Title
Towards a societal scale, mobile sensing system

Author
Honicky Jr., Richard Edward

Publication Date
2010

Peer reviewed|Thesis/dissertation

Towards a societal scale, mobile sensing system

by

Richard Edward Honicky Jr.

B.A. (University of Michigan) 1996
M.S. (University of California, Santa Cruz) 2004

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Eric A. Brewer, Chair
Professor John F. Canny
Professor Anna Lee Saxenian

Fall 2010

Towards a societal scale, mobile sensing system

Abstract

Towards a societal scale, mobile sensing system

by

Richard Edward Honicky Jr.
Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Eric A. Brewer, Chair

With the proliferation of sensor equipped smart phones, and augmented reality applications fast appearing, the mobile phone is becoming something much more than a scaled-down, connected IO and processing device. In addition to these standard PC traits, a cell phone is situated in an environment, mobile, and typically co-located with a user. These traits make the cell-phone ideally suited to track and understand the impact that the environment has on individuals, communities and cities, as well as to understand how humans effect their environment.

In this dissertation, I explore the possibility of building a societal-scale, mobile sensing system to monitor pollution and other environmental factors, using low cost-sensors embedded into mobile phones. I will discuss several hardware platforms we used to study mobile sensing over the course of three field campaigns, models of pollution dispersion, sensor characterization and its impact on model parameters, automatic calibration and increasing precision in densely sampled regions.

To my friend, mentor and adviser,

Dr. A. Richard Newton,

who knew how to change the world.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my family, particularly my wonderful children, Nana and Rhythm Porter-Honicky for their boundless faith and patience in me, and my parents, for their unwavering commitment. I would also like to thank Professors Eric Brewer and John Canny and Dean Shankar Sastry for taking me under their wing when Professor Newton passed away. Finally, I made so many wonderful friends in the TIER group and elsewhere at Berkeley, whose support I could not have finished without. Among these, Omar Bakr deserves special mention, for his constant and limitless friendship, loyalty, support and inspiration during my toughest times.

Chapter 1

Introduction

With the proliferation of GSP and accelerometer equipped smart phones, the mobile phone has the potential to become something much more than a scaled-down, connected IO and processing device. In addition to these standard PC traits, a cell phone is situated in an environment, mobile, and typically co-located with a user. Indeed, augmented reality applications are fast appearing on smart phones, already. These traits also make the cell-phone ideally suited to track and understand the impact that the environment has on individuals, communities, cities, as well as understanding how humans effect their environment.

By attaching sensors to GPS-enabled cell phones, we can gather the raw data necessary to begin to understand how, for example, urban air pollution impacts both individuals and communities. While integrating a sensor into a phone and transmitting the data that it gathers to a database is not very difficult, doing so at low cost, on a societal scale, with millions of phones providing data from hundreds of networks spread throughout the world makes the problem much more tricky.

On top of the systems challenges, understanding the raw data gathered from a network of cell-phone-attached sensors presents significant challenges as well. Cell phone users are mobile, are unlikely to calibrate their sensors, typically put their phone in their pocket or handbag (thus obstructing the sensor from airflow), spend significant time indoors or in cars, and typically charge their phone at most once per day, often much less frequently. Even if users did calibrate their sensors, the very low-cost sensors we intend to use drift over time and environmental conditions. Without knowing the location of a sensing event, automatically calibrating the sensors in the phone,

---

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7]

detecting the environment of the phone, and intelligently managing power (by sampling at the right times) the data gathered by the phones will be next to useless.

Integrating sensors into mobile phones, however, has several practical advantages. For many applications, the most significant challenges that face tradition wireless sensor networks are power management and network formation and maintenance. Of these, power management is greatly simplified (since users charge their phones regularly), and network formation is largely solved. Also, a dearth of real-world, practical applications has limited the number of "motes" (wireless sensor network nodes) that get manufactured, and thus the price of a mote remains relatively high. With the number of mobile phones sold in 2010 on track to surpass 1.2 billion [22], cell phones obviously have enormous economies of scale that will be hard to replicate in the near term. Thus the mobile phone platform has several significant advantages as a sensor that will allow relatively simple, and massive deployments.

The economics of mobile phones also provide a unique opportunity for developing countries in particular. Since mobile phones tend to first find markets in the highly industrialized world, and then secondary markets in less industrialized areas (either in the form of used devices or low priced overstock), if devices are manufactured with sensors integrated into them, they are almost certain to find their way to all corners of the globe.

Even today, the low cost of mobile phone-based computing offers the opportunity for scientists in developing regions with modest budgets to deploy sensing in their communities or areas of study. Integrating sensing into mobile phones is increasingly straightforward and commonplace, and an increasing number of examples abound.

The mobility of the phone also provides some important opportunities. At the expense of sampling a given location continuously, a sensor in a user's phone can provide significant geographic coverage. Also, mobile sensors will be heavily biased towards locations in which people congregate, so for human-centric applications, sensing in mobile phones will often provide coverage exactly where it is needed most. In over-sampled locations, the precision of the sensing system can be increased by carefully averaging the readings from several nearby sensors (see Section 7). Also, sensors close to one another can be automatically calibrated, especially if there are some "ground truth" reference sensors also situated in the environment (see Section 6).

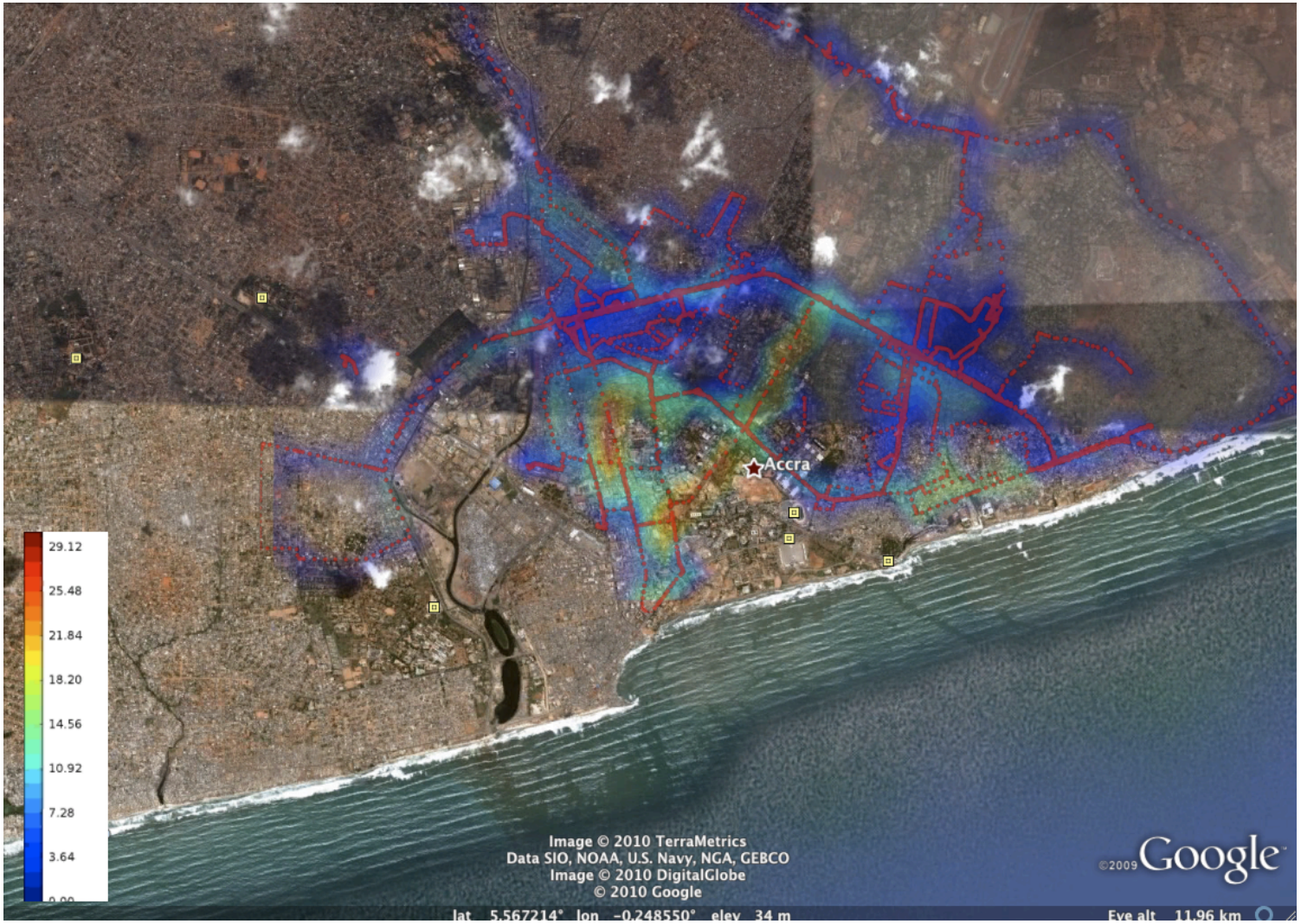


Figure 1.1: Carbon Monoxide data (red points) and interpolation (color heat map) gathered using sensors carried by taxis and students in Ghana, West Africa, on March 22, 2007. Images are overlaid on satellite imagery using Google Earth.

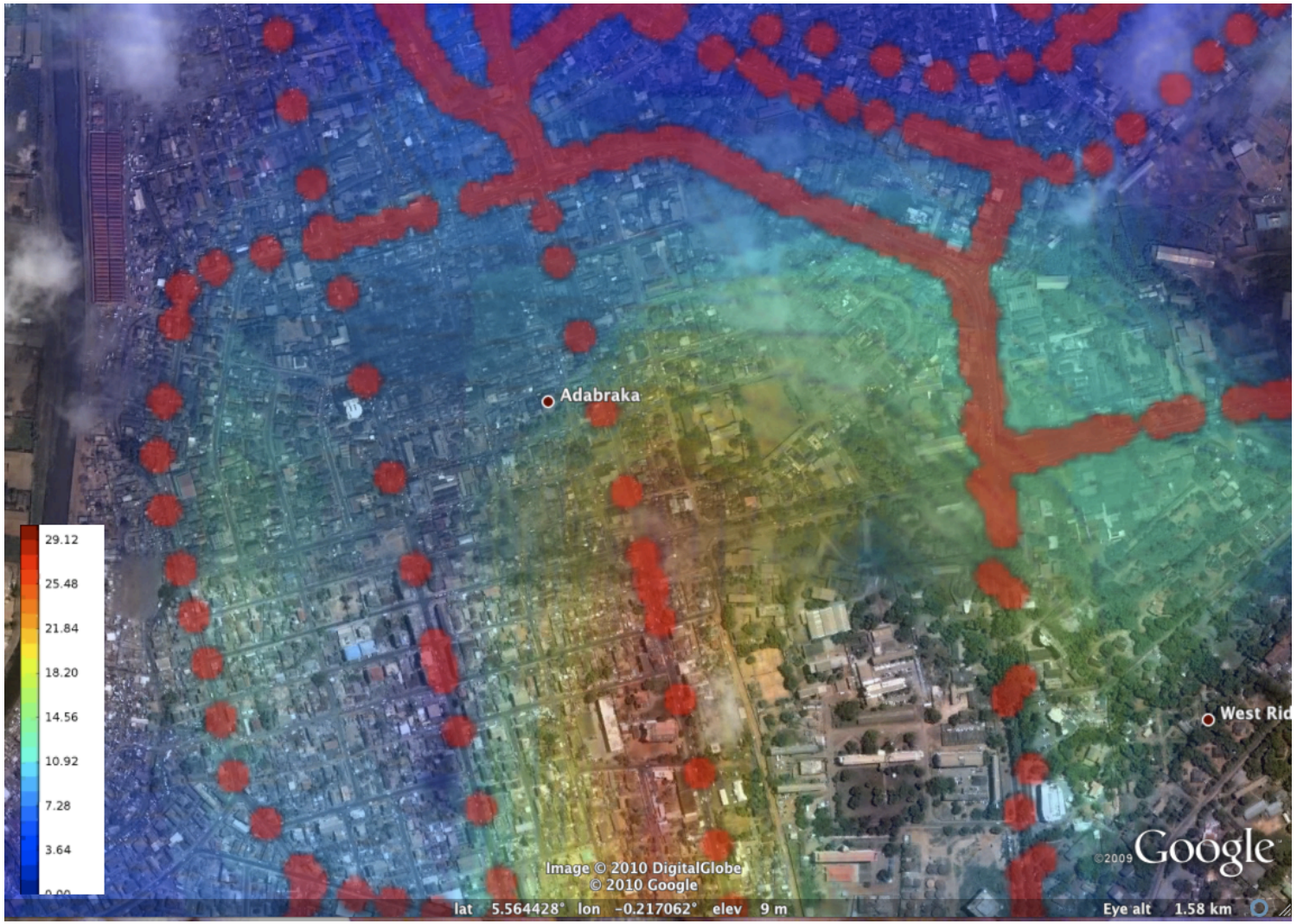


Figure 1.2: A close up view of the same data as in Figure 1.1.

Figure 1.3: The same view data as Figure 1.2, seen without the interpolation data.

All these advantages point to a new opportunity: to build the largest scientific instrument ever built, consisting of millions or billions of sensors, aggregating data on an unprecedented scale. This instrument could be truly societal scale, reaching across economic, social, geographic and political boundaries, and illuminating the corners of human activity, how our environment affects us, and how we affect our environment.

This thesis begins to explore some of the primary challenges to building a societal scale scientific instrument. It is by no means exhaustive. Instead, we discuss our own work in the context of the larger problem of societal scale sensing. We intend to provide a starting point for researcher, developers, activists, scientists, health workers and politicians, to consider how to start collecting data now, how to interpret and manage those data, and to hint at what might be on the horizon.

1.0.1 Active vs. Passive Sensing

In some cases, when using a sensor in a phone, we can imagine the user being explicitly involved in the sensing process, for example when using a micro-array connected to a phone in order to detect a pathogen in a blood sample. In other cases, for example when sensing ambient pollution of a large geographic and temporal area, the user might not even be aware of when the sensor is taking readings.

These two examples illustrate active vs. passive sensing, sometimes referred to as human-in-the-loop and human-out-of-the- loop sensing. Data gathered using active sensing tends to be very focused and specific to the user, whereas data gathered using passive sensing tends to be more global, and large scale. In this dissertation we will focus on passive sensing, since this dissertation focuses on achieving scale.

1.0.2 Users and Stakeholders

When we design a system, it is often useful to identify the users and stakeholders, in order to determine whether a design meet their needs. In this section we look at some of the primary users and stakeholders in a mobile-phone based sensing system.

Figure 1.4: A building housing generators for an ISP in Guinea Bissau. The closest generator was poorly tuned and exhausted into the room, causing dangerous levels of carbon monoxide to accumulate in the work area.

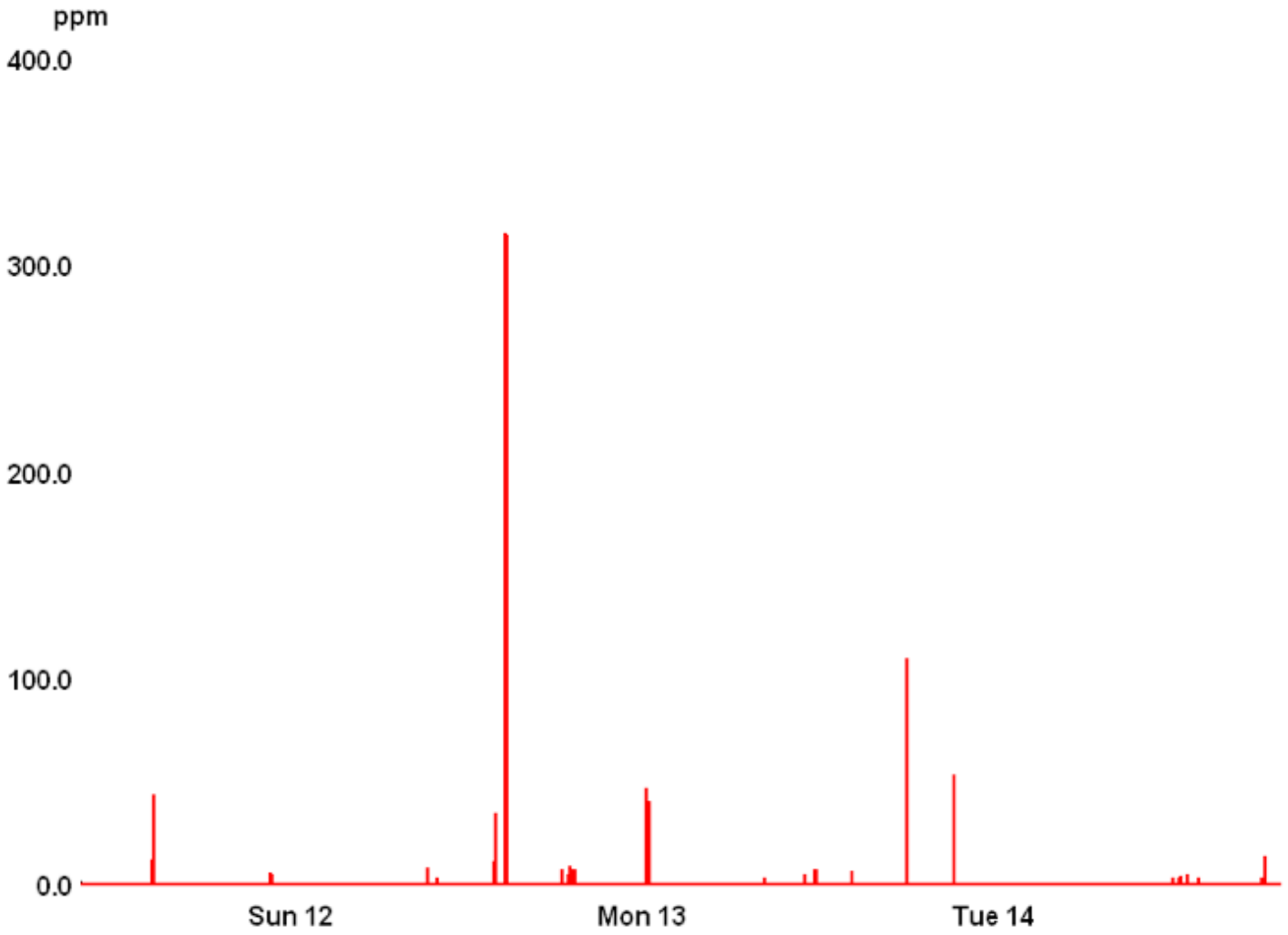


Figure 1.5: A personal log of carbon monoxide exposure during October, 2006. The 300ppm spike corresponds to standing in the doorway of the room depicted in Figure 1.4.

We should note that although some of the stake-holders might use the data we produce for persuasive purposes, our goal in building a societal scale sensing system is explicitly not persuasive. We believe that the underlying information should be available to society, under an open and equal access system, and that the various stake holders can utilize those data for their own purposes. A microscope is intrinsically exploratory, rather than persuasive, and so too should a "mobiscope." On the other hand, those with microscopes have more power in the form of information than those who do not, and so too with information from mobile sensing. Thus our commitment to open access.

Individuals: "What pollution am I being exposed to?"

Personal exposure to pollution has increasingly become a primary health concern. For example, many urban areas publish pollen counts and air quality indicators for asthmatics, and other people with breathing difficulties, because of the direct correlation between atmospheric pollution and asthma related Emergency Room visits. The trend for people with older gas-powered appliances to install Carbon Monoxide sensors in the environment is another example of personal interest in exposure to pollution.

Figure 1.4 and Figure 1.5 illustrate the important impact that personal sensing can have on health and safety. While working in Guinea Bissau at an ISP data-center, one of the authors briefly looked into a generator housing, where a poorly tuned generator was exhausting into the housing. Examining the log from a personal carbon monoxide sensor afterwords, we discovered that the carbon monoxide levels outside the room exceeded 300ppm, dangerously high even outside the room. For the safety of the technicians, the ISP immediately stopped using that particular generator until it could be tuned and properly exhausted. This is a perfect example of how personal monitoring leads to positive action for the individual and the community.

Similarly, in follow-up interviews to the sensing campaign in Ghana (see Section 4.1.1), one user, who had been using the sensing platform to monitor his own exposure, had his car tuned up at his own expense in order to "not be part of the problem." At least in that case, information lead directly to action. We believe this is not anomalous. Many of the users, in fact, indicated that understanding the amount of pollution they were being exposed to increased their level of concern and desire to take action.

A project out of Dartmouth University, BikeNet, aims to provide these type of data to users with heavily instrumented bicycles. BikeNet bicycles use a 802.15.4-based local network to aggregate

Figure 1.6: A sensing device mounted on a street sweeper in San Francisco as part of the Intel CommonSense project.

sensor data in a mobile phone, and then upload the data to a server. These data are analyzed, and can provide information about the "health" of a ride (including pollution exposure), and other computed metrics such as "fitness/performance" and also user-defined metrics such as "enjoyment" [17]. These data provide an important glimpse into what kind of in-context data users might be able to see about themselves, and suggests what users might do with them.

Policy makers and community activists: "What are the societal impacts of atmospheric pollution, and how can we mitigate them?"'

Individual interest in pollution extends, however, into the community, as well. Environmental justice groups are increasingly looking for ways in which to bring primary data to bear in their negotiations and confrontation with polluters and other stakeholders. Indeed, Jason Corburn asserts that a community's "political power hinges in part on its ability to manipulate knowledge and to challenge evidence presented in support of particular policies" [12].

The CommonSense project at Intel aims to provide a technology platform for personal sensing

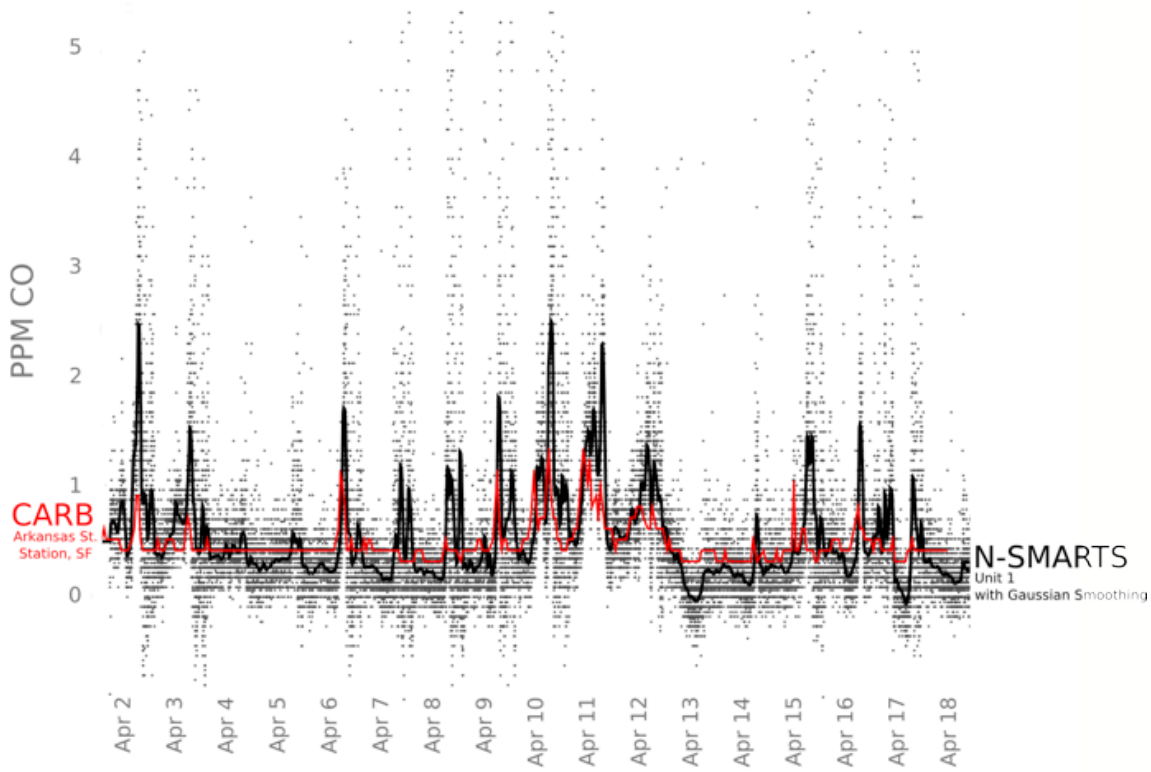


Figure 1.7: Although the CommonSense street sweeper sensor is far less sensitive than the CARB CO sensor on Arkansas St. in San Francisco, it still tracks the ambient pollution levels recorded by the CARB sensor.

with the explicit goal of enabling community action, and supporting local policy-making. The CommonSense project works with community activists in West Oakland, near one of the country's busiest ports [67], and with the City of San Francisco to document pollution throughout the city using sensors mounted on street sweepers [4] (Figure 1.6), some data from which can be seen in Figure 1.7. Figure 1.7 also shows the value measured by the California Air Resource Board's sensitive measurement station at Arkansas St. in San Francisco.

Atmospheric chemists: "How is pollution distributed, what are its sources, and how does it evolve"

Atmospheric chemists study how pollutants and other chemicals evolve in our atmosphere. For pollutants like the nitric oxides (NOX) and ozone (O3), the life cycle of pollutants is very complex (these pollutants change from one to the other in a process that is poorly understood "in the wild"). For other pollutants, like carbon dioxide (CO2), the chemical itself is relatively inert, but the presence of other pollutants can give us a hint as to the source of the CO2. The work of an atmospheric chemist is to understand and classify the different chemicals in our atmosphere, how they get there, and how they inter-relate with one another.

Atmospheric chemists use a wide variety of techniques and sensing mechanisms, from ice-core samples, to satellite imaging. Much of the work done by atmospheric chemists has used relatively high precision, high accuracy instruments, sparsely positioned throughout the environment. For locations not directly observed, chemists use models of dispersion and reaction in the atmosphere to predict the concentration of the chemicals.

Another common mechanism that gets more spatial coverage is to use vehicles instrumented with sensors, and move them through the environment. NASA has a special purpose DC-8 that allots "seats" for sensing equipment (Figure 1.8). The researchers provide an intake mechanism for their sensor that fit into a window frame in the plane and sample the air at high altitude. This mechanism allows researchers to observe the concentration of chemicals at high altitudes over a very large area. Because of the speed of the aircraft, the large area is observed over a relatively short period of time.

