An Exercise in Selecting Low-Cost Air Quality Sensor Placements within an Urban Environment

by
Adrianna J. Boghozian

B.S. Mathematics & B.A. Economics, University of Tennessee (2013)
Submitted to the Institute for Data, Systems, and Society and Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degrees of Master of Science in Technology and Policy and Master of Science in Electrical Engineering and Computer Science at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

© Massachusetts Institute of Technology 2021. All rights reserved.
An Exercise in Selecting Low-Cost Air Quality Sensor Placements within an Urban Environment

by

Adrianna J. Boghozian

Submitted to the Institute for Data, Systems, and Society and Department of Electrical Engineering and Computer Science on January 27, 2021, in partial fulfillment of the requirements for the degrees of Master of Science in Technology and Policy and Master of Science in Electrical Engineering and Computer Science

Abstract

Air pollution poses the most important environmental health risk to citizens of major cities all over the world. The high cost of current monitoring programs means that enforcement of current regulation, such as the United States Environmental Protection Agency’s ambient air quality standards, can be lacking at the individual level. Because of their low cost, sensor networks offer the benefit of providing detailed, high resolution pollutant exposure maps which can inform a number of community and government initiatives aimed at tackling air pollution.

The question then arises, what is the optimal configuration of low-cost sensors to measure air pollution within an urban environment? Due to the large number of potential locations in which to measure data, there are difficulties in defining where to place a limited number of sensors. This thesis outlines a proven decision method from spatial statistics: optimal experimental design, and applies the method to a test case in the city of London.

Thesis Supervisor: Youssef Marzouk
Title: Professor of Aeronautics and Astronautics

Thesis Supervisor: Stefanie Jegelka
Title: Associate Professor of Electrical Engineering and Computer Science
Acknowledgments

I am so very grateful to both of my research advisors, without whose guidance and support this thesis could not have happened. Youssef Marzouk, who provided immense computational advice – including helpful reviews of matrix algebra – and who was always willing to fit meetings in, even when his plate appeared too full. And David Hsu, whose vision and determination allowed this project to flourish.

I am also thankful to my EECS reader, Stefanie Jegelka, who advised a much earlier version of this project in the spring of 2019 and provided crucial feedback, without which this thesis would not be what it is today.

I could not have successfully managed all of the bureaucracies of MIT without the brilliant and compassionate Barb DeLaBarre. Nor could I have ever mastered the difference between ‘which’ and ‘that’ without the thorough editing of Frank Field. I owe both of them so many more thanks for making my overall experience at MIT engaging and enjoyable.

I am grateful to MIT’s Tata Center for the financial support and the amazing opportunity to travel to India... twice! I also have to thank the Tata Center’s teaching staff – specifically Chintan Vaishnav and Diane Rigos – for providing a once-in-a-lifetime tour of India as well as leading an enriching and thoughtful seminar throughout my first year at MIT.

I can’t thank Sidhant Pai enough for answering any and all of my air quality questions (sometimes more than once) and for being the best international travel companion. Thank you for sharing some of your favorite parts of your country with me and for teaching me to how to properly haggle for an auto.

I have been extremely lucky to be part of a thriving and supportive TPP and Tata student community while at MIT and am so grateful to my cohorts for making my experience one at which I will look back fondly.

I also have to thank all of my friends and family – especially those who tolerated me moving (again) to pursue graduate school (thankfully Zoom sessions have now become normalized for everyone). Thank you to my parents, who both went back to
school for their master’s degrees after having children, and who taught me that the most important things in life are not actually fancy degrees but spending time with family. And taking tequila shots.

Last but certainly not least, I want to thank my partner of over eleven years: Brian. Specifically for encouraging me to come to graduate school, even if it meant us initially managing another long-distance relationship. Your continued unwavering support, excitement for cross-country road trips, and love of explorin’ new places has made this New England chapter a memorable one.
5 Selecting Sensors using Mutual Information

5.1 Training the Gaussian Process Model ........................................... 51
5.2 Applying Krause’s OED Algorithm ............................................. 52
5.3 Discussion and Next Steps ......................................................... 58

6 Exploring the Political Usefulness of Environmental Data

6.1 Citizen sensing initiatives through the lens of AQ alerts ............... 62
6.2 Bringing behavioral change to bear .......................................... 65
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Great Smog of London, 1952</td>
<td>Britannica [44]</td>
</tr>
<tr>
<td>1-2</td>
<td>NO(x) emissions in kilotonnes in the UK by source, 1970-2020</td>
<td>UK Department for Environment, Food and Rural Affairs [21]</td>
</tr>
<tr>
<td>2-1</td>
<td>Posterior mean and variance of a temperature GP; (a) Predicted temperature; (b) predicted variance</td>
<td>Krause <em>et al.</em> 2008 [26]</td>
</tr>
<tr>
<td>2-2</td>
<td>Example of placements chosen using entropy (diamonds) and mutual information (squares) criteria</td>
<td>Krause <em>et al.</em> 2008 [26]</td>
</tr>
<tr>
<td>2-3</td>
<td>Krause’s approximation algorithm for maximizing mutual information</td>
<td>Krause <em>et al.</em> 2008 [26]</td>
</tr>
<tr>
<td>2-4</td>
<td>Mutual information of greedy sets of increasing size</td>
<td>Krause <em>et al.</em> 2008 [26]</td>
</tr>
<tr>
<td>2-5</td>
<td>Comparison of Krause’s greedy algorithm with the optimal solution on a small problem. Due to complexity of the problem in most cases, we cannot expect to find the optimal solution in polynomial time. Thus, Krause’s greedy algorithm is a good alternative to the optimal solution. The upper bound in this figure was able to be found by Krause <em>et al.</em> due to the small scale of the example problem</td>
<td>Krause <em>et al.</em> 2008 [26]</td>
</tr>
<tr>
<td>3-1</td>
<td>Coverage map of KCL air quality monitoring sites</td>
<td></td>
</tr>
<tr>
<td>3-2</td>
<td>Plot depicting the number of KCL monitoring sites across the time period for each pollutant</td>
<td></td>
</tr>
</tbody>
</table>
3-3 Coverage map of KCL monitoring sites NO\textsubscript{x} only, points are sized by the number of years in which data is reported \hspace*{1cm} 39

3-4 Raw time series of NO\textsubscript{x} Concentration data, 2005–2019 \hspace*{1cm} 40

3-5 Decomposed time series of monthly NO\textsubscript{x} data, 2000–2019 \hspace*{1cm} 41

3-6 Monthly NO\textsubscript{x} concentration data, averaged over all geographic sites and across the time series: 2005–2019 \hspace*{1cm} 41

4-1 Examples of structures expressible by some basic kernels (Source: Duvenaud 2014 [15]) \hspace*{1cm} 44

4-2 Examples of structures expressible by some combination of kernels (Source: Duvenaud 2014 [15]) \hspace*{1cm} 44

4-3 Comparison of the predicted versus actual NO\textsubscript{x} concentration levels during November 2008. The relative similarity suggests that the model does a pretty good job capturing the relative difference among sensor sites. \hspace*{1cm} 49

4-4 Time series model fit for three randomly sampled sensor locations. Points represent the NO\textsubscript{x} concentration across time (month-year). Red-colored points represent the observations used for model training, where black-colored points represent the observations for validation. The blue line represents the model’s prediction and the light blue shaded region represents the 95\% confidence interval. \hspace*{1cm} 50

5-1 Set of \( S \) locations where we can place sensors (purple squares). Set \( V = S + U \), where \( U \) (teal triangles) are where we wish to measure but cannot place a sensor, and \( S \) (purple squares) are the eligible sensor locations. \( S \) and \( U \) are disjoint. \hspace*{1cm} 54

5-2 The first 50 sensor locations selected by the OED algorithm. The above plot displays sensors 1-25 and the bottom plot overlays the next 25 selected sensors. \hspace*{1cm} 55

5-3 All sensor locations selected by the OED algorithm. \hspace*{1cm} 56
Calculated mutual information values for the sensors placed by the OED algorithm.

U.S. EPA’s Air Quality Index Guide. The AQI consists of six levels that range from Good to Hazardous (with a corresponding color scheme), depending on the potential health concern. State and local agencies are required to report daily air quality indexes for all U.S. cities with populations over 350,000. (Source: US EPA [2]).
List of Tables

3.1 High-level summary of pollutant representation in the KCL dataset. . 37
Chapter 1

Introduction

In the next century, planet earth will don an electronic skin. It will use the Internet as a scaffold to support and transmit its sensations. This skin is already being stitched together. It consists of millions of embedded electronic measuring devices: thermostats, pressure gauges, pollution detectors, cameras, microphones, glucose sensors, EKG’s, electroencephalographs. These will probe and monitor cities and endangered species, the atmosphere, our ships, highways and fleets of trucks, our conversations, our bodies—even our dreams.

Neil Gross, 1999, on the Internet of Things

As the proportion of the global population living in urban areas continues to increase, air pollution is posed to be the most important environmental health risk, particularly in the developing world [27]. Today’s air pollution monitoring programs throughout the world are managed by environmental or governmental authorities that rely on fixed, regulatory-grade stations that collect concentration levels of a number of pollutants proven to be hazardous for human health (i.e. carbon monoxide (CO), nitrogen oxides (NO, NO₂), sulfur dioxide (SO₂), ozone (O₃), and particulate matter (PM)). Advances in portable, low-cost sensors have been shown to “reduce initial investments and operating costs by 3- and 5-fold in the US, respectively” [28]. Because of their low cost, sensor networks offer the benefit of providing detailed high resolution
spatial-temporal data on the dispersion of air pollutants in urban environments. And since it has been shown that pollution concentrations can vary over relatively small scales [7], these low-cost sensor networks have the potential to provide real-time, high resolution pollution exposure maps (<1m) that are crucial to understanding population exposure in cities [65].

