old results and the baseline mostly lie on top of each other. The baseline is competitive on the distance metric.

Strong	RBF-RBF **0.987-0.997**	Sum **0.993-0.999**	RBF-Product **0.990-0.999**
	RBF † 0.940-0.966	Sum † 0.945-0.969	Product † 0.907-0.941
Selection	RBF-RBF **0.797-0.866**	Sum **0.816-0.886**	RBF-Product **0.826-0.894**
	RBF † 0.755-0.804	Sum † 0.756-0.802	Product † 0.750-0.810

Table 1: Confidence intervals for the means of the maximum ratio at the final iteration for the satellite data. Given are means ± one standard deviation of the mean. The best values are given in bold. This confirms the story from Figs. 4 and 5 that improved results are obtained on the strong and selection subsets. † indicates results from Hellan, Lucas, and Goddard (2020).

Strong	RBF-RBF **2.381-8.229**	Sum **1.290-4.695**	RBF-Product **0.778-6.040**
(km)	RBF † 11.386-19.010	Sum † 8.445-16.976	Product † 14.507-23.944
Selection	RBF-RBF 44.456-62.301	Sum 36.960-56.178	RBF-Product 36.126-54.977
(km)	RBF † 44.145-55.010	Sum † 40.551-51.918	Product † 35.041-45.519

Table 2: Confidence intervals for the means of the maximiser distance at the final iteration for the satellite data. Given are means ± one standard deviation of the mean. The best values are given in bold. This confirms the story from Figs. 4 and 5 that similar results are obtained on the selection subset and improved results on the strong subset. † indicates results from Hellan, Lucas, and Goddard (2020).

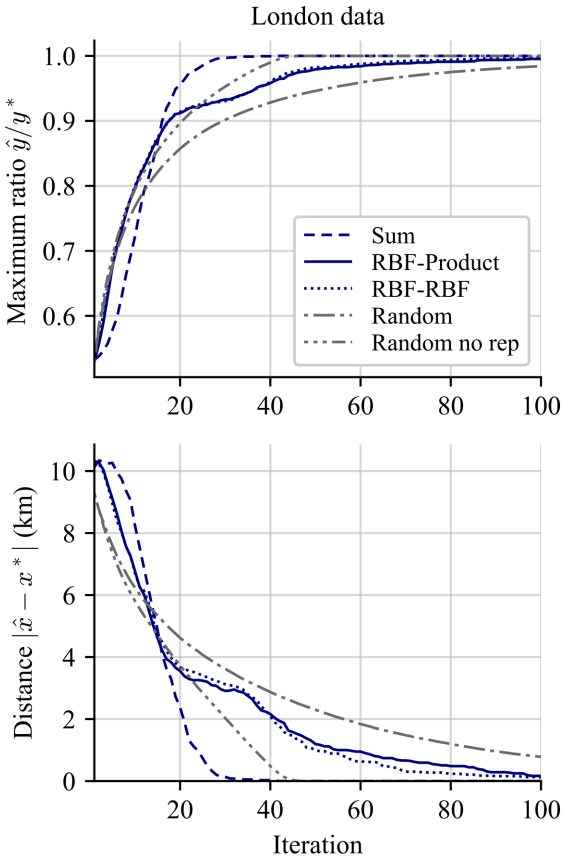

Figure 6: Results on London data. The variable definitions are given in Fig. 4. The RBF-Product and RBF-RBF lines follow each other closely and sometimes overlap. The Sum model is slower to start learning, but then finds the maximum much faster than the other models and the baselines.

Maximum ratio	RBF-RBF 0.925-0.938	Sum **0.996-0.999**	RBF-Product 0.927-0.939
	Random 0.904-0.905	Rand. no rep 0.957-0.958	
Distance (km)	RBF-RBF 2.841-3.330	Sum **0.042-0.133**	RBF-Product 2.694-3.148
	Random 3.510-3.562	Rand. no rep 1.855-1.896	

Table 3: Confidence intervals for the means of the maximum ratio at the 30$^{\text{th}}$ iteration for the London data. Given are means $\pm$ one standard deviation of the mean. The best values are given in bold.

Discussion

The results on the satellite data, Figs. 4 and 5 and Tables 1 and 2, show that our method gives improved results even on the challenging selection subset. The exception is the distance metric on the selection subset. An example of when a good ratio score corresponds to a bad distance score is an image that is zero everywhere except at the bottom right where it is 1.000, and at the top left where it is 0.999. The distance metric will be disproportionately high (bad) if only the marginally worse local optimum is found, but the ratio score will be good regardless. By inspection of Fig. 2 one can see that the selection examples have many local optima, and so would suffer from this. On the London data one of the GP models does much better than either baseline, but the other two perform worse. Why this happens is examined in the Exploration subsection. But it shows that BO can be much faster, even if care must be taken when choosing the models and priors. BO has the added benefit of providing models of the underlying process, with interpretable hyperparameters. An example of this is given in the Interpretability subsection.