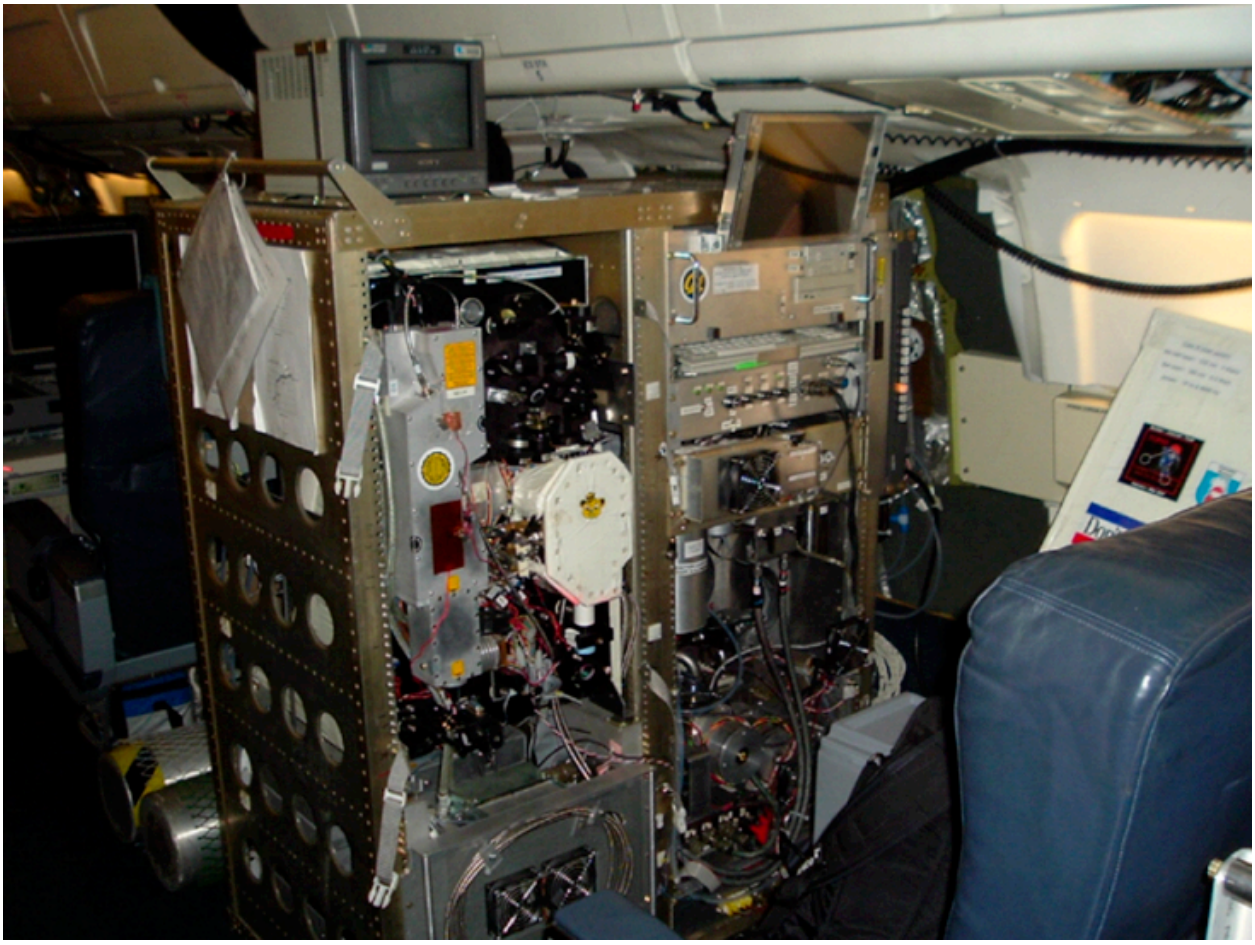


Figure 1.8: An extremely high precision spectroscopic NO2 sensor mounted with an intake out the window of a DC8.

Epidemiologists: "Who is being exposed, at what levels and how does that effect their health?"

At the end of the day, one of our most important concerns is how pollution impacts humans. Mobile sensing has the unique characteristic that it measures the pollution levels exactly at the locations where people are located. Not only does this enable us to track an individual's exposure, but this also means that samples will be concentrated exactly where people are concentrated. So if humans are our primary interest, then mobile sensing, in some sense, provides optimal spatial coverage.

It is also interesting to note that in conjunction with activity inference, an important ongoing research topic (see Section 2.2), we can begin to correlate activities (e.g. bike riding), with pollution exposure. Epidemiologists will also have a tool by which to correlate health outcomes to pollution exposure more directly, assuming the parties are willing to provide epidemiologists with identifying information.

Network operators: "How can we benefit from mobile sensing?"

Network operators, of course, need to be considered. Operators traditionally have taken a tightly controlled approach to the devices, applications and services, and have essentially viewed revenue going to third parties for phone based applications and services as "leaking revenue," particularly in American markets where devices are typically marketed and subsidized by operators.

American operators have grudgingly ceded control of the phone in various ways as smartphones have generated huge revenue jumps for mobile operators and manufactures alike. Even as "unlimited" data plans become more prevalent, thus removing one important obstacle to getting data from the phone into the cloud, however, operators still exert significant control over what devices operate on their networks.

European and other global operators that primarily use GSM based networks, however, don't typically exert as much control over the devices that use their networks, so a hit product in Europe which has built-in sensing might be effective in persuading operators elsewhere to support phones with integrated sensors.

If operators are interested in some of the data that are collected for their own purposes (geographic signal strength information, for example), they might even be persuaded to subsidize sensor devices. Governments might also provide subsidies or legislative encouragement for

operators to participate in building a societal scale sensor network. We discuss one potential opportunity for subsidization in Section 9.7.

Phone manufacturers: "How can we increase margins or revenues using mobile sensing?"

Manufacturers are also key industrial players in building a large- scale sensor network. With razor thin margins on highly competitive and commoditized mobile phones, and a market that moves at a very fast pace, even minor increases to the bill-of-materials or engineering costs of a device must be offset by a clear increase in revenue and profit. This is a particularly difficult challenge for scaling a mobile sensor network.

One approach would be to pitch sensor-phones as a green initiative for the company. With increasing focus on environmental degradation and the accompanying health impacts, a sensor-phone might be seen as a way for a manufacturer to contribute towards improving environmental conditions, and thus as good marketing material.

Another approach would be to sell context awareness as an advanced feature in the top model phones, including pollution, humidity, temperature, noise levels, location, proximity to friends, and many other context-related measures. Considering the increasing importance of context awareness in general, monitoring ones own behavior and environment might be a good selling point for a mobile phone.

These ideas notwithstanding, there are many questions to be answered about how to convince manufacturers and ultimately consumers to bear the increased cost of integrating sensors into phones, regardless of how marginal those costs are.

1.1 The goals of mobile sensing

Sensing in mobile phones has several aims. Of course, the primary aim is to provide data to the users. Part of this is to provide the raw data and/or aggregate data to the stake holders in a way that protects privacy while permitting collaboration and cross referencing as much as possible.

The raw data, however, are difficult to interpret for several reasons, outlined below (see Section 3), and so part of our aim needs to be to provide stakeholders with tools and allow them to interpret the data in a way which reflects the ground truth as much as possible. Our tools must also quantify our confidence about the data and our interpretation.

Finally, sensing in mobile phones, in particular, is unique in that it can amortize the cost of

monitoring our pollution exposure and our environment over millions or billions of people. This is a critical distinction from other wireless sensor network technologies, which might be highly scalable, but lack capital because they are not driven by consumer demand.

Another point which cannot be emphasized enough; pollution monitoring and other sensing in mobile phones is almost optimally capital efficient. That is to say, everything we need to do mobile sensing is in the phone, except the sensors themselves. The marginal cost of adding the sensors themselves is essentially the only cost.

Thus, one of the primary goals of sensing in mobile phones is to amortize cost over millions or billions of people, and to maximize the capital efficiency of deploying sensors.

1.2 The challenges for mobile sensing

Sensing in mobile phones confronts several important challenges. If these challenges are not addressed, we cannot hope to realize the benefits of a societal-scale sensing instrument. Instead, we will be limited to small, expensive deployments.

1.2.1 Energy

Often, in one way or another, challenges in the mobile phone domain boil down to energy. Sensing is no different. As phone processor speeds, network data rates and screen sizes push faster and bigger, the energy and power budgets in mobile phones becomes increasingly constrained. Not only are the increasing energy requirements of the hardware running up against the physical limitations of batteries, but the mobile phone chassis can typically dissipate a maximum of 3 watts of power, limiting the peak performance and peak concurrency for the mobile phone form factor. In these highly constrained circumstances, passive sensing must have a negligible impact on operating time1. 2

This has several impacts on our sensing protocol. First, and most obviously, the act of sensing itself must take minimal energy. This includes operating the sensors as well as the sampling mechanism (the mechanism for exposing the sensors to the atmosphere in a controlled way). Secondly, we must be able to wake up at the [1] appropriate time, take a sample and go back to sleep quickly. This means that we will need to carefully control long start-up times to limit energy

[1] "Operating time" is the time a device can operate before needing to recharge its battery. The term "battery life" is often used in this way, but actually refers to the time until a battery can no longer effectively hold a charge.

consumption, if they are necessary. Finally, following directly from the second point, we must be able to sample our location quickly: the long signal acquisition times for GPS are unacceptably expensive because they require significant processing power that will limit the reasonable duty cycle of our sensors to an order of once per hour or less. Fortunately, there is an elegant solution to this problem that we will discuss in detail in Section 3.5.1.

1.2.2 Cost

Of course, the cost of the sensors must also be low enough to have a minimal impact on the overall cost of the device. This must include any circuitry required to run the sensor, as well as the sensor itself. Fortunately, any sensing mechanism with a short bill of materials is likely to be affordable when manufactured in units of millions, assuming that the manufacturing process can be scaled that large. This includes almost any semi-conductor based technology, so we believe that mobile phones integrating atmospheric sensors will probably use semi-conductor based sensors. Since this is a relatively simple challenge that can most likely be addressed by large scale manufacturing, we will not discuss cost again, except in the context of other discussions.

1.2.3 Physical space

The mechanical design of a phone has an enormous impact on its desirability, and therefore a direct impact on the profitability of a company. The size and weight of a phone will have a direct impact on a phone's desirability. A practical implementation of atmospheric sensing using mobile phones must therefore have a negligible impact on the devices' form factor. This means that the sensor itself, as well as the sampling mechanism, must be extremely small and light-weight. We will discuss the size of different sensor technologies in Chapter 3.

1.2.4 Calibration

All sensors require calibration, and cheap sensors (which are the only viable option for a mass produced sensor-phone) usually require relatively frequent calibration (e.g. every few weeks or months). Traditionally, calibration happens by exposing a sensor to at least two (for linear-response sensors) known concentrations of a substance. The sensor's response to these substances allow us to determine the bias (the sensor's response to clean air), and gain (the change in the response of the sensor as the concentration changes).

These factors are complicated by the fact that sensors are sometimes "cross-sensitive" to other chemicals, the humidity, temperature, air-flow rate and air-pressure of the environment, and that sensors may not respond linearly to changes in concentration. When a sensor is sensitive to one or more of these factors, calibration usually requires that the sensor be exposed to the cross product of these factors: e.g. we must record the response of the sensor at different humidities, temperatures, etc.

This type of calibration must be done in a highly controlled setting (e.g. a chemistry lab), by a trained specialist. It is not reasonable to expect millions of users to do this on their own or incur expense or effort to calibrate their sensors. Rather, we must figure out how the system can figure out the calibration of the sensors automatically, without user intervention. We will discuss calibration in Chapter 7.

1.2.5 Precision

Another consequence of using cheap, mass produced sensors is that they typically have low precision. On the other hand, in many popular locations, we will have highly dense sampling: lots of people will be in the same place, at the same time. If we are smart, we will be able to increase the precision of our system by averaging readings in an appropriate way. We will discuss increasing the precision of the system by super-sampling in Chapter 8.

1.2.6 Security and Privacy

Whenever information about people gets collected, security and privacy immediately become an important issue. If we want to convince users to provide us with their data en mass, they must trust that we will use the data for its intended purpose, and not to exploit them. They must also trust that we will handle their data in such a way that it won't inadvertently or intentionally fall into the hands of people who will try to exploit them. We will need to prove to users that their data will be properly handled, this thesis does not security and privacy, but we offer a brief survey of the state of the art in Chapter 2.

1.3 Thesis statement and outline

With the considerations above in mind, this thesis will argue that we can and should build a high-precision, high-accuracy, societal scale atmospheric sensing device using sensors integrated

into mobile phones. We will define precision and accuracy more rigorously in Chapter 6.

By societal-scale, we mean on a scale suitable to impact society at large, and also on a scale that includes and encompasses much of society. While we will not investigate further the extent to which mobile phones in general, and perhaps sensor-phones in particular, will penetrate the "far reaches" of our global society, we do know that there were roughly 4.6 billion mobile phones in use at [61] (corresponding to 67% of the population), and many of these are in the hands of the economically poorest among us. The distribution of mobile phones throughout the world is hardly uniform, but we are optimistic that integrating sensors into phones on a large scale would have an impact on most people in the world.

After surveying related work and background information in Chapter 2 and relevant sensor technologies in Chapter 3, we will describe the hardware platforms that we have used to gather data (Chapter 4) and the models we use to describe and analyze those data (Chapter 5). Next we will do a detailed characterization of the sensors that we use in the study in Chapter 6, partly to aid in tuning and validating the parameters of our models, and partly as an example of the process of designing a mobile sensing system.

We then move to the two main areas of focus: maintaining calibration in a societal scale sensing system (Chapter 7) e.g. maintaining accuracy, and increasing precision (Chapter 8). Finally, we will describe the many interesting questions that our research, thus far, has raised, and the ways in which we are considering tackling them (Chapter 9). We will conclude with a brief analysis and synopsis of our thesis. It is our sincere hope that our work may contribute to a cleaner and healthier society.

Chapter 2

Related work

2.1 Primary related work

The MetroSense project out of Dartmouth University and Columbia University bears a great deal of similarity to our work, in that they also focus on opportunistic sensing on a large scale [15]. The previously mentioned BikeNet project explores mobile sensing on bicycles using a personal area network (PAN) of sensors affixed to bicycles, and a mobile phone to upload data to a database [17]. Calibree is a system for automatic calibration of sensors during opportunistic rendezvous [39], and will be discussed further in Chapter 7. Halo is a system for managing rendezvous [16], and important consideration in opportunistic sensing. We provide some analysis of rendezvous in 8, although our system has no need to explicitly manage rendezvous, since data aggregation is done at the server. Finally, Quinet is a mechanism for utilizing sensors of devices in close physical proximity to another device, in order to increase the amount of information a user or system has about their environment [15]. This bears some resemblance to our super-sampling mechanism, in that collaboration between devices is used to increase the fidelity of the system. In the case of Quinet, however, information between sensors is not combined to increase precision.

The CENS project out of UCLA has several mobile sensing related projects, including the Personal Environmental Impact Report (PIER) project, in which location traces of users are correlated with the pollution footprint of various modes of transportation in order to help a person

---

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7]

reduce their environmental impact [41]. PIER shares the common goal of providing information to the user to enable them to make responsible and healthy decisions.

Various other mobile sensing projects (e.g. [2, 30]) focus on automotive based sensing to monitor various aspects of society, including traffic (e.g. [26, 60]) and road surface conditions [18].

The CommonSense project at the Intel Research lab at Berkeley focuses on the social dynamics of participatory sensing [3, 67]. The CommonSense project provided significant data for this dissertation, including the San Francisco Street Sweeper data set [4], and the West Oakland data set [13]. Ethics in personal mobile and participatory sensing is also a continuing area of research [53].

There are also various other mobile sensing projects that focus on hardware and sensing devices. For example, Jing Li at NASA interfaced a solid state chemical sensor with the iPhone [38] and Wuest et al. mention a LLNL project to develop a low cost (around 1000USD) radiation sensor integrated with a mobile phone [68] for detecting and responding to WMD threats. These projects have not yet studied their sensors in large deployments, however.

Our collaborators at Berkeley are developing a MEMS aerosol pollution sensor and microfluidic sampling mechanisms suitable for integration into a mobile phone [45]. We discuss the MEMS device in more detail in Section 3.4.

2.2 Inferring Activity

An important way in which we can add value and better understand the data we collect with sensor phones is by labeling the samples from sensor according to the activities and context in which the user takes the samples. By tagging samples as having been taken indoors or out-of-doors, while walking, running, riding a bike or driving, etc., we can answer questions about peoples exposure to pollution during these various activities and contexts. These labels will also assist in determining, for example, if a sample should be included in a map of outdoor air pollution.

There exists a large body of research on techniques and applications of mobile context and activity inference. Many researchers have examined utilizing accelerometers positioned in various locations on the body [5], and using a wide variety of inference algorithms [37, 48]. Studies have also focused on integrating data from other sensors such as audio and barometric pressure [20], or location [64] to reduce the inference algorithm's reliance on sampling acceleration from multiple points on the body in order to give high precision, recall and accuracy. Accuracy rates are now

typically reported at above 90%.

As mobile phones have increasingly integrated accelerometers and other sensors, the focus of activity inference research has shifted towards inferring activity and context using phones or phone-like devices [20, 64, 49, 55]. Transportation mode inference, in which an algorithm typically distinguishes between some subset of walking, biking [49, 55, 62] , running, riding a bus, and driving, has become a common focus. Fortunately, these modalities are relatively easy to distinguish between, using acceleration data augmented with other sensor date, particularly location.

Energy consumption during activity inference is an important consideration, since some algorithms require high frequency sampling in order to be effective. Sohn et al. examined using changes in GSM radio finger-prints to determine velocity, rather than energy-intensive GPS signal-tracking. Energy was the explicit focus for Wang et al. [63], who found that by turning off sensors when they are not needed for a particular inference, they could extend the operating time of a Nokia N95 by up to almost 3 times versus leaving the inference sensors operating all of the time without sacrificing inference quality.

Finally, we should note that there are already many applications deployed that take advantage of activity and context inference in one respect or another. The Personal Environmental Impact Report uses location traces and transportation mode inference to give the user a report about the impact of their transportation choices [1]. Several health-monitoring applications use transportation mode as proxy for the level of activity of the user. For example, Ambulation helps monitor the activity level of elderly and disabled people [50], and UbiFit Garden is a persuasive application that encourages people to monitor and improve their activity level [11]. Finally, SenSay is one example of an application that tries to adjust settings like ringer volume on the phone based on the inferred context of the user.

The quantity of studies, application and research in the area of activity inference, along with the generally high accuracy of activity inference algorithms suggests that we can reasonably expect to label sensor readings with the users' activities.

2.3 Security and privacy

Obviously, if we are asking users to submit data to the cloud for analysis, we need to ensure that users' data will not be used in an unauthorized or undesirable way. If users do not have confidence that their data will be used appropriately, they will not submit it. This is particularly

true because we are asking them to submit location information: misuse or mishandling of location information can result in severe political, criminal and social consequences for users. For example, a government might arrest or otherwise intimidate everyone who was at a opposition political rally or protest, based on their location histories. Kapadia, Kotz and Triandopoulos provide a good overview of privacy issues in opportunistic sensing systems [31].

Besides the appropriate use of encryption and other standard information security practices, the dominant mechanism for ensuring privacy of users who publicly reveal information about themselves is called k-anonymity. k-anonymity is a mechanism for ensuring that at least k other individuals have the same publicly available identifiers [57]. The original work focuses on identifiers such as birth date, gender and zip-code, but it has been extended to location histories as well by many researchers [23].

The concept of k-anonymity is both important for plausible- deniability, in which a person can plausibly claim that a location history does not belong to them, and also for foiling tracking attempts.

The k-anonymity mechanism has been extended in various ways in order to provide a more practical privacy mechanism. Xu and Cai propose a mechanism for setting k to a level that corresponds to the amount of anonymity a person requires by having them chooses locations in which they would feel sufficiently anonymous, and then inferring k based on density of people in those locations.

Hoh, et al. use "virtual trip lines," in which traversals across pre- determined, non-sensitive locations are reported in lieu of location traces, for privacy enhanced traffic flow monitoring. This mechanism also allows the traversals to be aggregated semi- anonymously before reporting to ensure that at least k traversals are reported from the same location simultaneously, providing k-anonymity. The authors also increase privacy by distributing the aggregation process across two servers which would force attackers to compromise multiple systems in order to access sensitive information [26].

Although traffic flow is a linear process, whereas mobile sensing deals with unconstrained movement, the virtual trip line concept could be extended to regions of interest, where the aggregating servers would be informed when a device entered that region of interest.

Another important privacy protecting sensing system is AnonySense [32]. AnonySense provides an architecture for implementing privacy aware collaborative sensing applications, focused on real-time collaboration. AnonySense also uses k- anonymity as one of its privacy preserving techniques. AnonySense provides an interesting mechanism by which sensing tasks,

which can require various levels of identification from the sensor, can be submitted to the mobile device. The mobile agent then chooses whether to accept the task or not, based on its own privacy requirements.

Researchers have made significant progress towards ensuring privacy of data by providing a degree of intrinsic anonymity in the way that location information is transmitted and aggregated, these mechanism all rely on a degree of trust in the infrastructure. Homomorphic encryption schemes allow un-trusted entities to perform arithmetic operations on encrypted data without access to the unencrypted data. Thus the data can be aggregated by un- trusted entities before the aggregate values are decrypted by a trusted entity, and can be made resilient to collusion by a subset smaller than a quorum. Cramer and Damgrd provide a good overview [9]. These mechanism have been proposed for resource constrained wireless sensor networks, and would be directly applicable [8]. This brings an additional degree of security to the data collection system.

Chapter 3

Sensor technologies

In thus chapter, we briefly discuss the main types of atmospheric sensors that we might consider for integration into a phone. We will discuss electro-chemical sensors, spectroscopic sensors, silicon and metal oxide sensors, and MEMS devices. There are also several nano-technology based sensors that may be feasible in the future, but which of these will pan out is still not clear, and is out of the scope of this dissertation. Rather, we focus on technologies that are already commercially available, or are close to commercialization.

3.1 Electro-chemical

Perhaps the easiest to work with of the currently available technologies, electro-chemical sensors typically function by creating an electric current from a chemical reaction with the substance being sensed. They are often a type of fuel cell (similar to the high power hydrogen fuel cells being researched to replace batteries) that uses an electrolyte and catalyst specific to the substance being sensed. The sensor typically generates a very minuscule electric current (on the order of a few tens or hundreds of nano-amps), which can then be measured with sensitive electronics [54].

Electro-chemical sensors are reasonably well understood, and are commercially manufactured in a variety of shapes and sizes. Because the principle of operation relies on measuring an electric current generated by a chemical reaction, the more reaction that takes place for a given

---

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7]

concentration, the more sensitive the sensors can be. Since the current generated by the chemical reaction is so small, the noise and drift in the electronics measuring the current can impact the precision and accuracy of the measurements.

The consequence of this is that the larger the sensor, the more precise and accurate it can be. This obviously must be balanced with the size constraints in a mobile device. Unfortunately, the smallest practical sensors are typically around 1 cubic centimeter, which is rather large to integrate into a phone.

On the other hand, electro chemical sensors typically have a linear response over a wide range, are relatively sensitive, and require very little power to operate, since they are self powering. The device only needs to use power to amplify the signal from the sensor and actually take the measurement, and this circuitry can typically start up in a matter of microseconds, allowing for very fast sampling.

Although they are sensitive to extreme humidity, since the fuel cell can dry out or saturate with water (between 20% and 95% relative humidity is a typical operating range), electro-chemical sensors typically reasonably insensitive to humidity in the typical humidity range of air (about 25% in the desert to 100% during rainfall). They also can be somewhat sensitive to temperature, but this can be calibrated reasonably easily, and their response to temperature tends to be consistent. Similarly with air pressure.

These characteristic make electro-chemical sensors a very attractive option for researching mobile sensing, but an unlikely candidate for ultimate integration into a mobile phone.

3.2 Spectroscopic

Spectroscopic sensors typically operate by observing which frequencies of light are absorbed by a particular gas in a controlled chamber. Since different types of molecules have different absorption bands (e.g. the molecule resonates at particular frequencies, thus absorbing a lot of the energy of at that frequency), we can use a laser tuned to one or more of the absorption bands specific to a particular gas, and observer how much light in the absorption band transmits through the gas. The higher the concentration of the absorbing gas, the less light that gets transmitted.

Additionally, some spectroscopic sensors (sometimes denoted fluoroscopic sensors) operate by exciting a gas using a laser tuned to one or more of the gas' absorption bands. The gas then emits a photon at a lower wavelength and this fluorescent band of radiation can be detected, possibly in conjunction with the absorption band [27].

Spectroscopic sensors can be extremely accurate, and have the advantage that since they do not rely on a chemical reaction, they work for relatively inert compounds, such as CO2. While some spectroscopic sensors are available off the shelf, they are also a popular choice for high precision sensing, since their principle of operation is relatively simple and they can be made extremely precise when precise components are used. Figure 1.8 shows one such sensor, mounted in a DC8 for high altitude sensing of very small concentrations of NO2.

On the other hand, while spectroscopic sensors can be made somewhat compact by using mirrors to increase the distance over which the laser travels, the size is still constrained to be large enough to permit the laser to travel over sufficient distance to excite a sufficient amount of the measured gas. These sensors are typically a minimum of a few centimeters long.

3.3 Metal oxide/silicon

Another common technology for portable sensing is metal oxide sensors. In this technology, a thin film of a metal oxide that is sensitive to a particular gas is deposited on a silicon substrate. While there are many ways in which to construct a sensor using a metal oxide, a common one is to design the sensor in such a way that the resistance of the sensor changes as it is exposed to a particular the gas. The change is resistance can be measured and translated to a change in concentration. This typically works best for reasonably reactive gasses such as NOx, or Ozone, and less well for relatively inert gasses such as CO2.

This method typically requires that the surface of the metal oxide be heated to a high temperature (say, 500C), so it requires a lot of power to perform a measurement. Fortunately, the heater can be pulsed on an off quickly, and the way that the resistance changes as the temperature changes can give more information about the gas that is causing the change of resistance. Gasses to which the sensor is cross-sensitive may respond in different ways to the change in temperature than the primary gas [19].

Metal oxide sensors tend to be very sensitive to other environmental factors such as pressure, and flow rate (since flow rate can affect the surface temperature of the sensor), and even humidity.

Metal oxide sensors are available of-the-shelf, and are very small, but require extensive calibration. Accuracy as an absolute reference of gas concentration is difficult to attain, so they tend to be used in application where it is more important to detect a relative change in concentration. One such application is in automotive emission control.