One question that arises when attempting to measure pollutant concentrations in a given environment is: what is the optimal location for the next sensor within a network? Because of the large number of potential locations where we can capture data, it can be overwhelming to determine the best placement of individual sensors within a given urban environment. A method from spatial statistics utilizes a pilot deployment or expert knowledge to train a Gaussian process (GP) model for air pollution dispersion [26]. The GP model can then be used to optimize the placement position of sensors that are selected iteratively. This thesis explores applying these methods to a real urban environment: London. To this end, we utilize historical air quality measurement data from London’s largest monitoring program to (1) attempt to build a rich GP model that can offer reasonably accurate predictions of pollutant dispersion within the urban network, (2) train a similar (albeit simpler) GP model to deploy a test case of Krause’s optimal experimental design (OED) algorithm to explore its potential utility for future sensor placement strategies. We conclude with thoughts on the potential for environmental citizen-sensing initiatives in influencing citizen behavior.

1.1 thesis Organization

This thesis aims to explore low-cost air pollution monitoring in six parts. In the following subsection, we walk through the history of urban air quality monitoring. We summarize the increased attention paid to the environment throughout the second half of the twentieth century. We also outline the mixture of gases and particles that constitute our atmosphere and describe the proven harmful effects of air pollutants found within it. The subsection ends by investigating how our government
has approached regulating these harmful pollutants in the air. The next subsec-
tion describes the rise of low-cost air pollution monitoring alternatives thanks to the
miniaturization of computing technology and the open science movement. We also
recognize recent community sensing initiatives and review a few of the current city-
wide, low-cost monitoring networks. In the final subsection of Chapter 1, we turn our
attention to London’s air pollution problem, since our project will utilize data from
London’s monitoring network. This section provides necessary background on NOx – a pollutant of particular interest to this project. We also review seasonal pollution
patterns within the city.

Chapter 2 focuses on three key methods which are used within this project: Guass-
sian processes, mutual information criterion, and Krause’s optimal experimental de-
sign algorithm. We provide a high-level review of each method as necessary back-
ground for our analysis in later chapters.

In Chapter 3, we describe the dataset used in the project: a collection of pollution
concentration measurements from over 400 monitoring sites around London collected
from 1993–2019. We describe how NOx emerges as a compelling modeling candidate
after initial investigation of the data. We also present our time series analysis on the
NOx concentration data that unveils underlying patterns we attempt to model in the
following chapter.

Chapter 4 outlines the spatial and temporal Gaussian process (GP) model we
build to attempt to model the pollution dispersion of NOx. The chapter also provides
a detailed review of a few specific kernels used in the model and ends with interpreting
the model results.

In Chapter 5, we apply a test case of the optimal experimental design algorithm
to a subset of the London data. We describe the simplified GP model that is then
used to inform the sensor selection ordering. The chapter concludes with a discussion
on the feasibility of applying the algorithm in the real world and possible future next
steps.

In our final chapter, we end the thesis by exploring not only the intended purpose
of environmental monitoring and alert systems, but the potential gap between their
presumed and actual effects on human behavior. Due to the dearth of high quality studies on the effectiveness of citizen sensing initiatives, we focus on recent evaluations of federal air quality alert programs and their effect on human health. We then attempt to expand the insights from these evaluations to citizen science initiatives, like low-cost monitoring, while highlighting potential areas of improvement.

1.2 Motivation and Related Work

1.2.1 A Brief History of Urban Air Quality Monitoring

People started widespread environmental monitoring in the nineteenth century during the establishment of scientific observatories. Initially focused on astronomy, the observatories expanded their scope of work to include “precision measurement, numerical data processing, and the representation of scientific information on a global or cosmic scale” [5]. Air became a topic of increased attention as a consequence of the expansion of coal-fired manufacturing, culminating in catastrophic air pollution events in the middle of the twentieth century [5, 16]. Western governments responded to the resulting public health fears by drafting Clean Air Acts, such as those in the United Kingdom (1956) [12] and the United States (1963) [4], establishing national programs aimed at regulating and controlling hazardous air pollutants. These laws further legitimized environmental monitoring efforts and drew heightened attention to the potential environmental dangers that citizens interacted with daily.

There are many different chemical species that constitute “air pollution”, each with their varying effects on human health. In order to better understand air quality monitoring, it may be helpful to review what is actually in the air. Our atmosphere includes a combination of gases: nitrogen (78%), oxygen (21%), argon (0.93%), and carbon dioxide (0.04%) as well as other trace compounds [60]. Among those trace compounds are pollutants that are harmful to human health and usually take the

---

1Parts of this section are extracted from a previous essay I wrote during a course at MIT, which can be found as part of a larger collection here: A City is Not a Computer: The Data plays for a year.
form of ppm (parts per million) or ppb (parts per billion) molecules of air. Common harmful pollutants can be generally characterized as a mixture of small particles such as particulate matter (PM) – like dust, soot, or drops of liquid – that are further broken down into coarse (PM_{10}, < 10 microns in diameter) and ultrafine (PM_{2.5}, < 2.5 microns in diameter). Pollutants such as carbon monoxide (CO), nitrogen oxides (NO, NO_{2}), ozone (O_{3}), and sulfur dioxide (SO_{2}) have also been shown to harm humans [18].

Now that we know the general makeup of what is in the air, we can discuss the harms that derive from them. The World Health Organization (WHO) labels ambient (outdoor) air pollution as a major cause of death and disease globally – attributing 4.2 million annual premature deaths from exposure to heart, lung, and cardiopulmonary disease, with the strongest evidence of public health concern linked to PM_{2.5} [50]. Similarly, The Lancet’s 2015 Global Burden of Disease study found that exposure to outdoor fine particulate matter is the fifth leading risk factor for death worldwide [13]. Jennifer Gabrys, in her book Program Earth: Environmental Sensing Technology and the Making of a Computational Planet, states: “[p]ollutants, in other words, are often present in seemingly miniscule quantities and yet are able to disrupt and remake environments, bodies, and ecological processes on local and global scales” [19]. Thus,
environmental monitoring efforts are vital to understanding the prevalence of these particularly dangerous particles that in most cases are invisible to the human eye.

Nowhere are the effects of air pollution felt more than in densely populated cities. Since the initial drafting of the Clean Air Acts, the major pollutant sources have evolved from predominantly coal-driven sources to automobile traffic and energy generation that directly release the colorless and odorless (and dangerous) PM and NO\textsubscript{x} \[19, 50\]. The WHO estimates that over 90\% of the world’s population lives in places where air pollutant concentrations exceed their limits \[50\]; though while the burden is not equal among developed and developing countries, ambient outdoor air pollution is a consistent threat in many countries \[50\]. Furthermore, air pollution exposure disproportionately affects those living in proximity to urban roadways and highways \[55\].

Today’s air pollution monitoring efforts are led by environmental and governmental authorities and use fixed, regulatory-grade stations throughout the country. Institutional and governmental bodies set standards for air pollution monitoring based on what health research determines as evidence of the level of harm caused by particular pollutants \[19\]. Thus, a subset of pollutants categorized as dangerous to human health are the focus of monitoring, as well as management and mitigation efforts. The U.S. Environmental Protection Agency (EPA) sets ambient air quality standards for six criteria air pollutants that it deems harmful to human health: ground-level ozone (O\textsubscript{3} – which is not directly emitted but increases when NO\textsubscript{x} and volatile organic compounds are exposed to sunlight), carbon monoxide (CO), sulphur dioxide (SO\textsubscript{2}), particulate matter (PM\textsubscript{10}, < 10 microns in diameter; PM\textsubscript{2.5}, < 2.5 microns in diameter), lead, and nitrogen dioxide (NO\textsubscript{2}) \[3\].

Despite the fact that air quality standards are determined by the effect of air pollution on human health, monitoring and enforcement of those standards can be lacking at the individual level. The previous high cost of sensor placement and monitoring has necessarily led to strong assumptions about the homogeneity of pollutant dispersion in urban areas despite the science that shows that far more granular sampling would be required to get an accurate measure of human exposure.
1.2.2 Promise of Low-Cost Monitoring Alternatives

Even though air quality standards are set according to the effect of air pollution on human health, monitoring and enforcement of those standards can be lacking at the individual level. Currently, regulatory-grade (or “reference grade” i.e. highly accurate) monitoring stations can cost tens of thousands of dollars, and also require significant infrastructure and trained personnel to operate – meaning that only small numbers are scattered throughout major cities [55], despite the fact that pollution levels can vary widely due to the effect of buildings, wind patterns, and weather on pollutant dispersion [55]. Furthermore, as individuals move throughout the city in a given day, their cumulative exposure levels can vary drastically, even though a monitoring station may find that pollution emitted from a specific source may be within an allowed limit.