Given the relatively small difference in performance between the GP models and the best random baseline one might be tempted to ask whether the extra effort needed for the former is justified. Therefore, it is worth keeping in mind the application. When deployed, each extra iteration will mean one extra sensor having to be procured and installed, requiring both money and labour. In that context, using more resources when planning the placement is preferable. For other domains, it becomes a trade-off between the cost of implementation and the cost of acquiring extra samples. If new samples can be acquired cheaply enough, then the random selection might be preferable.

Throughout this paper, several simplifying assumptions have been made. Firstly, although the London data is temporal, this has been simplified by constructing spatial problems and treating them as independent. Secondly, the measurements are treated as ground-truth readings. This simplifies the error metric calculations, but also makes irrelevant one of the advantages of probabilistically modelling the data, as this allows that uncertainty to be modelled directly. Another limitation is that the results on the satellite data are based on iteratively improving the method, and so could be considered validation results and not test results. The subsets are kept because there are published results to compare to. Because of how the data was split up it is not straightforward to generate a new Strong subset, but a new Selection subset was generated and the method evaluated on it. The results are given in the appendix, and show a strong improvement over the random baseline on both metrics.

Exploration

The results in Fig. 6 can be examined in light of the the exploration-exploitation trade-off. An exploration score is used to do so, calculated as the minimum Euclidean distance between the new sample and any one of the already collected samples, $e_{i,j} = \min |x_{i,j} - x_{a,j}|, \quad a < i$. This is done for each iteration i and data snapshot j and the average taken across the data snapshots; see Fig. 7. It shows that the reason the Sum model improves more slowly in the beginning is that it emphasises exploration. Then, in the middle stage (5 to 20 iterations) it is able to learn faster, as it has already narrowed down what area to explore further, and manages to find the maximum faster. Conversely, and by chance, some of the random baselines' early iterations will be checking

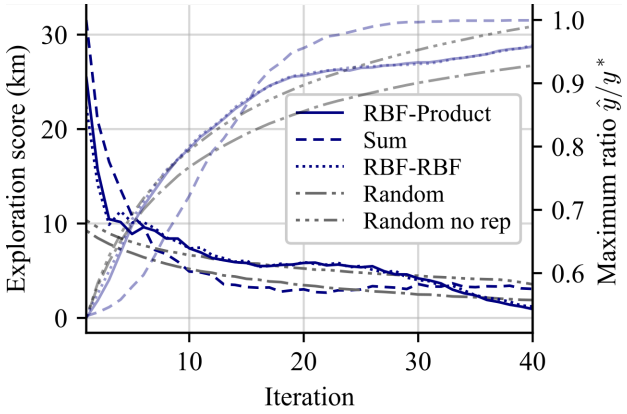

Figure 7: Exploration score of different methods on London data. Overlaid and paler are the ratio scores from Fig. 6. Note that Sum needs longer for the maximum ratio score to increase, and has a higher exploration score at the start.

neighbouring points of those already seen, which can give small-scale improvement. The two other BO kernels start off with an exploration score between Sum and the first random baseline, and the performance is also between the two.

Interpretability

Part of our rationale for using comparatively simple GP kernels is to facilitate interpretation of hyperparameters, and analysis of the distribution patterns of pollution in an area, e.g. its smoothness and variability in space and time. This lets us perform "at-a-glance" comparisons of different areas, discern whether extrema are more attributable to local sources or long-range patterns, and importantly lets domain experts assess the plausibility of a model's hyperparameters.

To illustrate, we consider the lengthscales $l_{\mathrm{r},i}$ of the RBF-RBF model in Eq. (5) on the London data. For simplicity, we take the mean of the samples generated for the prior, as that shows the trends better than the posterior on a single day. Converting the values to the standard RBF expression gives $l_{\mathrm{r},1}$ = 2.00 km, $l_{\mathrm{r},2}$ = 241 km, $\sigma_{\mathrm{r},1}$ = 2.05 and $\sigma_{\mathrm{r},2}$ = 2.04. This corresponds to one local component and one regional component contributing roughly equally to the overall variability of NO_2 concentrations. At a distance of 100 metres the expected correlation between NO_2 concentrations will be close to 1, driven by both local and long-range effects, whereas at a distance of 10 km it falls to 0.5, and local effects play a negligible role. The correlations are also visualised in Fig. 8. The LAQN provides a pollution map online

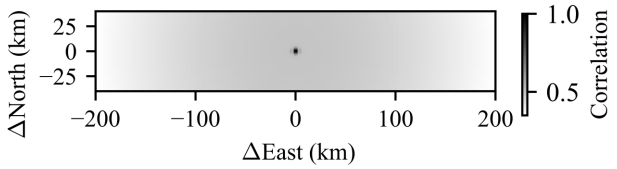

Figure 8: Expected correlation as function of distance for the mean of hyperparameter prior samples for the London data.

at their website (Imperial College London 1993). It shows pollution across all of London, but with more concentrated in central London radiating outwards and around Heathrow airport, as well as along major roads. This supports the modelling hypothesis of a local and a regional component.