One project in Berkeley affiliated with our research is exploring how to attain reasonable

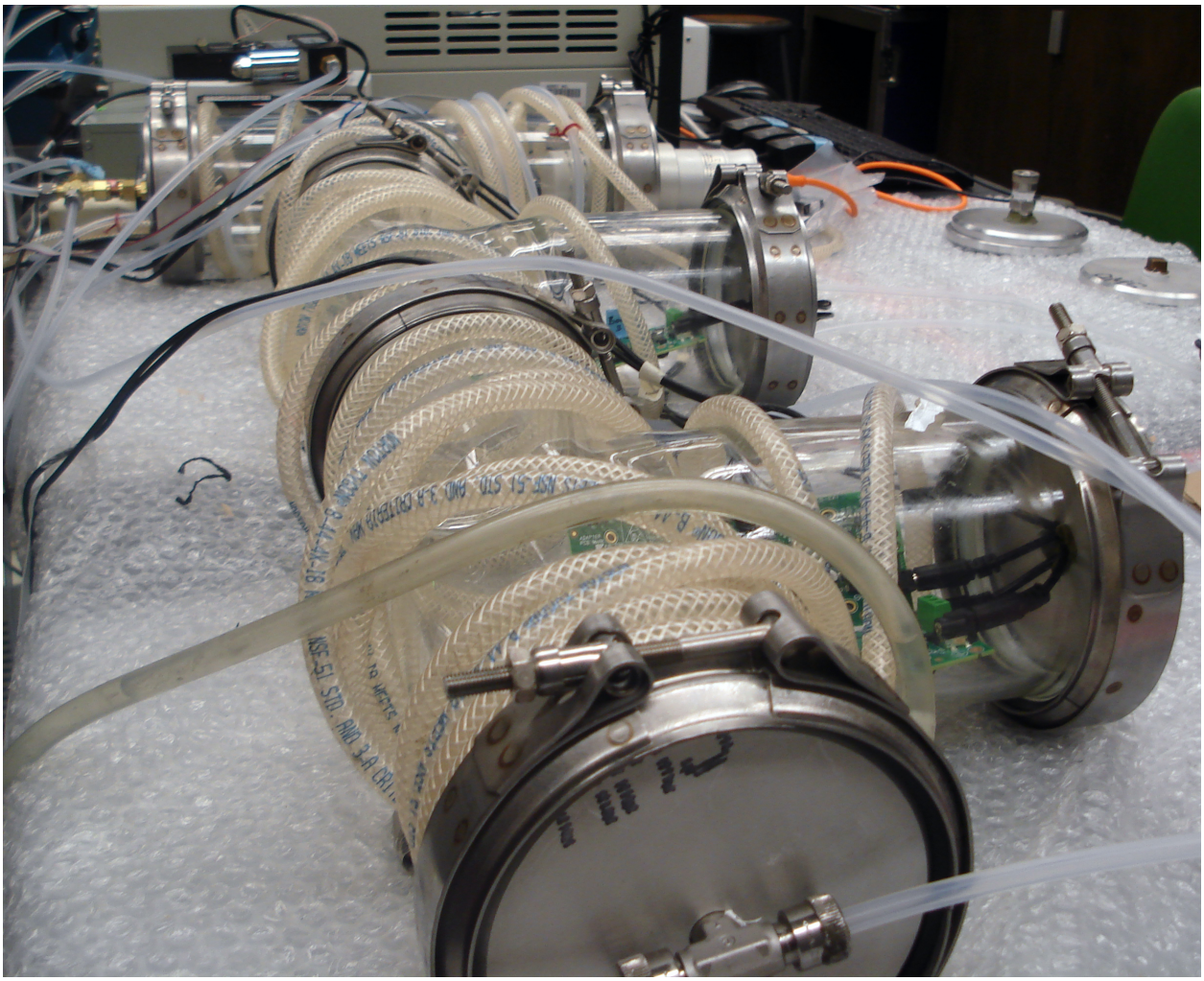


Figure 3.1: Another setup in our lab has a large chamber that allows us to test multiple large sensors simultaneously (Photo: Virginia Teige).

accuracy using very small, low-cost metal-oxide sensors. In this project, the researchers use a chamber in which they can carefully control gas concentrations, humidity and flow rate (see Figure 3.1). They are examining how to exploit temperature measurements of the surface of the metal oxide sensors using a precise infrared sensor to detect and respond to changes of surface temperature, in order to reduce sensitivity to flow rate and ambient temperature. These sensors will be cross calibrated against humidity, air pressure and a variety of gasses to try to develop a clear picture of the response characteristics of the sensor.

3.4 MEMS

Micro Electro-Mechanical Systems or MEMS-based sensors mechanical devices built using photo-lithography, the same technology used to build microchips. MEMS-based sensors share the advantage of extremely small size with other silicon/photo-lithography based approaches. There are many MEMS based sensor out there, and we will not try to survey the field, but rather discuss one technology affiliated with this research.

In this technology, a thin-film, bulk-acoustic resonator, or FBAR, is used to measure the concentration of particulate matter (PM or aerosol pollution) in the air. This is a particularly important application, since existing technologies for measuring aerosol pollution are very large, typically on the order of tens or hundreds of cubic inches. Manufacturing FBARs is a well understood process, because FBARs make good notch filters for high frequency radios, and so they commonly appear in mobile phones. Thus the manufacturing process has already been well tuned for high volume and low cost.

An FBAR resonates at a very high frequency, say 1.6 GHz. If we can cause particles in the air to deposit on the FBAR, the weight of the resonator will change slightly, thus causing a change in its resonance frequency. We can measure that change of frequency and translate it into a concentration of aerosol pollution [6].

One way to cause a particle to deposit on the FBAR is by inducing a thermal gradient over the FBAR using a heater. This technique is called thermal-phoresis, and is relative insensitive to the composition of the aerosol pollution. Other deposition methods (e.g. electro-phoresis) have greater sensitivity to the composition of the particles, but lower power, and thus might be used to create a lower-power sensor that is sensitive to only certain types of aerosol-pollution (e.g. pollen) [25].

In order to create a PM sensor specific to particles of a particular size (e.g. smaller than 2.5

microns), we can use various filtration mechanisms. One mechanism is inertial impaction, in which larger particles have greater inertia with respect to the viscosity of air, and thus follow a different trajectory than smaller particles. By carefully controlling the flow rate of the air, we can select for a particular size of particles (see Figure 4.11) [45].

A controlled flow-rate, however, means a relatively high power air pump, which will have an impact on the total power of the sensor. Optimizing the power of the deposition and airflow mechanisms are therefore a subject of ongoing research.

3.5 Energy considerations

When optimizing a system for energy consumption, we can either decrease the power used by the system, decrease the time that the system uses power, or both. The three-watt limit on power dissipation by a phone chassis not withstanding, instantaneous power is not a central issue for sensing in mobile devices, since sensors that could reasonably be integrated into a phone typically require orders of magnitude less power to operate. Nonetheless, if a sensor's power draw is reasonably high, then it needs to be operated very briefly in order to remain within an acceptable energy budget.

Sensors such as the metal oxide sensors mentioned in Section 3.3 fall into this latter category: since metal oxide sensors require a high power heater to operate, the heater should be pulsed just briefly enough for the temperature on the surface of the sensor to reach its operating temperature. Even at 70 mW, a 500 ms pulse would use about 2.6uAh, or 3.8mAh per day if we sample once per minute, 24 hours per day. Since batteries on phones typically have a capacity of 800-2000mAh, sampling the metal oxide sensor at a reasonably aggressive duty cycle would reduce the operating time by about 0.5%, or about 1.2 minutes for a phone with 4 hours of talk time. Of course, if we were smart about when we sampled, we would reduce that even further.

So reducing sample acquisition time is a key component of realistic energy consumption. One problematic sensor is the GPS. For a GPS to determine its position, it must sample the radio frequency at which the GPS satellites transmit (quick and low power), and then perform a search over a large parameter-space to lock onto the satellite signals (slow and high power). In order to sample quickly and reduce energy requirements for localization, the GPS can simply record the signal it receives from the GPS satellites, possibly pre-process them minimally, store them, and upload the recorded signal with the data samples. These data can then be processed in the cloud, and samples can be localized at minimal energy cost on the phone [33].

Of course, there is an energy cost to transmitting raw signal data as opposed to concise position information, so we must be careful to transmit those data at an appropriate time. For example, GPS signal data, or any other sensor data that is not time critical, can be stored until the phone is plugged into a charger, at which time, transmission energy is "free." If we have the flexibility to time shift in this way, then we can save significant energy. We believe that pollution data will generally not be time critical, and therefore can be uploaded at the energy-optimal time.

Another obvious, but important and practically difficult energy optimization is to actually turn off the sensing and support mechanisms when they are not being used. With multiple sensors sampling at different intervals, duty cycling the sensors and supporting hardware can become complicated. Wang et al. present a framework for managing that complexity, and report significant energy savings over both leaving sensors on continuously, and also over less careful control over sensor activity [63].

Finally we should note that for pollution sensing in particular, we will most likely need an active sampling mechanism. In order to ensure that the sensors in the phone quick get exposed to the atmosphere, we will probably need an air pump or another mechanism to create air flow around the sensor. Again, we will need to duty cycle the sampling mechanism to avoid significant energy consumption.

3.5.1 Energy and location

Mobile sensing has two characteristics that make in ripe for energy optimization in the localization mechanism. First of all, the device itself does not need to know the location of the sample. The location of the sample is primarily of interest when the data are aggregated. Secondly, the mobile sensing does not normally require real-time data. Rather, we can delay the transmission until it makes sense to do so from an energy perspective.

This suggests an extremely effective optimization. A GPS operates by sampling the frequency on which the GPS satellites transmit, and then performing a multi-dimensional parameter search on the signal to find the time-of-flight of the RF signal from the GPS satellites. By determining the time of flight of the satellites, the device can triangulate its position.

Typically, this multi-dimensional search takes place on the device, sometimes with assistance from devices in the network to aid with fast acquisition and to offload some of the computational work. Even with this network assistance, however, there is still significant processing and communication that must be done on the device. Obviously, this computation has an energy cost.

In the case of mobile sensing, however, we do not need to actually perform the multi-dimensional parameter search at all, until the location is needed during aggregation. If we simply sample the base-band signal, we can store it and upload the signal data when the device is plugged into charger or connected to a relatively low-power radio such as WiFi.

This mechanism is known as snapshot-GPS, and is used in wildlife sensing applications [66].

3.6 Conclusion

In this chapter we have briefly examined the four main sensor technologies that might be considered for integration with a phone. Of these, metal oxide sensors are quite promising for sensing reactive gasses, if they can be well characterized and calibrated. Electro chemical sensors are easy to work with, but are restrictively large. Spectroscopic sensors tend to be too large for large scale sensing in a mobile phone, but are the only viable option for non-reactive gasses such as CO_2. We hope that manufacturing advances might allow for further miniaturization. MEMS devices promise to make aerosol sensing in a phone feasible.

Chapter 4

Hardware platforms and mobile sensing data sets

In this chapter I will describe the various data sets we have utilized in our research, the hardware used to gather those data, and the rationale for choosing the various hardware platforms we used or built.

The California Air Resource Board (CARB) gathers hourly data on the concentration of several pollutants in the atmosphere at 35 stations located in California and northern Baja California, and they also aggregate data from over 250 stations operated by other agencies statewide. These data provide highly vetted, long-term information about many airborne toxins and pollutants.

In particular the CARB data provide us with information about the temporal evolution of pollution concentration at a variety of fixed locations. This allows us to both calibrate our algorithms and cross-validate our results (with appropriate hold-outs of course).

Of course, the CARB data also give us very sparse spatial coverage, and they are not gathered with mobile sensors, so they are of little use in determining the medium or small-scale spatial characteristics of pollution. They are also published hourly, so they provide limited temporal resolution as well.

So, in order to gain an understanding of real-world mobile sensing, we have designed and built several platforms for sensing. The first platform is a commercial, off-the-shelf (COTS) platform

---

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7] and “N-SMARTS: Networked Suite of Mobile Atmospheric Real-time Sensors” [28]

based on handheld pollution sensors, a handheld GPS, and two custom enclosures. This platform is very robust, easy to deploy, commercially supported and requires little engineering effort to build or maintain. It has allowed us to quickly gather data with which to get an initial understanding of mobile pollution sensing data. See Section 4.1 for details.

The COTS platform has the disadvantages, however, that the data are batch uploaded rather than uploaded over the air, the devices are too bulky to expect people to carry with them for long periods of time, they are expensive, and they have limited precision vs. a calibrated custom sensor deployment.

We have therefore also built two integrated sensor boards with an integrated micro-controller, and various sensors. These flexible platforms allow us to sample at relatively high spacial and temporal resolution, and to integrate the sensing and data transport into a variety of form factors. See Section 4.2 for details.

We have also designed a MEMS based $PM_{2.5}$. Current commercially available PM sensors are quite large (usually several pounds and several inches in each dimension). MEMS technology allows us to reduce the size of the device to the size where it might be integrated into a phone. Although this was largely other people's work, it merits mention because it is closely related to my work. See Section 4.5 for details.

The data real-world data collection campaigns were funded and performed by Intel. We served to advise them, provide technical assistance (including access to our labs), hardware design and implementation, and analysis.

Finally, we have designed and built a testing chamber that allows us to measure a sensor's response (at high frequency) to exposure to various poison gases while controlling and measuring the humidity and temperature in the chamber. This testing chamber allows precise control of poison gas concentrations and humidity, as well as very high resolution sensor sampling. See Section 6.2 for details.

4.1 The COTS platform

While researchers have increasingly devoted effort to understanding movement patterns, and some pollution data exists, when we started there were no publicly available, personal-mobile pollution data that we were aware of. Even to this day, the available data sets have very limited sample size and scope. We clearly needed to gather our own data.

In order to start working with data quickly, without a long engineering cycle we designed a

Table 4.1: Sensor data loggers used in the COTS platform

Gas	Manufac.	Device Name	Range	Prec.	Acc.	Humid.
CO	Lascar	USB-EL-CO[34]	0-1000ppm	0.5ppm	6% of read	15-90% RH
NO2	BW Technology	Gas ALert Extreme NO2[59]	0-100ppm	0.1ppm	5%	15-90% RH
SO2	BW Technology	Gas ALert Extreme NO2[59]	0-100ppm	0.1ppm	-	15-90% RH
GPS	Garmin	Quest[21]	-	3cm	10m	-

simple system based on off-the-shelf data-logging pollution sensors, and a GPS. By synchronizing the clocks on each of the devices, we can correlate the recorded sensor readings with the location recorded by the GPS.

Each sensor ensemble contains a Carbon Monoxide data logger, a GPS, and either a NO2, SO2 or O3 data-logger. See Table 4.1 for details on the data loggers.

We packaged each ensemble into one of two types of "kit:" an "automotive platform" that can be mounted near a car window or externally (Figure 4.1), and a "personal platform" that can be worn on a user's belt (see Figure 4.2).

4.1.1 The Accra study

The Intel Research lab in Berkeley ran a study which utilized this platform. We deployed six automotive platforms on taxis and four personal platforms on students in Accra, Ghana, West Africa, for two weeks in March, 2007. Accra was chosen, rather than a local study, because pollution in Accra (a dense urban area) is much more severe than in the San Francisco Bay Area. We employed an assistant in Ghana who managed all of the logistics for the project. Each participant was paid ten dollars per day for their participation.

The taxi drivers were members of a co-op, which helped us to ensure that our equipment would be returned at the end of the study. Each morning, they came to get the data-logging equipment, and installed it in their car, with the help of our assistant. Each evening, they returned to our assistant, who downloaded the data from each logger and GPS, charged the devices, replaced

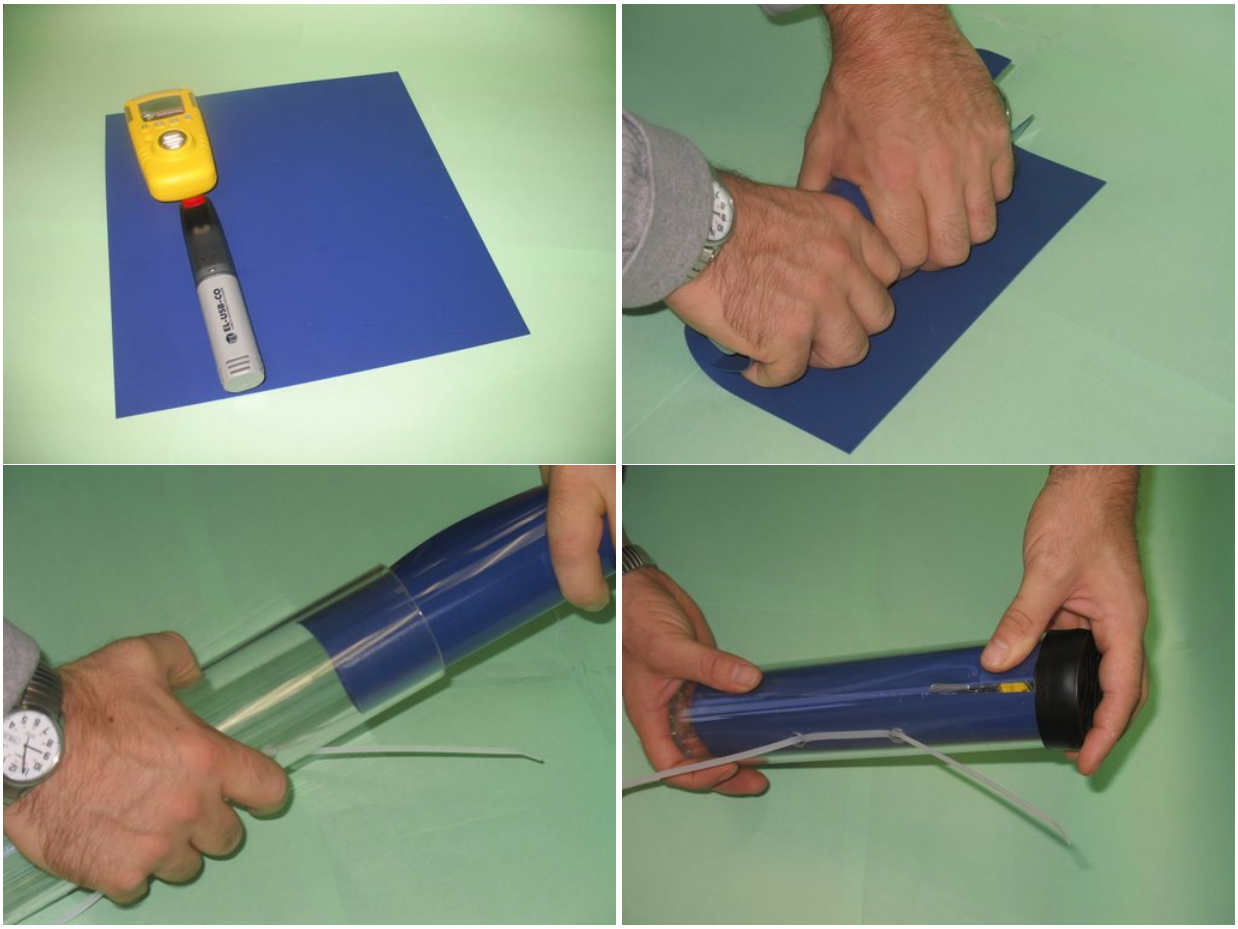


Figure 4.1: Assembling the automotive enclosure with CO and NO2 sensors inside

Figure 4.2: The "personal" version of the data-logging sensor platform

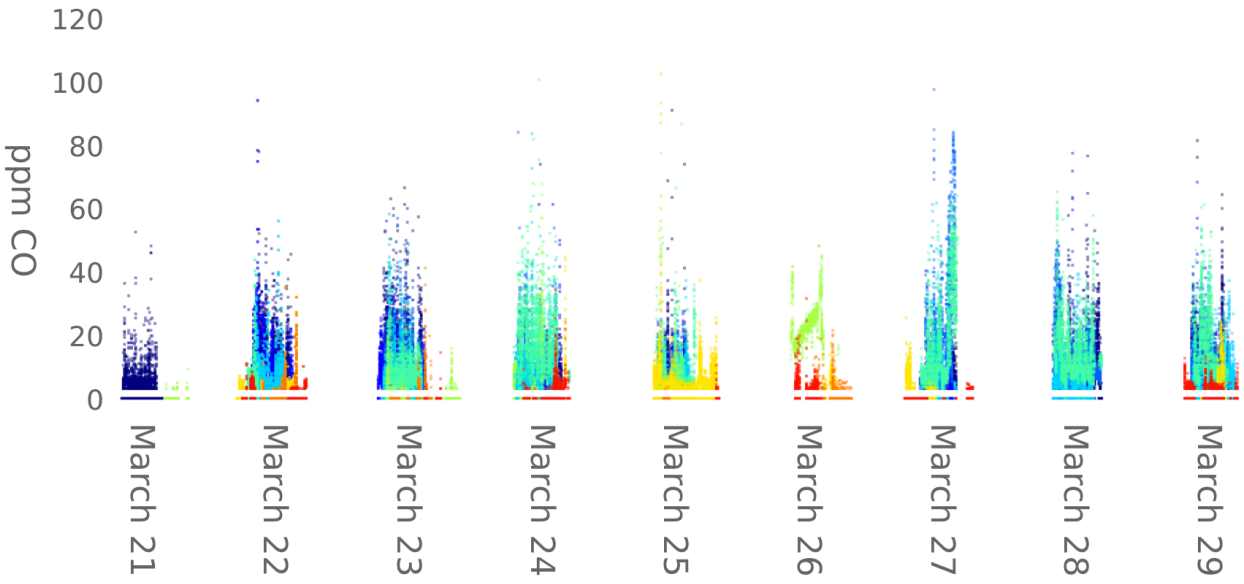


Figure 4.3: Data from the Accra sensing campaign. Different colors show data from different users. There is a gap between 0ppm and 3ppm because the sensor reports values less than 3ppm as 0ppm, presumably to avoid inaccuracy due to drifting bias by reducing precision near 0ppm.

batteries and ensured that the equipment was functioning properly. One NO2 logger and one CO logger malfunctioned during the study. Our assistant emailed us the data each evening.

The four students at Ashesi University wore the devices with them, and took responsibility for their maintenance. All of the devices that the students carried functioned properly throughout the study. The students charged the devices and emailed us the data on a daily basis. See Figure 4.3.

The sensors appear to report any concentration sensed at lower than 3ppm as 0ppm, presumably because the sensor can then be relatively robust to drifting bias (see Chapter 7). Unfortunately this makes the sensor data relatively useless for determining bias miscalibrations, and also increasing precision, as described in Chapter 7 and Chapter 8.

Despite this limitation, the data do give us a general idea of where pollution occurs (e.g. at busy intersections), and its magnitude (e.g. very high), and its temporal behavior.

The data from these loggers were uploaded into a database and can be viewed in a variety of formats, including an overlay on Google Earth (Figure 1.1, Figure 1.2, and Figure 1.3) [46].

4.1.2 Qualitative Results

In addition to gathering data for study, we asked participants in the study to take an exit survey. The students involved in the study viewed and shared their data with one another before uploading them, and seem to take a great deal of interest in figuring out which areas had the most pollution. The audible alarms also gave the students acute awareness of their own exposure, and one student indicated that he changed his routes, and even had his car tuned "to ensure I don't contribute to the problem."

The taxi drivers found the pollution monitors interesting, but were (predictably) most excited about the GPS, which gave them insight into their routes. The students were also excited about the GPS, and their ability to track their own behavior [46].

While these notes suggest that the data we gathered are not necessarily representative of typical user behavior, they do elucidate the fact that latent interest in pollution and people's environment can be tapped to influence people's individual behavior and awareness.

4.2 The integrated platform

We have also developed an integrated platform that more closely approximates a phone manufactured with sensors integrated directly into the phone itself. This design will allow

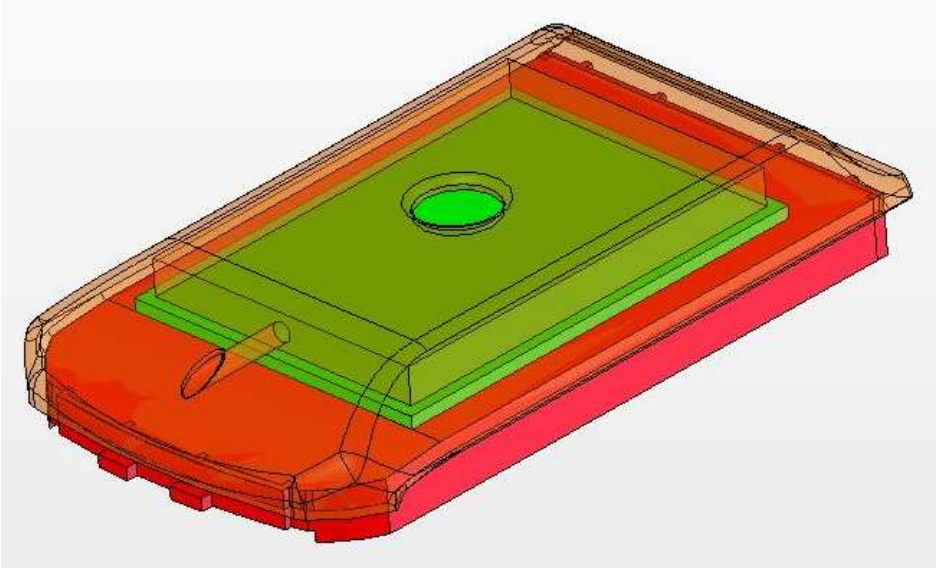


Figure 4.4: The battery of a LG VX9800 with a PCB mounted on top and covered by a new enclosure (outline shown for the new enclosure only). Mechanical design by Kyle Yeates.

Table 4.2: Major components of the N-SMARTS board

Component	Type
CO Sensor	Electro-chemical
CO/NOx Sensor	MOS
Temperature	Surface mount, silicon
Accelerometer	3 axis, low G
Bluetooth Radio	BlueCore
Flexible Power System	Linear regulated

significant cost reduction with respect to less complete integrations. For our testing platform, however, rather than actually manufacturing a new phone and enclosure, we can simply replace the battery pack of the phone with a module that clips in to the battery well of the phone, and contains both a battery and the sensor module (Figure 4.4). The mechanical design was done by Kyle Yeates). I will call this and the related devices "the N-SMARTS board."

Table 4.2 shows the components in the N-SMARTS platform. We chose to use Bluetooth to communicate with the phone to avoid mechanical problems with a physical serial link, and to make the design and software more generic (Figure 4.5).