All of this is to say that air pollution is bad for humans; the concentrations of the pollutant species are highly variable and are influenced by many local factors; but the cost of monitoring has led to a regulatory strategy that grossly misstates human exposure because of budgetary constraints. Fortunately, advancements in miniaturization of computing technology means we can embed a computer chip almost anywhere – turning anything into a potential sensor [54]. For instance, the Copenhagen Wheel, developed by MIT’s Senseable City Lab, can be added to any regular bicycle to transform it into an electric bike with internal sensors that allow it to map “pollution levels, traffic congestion, and road conditions in real-time” [29]. The advancement of cheap, small sensors has been a particular asset to the environmental monitoring community, where sensor networks with low-cost components have allowed for greatly increased spatial coverage in order to better understand environmental processes [43]. A 2019 review article in *Frontiers in Earth Science* on low-cost environmental sensor networks found that a large number of atmospheric studies were focused on air pollution rather than meteorology [43]. The authors found this finding unexpected, given that historically air pollution monitoring has relied on expensive laboratory equipment for passive sampling but that the “public awareness of health risks” and “the prolifera-
tion of low-cost *in situ* sensors” have driven the rise in using cheaper sensors for air pollution monitoring [43].

Low-cost technology is not the only factor that has driven the growth of these sensor networks. The open science movement [64] has been credited with enabling the sharing of sensor designs, underlying software, and data – making large swaths of it available to the public [43]. In general, urban science has seen a remarkable influx in the collection of big data [24] particularly in smart cities. The London Air Quality Network (LAQN) [38] is one of the largest and longest running networks and includes a mix of regulatory-grade, fixed sensors as well as low-cost sensors. A research project funded by the National Science Foundation called SAGE: A Software-Defined Sensor Network (previously entitled “Array of Things”) involves a collaboration among multiple universities including Northwestern University and Argonne National Laboratory as well as the City of Chicago to distribute sensor nodes across the city in an effort to collect “real-time data on urban environment, infrastructure, and activity for research and public use” [17]. Both LAQN’s and Chicago’s sensor data is easily accessible to the public in an effort to engage with local community groups, concerned citizens, industry, and the broader research community.

Air pollution mitigation efforts have been focused on two primary strategies: emissions reduction (decreasing pollution from fixed sites) and exposure avoidance (encouraging individuals to monitor their exposure and take alternative routes through cities) [19]. The latter strategy, in particular, offers an opportunity for “citizen sensing” initiatives. These initiatives utilize the proliferation of low-cost digital sensor technologies and smartphones to crowdsource environmental data, making it more immediate and relevant to citizens as they move throughout the city. Networks of digital sensors deployed across an urban area allow for much richer spatial and temporal data than that provided by stationary, high-cost monitoring stations, and proponents argue that this democratization of data enables effective modes of action on part of the citizen. Smartphone apps and personal air quality sensors allow citizens to view their exposure levels in real time as well as receive alerts whenever the air quality around them falls below a certain level [30]. While community-based air
quality monitoring initiatives such as hackAIR [1], Safecast [57], and PurpleAir [52] aim to tackle air inequality by providing open-source air quality data to researchers collected by low-cost sensors distributed throughout communities. The ubiquity of these low-cost digital technologies has allowed for democratization in the collection of environmental data; enabling everyone from data scientists to concerned citizens the ability to monitor what was previously only in the purview of the government.

Since most of these air quality monitoring initiatives are community-led or are built upon existing monitoring networks, the decision of where to place sensors within a city can be driven by a variety of factors. For instance, the London Air Quality Network is owned and funded by local authorities, Business Improvement Districts, Transport for London, and the city’s Department of Environment, Food, and Rural Affairs [31]. The decision of establishing a new monitoring site lies with the individual entity or borough, meaning that coordination among the entire network is lacking. Additionally, Chicago’s SAGE is interested in measuring a variety of urban data points, including vehicle and pedestrian traffic, so the placement of sensors is not necessarily optimized for measuring air quality [17]. Lastly, community-based air quality networks generally rely on grassroots distribution, whether through the agreement to share data collected by privately owned personal sensors [52] or by concerned citizens and/or volunteers installing them within their communities [57, 1].

Furthermore, while there are isolated examples of studies determining placement of low-cost air quality sensors [26, 61, 47], they are constrained to a smaller geographic area and/or a smaller number of fixed sensors. More commonly, studies regarding low-cost air quality sensors are more concerned with the accuracy of the network within an urban environment, and take the placement of individual sensors for granted [51, 58]. Thus, to our knowledge, this thesis presents the first attempt to directly apply Krause’s optimal experimental design algorithm to an urban city as a test case. Since we will utilize data from the London Air Quality Network for our project, the following section provides a necessary background on air pollution in the city.
1.2.3 Air Quality in London

Air pollution is not limited to developing countries. Citizens of Western countries commonly experience air quality events that can trigger serious adverse health effects [11]. In the UK (the primary focus of this research project), air pollution remains a pervasive threat. Most of London’s air pollution comes from traffic sources, as is the case in many European cities, due to the persistent use of diesel cars. London, in particular, is known as one of the most polluted places in the UK due to its highly concentrated road network and tall buildings, which can trap air pollution [35]. In 2017, researchers found that 7.9 million Londoners (representing close to 95% of the city’s population) live in areas that exceed the global limits for PM$_{2.5}$ pollution by 50% or more annually [62]. In fact, the residents along Brixton Road in South London reached their annual NO$_2$ exposure limit after the first five days of 2017 [8].

One of the most common and dangerous pollutant categories (and the focus of this project) are nitrogen oxides (NO$_x$). This group of gases commonly includes nitrogen monoxide (NO) and nitrogen dioxide (NO$_2$). Diesel cars are by far the largest contributor of NO$_x$ in the city, representing about 50% total emissions [21]. Figure 1-2 displays the sources of NO$_x$ pollution throughout the years. NO and NO$_2$ are produced when nitrogen (released from fuel combustion) reacts with oxygen in the air [48]. NO$_x$ gases can react to produce smog and acid rain – and the gasses are instrumental in the formation of ground level ozone (O$_3$) and fine particulate matter (PM$_{10}$, PM$_{2.5}$) [48].

London experiences air pollution episodes in both the winter and summer that result primarily from changes in weather, rather than changes in local source output [39]. Winter smogs are formed through a phenomenon called temperature inversion: when the layer of atmosphere closest to earth becomes colder than the layer above it. Pollutants are then trapped closer to the ground until the weather changes. Spikes in (NO$_x$) are common during these winter events due to continuous diesel combustion [32, 37]. In the summer, nitrogen dioxide (produced from industrial and traffic pollution) reacts with hydrocarbon in sunlight to produce ground-level ozone [33].
High levels of $O_3$ can irritate and inflame the lungs, nose, and throat causing asthma and other respiratory consequences. Generally, ozone forming chemicals can remain in the air for many days and can travel across long-distances. Thus, a reduction in levels requires a much larger coordinated effort across Europe [40].
Chapter 2

Statistical Methods

This chapter provides a high-level overview of the methods used in this project. The primary objective is to build a Gaussian process model that can accurately predict a given pollution concentration across space and time. Then a simpler GP model is used to apply Krause’s optimization algorithm as a test case. Therefore, this chapter walks through Gaussian processes, mutual information (the criterion utilized by the algorithm for selecting sensors), and the optimal experimental design algorithm itself.

2.1 Gaussian Processes

Gaussian processes (GPs) are a very powerful machine learning model. They can be used for both regression and classification problems; in which they utilize Bayesian inference by expressing a prior on the function to be estimated and then allow the user to integrate observed data into a model via the posterior [53].

For a given set of data points, there may be infinitely many functions that describe the underlying relationship of the variables. Instead of immediately ruling out any one functional form, GPs search over all possible functions that are consistent with the observed data and assign probabilities based on the likeliness that they fit the data. The process is similar to all other Bayesian methods: it begins by assuming a prior distribution and updates that distribution as more data points are observed. This training process results in a probability distribution over possible functions where the
mean of the distribution represents the most probable characterization of the data.

Gaussian processes get their name from the distribution on which they are based. The Gaussian (or normal) distribution makes up the fundamental building block of the process. Generally, Gaussian processes are made up of multivariate Gaussian distributions, with each random variable distributed normally and their joint distribution is also Gaussian. This joint distribution is defined by some mean, $\mu(x)$, and covariance matrix, $\Sigma$. While the mean is usually centered around zero (for simplification), the covariance matrix is determined by its covariance function $k$, also known as the kernel of the GP. Let $f(x)$ represent some property of interest that varies over a spatial domain indexed by the coordinates $x$. Then we model $f$ as a Gaussian process

$$f(x) \sim GP(\mu(x), k(x_i, x_j))$$

where the covariance between the property values at any two points, $\text{Cov}(f(x_i), f(x_j)) = k(x_i, x_j)$, and the mean $E[f(x)] = \mu(x)$.

The kernel is where all of the interesting work of the Gaussian process happens, as it ultimately describes the shape and “smoothness” of the distribution. Thus, the kernel is carefully chosen prior to the learning process – based on the domain knowledge of the data set. The covariance matrix is then generated by evaluating each sample point pairwise, $k(x_i, x_j)$, and returning a similarity measure for each pair of points. For further clarification, $\Sigma_{ij} = k(x_i, x_j)$ describes how much influence the points $x_i$ and $x_j$ have on each other. If $x_i$ and $x_j$ are determined by the kernel to be similar, then the output of the function at those points is expected to be similar too.