Conclusion

We have shown that Bayesian optimisation with hierarchical models can be successfully applied to ground-level urban pollution data. We also presented a pragmatic method for approximate inference when some of the work can be precomputed, which we believe can be useful in other applications. In Bayesian optimisation the GP tuning step needs to be repeated many more times than in standard regression, as new data points are added iteratively. Therefore, using standard MCMC on the full hierarchical model for each new data point would slow down the process. By using importance weighting in the BO tuning step the process can be sped up. This is even more useful in other applications where collecting samples is less time consuming than for pollution.

As expected, the models performed worse on the ground-level data. This is due in part to the data snapshots having fewer available samples. Some exploration is needed before the model can usefully discern the areas of interest, and where the data is sparse compared to the local variations this is harder. The strength of BO lies in being able to exploit structure in data, so it works better for slowly varying processes. While the results show that the approach has promise, more evaluation is needed. In particular, due to data constraints, London was used both for developing the prior and for evaluation. To test whether the prior can be constructed from other cities additional data needs to be used. It would also be beneficial to test the methods as they would be used, i.e. iteratively placing sensors, but this would be much more resource-intensive. There might be easy performance gains available by experimenting with kernel families (e.g. Matérn) and combinations, such as more complex compositional kernels to model correlations at different scales. The benefits of modelling seasonal and weekly patterns should also be explored. Additionally, the results showed the benefit of early exploration, so the methods could be biased towards this by modifying the acquisition function. Berk et al. (2018) present a modification to expected improvement that increases early exploration without needing manual tuning.

There are two main extensions needed before deploying the method: temporal modelling and accommodating sensors with varying precision and reliability. Temporal modelling is important as pollution levels vary temporally, with some days bringing high pollution levels throughout a city. The temporal aspect can be avoided by only considering averages, but this ignores the legal limits on hourly averages (World Health Organization 2006, p. 174-175) and makes adding data from new sensors harder. The uncertainty of a pollution reading is important when using low-cost sensors for monitoring to avoid misleading statements. Smith et al. (2019) present a solution to this, but do not discuss how to place the sensors. Unifying the active placement with the uncertainty treatment is necessary for successfully applying it using a network of low-cost sensors.

Ethics Statement

We believe the risks are very low from our proposed method. The only data used are NO_2 concentrations, and the latitude and longitude of the stationary sensors. No images or personal data of any kinds are used. This is an advantage over other proposed solutions, e.g. using taxis to collect data. The potential benefits far outweigh the harms, by allowing local communities to keep authorities accountable.

Acknowledgements

The authors would like to thank Dr Douglas Finch at the University of Edinburgh for extracting the data and helpful correspondence, and Professor Iain Murray, also of the University of Edinburgh, for helpful advice and discussions. The authors also thank the anonymous reviewers for their time and thoughtful comments.

This work was supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Ainslie, B.; Reuten, C.; Steyn, D. G.; Le, N. D.; and Zidek, J. V. 2009. Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *Journal of Environmental Management*, 90(8): 2715–2729.

Asenov, M.; Rutkauskas, M.; Reid, D.; Subr, K.; and Ramamoorthy, S. 2019. Active Localization of Gas Leaks Using Fluid Simulation. *IEEE Robotics and Automation Letters*, 4(2): 1776–1783.

Berk, J.; Nguyen, V.; Gupta, S.; Rana, S.; and Venkatesh, S. 2018. Exploration Enhanced Expected Improvement for Bayesian Optimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, 621–637. Springer International Publishing. ISBN 978-3-030-10928-8.

Carminati, M.; Ferrari, G.; and Sampietro, M. 2017. Emerging miniaturized technologies for airborne particulate matter pervasive monitoring. *Measurement*, 101: 250–256.

Cats, G.; and Holtslag, A. 1980. Prediction of air pollution frequency distribution—Part I. The lognormal model. *Atmospheric Environment (1967)*, 14(2): 255–258.

Chib, S.; and Greenberg, E. 1995. Understanding the Metropolis - Hastings Algorithm. *The American Statistician*, 49(4): 327–335.