Researchers at Intel also modified this design to create a dual-board version that splits the

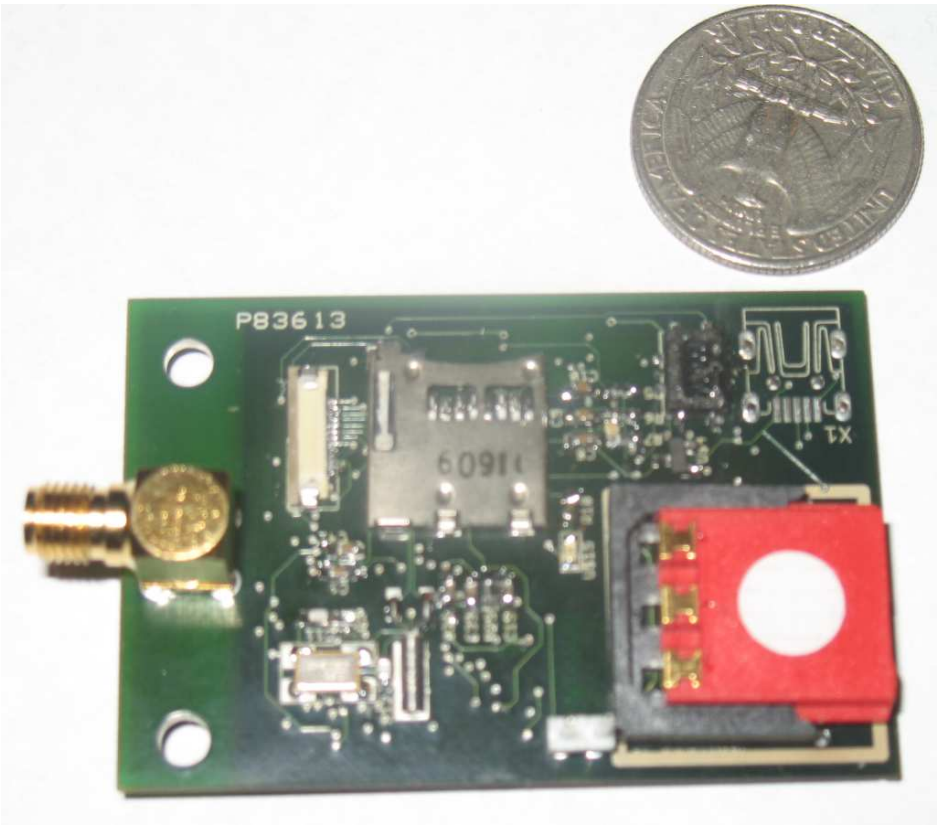


Figure 4.5: A prototype of the Bluetooth board with a CO sensor, a NOx/CO dual sensor, a temperature sensor and an accelerometer. This board will integrate directly with the phone.

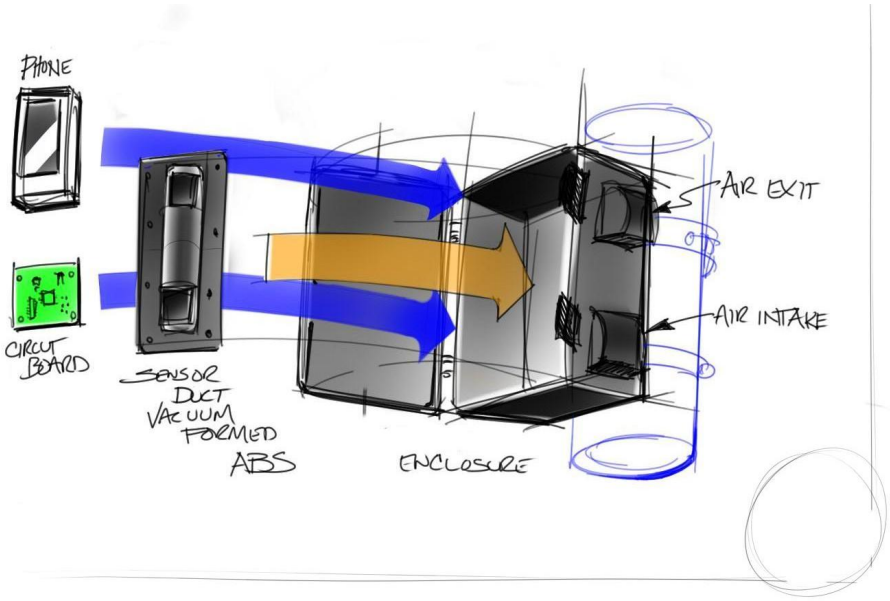


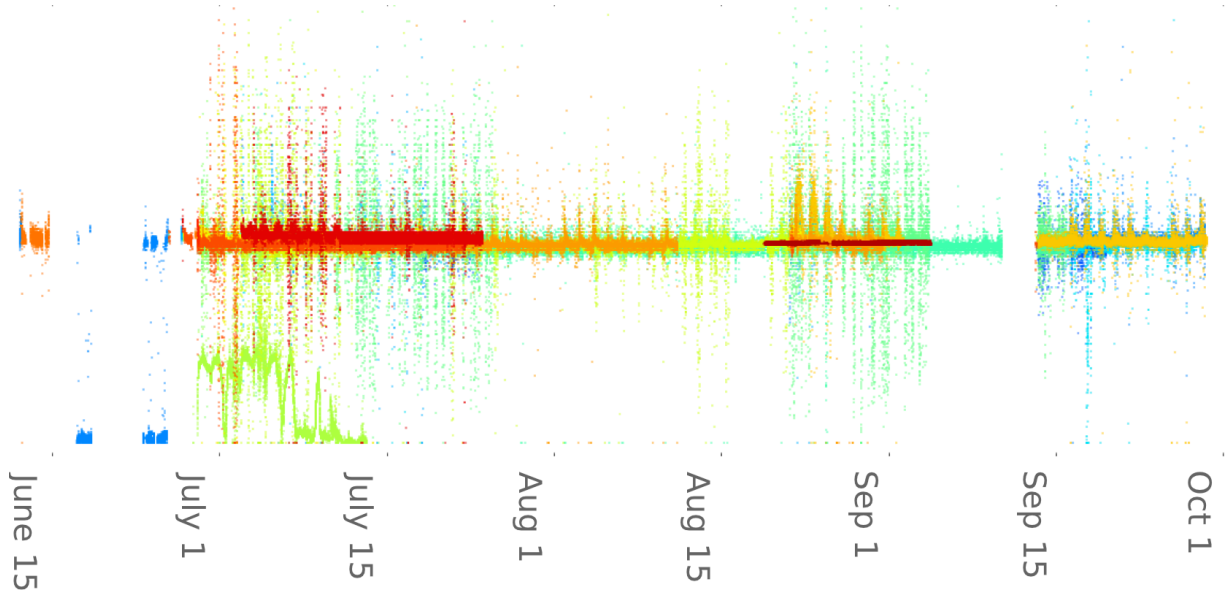
Figure 4.6: The large enclosure with fan and vent for automotive and stationary deployments (drawing by Christopher Myers) .

sensors onto a daughterboard. This allows us to physically separate the control circuitry from the sensing apparatus. This is convenient in deployments in which the platform is in a larger enclosure attached to a vehicle, for example (Figure 4.6 and Figure 1.6). Unfortunately, this decoupling also introduced significant into the sensor hardware. See Section 4.2.1.

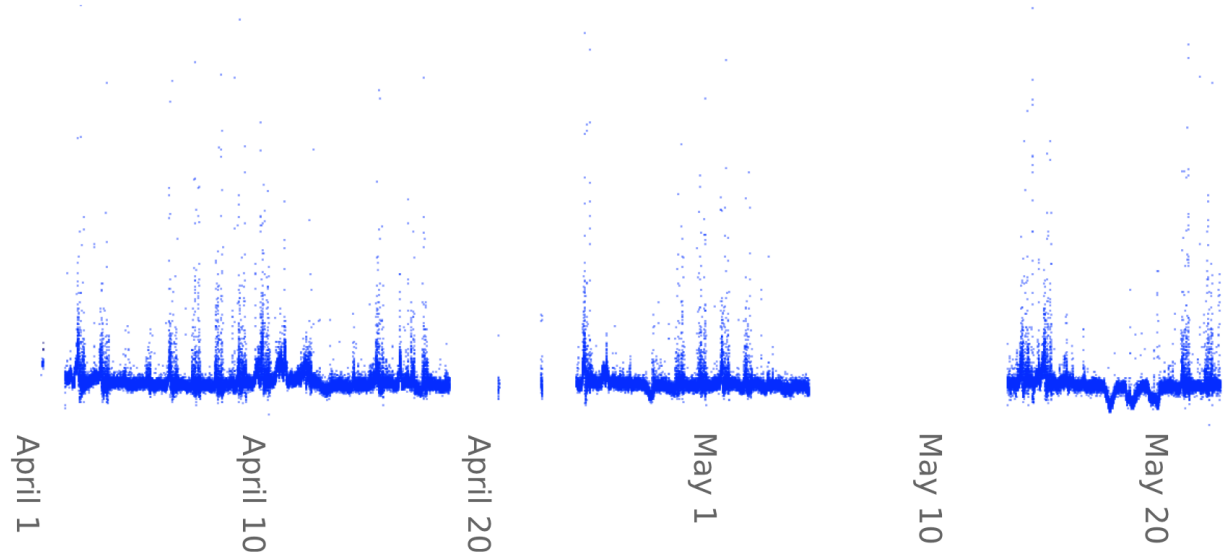
4.2.1 The San Francisco street sweeper deployment

We outfitted fifteen street sweepers with our boards, in conjunction with GPS-equipped Nokia N95 phones. These sweepers collected data on pollution levels over the course of about 9 months, with varying numbers of boards in operation at a given time. [4]

Unfortunately, the link between the motherboard and daughterboard in the modified board is made using analog lines between the sensors and the ADC. These analog lines seem to be very sensitive to mechanical noise, and have introduced significant noise into the data sets we have collected with them. We have worked on various algorithms to detect or compensate for the noise, but the power of the noise is very high, and we have met with limited success (Figure 4.7).



(a) Noisy data from the San Francisco street sweeper campaign



(b) Clean data from the San Francisco street sweeper campaign

Figure 4.7: Data from the San Francisco street sweeper campaign. One device had a fully integrated design, and produced relatively clean data (4.7(b)). The other devices in the deployment were plagued by noise which was probably caused by mechanical vibrations (4.7(a)). These devices were more sensitive because the sensors were separated from the analog-to-digital converter with analog lines.

Table 4.3: Major components of the Intel badge

Component	Type
CO Sensor	Electro-chemical
O3 Sensor	Electro-chemical
CO/NOx Sensor	MOS
O3 Sensor	MOS
Temperature	Surface mount
Humidity	Surface mount
Accelerometer	3 axis, low G
Bluetooth Radio	BlueCore
Flexible Power System	Linear regulated
GPS	N/A
GPRS Modem	N/A

4.3 The Intel "badge" platform

The Intel "badge" platform uses a similar design to the integrated platform (Section 4.2), but it also integrates a GPS and GPRS modem, plus an electrochemical ozone sensor. The addition of the GPS and modem eliminate the need for a separate phone in addition to the device. Rather than force the user to use a particular phone with an large prototype form-factor, the "badge" platform instead allows users to simply clip the device on their backpack, or belt, or otherwise wear it on their body (Figure 4.8).

Not only does the independent design of the device get around the difficulty of designing a device small enough to embed in a phone, but it also allows us to avoid issues surrounding obstruction of airflow to the sensor when the phone is in a pocket, purse or bag.

In addition to the components in the integrated platform, the device features both metal oxide and electro-chemical ozone sensors, a built and GPS and a built in GPRS radio. See Table 4.3. The device periodically uploads data via SMS to a database.

4.3.1 The West Oakland deployment

West Oakland is directly adjacent to the port of Oakland, the fourth busiest container port in the United States [43]. This makes it a major shipping hub, where trucks and trains rendezvous

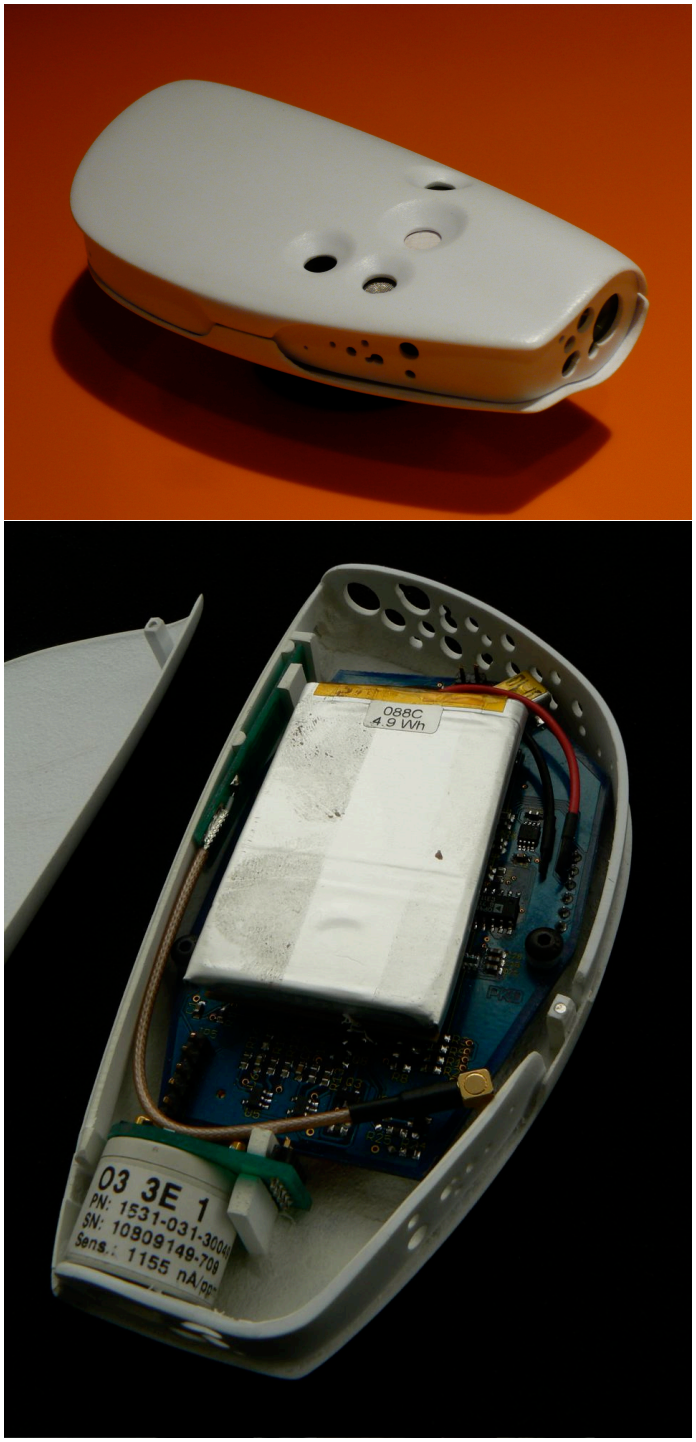


Figure 4.8: The Intel badge device, with the battery and electrochemical ozone sensor visible in the bottom image.

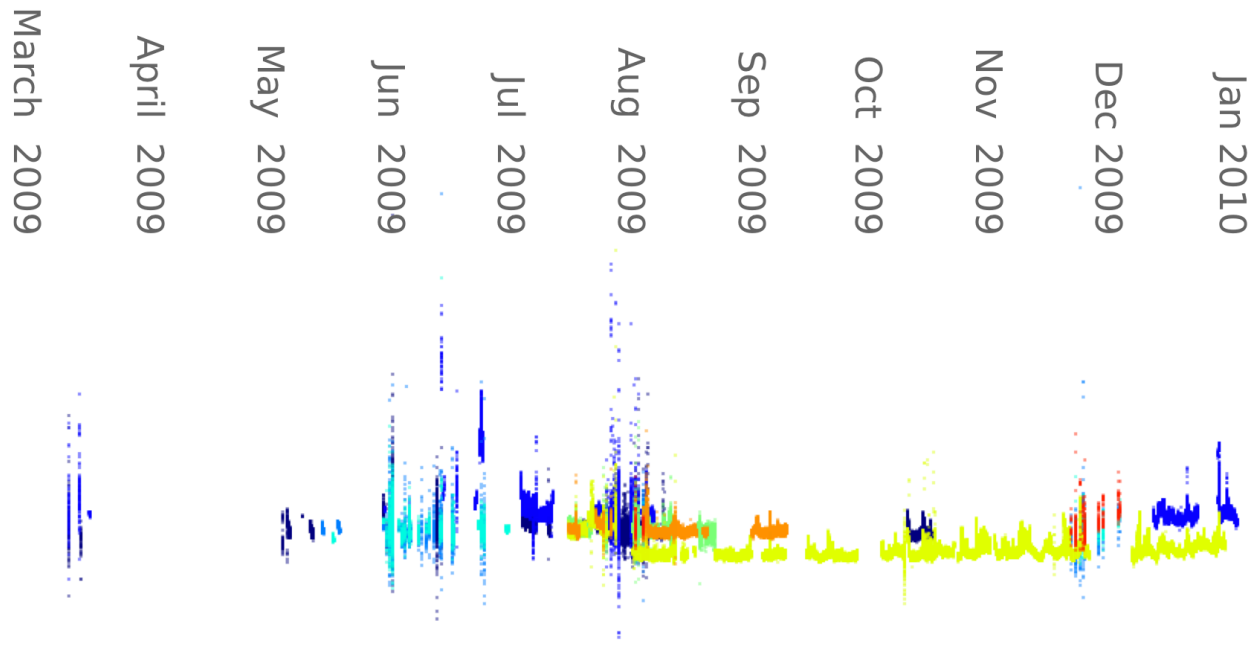


Figure 4.9: Data from the West Oakland sensor campaign. These devices were also sensitive to noise. In this case, however, the noise was eliminated by introducing vibration isolators in between the circuit board and the enclosure, in August 2009.

with ships to exchange cargo. All of these vehicles are typically diesel powered, and often idle their engines to when waiting to exchange cargo.

These factors make it an area of concern for residents in West Oakland. The CommonSense team, identified participants to carry the badge sensors through a process of formal and informal interviews under various circumstances. CommonSense team deployed a varying number of devices varying from five to ten from March 2009 until January 2010 (see Figure 4.9 and Figure 8.5).

Even with the redesigned board, which minimizes the trace length of analog sensing lines, there was significant noise in some of the sensors during the initial deployment. These issues were addressed by introducing vibration isolators between the board and the enclosure (Figure 4.9).

4.4 Phone considerations

We have designed the N-SMARTS board to work with any phone that allows programmatic control of the Bluetooth radio. If location sensing is also required, as it is for our application, then the phone should also have an integrated GPS. For our initial deployment, we are using the

relatively expensive but easy to use Nokia N95 smartphone.

The GPS integrated into GSM and some WCDMA-based phones, however has a fundamental limitation that Qualcomm MSM chipset-based phones have overcome. Qualcomm's chipset tightly integrates a GPS radio that takes advantage of the tightly synchronized clock that is necessary to make a CDMA radio work. Qualcomm CDMA networks are synchronized using the GPS clock. In doing so, it they effectively a dimension in the parameter space that the mobile phone's GPS radio searches when trying to lock on to satellite signals. This allows the radio to dwell on each parameter setting for about 1000 times longer than in a normal GPS, averaging the signal over a significantly longer amount of time, and significantly reducing noise. The end effect is a gain over a normal GPS radio, that Qualcomm quantifies as about 15–20dB [40]. Recent high-end WCDMA radios seem to have this capability as well.

Practically speaking, this means that MSM-chipset based phones, when combined with other Assisted-GPS technologies, are capable of getting very fast cold-start fixes as well as indoor fixes. Our experience has found this difference to be very significant, and for passive-sensing (human out of the loop) applications, this capability is doubly important. For this reason, we also are building BREW-based phone software to interact with the N-SMARTS sensor platform. Other A-GPS solution which use the same technique would also be effective.

Another possibility for lower power operation is to simply record the signal strength of all the mobile towers that the phone can hear, and possibly wifi and Bluetooth beacons as well. These data could be uploaded with the sensor data and interpreted in post-processing, offering a very low-power alternative localization technology, albeit one with less accuracy.

4.5 MEMS $PM_{2.5}$ sensing

Airborne solid (Aerosol) particulate matter smaller than 2.5 microns represents a serious health risk that has traditionally required large devices to detect. Richard White and several students and postdocs have developed a MEMS based device that measures particulate matter in the air by depositing aerosol particles on a thin-film bulk acoustic resonator (FBAR) oscillating at 1.6GHz. A thermal gradient is induced that causes airborne particles to deposit on the FBAR via thermal phoresis. This deposition changes the resonance frequency of the FBAR, which we can measured (Figure 4.10).

They have also designed several small inertial impaction filters [45] capable of long term filtration without replacement, and capable of selecting particles smaller than 2.5 microns

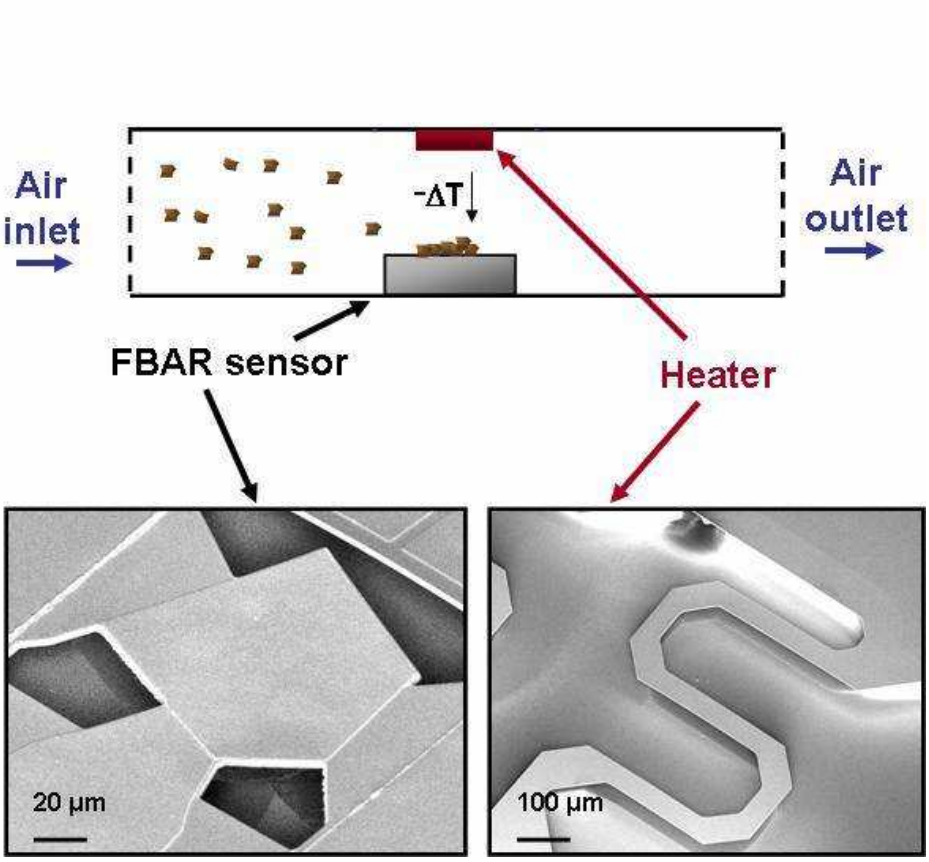
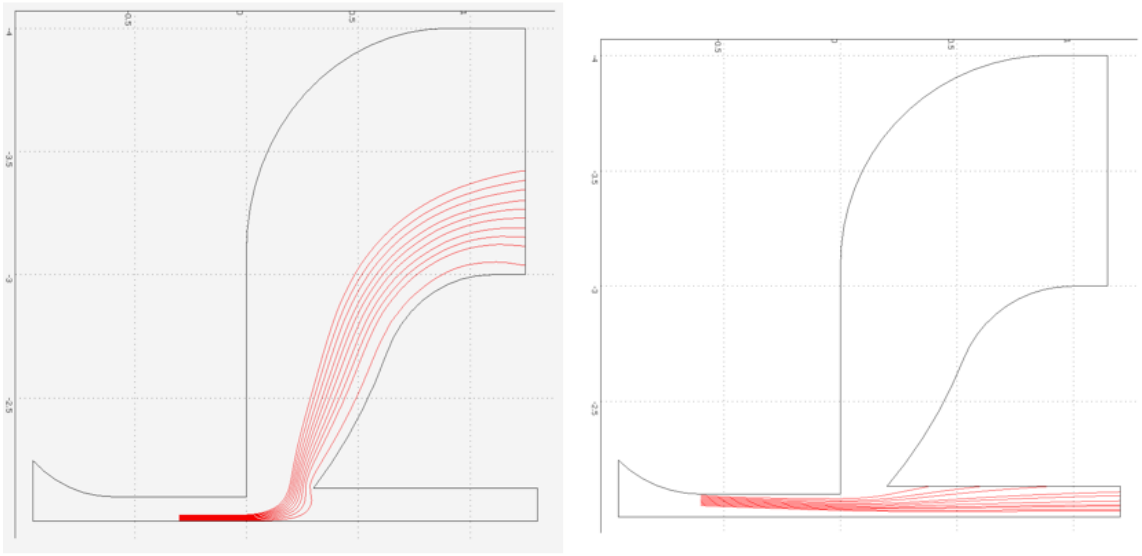


Figure 4.10: $PM_{2.5}$ particles are deposited on a resonating FBAR via thermal phoresis, changing the resonance frequency of the FBAR. Figure by Justin Black.



(a) (b)

Figure 4.11: Calculated particle trajectories through a simulated impactor for 0.5 μm (4.11(a)) and 5 μm (4.11(b)). Figure courtesey of Igor Paprotny [45].

(Figure 4.11).

The largest component of these MEMS devices is the air sampling mechanism, rather than the sensor itself, and significant effort has been put into miniaturizing them in order to make the entire device small enough to deploy in a mobile phone. Fortunately, the sampling mechanism could also be used to control the sampling process for other sensors, so the impact on size and energy could at least increase the accuracy of all of the sensors.

The MEMS $PM_{2.5}$ was not ready for deployment into a mobile phone at the time of writing.

Chapter 5

Basic models of pollution concentration and dispersion

There are several questions that we can ask about pollution in a given area, and which questions we are trying to answer will determine how we construct our model of pollution. One question that we can ask is, what is the concentration of pollution in a location that has not been explicitly sampled, but which for which there might be neighboring samples. This is essentially a problem of extrapolation or interpolation, and raises the related question of what our confidence in our estimate.

Another important question is, what is the total pollution being emitted inside of a region. Another question might be where are the sources of pollution? Another might be: where are the pollution "hot-spots," e.g. where are the areas of the worst pollution.

We have taken two complementary approaches to modeling pollution concentration and dispersion: statistical and generative. In the statistical model, the concentration of pollution is represented by a probability distribution. For every location in space-time, there is a probability distribution which represents *our knowledge* of the concentration of pollution in that location. In the generative model, we use our understanding of the propagation of pollution in the atmosphere to extrapolate what the concentration of pollution in other, unmeasured locations might be.