How does all of this relate to our problem? There are many advantages to using GPs to model the distribution of air pollutants. What makes Gaussian processes particularly beneficial is that they know what they don’t know. Unique among most other machine learning methods, GPs return the uncertainty surrounding their estimations. This is due to the fact that the posterior variance can be calculated for
unobserved locations – a helpful property of GPs [26]. So we can determine the values where the model has high uncertainty. Figure 2-1 displays the posterior mean (a) and variance (b) derived from a Gaussian process model trained on 54 sensors at Intel Labs Berkeley by Krause et al. in 2008 [26]. The two areas in red (in the center of the plot) represent the model’s high level of uncertainty, since that is where there were no sensors placed originally. For our purposes, knowing where the uncertainty is the highest is extremely useful when trying to determine the placement of sensors; especially since we want the capability to collect more information from a new sensor in an area the model is the most unsure about [26].

Another benefit of Gaussian processes is that they are commonly used for representing spatial relationships, which are inherent to geo-statistical problems such as air pollution [53]. Furthermore, GPs are extremely flexible, require relatively few parameters, and are generally easy to use [26].

2.2 Mutual Information Criterion

In a standard experimental design problem such as optimizing the placement of sensors, we need to define what a “good” potential placement is. Once we have a trained GP model, there are a number of criteria of which we could choose to characterize the quality of a placement. One strategy is to greedily place sensors at the points
of the GP’s highest posterior variance (also known as entropy). Yet, Krause et al. explains that this criterion is flawed for our purposes:

\[ E \text{ntropy is an indirect criterion, not considering the prediction quality of the selected placements. The highest entropy set, that is, the sensors that are most uncertain about each other’s measurements, is usually characterized by sensor locations that are as far as possible from each other. Thus, the entropy criterion tends to place sensors along the borders of the area of interest. . . . Since a sensor usually provides information about the area around it, a sensor on the boundary ‘wastes’ sensed information.} \]

Instead, Krause recommends an alternative criterion: mutual information. Initially proposed by Caselton and Zidek [10], this criterion finds recommended placements that are “most informative about unsensed locations”; by directly evaluating “the effect of sensor placements on the posterior uncertainty of the GP” [26]. Unlike the entropy criterion, mutual information is concerned with the entire space that we want to measure, not just the entropy of the possible sensor locations. Practically, this means that the criterion is searching for sensor locations that “most significantly reduce . . . the uncertainty about the estimates in the rest of the space” [26]. In Figure 2-2, Krause et al. demonstrates (with a simple example) that mutual information results in more intuitive sensor placements “since the sensors that are most uncertain about each other should cover the space well” without ‘wasting’ any information collected outside of the bounds [26].

Formally, the placement selection is done by breaking up the space into two discrete sets of locations: a set \( S \) of all possible positions where we can place sensors, and a set \( U \) of positions we would like to measure, but we are unable to place sensors. We can now define the entire space \( V \) as the union of the discrete sets: \( V = S \cup U \). We also define \( A \) as the set of selected locations that we will eventually choose. Then \( X_A \) is the notation for the vector of values of \( f \) at points in \( A \). The goal is to place a set of \( k \) sensors that will give us better predictions at all of the locations we will not be able to measure: \( V \setminus A \). Krause et al. formally define the criterion as:
Figure 2-2: Example of placements chosen using entropy (diamonds) and mutual information (squares) criteria. (Source: Krause et al. 2008 [26])

\[ A^* = \arg\max_{A \subseteq S : |A| = k} H(X_{V \setminus A}) - H(X_{V \setminus A} | X_A), \]

where \( A^* \) is the set that maximally reduces the entropy over the rest of the space \( V \setminus A^* \). Krause notes that the criterion \( H(X_{V \setminus A}) - H(X_{V \setminus A} | X_A) \) is equivalent to finding the set that maximizes the mutual information \( I(X_A ; X_{V \setminus A}) \) between locations \( A \) and the rest of the space \( V \setminus A \).

In their paper, Krause et al. showed that the mutual information criterion outperforms other commonly considered experimental design criteria choices (such as entropy, A-optimal, D-optimal, and E-optimal), “leading to improved prediction accuracies with a reduced number of sensors” [26]. The following section describes their optimal experimental design algorithm utilizing mutual information in greater detail.
2.3 Krause’s Optimal Experimental Design (OED) Algorithm

Krause’s OED algorithm is greedy, meaning that it chooses the sensors one-by-one, picking the next sensor that provides the maximal increase in mutual information. Figure 2-3 shows the greedy sensor placement algorithm, where \( y \) is the next sensor to choose, \( A \cup \{y\} \) denotes the set \( A \cup \{y\} \), and \( \overline{A} \) denotes \( V \setminus (A \cup y) \). The algorithm takes as inputs: (1) the covariance matrix from the trained GP model; (2) the number of sensors to place \( k \); and (3) the coordinates of the entire space broken out in places we can place sensors \( (S) \) and those we cannot but would like to measure \( (U) \). We will further explain the algorithm in greater detail when we apply it in Chapter 5.

**Figure 2-3:** Krause’s approximation algorithm for maximizing mutual information (Source: Krause et al. 2008 [26])

```
\textbf{Input:} Covariance matrix \( \Sigma_{y' y}, k, V = S \cup U \)
\textbf{Output:} Sensor selection \( A \subseteq S \)

\textbf{begin}
\hspace{1cm} A \leftarrow \emptyset;
\hspace{1cm} \textbf{for} j = 1 \textbf{ to } k \textbf{ do}
\hspace{1.5cm} \textbf{for} y \in S \setminus A \textbf{ do}
\hspace{2cm} \delta_y = \frac{\sigma_y^2 - \sum_{y \in A} \Sigma_{y A}^{-1} \Sigma_{y y}}{\sigma_y^2 - \sum_{y \in \overline{A}} \Sigma_{y \overline{A}}^{-1} \Sigma_{y y}};
\hspace{1.5cm} y^* \leftarrow \text{argmax}_{y \in S \setminus A} \delta_y;
\hspace{1cm} A \leftarrow A \cup y^*;
\textbf{end}
```

Mutual information, as defined in [26], is not monotonic in the size of \( A \), meaning that it does not increase forever as you place more sensors. There becomes a point along the curve in which you can place too many sensors and the mutual information will eventually start to decrease. Figure 2-4 displays this phenomenon. Krause argues that since MI is monotonic in the initial part of the curve, we can prove approximate monotonicity, meaning that the algorithm is valid with a small number of placements.

In their paper, Krause et al. show that their greedy algorithm is always within 95% of the optimal solution when picking a relatively small number of sensors [26]. This is important, since if you wanted to select \( k \) sensors out of size \( m \), then the brute-force
Figure 2-4: Mutual information of greedy sets of increasing size. (Source: Krause et al. 2008 [26])

alternative requires enumerating over all $m$-choose-$k$ possible subsets. This approach becomes highly impractical for even moderate values of $m$ and $k$. Figure 2-5 displays the comparison of Krause’s algorithm with the optimal solution – which they were able to derive since their problem was small enough.

In Chapter 5, we explore applying this algorithm to a test case of the London data.
Figure 2-5: Comparison of Krause’s greedy algorithm with the optimal solution on a small problem. Due to complexity of the problem in most cases, we cannot expect to find the optimal solution in polynomial time. Thus, Krause’s greedy algorithm is a good alternative to the optimal solution. The upper bound in this figure was able to be found by Krause et al. due to the small scale of the example problem. (Source: Krause et al. 2008 [26])
Chapter 3

Understanding the Data

This chapter describes the spatio-temporal dataset used in this project. The first section gives a brief overview of the data and how it was collected. The next section outlines how NO\(_x\) emerges as a compelling modeling candidate based on initial data analysis. The following section describes a time series analysis of the NO\(_x\) concentration data that is later used to inform modeling decisions outlined in the following chapter.

3.1 Data Source

The data for this project is procured from the OpenAir R package \[9\] and was created by the Environmental Research Group at King’s College London (KCL). The research group manages the London Air Quality Network (LAQN), which was formed in 1993 to “coordinate and improve air pollution monitoring in London” \[41\]. The monitoring program is distributed throughout London and its boroughs, with each borough funding the monitoring within their jurisdiction \[41\]. The package contains historical air quality measurements spanning 1993–2019 on a range of pollutants at stationary monitoring sites. The sites consist of automatic equipment in fixed cabins, drawing in samples about every 15 minutes \[34\]. There are four site types labeled in the data: Roadside, Suburban, Urban Background, and Industrial \[36\]. Pollutants monitored in the dataset include: carbon monoxide (CO), nitrogen oxides (NO\(_x\) and
NO\textsubscript{2}, sulfur dioxide (SO\textsubscript{2}), ozone (O\textsubscript{3}), and particulate matter (PM\textsubscript{2.5} and PM\textsubscript{10}). Hourly mean concentrations are reported in mass terms (micrograms per cubic meter – μg/m\textsuperscript{3} – for NO\textsubscript{x}, NO\textsubscript{2}, O\textsubscript{3}, SO\textsubscript{2}; milligram per cubic meter – mg/m\textsuperscript{3} – for CO) \cite{9}.

Overall, there are \(\approx 22\) million total air quality observations reported as hourly means across the 409 historical monitoring sites during the 1993–2019 time period. Figure 3-1 displays the spatial location of the sensor sites throughout the Greater London area.