Cocchi, D.; Greco, F.; and Trivisano, C. 2007. Hierarchical space-time modelling of PM10 pollution. *Atmospheric Environment*, 41(3): 532–542.

Cohen, A. J.; Brauer, M.; Burnett, R.; Anderson, H. R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; Feigin, V.; Freedman, G.; Hubbell, B.; Jobling, A.; Kan, H.; Knibbs, L.; Liu, Y.; Randall, M.; and Forouzanfar, M. H. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082): 1907–1918.

Dawkins, L. C.; Williamson, D. B.; Mengersen, K. L.; Morawska, L.; Jayaratne, R.; and Shaddick, G. 2020. Where Is the Clean Air? A Bayesian Decision Framework for Personalised Cyclist Route Selection Using R-INLA. *Bayesian Analysis*, 16(1): 61–91.

De Ath, G.; Everson, R.; and Fieldsend, J. 2021. How Bayesian Should Bayesian Optimisation Be? arXiv:2105.00894.

Hellan, S. P.; Lucas, C. G.; and Goddard, N. H. 2020. Optimising Placement of Pollution Sensors in Windy Environments. Presented at the NeurIPS 2020 Workshop on AI for Earth Sciences, arXiv:2012.10770.

Hoffman, M. D.; and Gelman, A. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623.

Imperial College London, E. R. G. 1993. London Air Quality Network. https://www.londonair.org.uk/LondonAir/Default.aspx. Accessed: 2021-08-31.

Jones, D. R.; Schonlau, M.; and Welch, W. 1998. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4): 455–492.

Katsouyanni, K. 2003. Ambient air pollution and health. *British Medical Bulletin*, 68(1): 143–156.

Kelly, K. E.; Whitaker, J.; Petty, A.; Widmer, C.; Dybwad, A.; Sleeth, D.; Martin, R.; and Butterfield, A. 2017. Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental Pollution*, 221: 491–500.

Lalchand, V.; and Rasmussen, C. E. 2020. Approximate Inference for Fully Bayesian Gaussian Process Regression. *Symposium on Advances in Approximate Bayesian Inference*, 1–12. PMLR.

Lelieveld, J.; Evans, J. S.; Fnais, M.; Giannadaki, D.; and Pozzer, A. 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569): 367–371. Number: 7569 Publisher: Nature Publishing Group.

Liu, X.; Jayaratne, R.; Thai, P.; Kuhn, T.; Zing, I.; Christensen, B.; Lamont, R.; Dunbabin, M.; Zhu, S.; Gao, J.; Wainwright, D.; Neale, D.; Kan, R.; Kirkwood, J.; and Morawska, L. 2020. Low-cost sensors as an alternative for long-term air quality monitoring. *Environmental Research*, 185: 109438.

Marchant, R.; and Ramos, F. 2012. Bayesian optimisation for Intelligent Environmental Monitoring. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2242–2249. IEEE.

Morere, P.; Marchant, R.; and Ramos, F. 2017. Sequential Bayesian optimization as a POMDP for environment monitoring with UAVs. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 6381–6388. IEEE.

Murray, I.; and Adams, R. P. 2010. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, volume 23.

Rasmussen, C. E.; and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Reggente, M.; and Lilienthal, A. J. 2009. Using local wind information for gas distribution mapping in outdoor environments with a mobile robot. In *2009 IEEE SENSORS*, 1715–1720.

Sahu, S. K. 2012. Hierarchical Bayesian Models for Space–Time Air Pollution Data. In Subba Rao, T.; Subba Rao, S.; and Rao, C. R., eds., *Handbook of Statistics*, volume 30 of *Time Series Analysis: Methods and Applications*, 477–495. Elsevier.

Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; and de Freitas, N. 2015. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1): 148–175.

Shiffrin, R. M.; Lee, M. D.; Kim, W.; and Wagenmakers, E.-J. 2008. A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science*, 32(8): 1248–1284.

Singh, A.; Ramos, F.; Whyte, H. D.; and Kaiser, W. J. 2010. Modeling and decision making in spatio-temporal processes for environmental surveillance. In *2010 IEEE International Conference on Robotics and Automation*, 5490–5497.

Smith, M. T.; Ssematimba, J.; Alvarez, M. A.; and Bainomugisha, E. 2019. Machine Learning for a Low-cost Air Pollution Network. Presented at the NeurIPS 2019 Workshop on Machine Learning for the Developing World, arXiv:1911.12868.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 25: 2951–2959.

Williams, C. K. I.; and Rasmussen, C. E. 1996. Gaussian processes for regression. *MIT*.

World Health Organization. 2006. *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. Copenhagen, Denmark: World Health Organization. ISBN 978-92-890-2192-0. OCLC: ocn162119780.