In this chapter, we will lay out the fundamental equations, models and assumptions that will be that basis for the algorithms described in future chapters. The purpose of this chapter is to

---

Parts of this chapter were previously published in "mScience: Sensing, Computing and Dissemination" [7]

describe the basic ideas and equations that underlie our work, for those unfamiliar with atmospheric modeling, and to document our assumptions. These approaches are "textbook," and are not innovative, as such, although their particular combination might be somewhat novel.

We acknowledge that these models are relatively simple, and do not take into account many recent advances in atmospheric modeling. The goal of this and subsequent chapters is to present a relatively simple model from which to begin analyzing real data. As more data are collected, insufficiencies in the models will become more apparent, and they can be revised accordingly.

Our approach permits us to test system and computational models. As such, we hope to remain relatively agnostic to the specifics of a given model, and examine the larger systemic approach to gathering and integrating data from a large number of mobile sensors. Indeed, models, no matter how carefully designed, must be validated by empirical observation. We therefore present the following naive models with the understanding that their refinement will be guided by data as it becomes available.

5.1 Models of pollution

The caveats expressed above not withstanding, our statistical model should reflect the physics of our environment as much as possible and practical. In this section, I will motivate and describe the statistical model we have adopted. These models were largely adapted from Seinfeld and Pandis' excellent book on Atmospheric Chemistry [51].

Consider for a moment, what might happen at a molecular scale as we, or a sensor are exposed to pollution. A particle is emitted by a pollution source somewhere upwind from us. This pollution particle travels through a lot of turbulence, and eventually meanders its way to our nose, or sensor, where it registers in our brain or in our sensor as some sort of current or voltage.

If the process of emitting pollution is relatively stable over time, then our sensor will continue to detect pollution. However, because turbulence is highly random, those particles will not arrive in a steady stream, even if they are being emitted in a steady stream. If the turbulence is so chaotic that there is no correlation between the inter-arrival time of particles, but they arrive at a mean rate related to rate of emission, then they would reasonably be model as arriving according to an exponential distribution. In that case, the concentration of pollution would follow a Poisson distribution.

In reality, of course, we typically don't count the number of molecules that our sensor gets exposed to (although we sometimes do count particles of aerosol pollution). Instead we look at

voltages or other average statistics that essentially indicate the number of particles to which our sensor has been exposed over time.

The Poisson distribution,

$$P_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{\lambda^k e^{-\lambda}}{\Gamma(k+1)}. \tag{5.1}$$

is well approximated by the binomial distribution as the k parameter gets large. As k gets very large, it makes sense to view the binomial distribution as a Gaussian distribution. In any event, for small values of k, the measurement noise due to thermal noise in the sensor, background radiation, etc. will swamp the signal. These noise components are typically Gaussian, and so a Gaussian approximation to the distribution of the concentration at a given location seems appropriate.

With that said, our earlier decision to model the inter-arrival times between particles of pollution as exponential was arbitrary. One common assumption in atmospheric chemistry, however, is that the velocities of particles from a pollution source to a given point are IID Gaussian [51]. This is not an unreasonable assumption if pockets of air move through a larger turbulent vortex relatively perturbed.

The independence assumption along with the stationarity assumption, regardless of the actual distribution, means that the probability that a particle arrives at our sensor in a given time window is a scalar function. In other words, arrivals at our sensor during a given window of time is a Bernoulli trial. The important implication of this is that in the limit, the distribution of the mean number of arrivals is just the Gaussian distribution, by the Central Limit Theorem.

With that said, it is not clear to what extent the Gaussian velocity distribution assumption is motivated by physics or rigorous empirical study versus analytical convenience. Nonetheless, we have little reason to prefer exponential inter-arrival times over Gaussian velocities, and in any event, it is not likely to make much practical difference.

The Gaussian assumption on the velocities has another important consequences. Besides inducing a Gaussian distribution on the concentration density at a given location, it also induces a Gaussian spatial diffusion pattern [51], essentially allowing us to use a Gaussian-like kernel function in the sections that follow.

5.1.1 Atmospheric dispersion models

There are two primary mechanisms for diffusion of pollution from a source into its surrounding environment: chemical diffusion and turbulent dispersion. Both of these can be modeled, in general, as systems of partial differential equations, and can be independently quite complex,

let alone in conjunction with one another. It is typical for atmospheric chemists to make several simplifying assumptions about the dispersion mechanism, for the purposes of analytic and computational tractability.

For outdoor models, the first of these assumptions is often to disregard chemical diffusion: turbulent dispersion is typically several orders of magnitude faster than chemical diffusion, except in extremely still air. Another important assumption common in dispersion models is that chemical reactions which alter the species concentration are negligible with respect to the rate of flow in the atmosphere [51].

After these two assumptions, if we make the assumption that the velocities of particles in the flow of pollution are Gaussian-distributed, then we induce a Gaussian-like kernel function (Section 5.1).

More specifically, if we define $\sigma_x^2 = a_x t$, $\sigma_y^2 = a_y t$, and $\sigma_z^2 = a_z t$, and $\overline{u}$ is the mean velocity of the wind, then the expression (derived using a "Lagrangian" approach) for the mean concentration from a continuous point source in an infinite fluid in stationary, homogeneous turbulence [51] is

$$\begin{aligned} \langle c(x,y,z,t)\rangle &= \frac{q}{2\pi \left(a_y a_z\right)^{1/2} r} \exp\left(-\frac{\overline{u}}{a_x}(r-x)\right) \\ r^2 &= x^2 + \left(a_x/a_y\right) y^2 + \left(a_x/a_z\right) z^2. \end{aligned} \tag{5.2}$$

The a_x, a_y and a_z parameters must be set depending on the turbulence in a given location.

It is important to note that t does not appear in the mean concentration formula (5.2). In other words, once we reach a steady state, the concentration of pollution in the area around a source is independent of time.

Since we are often concerned with sources and measurements that are very close to the ground, then we can ignore the z term, and (5.2) becomes

$$\begin{aligned} \langle c(x,y,t)\rangle &= \frac{q}{\left(2\pi a_y\right)^{1/2} r} \exp\left(-\frac{\overline{u}}{a_x}(r-x)\right) \\ r^2 &= x^2 + \left(a_x/a_y\right) y^2. \end{aligned} \tag{5.3}$$

See Figure 5.1 for an example plot.

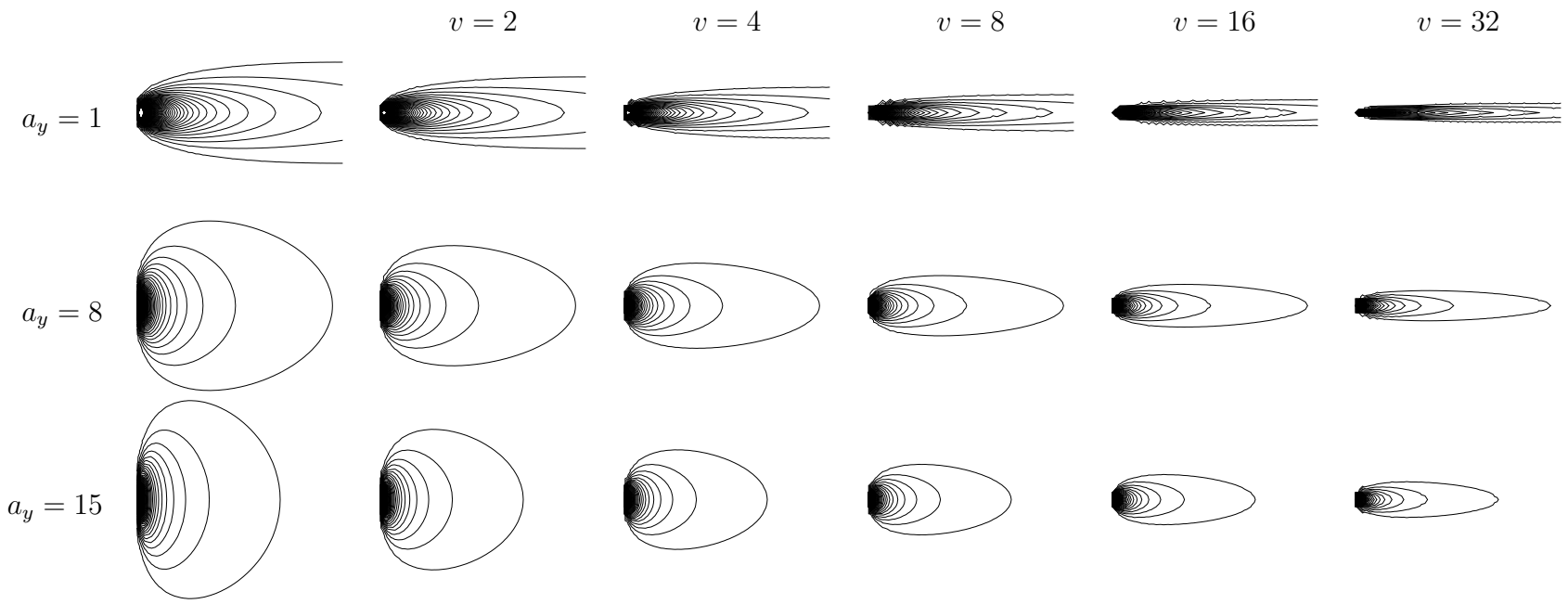


Figure 5.1: Contour plots of the pollution density in a two dimensional, continuous point source dispersion model in a steady state, described in (5.3) for various v and a_y parameter values.

If we are concerned with the relationship between a puff of pollution and its concentration later in time (e.g. in a non-static model), we can simply use a Gaussian formulation, with the caveat that the variance of the Gaussian is a function of time [51]:

$$\langle c(x,y,t)\rangle = \frac{q}{2\pi\sigma_x(t)\sigma_y(t)} \exp\left(-\frac{(x-\overline{\mu}t)^2}{2\sigma_x^2(t)} - \frac{y^2}{2\sigma_y^2(t)}\right) \tag{5.4}$$

Next we assume that the correlation between the velocities of particles decays exponentially with time, eg:

$$\langle (u(t)-\overline{u})(u(\tau)-\overline{u})\rangle = \sigma_u^2 \exp(-b\,|t-\tau|), \tag{5.5}$$

with τ being an initial time, and $t > \tau$, σ_u^2 being the variance of the Gaussian particle velocity distribution, $\overline{u}$ being the mean velocity of the particles, and $1/b$ being the characteristic decay time of the correlation in velocities. With this assumption, the function $\sigma_x^2(t)$ can be written as

$$\sigma_x^2(t) = \frac{2\sigma_u^2}{b^2}\left[bt + e^{-bt} - 1\right]. \tag{5.6}$$

Furthermore, when $t >> b^{-1}$, (5.6) reduces to

$$\sigma_x^2(t) = \frac{2\sigma_u^2 t}{b}. \tag{5.7}$$

When $t << b^{-1}$, then $\exp(-bt) \simeq 1 - bt + (bt)^2/2$, and $\sigma_x^2(t)$ reduces to

$$\sigma_x^2(t) = \sigma_u^2 t^2. \tag{5.8}$$

In other words, σ_x^2 is proportional to t^2 for small t and proportional to t itself for large t:

$$\sigma_x^2(t) \propto \begin{cases} t^2 & \text{for large } t \\ t & \text{for small } t \end{cases} \tag{5.9}$$

See Seinfeld and Pandis for a detailed analysis [51].

Of course, this means that at $t = 0$, the concentration of pollution is infinite, so if we use this model as a kernel function, we will need to to assume the pollution is concentrated over a non-zero volume, but that the model above holds as t becomes sufficiently large.

5.2 Generative models

In a generative approach, we use a model such as (5.3) to directly simulate the dispersion of pollution from a source to a field. If the mean concentration and emission of pollution is relatively

steady over a long time with respect to the time it takes wind to transport pollution over a region, then we can assume that a region with such behavior will reach a steady state over time. In other words, observations of pollution within a short window of time are essentially made at the same time. Over longer periods of time, the sources of pollution may change their location or intensity, in which case observations might be broken into epochs.

This is a reasonable model for typical dense urban environment. Primary sources of pollution include motorways, particularly intersections and highways, and factories. Traffic patterns and factory operations typically change over the course of a couple of hours, as do prevailing winds. Wind velocities, however, are typically on the the order of a few meters per second. Thus, over the course of a few hours, the pollution from a source could be dispersed several kilometers. For small areas of a few hundred meters, a steady state assumption seems reasonable. We will briefly address this again in Chapter 7.

In a steady state model, the mean concentration of pollution in a given location will be the sources of pollution, convolved with kernel function representing the dispersion of pollution from each source (e.g. (5.3)). We can use this model to infer the pollution levels at unobserved locations, and also, under limited circumstances (e.g. when the concentrations are fully observed), to infer the sources of pollution.

One steady state model-based algorithm will be presented briefly in Section 9.1.

5.3 Statistical models

This section presents basic statistical models that are informed by the dispersion models discussed in Section 5.1.

5.3.1 Gaussian processes

The statistical approach we have chosen is fundamentally Bayesian in nature. That is, we are modeling our knowledge about pollution. The mean (or mode) of the statistical distribution represents our best guess, and the variance represents our confidence in that guess.

By modeling pollution densities at a given location as Gaussian, we can take advantage of the statistical machinery of the continuous Gaussian process. As described in Section 5.1 above, the Gaussian assumption on the distribution of concentrations at a given location is appropriate not only because the *sensor noise* is Gaussian (see Section 6.2), but because the process by

which concentrations of gas mix and vary is also reasonably modeled as Gaussian (Section 5.1.1). Modeled this way, we have the sum of two Gaussians, which is itself a Gaussian.

Further support for this Gaussian assumption can be found in Figure 1.7. The fact that smoothing the signal from our sensor approximates the ambient pollution levels measured by the sensitive CARB sensors indicate that the fast changes in pollution concentration that our sensors are random enough to be averaged out by a Gaussian smoother. These fast changes are caused by transient effects in the immediate vicinity of the sensors (e.g. a truck driving by), and appear to be relatively uncorrolated on short time scales. While this assumption may not be exactly correct, it appears to be good enough for our algorithms to be reasonably effective (Chapter 7).

A Gaussian process can be thought of as a multi-variate Gaussian with an infinite number of variables, who's covariance is defined by a covariance function, often called a *kernel function*, rather than a covariance matrix. A Gaussian process can also be considered a Gaussian distribution over functions, with each point in the function represented by a Gaussian distribution.

Basic Gaussian process regression

Rasmussen and Williams provide an excellent introduction to Gaussian Processes [47]. I will summarize their well-considered analysis, and adopt their notation below. We will build and depend on this basic machinery in Chapter 7 and Chapter 8.

In standard linear GP regression, we consider functions of the form

$$f(x) = \phi(x)_T w + \varepsilon, \tag{5.10}$$

for some basis function $\phi : d \mapsto n$, and a Gaussian random vector $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. By putting a zero mean Gaussian prior over the weight vector

$$w \sim \mathcal{N}(O, \sigma_p), \tag{5.11}$$

we favor solutions with low weights. This is analogous to the complexity term in a regularization method, since it causes the algorithm to favor functions with a small weight vector.

As with any kernel method, there are two ways to view and analyze Gaussian process regression: the weight-space perspective, and the function space-perspective. We will focus on the function-space perspective, since for Gaussian processes, the function-space perspective seems far more intuitive. Readers interested in a more thorough treatment should read Chapter 2 of Rasmussen and Williams [47].

From a function space perspective, a Gaussian process can be specified in terms of its mean and covariance functions:

$$m(x) = \mathbb{E}[f(x)] \tag{5.12}$$

$$k(x_p, x_g) = \mathbb{E}[(f(x_p) - m(x_p))(f(x_q) - m(x_q))] \tag{5.13}$$

Plugging in $f(x) = \phi(x)^T + \varepsilon$, and recalling that w is drawn from a zero mean Gaussian, we have

$$\begin{aligned}
\mathbb{E}[f(x)] &= \mathbb{E}[\phi(x)^T w + \varepsilon] = 0 \\
\mathbb{E}[(f(x_q) - m(x_q))(f(x_r) - m(x_r))] &= \mathbb{E}[f(x_q)f(x_r)] \\
&= \mathbb{E}[(\phi(x_q)^T w + \varepsilon_q)(\phi(x_r)^T w + \varepsilon_r)] \\
&= \mathbb{E}[(\phi(x)^T w w^T \phi(x') + \varepsilon_q \phi(x_r)^T w + \varepsilon_r \phi(x_q)^T w + \varepsilon_q \varepsilon_r] \\
&= \phi(x)^T \mathbb{E}[w w^T] \phi(x') + \mathbb{E}[\varepsilon_q \varepsilon_r] \\
&= \phi(x)^T \Sigma_p \phi(x') + \sigma_n^2 \delta_{qr},
\end{aligned} \tag{5.14}$$

where δ_{qr} is the Kronecker delta.

Furthermore, since ϕ always appears as a quadratic term with σ_q (which is a covariance function, and thus positive definite), we can kernelize ϕ as

$$\begin{aligned}
k(x_q, x_r) &= \phi(x_q)^T \sigma_q \phi(x_r) + \sigma_n^2 \delta_{qr} \\
&= \phi(x_q)^T \sigma_q^{1/2} \sigma_q^{1/2} \phi(x') + \sigma_n^2 \delta_{qr} \\
&= (\Sigma_q^{1/2} \phi(x)) \cdot (\Sigma_q^{1/2} \phi(x')) + \sigma_n^2 \delta_{qr} \\
&= \psi(x) \cdot \psi(x') + \sigma_n^2 \delta_{qr},
\end{aligned} \tag{5.15}$$

where $\left(\Sigma_p^{1/2}\right)^2 = \Sigma_p$.

Now, lets consider how to make predictions using Gaussian process regression. If X is a vector of observations, then we have

$$\operatorname{cov}(y) = K(X, X) + \sigma_n^p I. \tag{5.16}$$

This, of course, assumes that the variance of all observations is constant. If we have different confidence in different observations, we can incorporate that into our regression by constructing a diagonal matrix from the vector of variances of the observations.

Sticking with our simpler formulation, we can write the joint distribution as

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Conditioning on X, y and X_*, we have

$$\bar{f}_* = \mathbb{E}[f_*|X, y, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y \tag{5.17}$$

$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*) \tag{5.18}$$

The computational complexity of this algorithm for regression is dominated by the need to invert the kernel matrix $K(X, X) + \sigma_n^2 I$. In the worst case, this is an $O(n^3)$ problem, where n is the dimensionality of the X vector.

There are, however, several ways to significantly speed this process. If the kernel function is sparse (for example, as in Section 5.3.3), e.g. it is exactly zero most of the time, then Cholesky solvers can ignore all zero element in the matrix. Even with dense matrices, conjugate gradient algorithms, and other iterative algorithms can terminate significantly faster than in $O(n^3)$ iterations, particularly when most of the elements of the matrix are nearly zero. Conjugate gradient can also exploit sparsity.

Finally, in this case, only the diagonal of the covariance matrix is actually needed to know the variance at each location. Nissner and Reichert explore and build on an algorithm originally proposed by Takahashi et al. to allow more efficient computation of the diagonal elements of an inverse using the LU decomposition [58, 42]. Since the LU decomposition can exploit sparsity, this algorithm is ideally suited to this problem.

5.3.2 A kernel function for atmospheric dispersion

As described in Section 5.3.1, Gaussian process regression uses a kernel function to determine the covariance between two points. That is, the kernel function is a function that takes two values in the domain of the function (in our case, locations in space-time) as inputs (say x and x') and maps to the covariance between the two values. In kernel methods, the kernels are often actually used un-normalized, with a scaling factor eventually applied as a parameter to the model.

It is not immediately obvious how to map from the dispersion model described in Section 5.1.1 to a co-variance function. Examining both (5.3) and (5.4) reveals that our models ascribe infinite concentration to the source of pollution. While this strictly makes sense, it doesn't match physical reality, since pollution is not emitted from an exact point.

Instead of using this model directly, we note that because the concentration falls off exponentially with space and time, the covariance will be dominated by points extremely close to one another. Thus, a sharp-peaked co-variance function, such as the Matérn covariance [47]

function might be appropriate. On the other hand, since the pollution from a puff quickly spreads out to a wide area, a high variance squared exponential kernel might also make sense. Finally, neither of these options will result in predictions that reflect the wind speed and direction accurately, so a "squashed" version of (5.3).

Our choice of kernel will largely depend on the purpose. A share covariance function will be favorable for automatic calibration, in which we must ensure a very close correspondence between the readings from different sensors (Chapter 7). A disperse kernel will work well for inferring large scale ambient levels of pollution. A more realistic kernel to incorporates wind parameters will work best for predicting the concentration of pollution downwind from sensor readings.

5.3.3 A kernel for computational efficiency

The analysis in Section 5.3.2 suggests that some composition or variation of the Matérn and Gaussian covariance function might match the data we have collected well. Unfortunately, these kernels have infinite support, so they will result in dense, extremely large matrices to invert. We can, however, replace these relevant portion of the infinitely supported covariance function with a similar complactly supported one, thus inducing a sparse covaraince matrix that can be exploited by sparse solvers. This section discusses how to replace the squared exponential kernel function with a sparsely supported appoximation: a practice we have used throughout this dissertation.

We begin our analysis with a squared exponential function kernel:

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{\sigma_x^2}\right) \tag{5.19}$$

As our data set grows, inverting the noisy covariance matrix of our Gaussian process $[K(X, X) + \sigma_n^2 I]^{-1}$ will take $O(n^3)$ time if we are not careful, where n is the number of observations. Since $[K(X, X) + \sigma_n^2 I]^{-1}$ must be positive definite, we can do the inversion using a Cholesky decomposition. The Cholesky decomposition can take advantage of sparsity in K, depending on the sparsity structure of K.

The squared exponential kernel has infinite support, so in order to reap the benefits of the Cholesky decomposition's efficiency, we must approximate the radial basis kernel. A naive solution to this problem would be to truncate a Gaussian kernel after, say, three standard deviations. Unfortunately, this leads to non-positive semi-definite kernels, which can not be inverted, and thus can not be used with our algorithm.

Wendland [65] discusses a method for constructing PSD kernels with compact support using

piecewise polynomials. One applicable example is the function

$$k(r) = \begin{cases} (1-r)^{j+1}\left((j+1)r+1\right) & 0 \leq r \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{5.20}$$

where $j = \lfloor \frac{D}{2} \rfloor + 2$ and D is the maximum number of dimensions for which the kernel is PSD. Obviously this can be scaled to any size.

This function behaves sufficiently similarly to a radial basis kernel to provide an efficient approximation to the radial basis kernel. This family of functions can also be used to approximate other common covariance functions such as the Materne covariance.

Of course, compactly supported kernels such as (5.20) will ignore data that is sufficiently far from its centroid, so data sets that contain very high dynamic range data will be poorly modeled with this kernel. The sensors we work with have limited dynamic range, and the data themselves are typically limited in their dynamic rang so it seems unlikely that this will present a problem for us.

If we need further computational gains, we can also use iterative methods such as Conjugate Gradient. Conjugate gradient is well suited to exploiting sparsity in the kernel matrix, and because it is iterative, we can also trade off accuracy for computational efficiency.

Figure 5.2 shows an example inference using (5.20), with a relative large support to illustrate the effect of smoothing, since the density of data is so low.

5.3.4 Non-uniform wind velocity

(5.4) and other equations above all assume that the wind velocity is uniform: that is, the (Gaussian) distribution that the wind velocity is drawn from does not change over time or space. If the rate of change (over time or space, e.g. the characteristic length scale) of the wind velocity is slow, this might be a reasonable assumption for an epoch. In the event that the characteristic time scale of the wind velocity is short, then we can retain our Gaussian assumption while allowing for the wind velocity to change by making the wind velocity itself drawn from a Gaussian process. The model then becomes a hierarchical model.

None the less, just as drawing the wind velocity from a Gaussian distribution induces a Gaussian process in (5.4), it also does so if the Gaussian is infinite dimensional (e.g. a Gaussian process). This is because the sum of correlated Gaussian random variables is itself a Gaussian. The mean velocity over a time interval is just the integral over all the possible paths based on the velocity distribution:

$$\overline{u}([t,\tau],X) = \int_t^\tau u(t',X')dt', \tag{5.21}$$

If the wind velocity's characteristic length scale (in time and space) is significantly larger than the characteristic length scale of the pollution concentrations, however, then we can assume that the mean velocity at a given location over time is approximately the mean velocity over time and space between two points. In that case, then we can simply use a Gaussian process to interpolate between wind measurements, and then just utilize the interpolated value as the mean wind velocity.

We believe that this simplification is reasonable since it is unlikely that we will have wind velocity measurements over the scale of a few minutes and in spacial density of tens of meters.

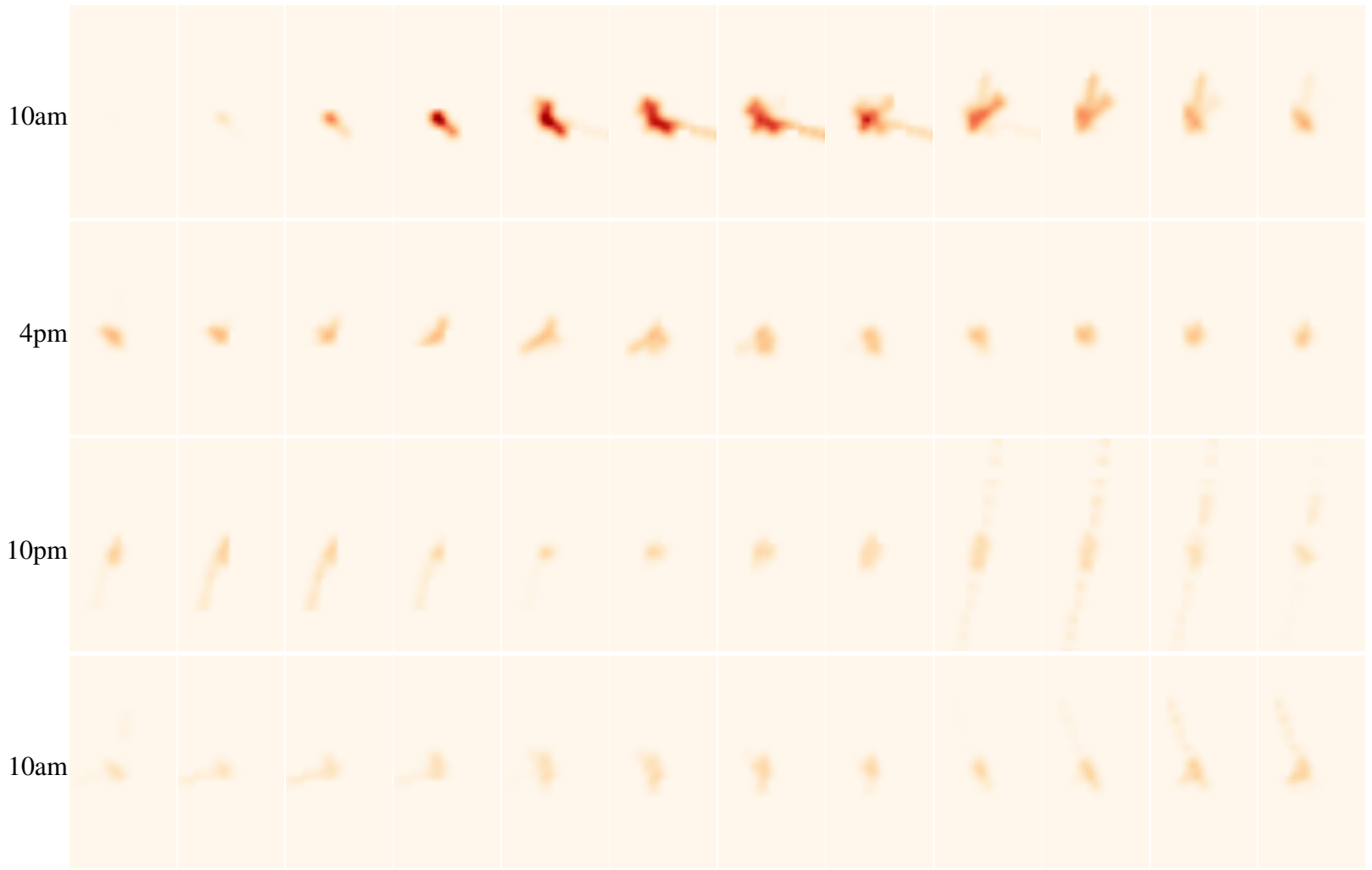


Figure 5.2: Inference of the CO concentration using two badges from the West Oakland study over a 24 hour period, using a sparse kernel (5.20) with a scale of 2 hours and 100m. The dimensions of each rectangle is 675m by 1375m. The lines in the early morning hours are an artifact of the GPS, rather than movement by the participants.