![KCL Sites in Greater London Area](image)

Number of Sites: 409

Figure 3-1: Coverage map of KCL air quality monitoring sites

The next section walks through how a further look into the data reveals NO\textsubscript{x} as a compelling modeling subject.
3.2 NO\textsubscript{x} as a Modeling Candidate

For simplification of our problem, we will focus on one pollutant to base our prediction model and sensor selection problem on. Therefore, we will investigate the raw data to determine the best pollutant candidate to be used as our proxy.

Upon initial analysis of the data, we find that there is a wide disparity in how often the different pollutants are measured across the time series. Figure 3-2 displays the number of sensor sites per year for each pollutant type. There is an evident split in the number of sensors that measure nitrogen oxides (NO\textsubscript{2} and NO) and particulate matter (PM\textsubscript{10}) after the year 2000 compared to the stagnant number of sensors collecting data on other pollutants like carbon monoxide (CO), ozone (O\textsubscript{3}) and sulfur dioxide (SO\textsubscript{2}). Additionally, Table 3.1 shows the high-level summary of the monitoring prevalence of each pollutant.

![Figure 3-2: Plot depicting the number of KCL monitoring sites across the time period for each pollutant](image)

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Total Number of Unique Sites</th>
<th>Number of Total Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO\textsubscript{x}</td>
<td>242</td>
<td>15,155,035</td>
</tr>
<tr>
<td>PM\textsubscript{10}</td>
<td>225</td>
<td>12,730,221</td>
</tr>
<tr>
<td>O\textsubscript{3}</td>
<td>82</td>
<td>5,392,232</td>
</tr>
<tr>
<td>SO\textsubscript{2}</td>
<td>73</td>
<td>4,743,417</td>
</tr>
<tr>
<td>PM\textsubscript{2.5}</td>
<td>68</td>
<td>2,054,504</td>
</tr>
<tr>
<td>CO</td>
<td>66</td>
<td>3,452,825</td>
</tr>
</tbody>
</table>

Table 3.1: High-level summary of pollutant representation in the KCL dataset.
For our goal of building a rigorous prediction model, the more robust the dataset, the better. Therefore, it becomes advantageous to choose a pollutant to model that has good spatial coverage in the dataset. Since there are two pollutants with a noticeable increase in the number of sites, we can easily reject pollutants that were not subject to an increase in monitoring efforts. For instance, NO\textsubscript{x} has an average of 88 sites across the time period versus O\textsubscript{3} that has an average of only 30 sites. When deciding between focusing on PM\textsubscript{10} and NO\textsubscript{x}, it is not necessarily obvious which pollutant would be a better modeling candidate – given that both are considered to be important pollutants that are hazardous to human health. Yet, further investigation into the monitoring of pollutants shows that while there have been great improvements in instruments utilizing electrochemical techniques, it is less straightforward to reliably monitor particulates \[14\]. This is due in part to the fact that real-time portable particulate monitors exhibit significant technical issues “related to calibration and equivalence with other particulate monitoring techniques” \[14\]. Therefore, measurements of ozone (O\textsubscript{3}), nitrogen dioxide (NO\textsubscript{2}), nitric oxide (NO), and carbon monoxide (CO) – all pollutants that are easily measured with electrochemical devices – have been reliable for a longer period of time. This is also confirmed in the OpenAir documentation, that states there have been a “variety of methods used to measure particle mass and changes in their use over time” \[9\]. Thus the PM\textsubscript{10} concentration levels reported in the dataset are a “current best estimate” given the measurement restrictions.

Therefore, due to its direct and indirect effects on human health (outlined in more detail in Chapter 1) and its robust spatial and historical coverage within the dataset, NO\textsubscript{x} emerges as an overall strong candidate as a proxy pollutant for our project. It is the second most monitored pollutant across the time period in both the number of sites and the number of observations. There also appears to be good spatial coverage of the monitoring sites that collect NO\textsubscript{x} data. Figure 3-3 presents the coverage map of NO\textsubscript{x} monitoring sites in the dataset.

Now having determined the pollutant that we will focus our modeling efforts on, we want to further investigate its historical concentration data to note any possible
Figure 3-3: Coverage map of KCL monitoring sites NO$_x$ only, points are sized by the number of years in which data is reported.

Trends. Since our goal is to design a spatio-temporal model, we want to be able to accurately understand the potential changes in NO$_x$ concentration as a function of time. The next section outlines our time series analysis.

### 3.3 Time Series Analysis

We want to investigate any potential trends in the data in order to better understand how NO$_x$ concentration levels may change over time. The decomposed time series analysis allows for us to disentangle any overall trend from both the seasonal and random components of the data. These insights will inform the eventual specification of our Gaussian process model in Chapter 4, where we will aim to represent each trend with a combination of kernels.
To begin, the hourly data is aggregated monthly in order to smooth out any potential outliers at the hourly or daily level. The dataset is further reduced to focus on the years 2005 to 2019 to ensure a stable number of monitoring sites of the pollutant. For instance, prior to 2000 there are as little as 26 NO\textsubscript{x} monitoring sites in the data. After 2005, there are an average of 103 sensors measuring NO\textsubscript{x} per year.

The raw time series data is plotted in Figure 3-4. Since the random fluctuations are roughly constant in size over time, we can deduce that the data can be most likely represented by an additive time series model. The relatively stable repetitive peaks suggest that there may be some seasonality to the concentration spikes. In fact, as reviewed in Chapter I, we know that NO\textsubscript{x} can exhibit spikes – particularly in the winter – due to temperature inversion. In order to confirm this seasonality, we can compute a Fourier transform on the data that returns the dominant frequency of the data: 12 months. This tells us that the average length of time between each successive peak is roughly 12 months. A linear regression run on the monthly data reveals that that the annual seasonality of NO\textsubscript{x} is driven by a spike particularly in the winter months (see Appendix).

![Monthly Average NOx Concentration Levels](image)

Figure 3-4: Raw time series of NO\textsubscript{x} Concentration data, 2005–2019

Figure 3-5 displays the decomposed time series, showing an obvious seasonal trend, repeated yearly, along with a potential underlying linear trend. We are also able to see random fluctuations in the concentration levels across time. Collapsing the data across all years confirms a spike in NO\textsubscript{x} concentration in the winter months (see Figure 3-6).

The next chapter utilizes these data and analysis insights to build a spatio-
temporal Gaussian process model in order to accurately predict NO\textsubscript{x} pollution concentration within the urban environment.
Chapter 4

Predicting Pollution Concentration

This chapter describes how we derive our spatial temporal Gaussian process model. As reviewed in Chapter 2, the key decisions in building the model revolve around the combination of several different kinds of kernels. We then use marginal likelihood to set the hyperparameters of our model, which end up providing a pretty good fit to the data.

4.1 Variable Selection and Data Pre-Processing

Since the goal of our modeling project is to model the NO$_x$ concentration as a function of time and space, several variables are immediately important. We want to represent the spatial relationship as well as the time series trends, so we focus our variable selection on the latitude and longitude of the sites where the concentration is measured, while also including a measure of time. For ease of computation on such a large historical dataset, the concentration is aggregated monthly and the time period is reduced slightly by dropping the earliest five years.

The resulting dataset spans from 2005-2019 with measurements at 208 sensor sites. Each site’s measurements are averaged monthly across the time period, resulting in $\approx 15,000$ data points.
4.2 Selecting the Kernels

As outlined in Chapter 2, the choice of the kernel function is key to Gaussian processes. The kernel evaluates the training data to produce the covariance matrix – which describes the characteristics of the function we want to predict. While there are many different choices for kernels, an additional benefit is that they can always be combined to create specialized kernels. Since the covariance matrix has to be positive semi-definite, then all types of combinations that preserve that feature are allowed, such as multiplication and addition.

![Figure 4-1: Examples of structures expressible by some basic kernels (Source: Duvenaud 2014)](image)

![Figure 4-2: Examples of structures expressible by some combination of kernels (Source: Duvenaud 2014)](image)

Figure 4-1 displays three basic kernels: squared-exponential, periodic, and linear,
as well as potential sample functions generated from the prior; while Figure 4-2 shows examples of possible kernel combinations. For our model, we will pick a handful of kernels to represent the different components of the trends we’ve seen in the data. The following subsections describe the three types of kernels we will use in our spatial-temporal model: squared-exponential kernel (radial basis function kernel), locally periodic kernel, and a white noise kernel.

**Squared-Exponential Kernel (Radial Basis Function Kernel)**

To model the spatial relationship of NO\textsubscript{x} dispersion, we start with a common kernel – the squared-exponential (SE) kernel. Also known as the radial basis function (RBF) kernel, it is considered the de-facto kernel for both GPs and Support Vector Machines due to the fact that it is seen as universal, relatively simple, generally a good approximator. The kernel measures the squared Euclidean distance between the two feature vectors \cite{15}. In general, this translates to the similarity between two points – a value of 1 if the points are the same \((x_i = x_j)\) and a value close to 0 if they are not the same.

Considering our dataset, we will use this kernel to represent both space and time, separately. For the spatial relationship, it is an appropriate choice since we assume that the NO\textsubscript{x} concentration is smooth relative to some defined distance between monitoring sites. Since our features for this kernel are the physical location of sensors (latitude and longitude), we can interpret the kernel output as the measure of how far you have to travel until the sensors differ in their levels of NO\textsubscript{x} pollution measured. Generally, we can expect that the farther apart in distance that the sensors are, the less likely they have similar measures of NO\textsubscript{x} pollution – since NO\textsubscript{x} concentrations fall off the farther away they are from point sources.