Chapter 6

Sensors and pollution characterization

In this chapter, we will discuss the basic theory of sensing, and how we have measured the basic properties of the sensors we used for this study. We explain terminology, describe and characterize the physical mechanisms we have used to understand our sensors, and the sensors themselves. In later chapters, particularly Chapters 5, 8 and 7, we will build on these characterizations to calibrate our model to real world data.

6.1 Precision and accuracy

Since manufacturers build mobile phones in very high volumes, typically with thin margins, cost will be a central issue. The question we need to ask is ”how can we build a societal scale sensor using mobile phones, and using affordable technologies, so as not to significantly impact the desirability of the devices to the consumer.”

Since the incremental cost of integrating a sensing into a phone is almost entirely in the sensing mechanism itself (including the mechanism for providing airflow), the cost of the sensor and airflow mechanism (given that they are suitably sized) will ultimately determine whether manufacturers will consider participating in building a societal scale sensor.

Unfortunately, there is often a trade off between the size of a sensor, and the precision and/or accuracy of that sensor. Since precision and accuracy ultimately determine the usefulness of data in a sensor or sensor system, the precision and accuracy of a sensing system become the fulcrum on which a sensing system pivots between feasibility and usefulness.

Precision and accuracy are two related but orthogonal concepts in sensing. Roughly speaking precision refers to the amount of information in a signal, (e.g. the number of bits that we need

to capture a signal), and accuracy refers to the correctness of those bits (e.g. how closely those bits reflect ground truth). We can also understand precision in the context of ”repeatability,” or the extent to which different measurements under the same conditions produce the same results. These two concepts of precision, e.g. measurement resolution vs. repeatability, are essentially the same.

We can have a very precise, but inaccurate signal, in which the signal is very steady (in the short term), but is improperly calibrated (or has drifted over time). The sensor thus provides ”bad” data, in which our information about ground truth is incorrect. We can also have an accurate signal that is imprecise, in which the noise of a signal makes it difficult to get a lot of information about ground truth, despite proper calibration. In the extreme, we might have a sensor that only gives us a single bit of information about ground truth, although we have a very high confidence that that single bit is correct.

Obviously, if the drift of a sensor can be bounded (as it often can), then we can make a sensor more accurate by reducing its precision, until the drift of the sensor falls within the bounds of the sensor's precision. Conversely, if we can characterize or somehow manage the drift of a sensor, then we can increase the precision of the sensor. Typically, however, we view drift as a longer term and somewhat deterministic process, and noise as a short term, and fundamentally random process, so our approach to mitigating problems with accuracy vs. precision are necessarily different.

Cross-sensitivity to other environmental factors such as humidity, temperature, pressure and other compounds in the air also contribute to inaccuracy. If we can characterize these cross-sensitivities, using calibration and modeling, then we can increase our accuracy without reducing our precision.

6.2 Noise model

Although the noise in a sensor depends on the particular sensor, and the underlying technology, many sensors are dominated by thermal noise, and possibly, to some extent, shot noise (for extremely sensitive sensors). Shot noise is most accurately modeled as Poisson distribution. For less sensitive sensors (as in our cheap sensors), however, a Gaussian noise distribution for both thermal and shot noise is sufficient, since a Gaussian distribution provides a good approximation of a Poisson distribution for large values of λ, the shape and scale parameter.

In order to study our devices and algorithms, we have built two testing chambers for exposing our devices to poison gasses at precise concentrations and under controlled humidity. Both chambers operate by diluting a known concentration of one or more poison gasses with clean, dry

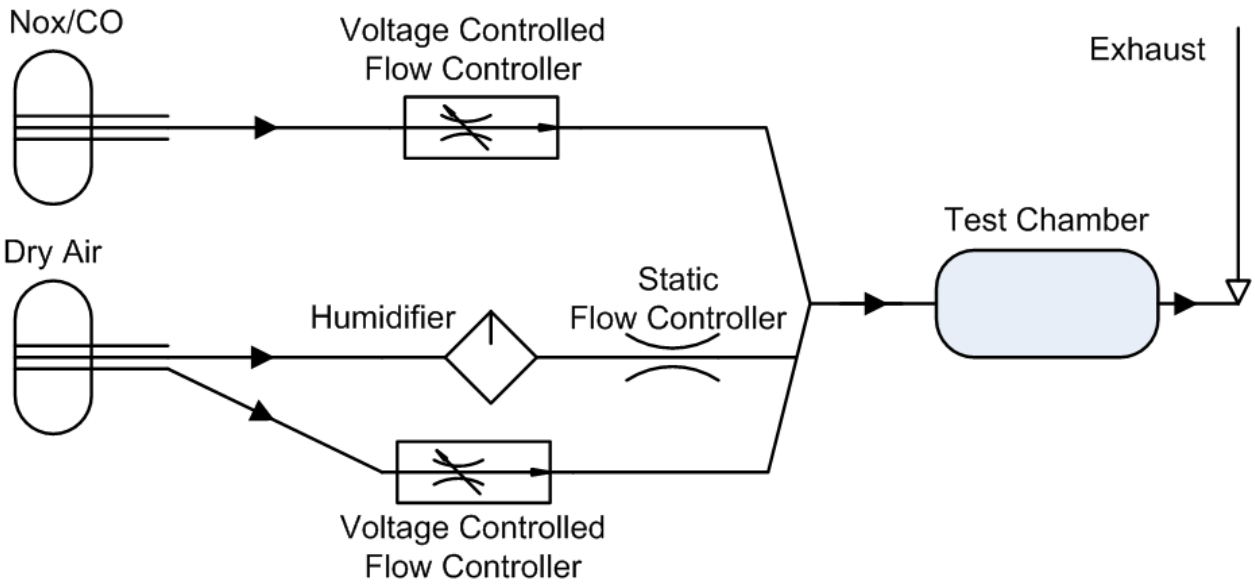


Figure 6.1: Dry air is humidified and mixed with a poison gas using voltage-controlled flow controllers to allow precise control of a poison gas in a test chamber at a stable humidity.

air, and clean humidified (at close to 100% humidity) air. To alter the concentration of the poison gas while keeping the humidity constant, the flow rate of the humidified air is kept constant while the rate of flow of the poison gas and clean air are simultaneously adjusted to keep the same total flow rate. Similarly, to adjust the humidity of the air while keeping the concentration of poison gas constant, the flow rates of the clean dry air and poison gas are both adjusted proportionately, while the rate of flow of the humidified air is adjusted to keep the total rate of flow the same (Figure 6.1).

The first chamber is a small chamber that permits the concentration in the chamber to change rapidly, thus allowing us to characterize temporal characteristics of the sensors such as response time (Figure 6.2). The second chamber is a larger chamber that has room for several large sensors, allowing us to characterize and experiment with larger sensors (Figure‘3.1) [1].

With these setups we can easily measure the noise in our sensor and circuitry, simply by placing the sensor in the chamber, flowing clean air, and examining the readings. As we can see in Figure 6.3, our sensor board output is very close to Gaussian.

[1]Thank you to Virginia Tiege for building and photographing this chamber

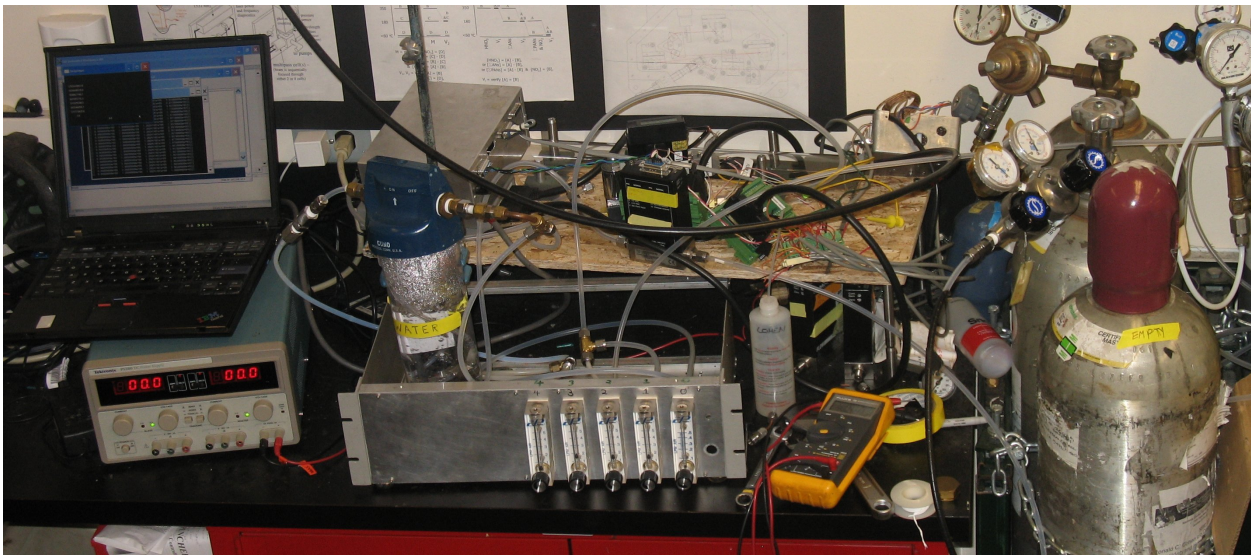


Figure 6.2: The test chamber we use to calibrate and test response while carefully controlling poisonous gas concentration and humidity. The small chamber size allows us to quickly vary the concentration of a gas in the chamber.

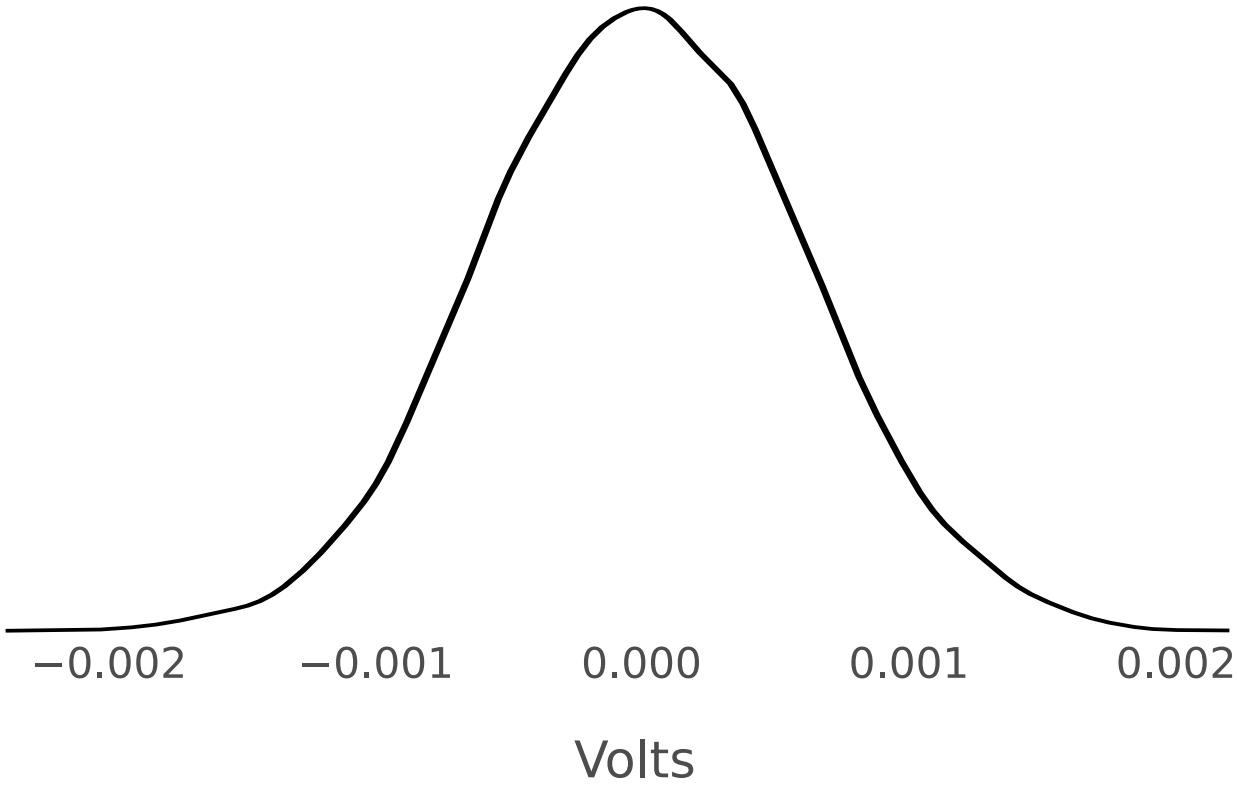


Figure 6.3: A kernel density estimate of readings taken while the sensor was exposed to clean air. This distribution closely approximates a Gaussian distribution.

6.3 Sensor bandwidth and system spectrum

The rate of response of a sensor to changes in the sensed medium will determine the maximum amount information that a sensor can reveal about the underlying system. Viewing the sensed medium as a signal, the sensor will act as a low-pass filter on the signal, integrating out high-frequency fluctuations. Thus, the sensor will determine the maximum frequency to which we can increase the sampling rate of the sensor and gain information about the system, e.g. the Nyquist rate [52].

We can measure the bandwidth of our sensor by examining the step-response of our sensor. Since the step response function is the integral of the Dirac-delta function, we can measure the step-response function, take the derivative, do a Laplace transform (in this case an FFT), and then examine the resulting frequency domain function to see the maximum frequency component of the system.

Analytically, for a signal $y(t)$ at time t, we can represent the step response as the convolution of the impulse response h and the step function H.

$$y(t) = h(t) * H = \int_{-\infty}^{\infty} h(\tau)H(t-\tau)d\tau = \int_{-\infty}^{t} h(t)dt \tag{6.1}$$

Thus we can take the derivative of $y(t)$ to get the impulse response $h(t)$.

The Laplace transform of the impulse response gives us the transfer function $Y(f)$ of the system [24].

$$Y(f) = \int_{-\infty}^{\infty} h(t)e^{-i2\pi ft}dt \tag{6.2}$$

In practice we calculate the Laplace transform using a Fast Fourier Transform on the impulse response vector. Figure 6.4 shows the step response of the sensor to a 25ppm step. Figure 6.5 shows the corresponding transfer function, revealing significant noise in the mid and high frequencies. Examining the transfer function more closely, we see that the meaningful signal is restricted to about 1/8 Hz bandwidth, so we can use a low pass filter to eliminate the noise at higher frequencies. Figure 6.4 and Figure 6.6 also show the results of using a Butterworth IIR filter with a 1dB maximum attenuation at 1/8 Hz and a 10 dB minimum attenuation at 1/4 Hz.

Also note that after filtering the signal in this way, the maximum sampling rate that we need to reconstruct this signal without aliasing is 1/4 Hz, or approximately every 4 seconds [52].

We can also examine the spectrum of samples taken from sensors “in the wild,” to get an idea of the rate of evolution of the underlying system. Examining the data from six of the sensors from

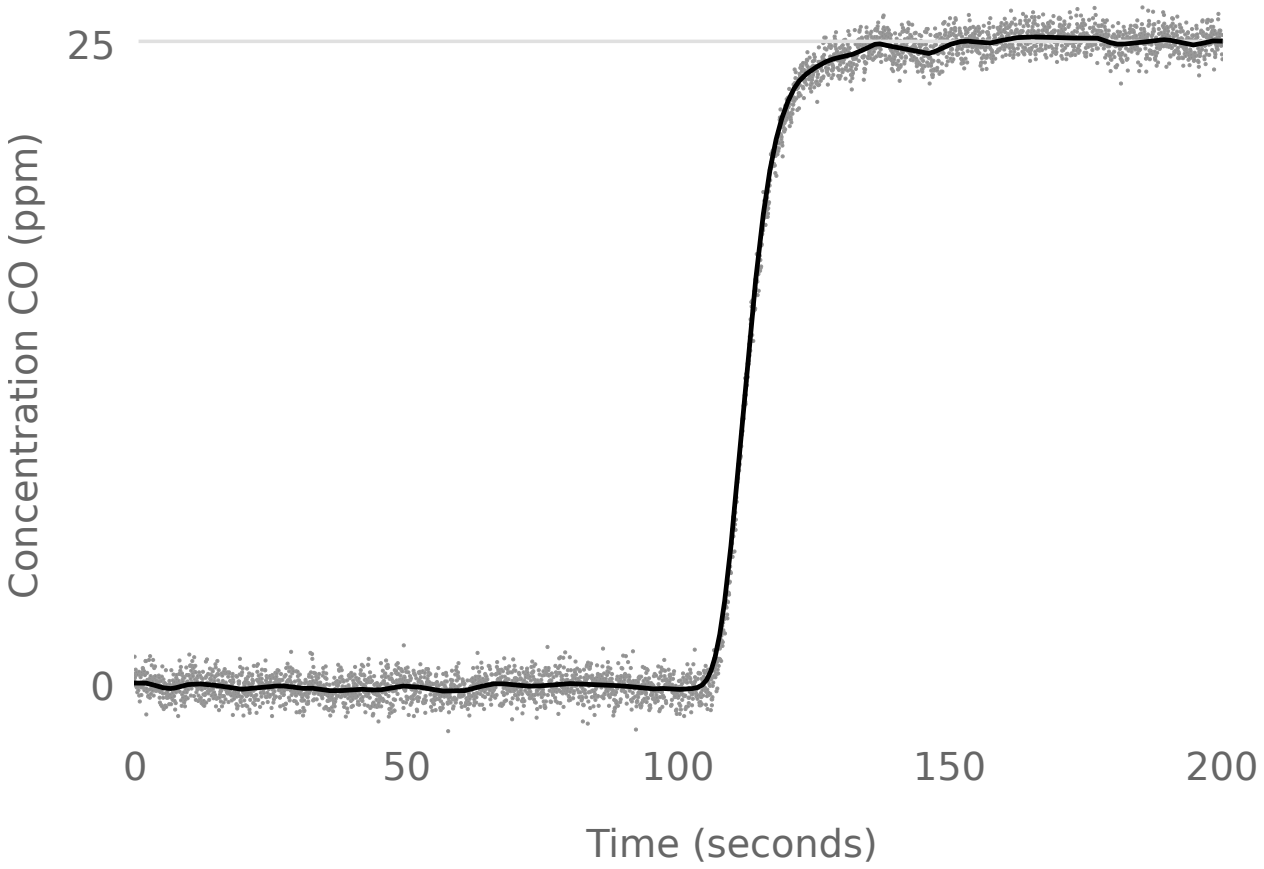


Figure 6.4: The step response of the MicroCell CO sensor, responding to a step from 0ppm to 25ppm, before and after low-pass filtering. The signal was filtered with a Butterworth IIR filter with maximum 1db passband until 1/8Hz and at least 10db of attenuation past 1/4Hz.

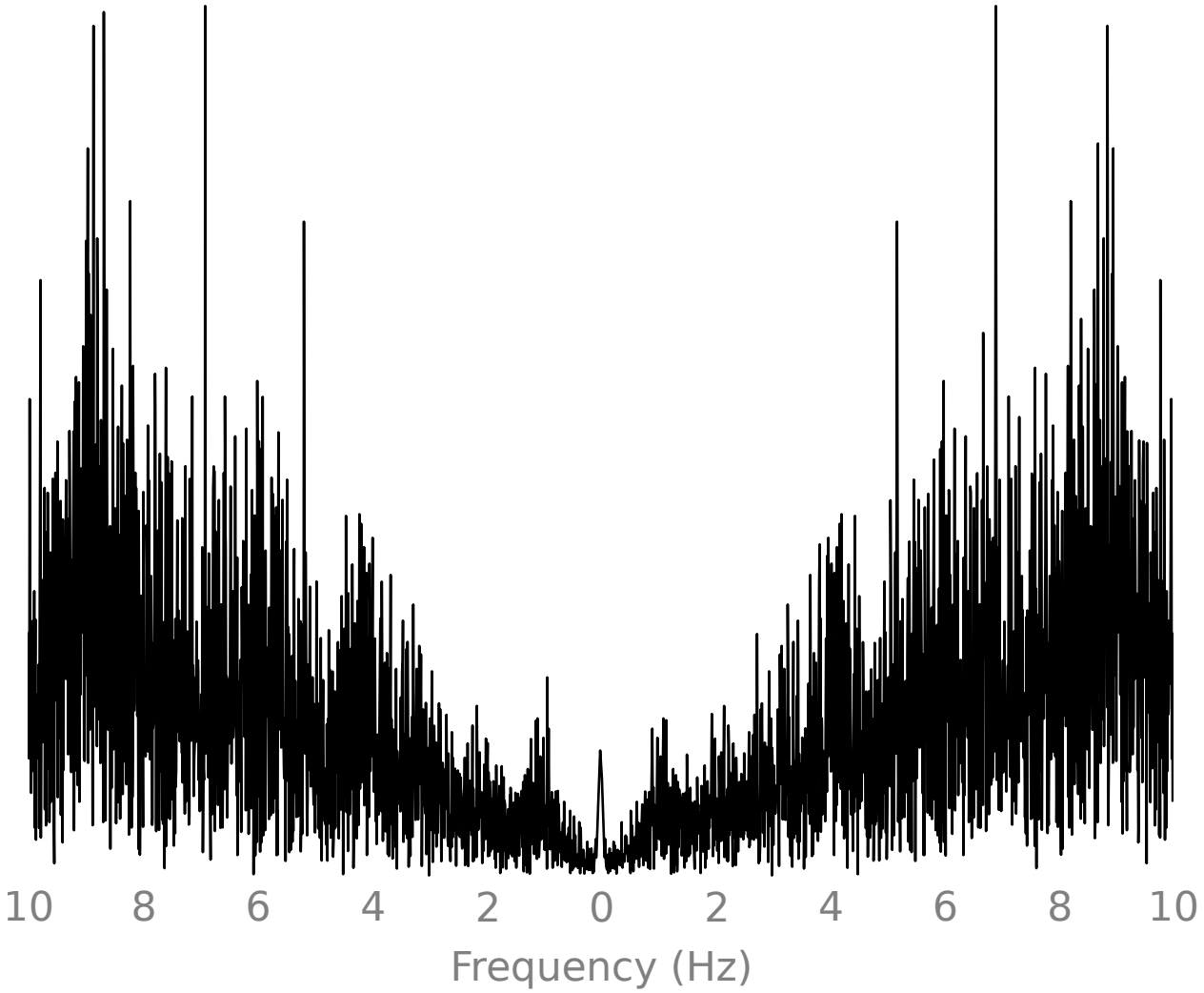


Figure 6.5: The transfer function of the MicroCell CO sensor, corresponding to the step response in Figure 6.4. The noise power in the higher frequencies dominates the signal at the low frequencies.

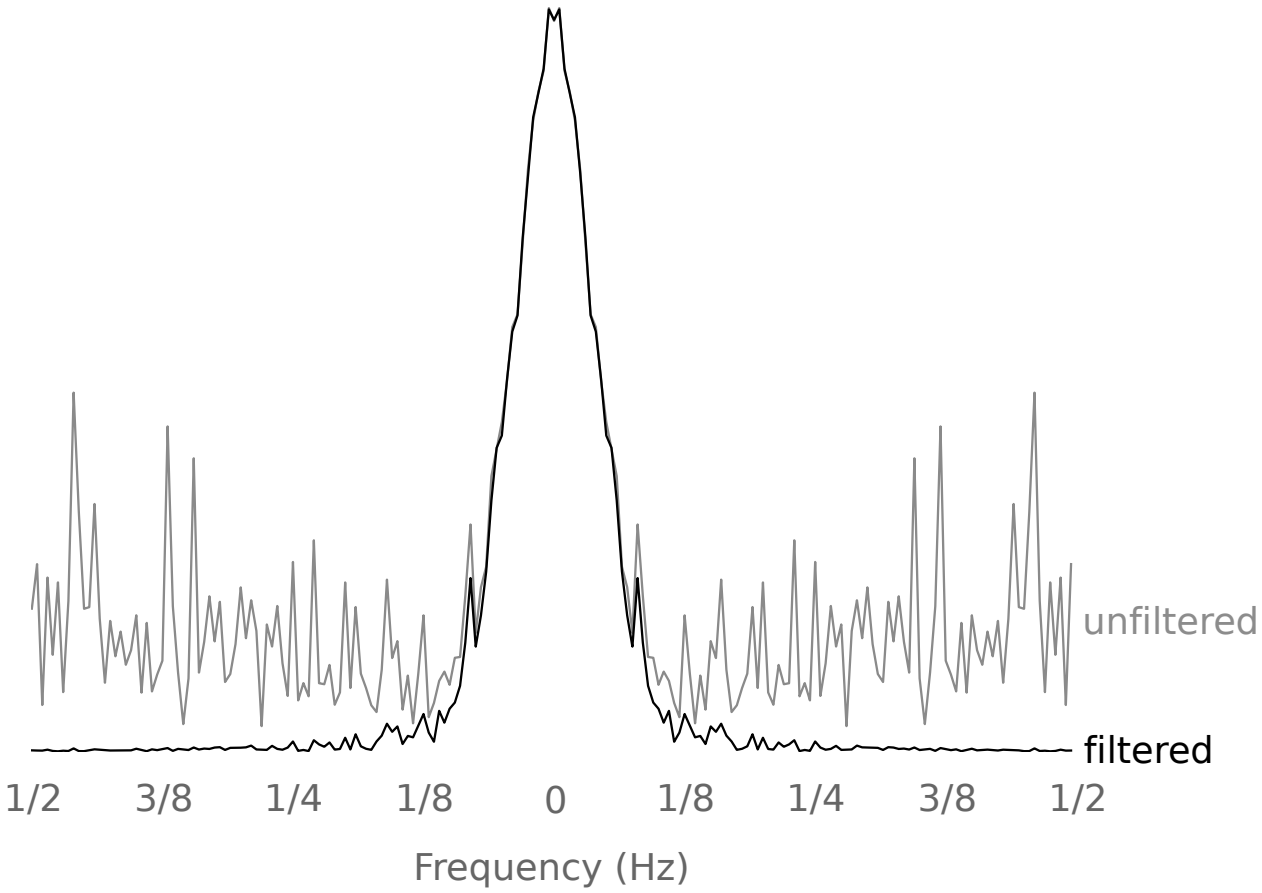


Figure 6.6: The transfer function of the MicroCell CO sensor, before and after filtering with a Butterworth IIR filter with maximum 1db passband attenuation until 1/8Hz and at least 10db of attenuation past 1/4Hz. Since there is no meaningful signal at frequencies greater than about 1/8Hz, we don't loose any information by filtering.

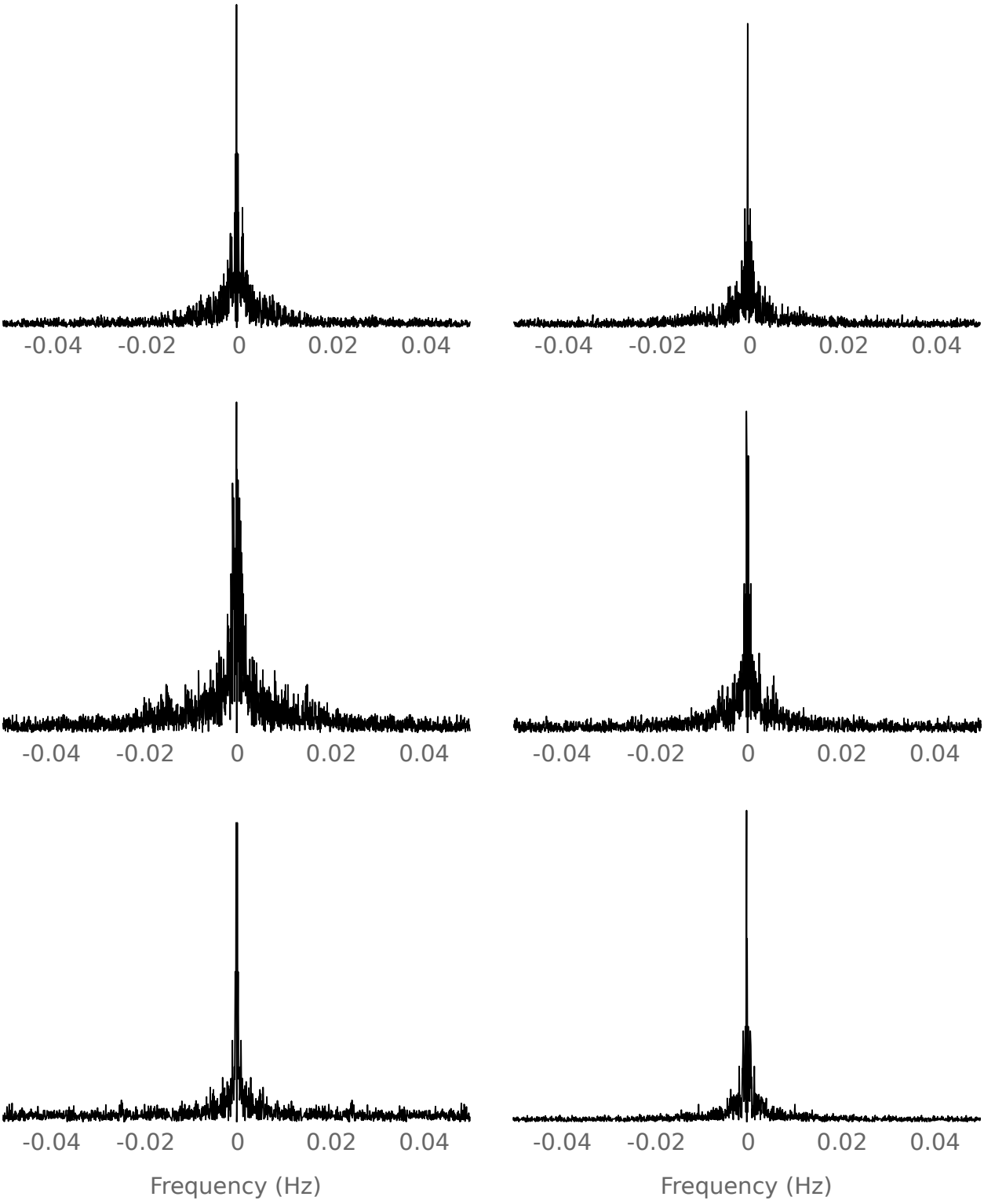


Figure 6.7: The spectrum calculated from data gathered during our Ghana sensing campaign. Each plot corresponds to data from a different sensor over an extended period of time. The mass of the spectrum is concentrated in the low frequencies with a signal bandwidth of about 0.01-0.02 Hz, and almost no mass beyond about 0.04 Hz. This suggests that concentration of pollution in this environment tends to evolve over the course of several minutes.

our campaign in Ghana (Figure 6.7), we can see that the bandwidth of the signal is typically less than about 0.04Hz, corresponding to a period of about 25 seconds, and the vast majority of the spectrum is within a bandwidth of about 0.02Hz, corresponding to about 50 seconds. We will consider this later, when we calibrate our models.

Chapter 7

Automatic sensor calibration

Section 5 briefly explained the concepts of accuracy, precision and bandwidth. From that discussion, it should be apparent that if the calibration of sensors drifts over time, then the sensing system will become inaccurate. Without a reasonable mechanism for manually calibrating sensors in a societal scale sensor network, we must rely on the data themselves to calculate the calibration of sensors in the system without explicit user action.

In this section we outline the background concepts pertaining to sensor calibration, and two mechanisms for automatic calibration of sensors. The first mechanism, called CaliBree, was developed by researchers at Dartmouth, and the second mechanism is our own research.

7.1 Gain and Bias

In the context of sensing, gain refers to the amount of change in the sensed value with respect to the change in the underlying, true value, and bias refers to the sensed value when the true value is zero. For linear sensors, gain and bias simply refer to the slope of line of ADC readings versus true values, and bias refers to the y-intercept of the line (Figure 7.1).

Because gain error is, by definition, a percentage of the sensed value, it does not impact readings very much unless the reading is large. Bias is therefore a more important consideration in a system that measures ambient pollution: typically, ambient pollution has a very low concentration, so in order to achieve any realistic accuracy we must reduce the bias of our readings

---

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7]

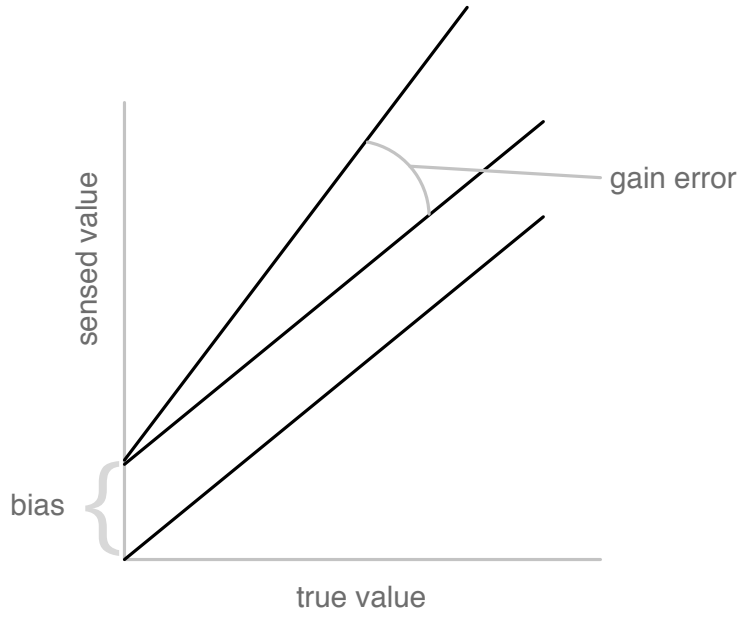


Figure 7.1: Linear sensors can be characterized by the additive bias and multiplicative gain.

to a minimum. Furthermore, the gain drift of a sensor can often be bounded to a small percentage, further limiting the impact of mis-calibrated gain. For example, the gain of the electrochemical sensors that we use can drift no more than 5% per year, according to the data sheet [69].

With that said, we will first discuss our techniques for automatic calibration of sensor bias in a network of mobile sensors, and then we will expand the techniques slightly to calibrate for gain as well.

7.2 Problem Setup

In a network of sensor phones, with potentially millions of devices, we cannot reasonably expect people to invest their time or money into calibrating the sensors. The human must be out of the loop, so to speak. The system needs to exploit characteristics of the data to provide accurate information to the end users of the data without requiring action by the user.

Fortunately, people tend to congregate, and when they do, our sensor phones are likely to be close to one another. In that case, they should sense roughly the same value. If they don't, then there is likely a mis-calibration in one or both of the sensors that results in the sensors reading drastically different values (although some variation is expected, since the sensors are not in exactly

the same location).

7.3 CaliBree

CaliBree is a self calibration algorithm for mobile sensor networks based on the idea of calibrating during very close rendezvous, in which the sensors are close enough to be exposed to a negligibly different environment [39]. CaliBree uses a weighted average of the difference between ”ground-truth nodes,” which have been recently calibrated, and mobile sensor nodes. The weight that gets assigned to the readings from a particular given rendezvous can be adjusted according to probability correspondence between the ground truth reading and the mobile node, the authors do not specify exactly how that correspondence is determined.

Although CaliBree makes an excellent start towards providing a calibration mechanism for a mobile phone-based sensing system, it makes some assumptions that we consider unnecessarily limiting. First, the assumption that the there are well-calibrated ”ground-truth” nodes located throughout the environment is not scalable, particularly for developing countries. Although it is likely that a few well calibrated sensors will exist in the environment, its is unlikely that many users will rendezvous with these sensors close enough for the mobile and ground-truth sensors to be in the same sensing environment.

Secondly, restricting the rendezvous region to an area in which the readings of the two sensors should be virtually identical discards useful information. Since readings from nearby locations should be correlated even if they are not close enough to be identical, we should be able use information from nearby locations, weighted by some distance metric.

Similarly, even if samples are not taken close enough in time to one another to be considered exactly the same, the correlation between the samples should be related to the time difference between the samples. We should be able to exploit that correlation.

Finally, CaliBree does not offer a good way to integrate heterogeneous sensors with varying accuracies, drift, etc. We imagine that many different phone models, with different sensors, plus higher accuracy instruments located in the environment will all provide readings to our societal scale instrument.

7.4 Automatic Calibration using Gaussian Processes

With these limitations in mind, we will now develop an algorithm based on Gaussian process regression, for inferring the bias of each sensor. This algorithm is capable of incorporate high precision and accuracy "ground-truth" nodes if they are available, and can exploit their increased accuracy to increase the accuracy of the system, but in no way relies on these "ground-truth" nodes to work. Instead, these algorithms rely on the opportunistic rendezvous between any sensors in the system, and the proximity of the rendezvous determines the amount of correlation that the system expects between two readings, and hence determines the amount of information that can be used for the sensors mutual calibration.

The algorithm relies on two important assumptions. First, that the bias of the sensors in the system is distributed as a Gaussian. Secondly, the algorithm relies on the calibration event being done with information about all of the sensors. Whereas CaliBree is completely distributed, with calibration happening on the mobile sensor itself, our algorithm works in the cloud, using global information about all of the sensors in the system.

7.5 Modeling Bias

In Section 5.3.1, we discussed basic Gaussian process regression. Rasmussen and Williams also provide an excellent introduction to Gaussian Processes, and their book is available online for free (although we feel that the hard copy is well worth the money for those who can afford it) [47].

Rasmussen and Williams describe a mechanism for incorporating explicit basis functions into our Gaussian Process. We adapt this mechanism to modeling bias by augmenting our x vector with indicators that indicate which sensor took a particular reading, and using a basis function corresponding to the indicator to model the bias of that sensor.

Following from Rasmussen and Williams, to model bias, we first augment our x vector:

$$x = \begin{bmatrix} x_d \\ x_b \end{bmatrix},$$

where x_d is the previous x vector, representing the a single sample location, and x_b is a vector of indicators which are 1 if the sample came for sensor i, and 0 otherwise.

Next we change our model slightly as follows:

$$g(x) = f(x) + h(x)^T \beta$$

where f is our original model, and $h(x) = x_b$. Now, if we assume a Gaussian prior over β:

$$\beta \sim \mathcal{N}(b, B)$$

then we can incorporate β, b and B into our model [44]:

$$g(x) \sim \mathcal{GP}(h(x)^T b, k(x, x') + h(x)^T B h(x'))$$

Plugging this into the prediction equations (5.17) and (5.18), we get

$$\bar{g}(X_*) = \bar{f}(X_*) + R^T \hat{\beta} \tag{7.1}$$

$$\text{cov}(g_*) = \text{cov}(f_*) + R^T (B^{-1} + H K_y^{-1} H^T)^{-1} R \tag{7.2}$$

where $K_y = K + \sigma_n^2 I$, $\hat{\beta} = (B^{-1} + H K_y^{-1} H^T)^{-1}(H K_y^{-1} y + B^{-1} b)$, and $R = H_* - H K_y^{-1} K_*$. $\bar{\beta}$ can be seen as trading off between the prior and the data.

In the event we have information about the bias of any of the sensors, we can choose b and B to reflect that information. This might be the case if we have already calibrated them in the factory or lab, or if previous iterations of this algorithm have made an inference about the values of b and B. In that case, we use the equations above directly.

Next, here is how we can specify ground truth sensors: we can set b with the calibration information for each particular sensor, and set B with the variance on our prior information about b. Furthermore, if we have very precise sensors, then we can set the variance on the readings from those sensors to reflect that precision. We set the variance for those readings by setting the corresponding entries on the diagonal that corresponding to readings from the each sensor to the variance of the noise for that sensor.

In the case where we have no apriori knowledge about each β, we can let the covariance of the prior on β go to infinity, and hence we have that $B^{-1} \to 0$ (thus specifying a diffuse prior). In that case we can simplify to

$$\text{cov}(g_*) = \text{cov}(f_*) + R^T (H K_y^{-1} H^T)^{-1} R \tag{7.3}$$

with $\hat{\beta} = (H K_y^{-1} H^T)^{-1}(H K_y^{-1} y)$. Thus b has no influence on our prediction, as we would expect from a diffuse prior [47].

Even if we don't have any "ground-truth" sensors in our system at all, we can still approximate the most likely bias value for all of our sensors. Assuming that the bias of the sensors are IID, then the maximum likelihood value of the "true" bias is the empirical mean of the inferred $\hat{\beta}$ values

$$\hat{\beta}_{ML} = \frac{1}{s} \sum_{i=1}^{s} \beta_i.$$

This algorithm adds little computational overhead to our regression algorithm. In fact, the computational complexity of the algorithm is dominated by the need to invert K_y, which must be done anyway in order to get the variance estimates for the regular regression problem.

In the worst case, inverting K_y is an $O(n^3)$ problem, although sparsity in K_y makes this tend towards $O(n)$. Fortunately, all of the data do not need to be used to construct K_y if we are only doing the calibration inference. Instead, we can just include the sample points which are within the scale of our kernel from one another, since the kernel has compact support. Depending on the geographic and temporal density of the data, and on the particular kernel used, this could lead to significant computational savings.

7.6 Inferring Gain

So far we have only discussed inferring bias, but we previously noted that gain is also sometimes an important contributor to sensor error. The algorithm presented above does not treat gain explicitly, a minor adjustment will allow it to do exactly that.

First, we should note that we want to calibrate gain and bias using different sensor readings. To calculate gain, we want to use the largest values possible, since large readings will be the least influenced by biased sensor readings, and thermal noise (in terms of a percentage of the sensed value). On the other hand, we want to use small readings to calculate bias, since small readings will be least influenced by gain error.

Thus, we can choose a cutoff, below which readings are used to calculate bias, and above which readings are used to calculate gain. Bias is calculated as above.

Next, we note that whereas in the case of bias, the per-sensor differential is additive, in the case of gain, the per-sensor differential is multiplicative.

$$g_{gain}(x) = f(x) \cdot h(x)^T \beta$$

Unfortunately, if we place a prior distribution over beta, then the distribution of g_{gain} will no longer be Gaussian. We could optimize the likelihood of g_{gain} numerically, but that would be extremely computationally intensive, since each sensors in the system will introduce a new dimension to the problem, thus making the problem exponential in the number of sensors.

Note, however, that if we consider $\log(g_{gain}(x))$ then the problem once again becomes additive:

$$\begin{aligned}\log(g_{gain}(x)) &= \log(f(x) \cdot h(x)^T\beta) \\ &= \log(f(x)) + \log(h(x)^T\beta)\end{aligned} \tag{7.4}$$

If we consider, for analytic convenience, $f(x)$ and $h(x)$ to be drawn from log-normal distributions, rather than normal distributions, then the problem devolves to the same problem as the bias problem.

This may appear to be a contrived solution to our problem. In practice, however, the mis-specification of the distribution may not turn out to make a significant difference on the gain inference. Since this (cheap) trick makes the gain inference computationally tractable (whereas numerical optimization over thousands or millions of dimensions is not), it is worth exploring.

7.7 Cross-sensitivity

If the sensor is cross-sensitive to other compounds in the atmosphere, then we need to be able to compensate for these cross-sensitivities. Gas sensors are often cross sensitive to temperature, humidity and air-pressure, in addition to other chemicals which react with the sensor.

If the response to the cross-sensitive factors does not change over time, then we can simply use a mean function, $m(x)$, to first normalize the response according to the concentration of the cross-sensitive factors, and then proceed as normal [47].

$$f(x) = \mathcal{GP}(m(x), k(x, x')) \tag{7.5}$$

We then predict using

$$\hat{f}_* = m(X_*) + K(X_*, X)K_y^{-1}(y - m(X)). \tag{7.6}$$

If we want to exclude the impact of the cross-sensitivities in our prediction, then Equation 7.6 becomes

$$\hat{f}_* = K(X_*, X)K_y^{-1}(y - m(X)). \tag{7.7}$$

If the cross-sensitivity function has a multiplicative factor, we can similarly "warp" the inference by first dividing out the factor from the observations, and then multiplying it back in to the inference.

$$f(x) = m(x)\mathcal{GP}(0, k(x, x')) \tag{7.8}$$

We then predict using

$$\hat{f}_* = m(X_*)K(X_*, X)K_y^{-1}(y/m(X)). \tag{7.9}$$

Again, if we want to exclude the impact of the cross-sensitivities in our prediction, then Equation 7.9 becomes

$$\hat{f}_* = K(X_*, X)K_y^{-1}(y/m(X)). \tag{7.10}$$

In the event that the sensor's response to the cross sensitivities changes over time, we can augmenting our x vector and expand our definition of $h(x)$. We can add any analytic function of the cross-sensitive factors into our calibration algorithm, and parametrize it using β, as we did for the bias.

$$h(x)^T\beta = f_1(x)\beta_1 + f_2(x)\beta_2 + \cdots \tag{7.11}$$

This could be used, for example, to fit an additive polynomial function of humidity or temperature. Obviously, we could also do this in the log domain for multiplicative factors.

7.8 Experimental results

This section illustrates the algorithms described in the previous sections in a variety of real and simulated contexts.

7.8.1 Laboratory and simulation results

Evaluating these algorithms under realistic circumstances is difficult, since they rely on a high density of mobile devices in some locations, and depend to some extents on user mobility patterns. The first step is therefor to observe the algorithms' effectiveness in laboratory and simulation scenarios.

Using the test chamber shown in Figure 6.2, we exposed four sensors to Carbon Monoxide over 25 minutes at concentration increments of approximately 25ppm, for five minutes at each, and keeping the relative humidity at approximately 50%. The raw readings from the sensors are show in Figure 7.2. Each of the four sensors has a slightly different gain and bias. We can see in Figure 7.3 that our algorithm effectively removes the bias when examining small values. In Figure 7.4, we can see that our algorithm also effectively removes gain error once the bias has been removed.

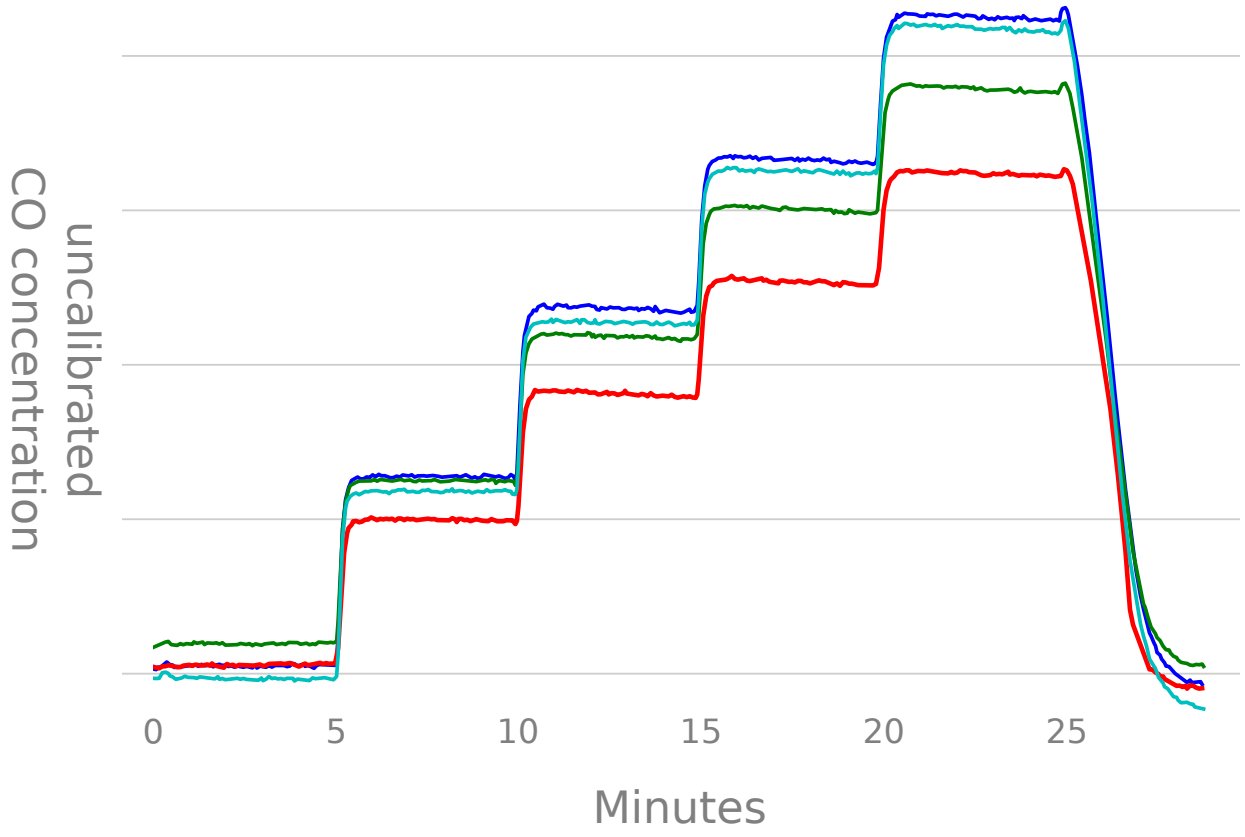


Figure 7.2: Raw data from four sensors in a test chamber showing 25ppm increments every 5 minutes, under semi-controlled humidity and temperature. The gain and bias of these sensors are uncalibrated.

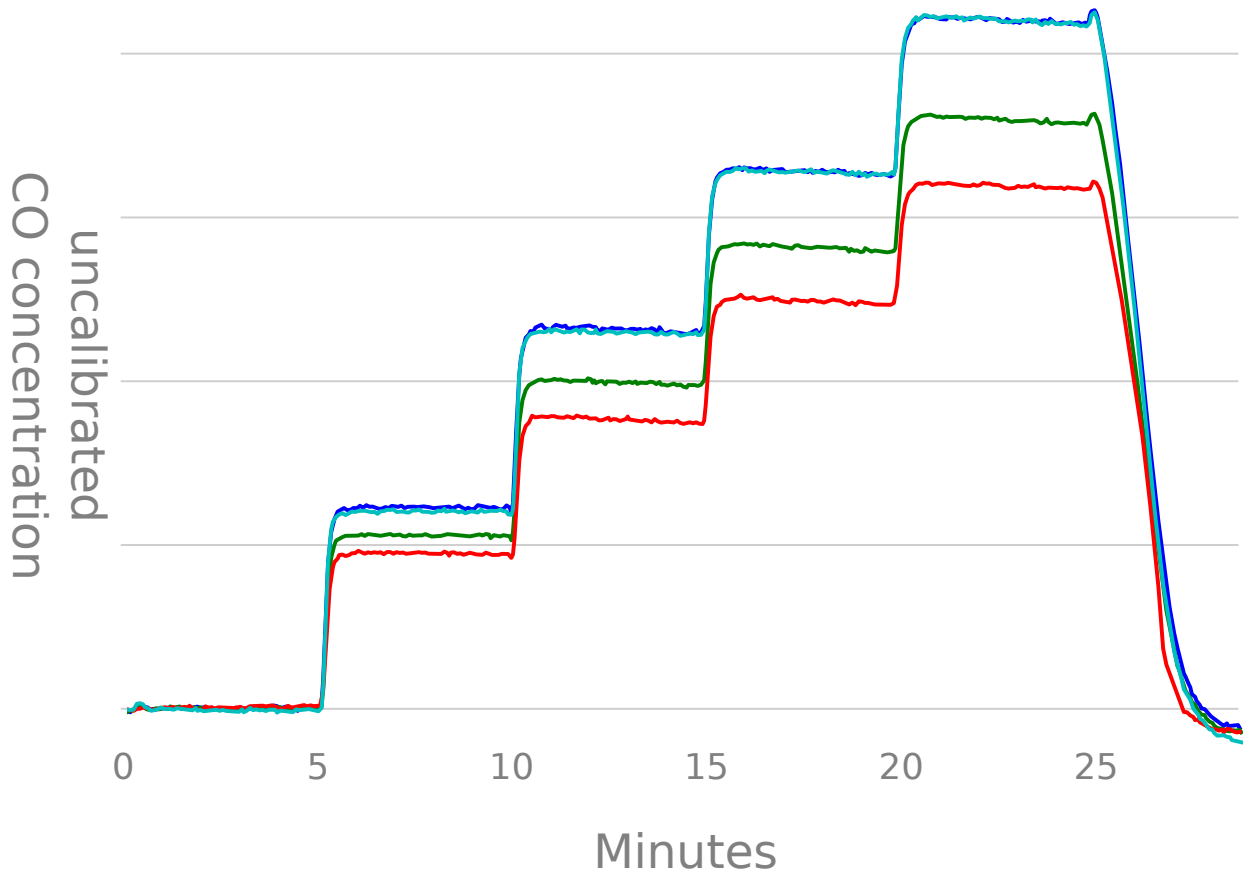


Figure 7.3: The same data from the same four sensors. The bias of these sensors has been automatically calibrated.