When representing the NO\textsubscript{x} concentration data across time, we can assume that there is a similar relationship to space. The month-years that are closer to each other are likely to have similar pollution concentration levels. For instance, there may be a given year where the pollution levels were particularly high, which could persist throughout many months of that year – with the data recording elevated
concentration levels compared to during the same months the previous year. This phenomenon can be seen in the decomposed time series plot in Chapter 3 (the trend line). Additionally, both SE kernels will also capture the effect of local, short-term weather phenomena across both time and space.

According to the literature, the SE kernel is universal and well-behaved (every prior function has infinitely many derivatives). It is also a stationary kernel, meaning that the value only depends on the difference $x_i - x_j$ and implies that the probability of observing a dataset remains the same even if the $x$ values are moved by the same amount \[^{15}\].

$$k_{SE}(x_i, x_j) = \sigma^2 \exp \left( -\frac{(x_i - x_j)^2}{2\ell^2} \right)$$

Another benefit of using the squared-exponential kernel is that it only has two parameters: $\ell$ and $\sigma^2$. The length scale ($\ell$) determines how closely related the function values of two points are. The smaller the length scale, the shorter the distance you have to travel to find similar points. Conversely, the longer the length scale, the farther you have to travel to find correlated points. The signal variance ($\sigma^2$) is a scale factor and determines the average distance the function is away from the mean.

**Locally Periodic Kernel**

Another kernel of interest is the periodic kernel, which allows us to model functions that repeat over a specified range \[^{42}\]. We will use this kernel to represent the seasonality that we witnessed in the data during our time series analysis in Chapter 3.

$$k_{Per}(x_k, x_l) = \sigma^2 \exp \left( -\frac{2\sin^2(\pi|x_k - x_l|/p)}{\ell^2} \right)$$

Similar to the SE kernel, the periodic kernel also has two parameters: $p$ and $\ell$. The period, $p$ represents the distance between repeating peaks of the data – in our case is 12 months, since we found in our time series analysis that the concentrations spikes repeated yearly. The other parameter is the length scale ($\ell$), which is interpreted the
same as in the SE kernel: how far you can extrapolate between points in your data \[15\].

We could leave the kernel as is in our model, but we know that most real world data does not repeat itself exactly. In order to add flexibility to our model, we can create a locally periodic kernel – where the function still repeats, but the shape of the repetitions can change over time \([15]\). We do this by multiplying a squared-exponential kernel to our existing periodic kernel.

\[
k_{\text{LocalPer}}(x_k, x_l) = k_{\text{Per}}(x_k, x_l)k_{\text{SE}}(x_k, x_l) = 
\sigma^2 \exp \left( -\frac{2 \sin^2(\pi |x_k - x_l|/p)}{\ell^2} \right) \exp \left( -\frac{(x_k - x_l)^2}{2\ell^2} \right)
\]

The parameters for this combined kernel are the four parameters we already reviewed – two per kernel (\(p\) and \(\ell\) for the periodic kernel; \(\ell\) and \(\sigma^2\) for the squared-exponential kernel).

**White Noise Kernel**

Lastly, we want to represent any additional noise that are in the measurements that are not correlated spatially or in time, since those aspects should be picked up by our SE kernels. Therefore, we can use a white noise kernel to capture any remaining uncorrelated noise.

\[
k_{\text{White}}(x_k, x_l) = \sigma^2 \delta_{kl}
\]

This kernel only has one variable: \(\sigma^2\) which is the magnitude of the independent noise component.

Now that we have walked through each individual type of kernel we are interested in, the next section outlines how we can combine them to create our final covariance function.
4.3 Final Model Specification

The full covariance kernel can be created by adding together all of the individual kernel components:

\[
k(x, x) = k_1(x_i, x_j) + k_2(x_k, x_l) + k_3(x_k, x_l) + k_4(x_k, x_l)
\]

Where the first kernel is the squared-exponential kernel in space:

\[
k_1(x_i, x_j) = k_{SE}(\text{long}_i, \text{lat}_i; \text{long}_j, \text{lat}_j) = \theta_1^2 \exp\left(-\frac{(x_i - x_j)^2}{2\theta_2^2}\right)
\]

The second kernel is the locally periodic kernel in time:

\[
k_2(x_k, x_l) = k_{LocalPer}(\text{month-year}_k; \text{month-year}_l) = \theta_3^2 \exp\left(-\frac{2\sin^2(\pi|x_k - x_l|/p)}{\theta_4^2}\right) \exp\left(-\frac{(x_k - x_l)^2}{2\theta_5^2}\right)
\]

The third kernel is the squared-exponential kernel in time:

\[
k_3(x_k, x_l) = k_{SE}(\text{month-year}_k; \text{month-year}_l) = \theta_6^2 \exp\left(-\frac{(x_k - x_l)^2}{2\theta_7^2}\right)
\]

The final kernel is the white noise kernel:

\[
k_4(x_k, x_l) = k_{White}(x_k, x_l) = \theta_8^2 \delta_{kl}
\]

In total, there are eight hyperparameters in our final covariance model: \([\theta_1, \ldots, \theta_8]\). In order to fit them, we normalize the data and optimize the marginal likelihood using a conjugate gradient optimizer, which is used to minimize the negative log-likelihood. The final section interprets the hyperparameters which optimize the marginal likelihood.
4.4 Results

After optimizing the marginal likelihood, we can examine the hyperparameters. The spatial effects modeled by the squared-exponential kernel have values of magnitude $\theta_1 = 411.1 \, \mu g/m^3$ and a length scale of $\theta_2 = 0.0000086$ degrees (representing 0.00086% of the space). We can interpret the magnitude as the variability due solely to the spatial dimension of the data. Figure 4-3 displays the predicted versus actual NO\textsubscript{x} concentration levels at each sensor location during a specific month. The model does a pretty good job capturing the spatial differences.

![Figure 4-3: Comparison of the predicted versus actual NO\textsubscript{x} concentration levels during November 2008. The relative similarity suggests that the model does a pretty good job capturing the relative difference among sensor sites.](image)

Looking at the time dimension, the values for the local periodic component are: magnitude $\theta_3 = 159.2 \, \mu g/m^3$, the decay-time away from exact periodicity is $\theta_5 = 1.05$ months, and the smoothness of the periodic component is $\theta_4 = 2.31$ months (relative to the period of one year). Similar to the spatial component, we can interpret the magnitude as the variability we would ascribe to the periodic time scale. The hyperparameter values for the short-term time effects modeled by the squared-exponential kernel are: magnitude $\theta_6 = 160.1 \, \mu g/m^3$ and length scale of $\theta_7 = 2.35$ months. The plots in Figure 4-4 display the time series fit for a few randomly selected sensor locations.

For the noise component, we get the magnitude of the independent noise to be $\theta_8 = 130.4 \, \mu g/m^3$ – meaning the remaining variability we cannot explain with the previous kernels is 130 $\mu g/m^3$. 

49
Figure 4-4: Time series model fit for three randomly sampled sensor locations. Points represent the NO$_x$ concentration across time (month-year). Red-colored points represent the observations used for model training, where black-colored points represent the observations for validation. The blue line represents the model’s prediction and the light blue shaded region represents the 95% confidence interval.

Overall, the model performs quite well on the data: 89% of the validation data falls within the 95% confidence interval and the MSE = 0.18. This modeling exercise displays how a few covariance kernels can infer pollutant concentration throughout both time and space. In general, this is a promising method of modeling pollutant dispersion within an urban environment.
Chapter 5

Selecting Sensors using Mutual Information

This chapter focuses on applying the Krause optimal experimental design (OED) algorithm as reviewed in Chapter 2. We reduce the Gaussian process model from the previous chapter to a squared-exponential kernel and a white noise kernel in order to represent the spatial relationship of the variables. We then use the model’s trained hyperparameters as inputs to the OED algorithm in which we generate a suggested ordering of sensor locations, from most important to least. The chapter concludes with a discussion on the implications of the placement algorithm within this example and highlights potential next steps for future work.

5.1 Training the Gaussian Process Model

In an ideal world, we would use the Gaussian process model trained in the previous chapter to inform the placement of sensors with Krause’s OED algorithm. However, practically adapting the algorithm to our spatial and temporal model in Chapter 4 is outside the intended scope of this thesis. So this section outlines the simpler GP we will be using as an input to the placement algorithm. By “simpler”, we mean that we are representing only the spatial component of the dataset in order to investigate sensor placement with the algorithm.
To hold the time dimension constant, we collapse the data by averaging all of the “winter” month concentrations (defined as September, October, November, December, January, February, and March) for each sensor location. We chose to focus on only the months with elevated NO$_x$ concentrations. Since our data does not include a time dimension, our model is much simpler: we will only use a squared-exponential kernel and a white noise kernel.

\[
    k_1(x_i, x_j) = k_{SE}(\text{long}_i, \text{lat}_i; \text{long}_j, \text{lat}_j) = \theta_1^2 \exp\left(-\frac{(x_i - x_j)^2}{2\theta_2^2}\right)
\]

\[
    k_2(x_k, x_l) = k_{\text{White}}(x_k, x_l) = \theta_3^2 \delta_{kl}
\]

So our full covariance kernel is a sum of the two kernels:

\[
    k(x, x) = k_1(x_i, x_j) + k_2(x_k, x_l)
\]

Similar to in [Chapter 4](#), we fit the data by normalizing the concentrations and then optimize the marginal likelihood using a conjugate gradient optimizer, which is used to minimize the negative log-likelihood. The optimized values are $\theta_1 = 169.2$ $\mu$g/m$^3$ and a length scale of $\theta_2 = 0.0015$ degrees (representing 0.15% of the space). For the noise component, we get the magnitude of the independent noise to be $\theta_3 = 130.8$ $\mu$g/m$^3$ – meaning the remaining variability we cannot explain with the previous kernel is 130 $\mu$g/m$^3$.

5.2 Applying Krause’s OED Algorithm

As reviewed in [Chapter 2](#), the inputs to the OED algorithm include: (1) the covariance matrix from the trained GP model; (2) the number of sensors to place ($k$); and (3) the coordinates of the entire space broken out in places we can place sensors ($S$) and those we cannot but would like to measure ($U$). For our placement exercise, we will limit $S$ to be the set of all 200 existing sensor placements within the London network in our
dataset. $U$ is generated by randomly sampling 1,000 locations within the geographic bounds to represent the points of interest that we would like to measure, but cannot place a sensor. Figure 5-1 displays $V = S + U$ site locations.

We should briefly review the selection algorithm. We touched briefly on the algorithm in Chapter 2 where Figure 2-3 displays the pseudo-code as written in Krause’s paper [26]. We start with no selected locations ($A = \emptyset$), then for each potential sensor location $y$, the algorithm approximates the mutual information $\delta_y$ (in line 1) as the information gained about the entire space $V$ when measuring at $y$. The sensor location with the highest mutual information is selected as the first location in set $A$ and the algorithm is repeated for all remaining possible locations until all $k$ sites are chosen.

As stated above, we ran the selection algorithm where $S =$ all existing NO$_x$ measurement sites in an attempt to determine which sites are the most important in the network. Figure 5-2 displays the first 25 sensors selected by the algorithm. We can see that the algorithm attempts to space out the selection of sensors to equally cover the area of interest, particularly since we distributed $U$ equally across London. The figure also shows the next 25 sensors selected by the algorithm which for the most part, follows the same pattern as the first 25. Figure 5-3 displays the entire selection order for all of the sensor locations.

We can calculate the mutual information $I(A_i; U)$ for the sensor selections using the following formula [25]:

$$I(A_i; U) = \frac{1}{2} \log \left( \frac{\det(\Sigma_{A_i,A_i})\det(\Sigma_{UU})}{\det(\Sigma_{A_i,U})} \right)$$

Where $i$ takes on values [1:200] and represents the placement order of sensors. $A_i$ represents the sensors placed up to the $i$th sensor and $\Sigma$ represents the covariance matrix of the random variables indicated in the subscript. We saw in Chapter 2 that the mutual information should monotonically increase with a small number of placements. We also saw that on a small problem, Krause’s greedy algorithm performed within 95% of the optimal solution. Plotting the calculated mutual information for
Figure 5-1: Set of $S$ locations where we can place sensors (purple squares). Set $V = S + U$, where $U$ (teal triangles) are where we wish to measure but cannot place a sensor, and $S$ (purple squares) are the eligible sensor locations. $S$ and $U$ are disjoint.
Figure 5-2: The first 50 sensor locations selected by the OED algorithm. The above plot displays sensors 1-25 and the bottom plot overlays the next 25 selected sensors.

the selected sensors (in Figure 5-4), we can see that we have a leveling off close to the first 50 sensors. This means that in our example, the network gains the most
information from placing the first 50 sensors and only receives very marginal gains from the next 150 sensors.

The following section expands on these results while addressing the feasibility of real-world monitoring groups deploying this method in the field.
Figure 5-4: Calculated mutual information values for the sensors placed by the OED algorithm.
5.3 Discussion and Next Steps

So, what can this exercise tell us about the nature of sensor placement within an urban area? Krause’s algorithm can be a powerful tool to help resource-strapped agencies and organizations tackle the problem of selecting where to distribute low-cost sensor nodes within urban communities. In our example, we can determine that the largest information gains regarding NO\textsubscript{x} in the existing monitoring network come from the first 50 selected sensor locations – with the rest of the network providing mostly redundant information. Rather than randomly dispersing hundreds of sensor nodes throughout a city, an agency could instead optimally select a much smaller subset of locations in which to collect information.

How feasible is this method for monitoring groups to employ? A great benefit to this approach, is that it is very flexible – thanks to the Gaussian process model at the heart of it. Instead of requiring a complicated pollutant dispersion model, all that is needed is to monitor the environment directly through an existing or temporary observation network. Those observations can then be used to train a GP, which then informs the selection algorithm.

This small exercise can be expanded in a number of ways in order to more easily represent a real-world scenario. First off, as mentioned in the beginning of this chapter, we’ve greatly simplified the GP model used to inform the selection algorithm. A more robust version would use a model similar to the one in Chapter 4 where you would be able to incorporate time series data to have a richer understanding of NO\textsubscript{x} dispersion throughout London. Another change to this example would be expanding the possible sensor locations to more than just the existing network. Additionally, if a monitoring organization wishes to pay particular attention to the effects of air pollution on a specific sub-population, \( U \) could represent geographic areas in which that population is concentrated – which tells the algorithm that we want more information about those locations.

Furthermore, this thesis has been formulated under the assumption of modeling one pollutant: \( \text{NO}_x \). Obviously, in the real-world monitoring groups care about mon-
itoring a suite of pollutants, so you would want to optimize the sensor selection to take multiple pollutants into account. This could be done through training a multi-pollutant GP and use that to inform the sensor selection algorithm.

In their paper, Krause et al. discuss sensor placements with non-constant cost functions [26]. Meaning that in the real world, the cost of placing a sensor may vary depending on the location. A version of their algorithm allows for specifying a total budget $L$ and select $A$ where the total cost $c(A)$ is within the budget. This could be another useful extension of this project.

More broadly, there are still a number of questions regarding urban sensor networks that are yet to be determined. For example, given that low-cost sensors are not of regulatory quality, who will pay to install them and use their data? Will it be regulatory authorities, research-funded bodies, commercial entities, or a mix? What about grassroot citizen-sensing initiatives? How might monitoring initiatives in the developing world (where air pollution is proven to be worse and monitoring budgets tend to be smaller [28]) utilize networks of low-cost sensors? The last chapter aims to address how the data collected by these low-cost networks may be utilized effectively to combat the hazards of urban air pollution.
Chapter 6

Exploring the Political Usefulness of Environmental Data

This chapter aims to explore not only the intended purpose of environmental monitoring and alert systems, but the potential gap between their presumed and actual effects on human behavior. Due to the dearth of high quality studies on the effectiveness of citizen sensing initiatives, we focus on recent evaluations of federal air quality alert programs and their effect on influencing human behavior and health. We then attempt to expand the insights from these evaluations to citizen science initiatives, like low-cost monitoring, while highlighting potential areas of improvement.

In the twenty years since journalist Neil Gross’ prediction regarding the Internet of Things, society has startlingly made significant progress towards bringing the highly connected world he illustrated into reality. The ubiquity of low-cost digital technologies has opened up entirely new ways in which we as citizens can interact with, understand, and exist in our environment. In her book Program Earth, Jennifer Gabrys explains, “it is through collecting data that everything from enhanced participation in environmental issues to changes in policy are hoped to be achieved” [19]. This follows from the belief that data enables more effective modes of action in response to environmental problems. “Data are intertwined with practices, responses

---

[1] Parts of this chapter are extracted from a previous essay I wrote during a course at MIT, which can be found as part of a larger collection here: A City is Not a Computer: The Data plays for a year.
to perceived problems, modes of materializing and evidencing problems, and anticipations of political engagement” [19]. Hence, it is through environmental monitoring and the dissemination of that data to the public that environmental problems are intended to be more readily and effectively addressed.

When considering the potential impact of citizen-sensing initiatives that further democratize air quality data, questions begin to arise regarding what specifically is the intended citizen behavior? Exactly how does arming citizens with more detailed information on air quality enable them to engage differently with their environment? Many of the data-driven initiatives mentioned in the beginning of this thesis lack specific instructions on what they advise citizens to do with the data they provide. And while there are not any obvious studies regarding the effect of open source air quality data on citizen behavior, there exists more information regarding an earlier (and still widely implemented) form of environmental information dissemination: federal air quality alert systems. The following section delves into the empirical effects of providing this information and attempts to apply conclusions to these new citizen-driven initiatives.

6.1 Citizen sensing initiatives through the lens of AQ alerts

In 1968, the EPA created the Air Quality Index (AQI) that is used to report daily air quality levels for cities across the U.S. For each city, the EPA calculates a value according to the levels of the five major types of air pollutants regulated by the Clean Air Act: PM, CO, NO2, O3, and SO2 collected from its monitoring stations. On particularly bad air quality days, the AQI information is disseminated through the media. There is also an option to automatically receive free air quality mobile alerts when high concentrations of dangerous pollutants are predicted in a citizen’s area. Additionally, cities may manage their own city-wide alert systems that along with air quality, cover many different types of alerts including traffic, police activity,
and weather. Many countries have very similar indexes and alert systems including Australia, Canada, China, India, Mexico, Singapore, U.K., and Europe.

<table>
<thead>
<tr>
<th>Air Quality Index</th>
<th>Who Needs to be Concerned?</th>
<th>What Should I Do?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good 0-50</td>
<td>It's a great day to be outside.</td>
<td></td>
</tr>
<tr>
<td>Moderate 51-100</td>
<td>Some people who may be unusually sensitive to particle pollution.</td>
<td>Unusually sensitive people: Consider reducing prolonged or heavy exertion. Watch for symptoms such as coughing or shortness of breath. These are signs to take it easy. Everyone else: It’s a good day to be active outside.</td>
</tr>
<tr>
<td>Unhealthy for Sensitive Groups 101-150</td>
<td>Sensitive groups include people with heart or lung disease, older adults, children and teenagers.</td>
<td>Sensitive groups: Reduce prolonged or heavy exertion. It’s OK to be active outside, but take more breaks and do less intense activities. Watch for symptoms such as coughing or shortness of breath. People with asthma should follow their asthma action plans and keep quick relief medicine handy. If you have heart disease: Symptoms such as palpitations, shortness of breath, or unusual fatigue may indicate a serious problem. If you have any of these, contact your healthcare provider. Everyone else: Reduce prolonged or heavy exertion. Take more breaks during all outdoor activities.</td>
</tr>
<tr>
<td>Unhealthy 151 to 200</td>
<td>Everyone</td>
<td>Sensitive groups: Avoid prolonged or heavy exertion. Move activities indoors or reschedule to a time when air quality is better. Everyone else: Reduce prolonged or heavy exertion. Take more breaks during all outdoor activities.</td>
</tr>
<tr>
<td>Very Unhealthy 201-300</td>
<td>Everyone</td>
<td>Sensitive groups: Avoid all physical activity outdoors. Move activities indoors or reschedule to a time when air quality is better. Everyone else: Avoid prolonged or heavy exertion. Consider moving activities indoors or rescheduling to a time when air quality is better.</td>
</tr>
<tr>
<td>Hazardous 301-500</td>
<td>Everyone</td>
<td>Everyone: Avoid all physical activity outdoors. Sensitive groups: Remain indoors and keep activity levels low. Follow tips for keeping particle levels low indoors.</td>
</tr>
</tbody>
</table>

Figure 6-1: U.S. EPA’s Air Quality Index Guide. The AQI consists of six levels that range from Good to Hazardous (with a corresponding color scheme), depending on the potential health concern. State and local agencies are required to report daily air quality indexes for all U.S. cities with populations over 350,000. (Source: US EPA [2])

These alert systems rely on similar assumptions as the citizen-focused air quality initiatives. Particularly, that by directing citizen attention to immediate environmental phenomena – whether it be through a smartphone app that is constantly monitoring the surrounding pollution or through a federal alert messaging system – citizens will adjust their behavior. This follows other digital information technologies such as electronic meters that modulate energy use or wearable fitness and sleep trackers that capitalize on the increasing trend of soliciting attention in which to bring about behavior change. While probing the nature of attention in digital media, Matthew Wilson draws on the philosopher Bernard Stiegler’s criticisms of these technologies. Stiegler suggests they can be thought of as powerful systems of care. By deciding what to draw citizen’s attention to, “contemporary power technologies no longer mainly aim at disciplining bodies or regulating life-processes, but at control-
ling and modulating consciousness” [66]. While Steigler is generally critical of these technologies and the new “economic system of attention”, his insights can inform the design of such systems. For a more productive take, Wilson references the work of Katherine Hayles: “It is far too simplistic to say that hyper attention represents a cognitive deficit or a decline in cognitive ability among young people. . . . On the contrary, hyper attention can be seen as a positive adaptation that makes young people better suited to live in the information-intensive environments that are becoming ever more pervasive” [66]. It is important then to understand how successful these technologies are at directing attention in order to stimulate some desired behavioral response.

Until recently, little evidence of the effectiveness of city-wide air quality alert systems existed. In 2018, Hong Chen et al. published what they determined was the first study to comprehensively evaluate the effect of air quality alert programs on a wide array of public health outcomes [11]. The study takes place in Toronto, Canada, a region where air pollution levels are on par with other high-income countries. The research team employed a robust regression discontinuity design to study the public health outcomes of Toronto’s city-wide air quality alert system on data from 2003 to 2012. The findings show that though the alerts were related to some reductions in asthma-related emergency room visits on high pollution days, the system as a whole had no effect on mortality or cardiovascular disease. Overall, the researchers claim the alert program “yielded inadequate protection of the public from air pollution”. Since Toronto’s air quality alert program relies solely on information campaigns – and thus emissions reduction is voluntary\(^2\) – this study concludes that “interventions relying on information campaigns alone to encourage exposure avoidance and voluntary emission control could yield little benefit if not accompanied by mandatory actions.”

Additionally, while studies in the U.S. have shown that smog alerts have reduced attendance in public parks in Los Angeles [46] and Atlanta [49], several other U.S. studies fail to link air quality alerts to behavioral change among the public [50, 6, 63].

\(^2\)Rather than additionally requiring government action such as reducing traffic or shutting down factories like in Santiago, Chile or Delhi, India [11]
In fact, there is sufficient evidence that behavior response decreases as successive alerts are issued [67]. A narrow study evaluating the behavioral impacts of air quality ozone alerts in Sydney, Australia on cycling behavior found that bicycling trips decreased by 25 percent in response to smog alerts. However, when alerts were issued over two consecutive days, there was no decrease on the second day. The study found larger impacts on the weekends compared to weekdays, suggesting that if cycling was utilized for commuting, there was less likelihood of substitution [56].

What do these studies suggest for the future applications of citizen-sensing initiatives? The final section aims to draw conclusions from the empirical results in order to highlight how they may be best utilized in regard to citizen behavior.

### 6.2 Bringing behavioral change to bear

In an attempt to shed light on the potential of low-cost sensing initiatives, these evaluations provide possibly enough evidence to temper expectations related to changes in citizen behavior. If we are to apply the results of national air quality alert programs to citizen-driven initiatives, then we can deduce that there is some evidence to suggest that alerts trigger immediate citizen compliance, particularly in regards to more at-risk populations. Though, overall it appears that large potential changes in behavior are less likely to occur, particularly when citizens are asked to make more significant changes that may affect their daily lives, like changing their mode of commuting.

Yet, it is not entirely realistic to expect these new initiatives to function exactly the same as the government alert programs and therefore, the potential impacts could be more effective. With the ability to collect more fine-grained data and give personalized suggestions, low-cost monitoring offers a more nuanced approach to information disclosure. It may be interesting to consider an extreme example, such as Delhi, India. Delhi is consistently ranked by the WHO as one of the world’s worst cities for air pollution [23], with the capital city routinely setting mandatory bans on high polluting industry, construction activities, and heavy vehicles for days at a time when the air pollution reaches extreme levels [22]. Since air pollution is a regular aspect
of daily life to Delhi’s citizens, what does a city-wide alert system offer? In a city
overcome with severe levels of air pollution that last up to months at a time, is it
realistic to assume that citizens would solely adhere to city-wide recommendations
that they stay indoors? At what point does the air pollution become deprioritized
by the citizens who have to get on with living their lives? This may highlight the
potential advantage of low-cost monitoring. Would citizens be more willing to abide
by government suggestions if they were given realistic, catered options in which to
alter their behavior? What about if these systems enabled citizens to actively care
more about their environment?

There is a broader opportunity for these initiatives to embrace the strategy of
directing attention to what is happening in a citizen’s environment. Where sensing
up until this point has focused on the passive monitoring of the environment, these
applications offer the opportunity to take a more active role in bringing about real
behavioral change. As Wilson points out:

The point is to pay attention to attention as an object, to cultivate atten-
tion as care through technological engagements to confront what Stiegler
considers a “systemic carelessness,” or, more profane, where “I don’t give
a fuck” (je-m’en-foutiste) has become a persistent affect toward societal
(human, environmental, cultural) challenges [66].

We can imagine a system in which caring about one’s environment is the ideal be-
behavioral change, specifically when advocacy is a vital component to addressing such
sticky, complicated societal problems as air pollution. Wilson again: “perhaps tech-
nological engagement requires an awareness of the conditions of thought-action, to
better frame interventions with technology by being aware of the tendencies toward
attention craft and control” [66]. In order to bring about real action, these initia-
tives offer the ability to modify existing behavior by exploiting hyper attentiveness.
William Connolly, professor and chair of political science at the Johns Hopkins Uni-
versity, writes “[t]hinking is not merely involved in knowing, explaining, representing,
evaluating, and judging,... To think is to move something. And to modify a pattern
of body/brain connections helps to draw a habit, a disposition to judgment, or a capacity of action into being” [66].

Overall, there exists potential for citizen-based, low-cost monitoring to fill the gap left by city-wide alert programs. With the rise of urbanization increasing the percentage of the world’s population exposed to extreme health consequences of air pollution, networks of sensors and personalized applications allow for promising improvement for tackling our most pressing environmental problems.
Bibliography


[52] PurpleAir. URL: [https://www2.purpleair.com/](https://www2.purpleair.com/) (Accessed: 01-12-2021).