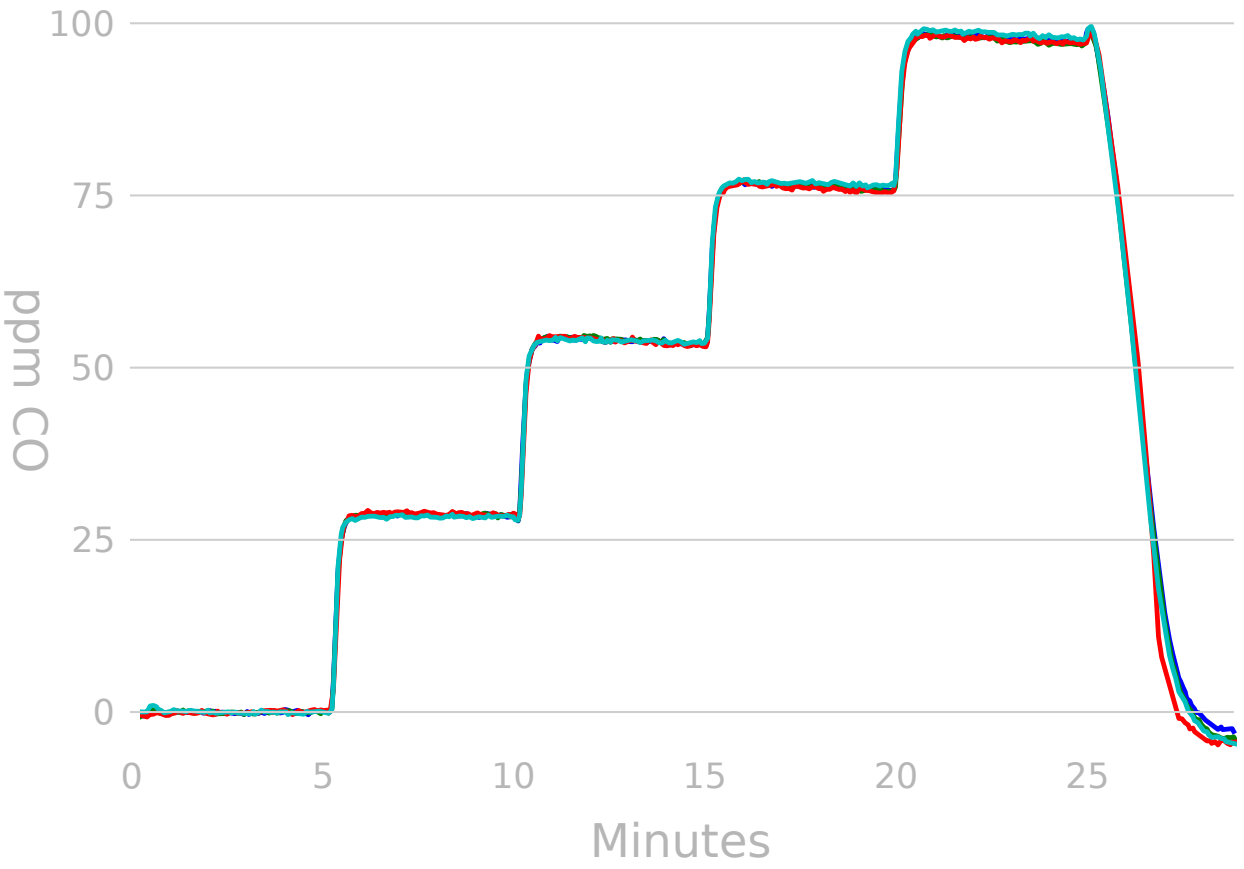


Figure 7.4: The same data from the same four sensors. The gain and bias of these sensors have been automatically calibrated.

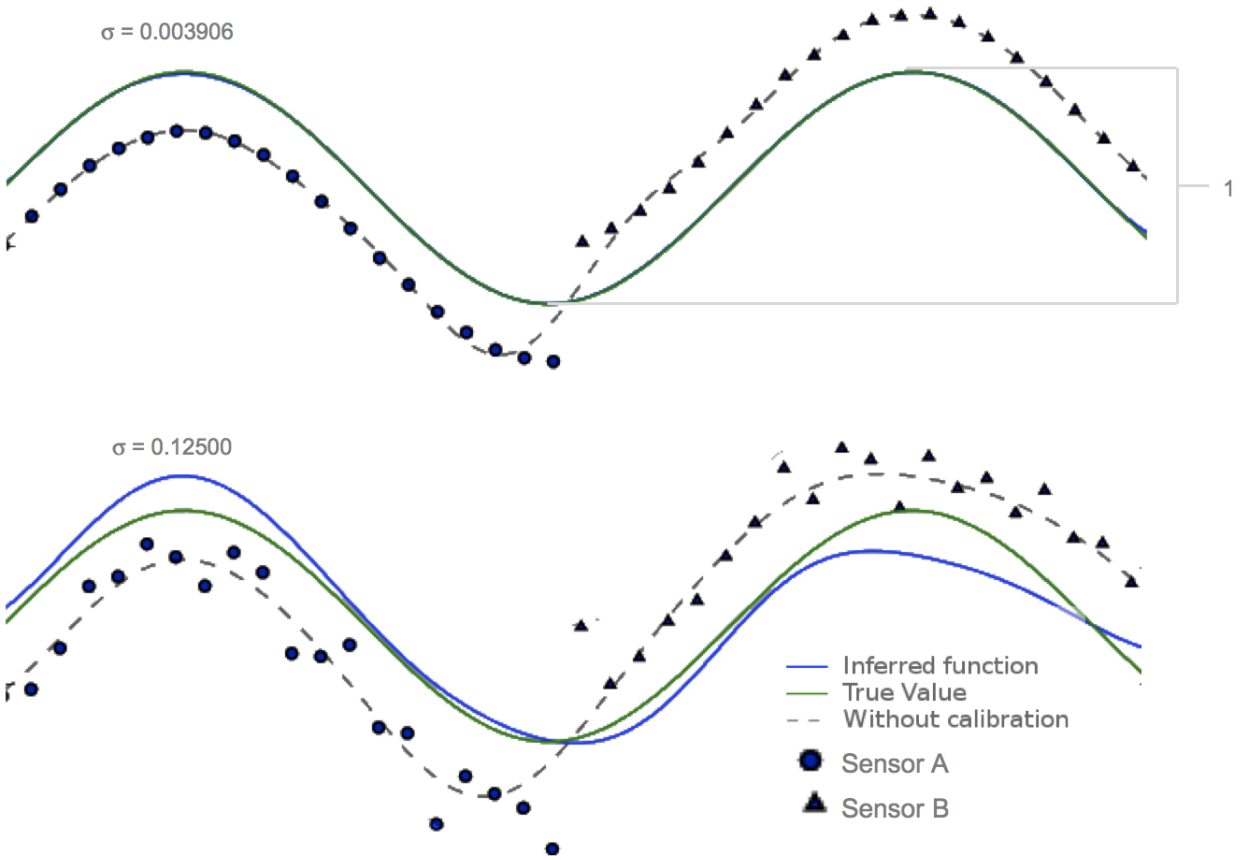
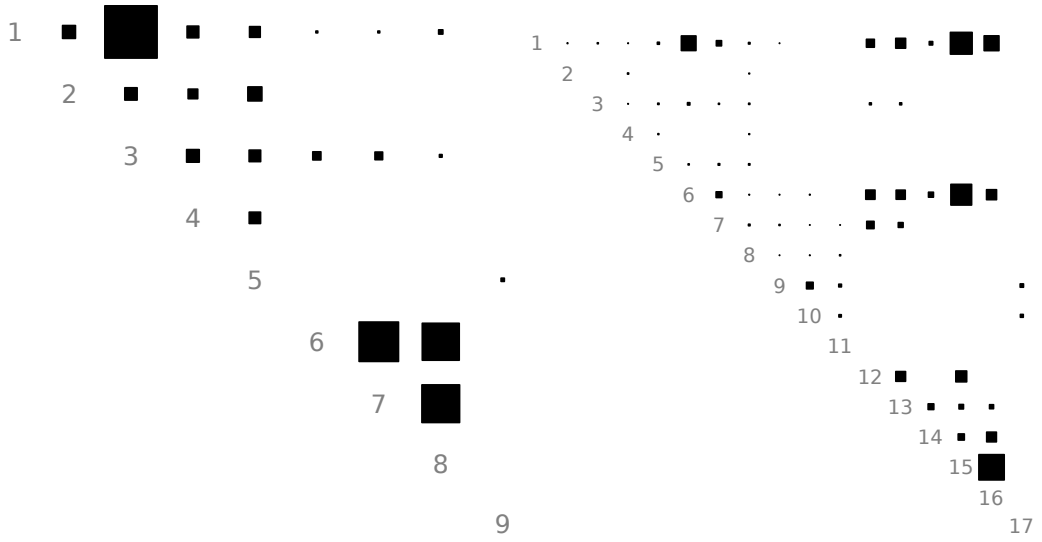


Figure 7.5: A simple simulation in which a function is sample by two biased sensors, under low and moderately high noise conditions. Even with moderate noise, our algorithms can infer the bias of both sensors with reasonable accuracy.

One thing to note in Figure 7.4 is that when the sensors return to zero, they overshoot, and that the steps are not evenly spaced at 25ppm increments. This is probably due to their slow response to the sudden change in humidity and temperature versus the humidity and temperature of the ambient air. We can correct for this effect, and it will not impact the effectiveness of our algorithms, but it will require further sensor characterization. Regardless, the gain and bias calibration algorithm is clearly effective in making results between sensors repeatable.

Of course, one of the important features of our algorithm is to allow for automatic calibration even when sensors are not exactly co- located. Figure 7.5 shows a simple experiment in which a function is sampled by two different sensors that are never co-located. The algorithm is able to infer their bias with reasonable accuracy, even with the signal to noise ratio is relatively low.



(a) Ghana (b) Oakland

Figure 7.6: The rendezvous matrices for the Ghana and West Oakland studies. The area of each square indicates the number of rendezvous between two sensors. A rendezvous takes place when two sensors are in the same 20 meter by 20 meter by 2 minute space-time box.

7.8.2 Field trial results

Users in both the Accra study (Section 4.1.1) and the West Oakland badge study (Section 4.3.1) co-locate with one another reasonably often. In Figure 7.6, we can see the relative frequency at which users are in the same 20 meter by 20 meter by 2 minute space-time box.

Unfortunately, the COTS data-logging CO sensors that we used for the Accra study do significant pre-processing of the data before logging them. In addition to integrating the data over about 10 seconds (which is not a big deal, since we plan to do the same), any readings less than about 3 ppm are truncated to 0 ppm. The sensor presumably does this to avoid exposing minor mis-calibrations. Unfortunately, this makes the data unusable for studying our automatic calibration algorithms.

While the noise issues that plagued the sensor badges will interfere with our calibrations (since we don't have a mechanism for rejecting outlier data), the relatively noise-free second half of the West Oakland data set has sufficient rendezvous to calibrate several of the sensors in that part of the study.

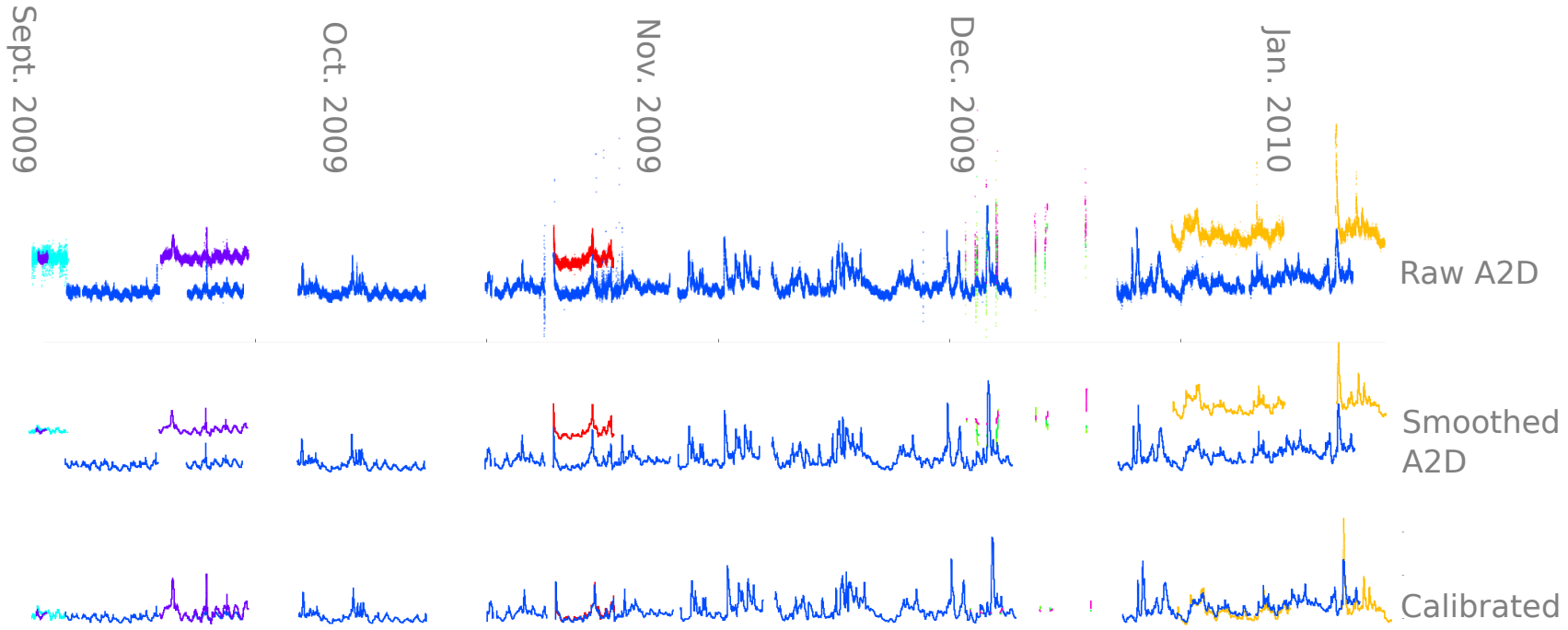


Figure 7.7: The raw data from the badges' analog-to-digital converter is show vs. time on the top. The bottom shows the same data after each sensor is calibrated using our algorithm. The data on the bottom are also smoothed using the low pass filter described in Section 6.3 so that the data from the different sensors are discernible. The middle plot shows the uncalibrated data, after filtering, for comparison.

In Figure 7.7, we can see the raw A2D converter data, smoothed A2D converter data (using the filter described in Section 6.3), and the data after automatic calibration. Since there is no absolute reference sensor in the data, the y-axis scale is unlabeled: we only know the relative calibration unless one of the sensors has absolute calibration information.

Figure 7.7 illustrates that our calibration algorithm is quite effective at inferring the relative bias between two sensors and compensating for it (Figure 7.8). The gain calibration is effective as well, but might be better if there were more rendezvous in higher PPM areas.

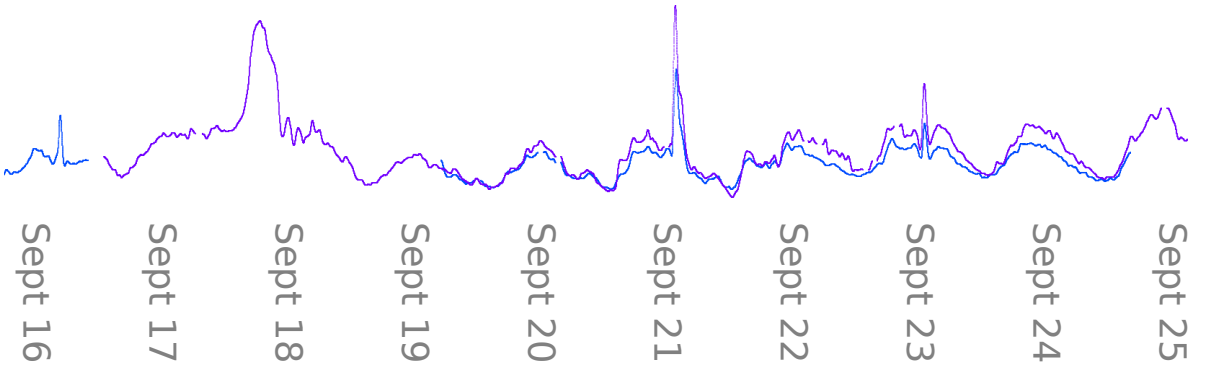
Unfortunately, these sensors were not calibrated in the lab before they were deployed, so that needs to be done before we can quantify their correctness. It is clear, however from the the plots that the calibration has been quite effective.

7.9 Density and time between rendezvous

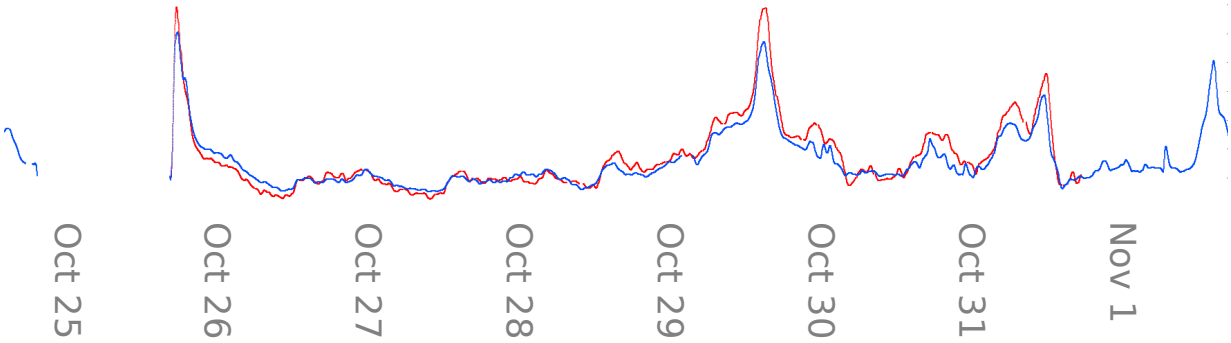
Finally, we need to understand the distribution of time between rendezvous. As the density of users increases, we would expect the time between rendezvous to decrease, since each user will have more opportunities to rendezvous with others as more users move in a given area. The maximum time between rendezvous is of primary concern, in this case, since we want to understand how much our sensors can drift between calibration events.

To determine the impact of density on the rendezvous frequency, we used data from the Reality Mining data set (see Chapter 8) for a complete description of these data). We examined the time between rendezvous for successively larger subsets of the study participants. For each number of users $n \in (2, 97)$ (three users did not participate sufficiently to be included in the simulation), we selected 1000 subsets of the users of size n, and examined the maximum time to rendezvous for all the users in each subset. We then calculated the maximum, 99th percentile, 95th percentile and median subset, of the 1000 subsets. Thus, the maximum subset approximates the worst case maximum time between rendezvous for subsets of size n. The median represents the average case maximum time between rendezvous for subsets of size n.

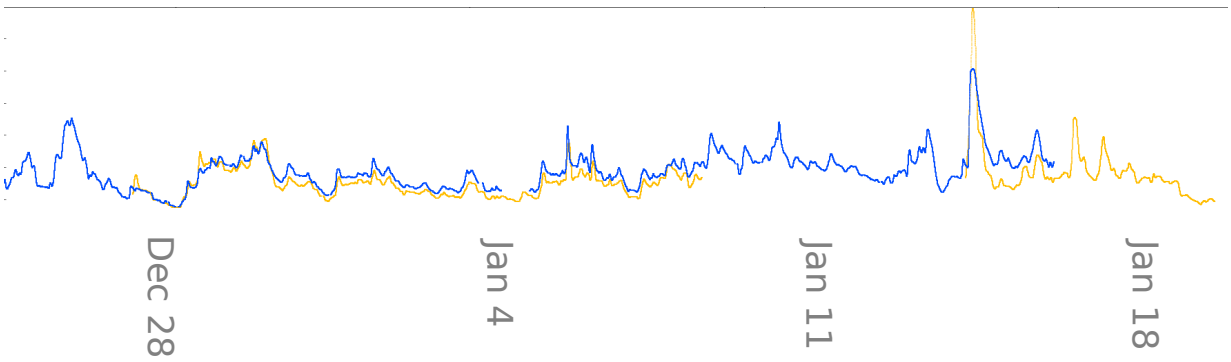
Figure 7.9 shows the results our simulation. We can see that in the worst case, for a small group of sensors, one of the sensors in the group would have to wait over 200 days for a calibration event. Considering that the drift of the CO sensors we characterized in Chapter 6 is less than 5% of the signal per year [69], even this extreme case is probably acceptable. Interestingly, once the size of the subset reaches around 5 users, the median of our samples of the maximum time until rendezvous does not decrease significantly with more users. This is probably because some



(a)



(b)



(c)

Figure 7.8: Three close ups showing the calibration of different sensors. The bias calibration is very accurate. The gain calibration is reasonable, but would probably be better if there were more rendezvous in high PPM areas.

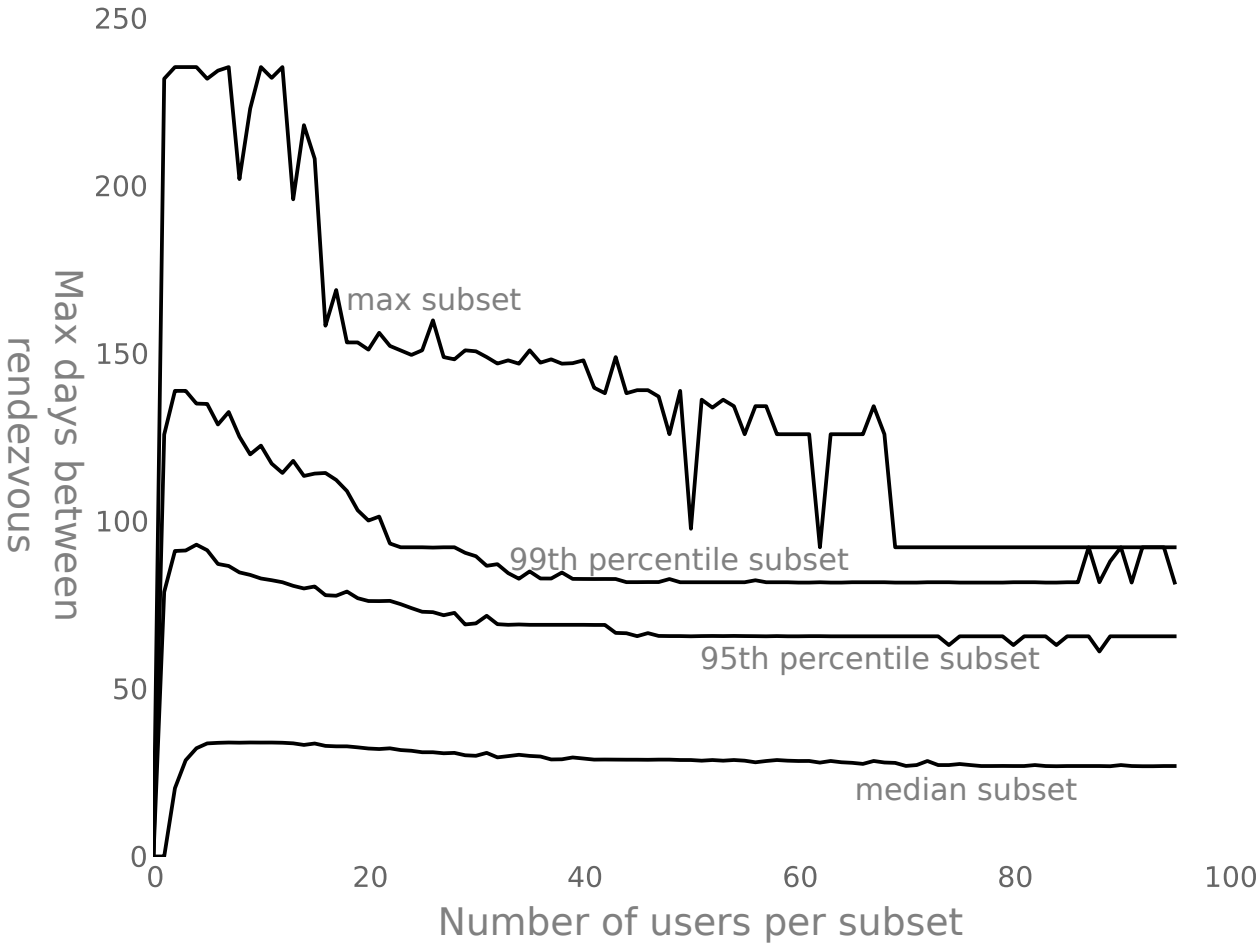


Figure 7.9: The maximum number days between rendezvous as the number of users in the area increases. The quantiles show the maximum time between rendezvous during the study for most reclusive subset of users out of 1000 randomly selected subsets of a given size (marked max subset), the 99th percentile subset (e.g. 99% of the subsets had a shorter maximum interval between rendezvous), the 95th percentile subset and the median subset.

users tend to have only a few associates, so adding more users to the subset does not significantly increase the chances of them rendezvousing with someone else.

7.10 Conclusion

Although our algorithms remain sensitive to bad data, since we have not yet developed a mechanism for rejecting bad data, they are quite effective when the data are relatively clean. We have reason to believe that rendezvous will be frequent enough to keep sensors well calibrated. By maintaining good calibration of our sensors (e.g. accuracy), we can also increase the precision of the system in locations that are densely sampled. This is the subject of Chapter 8.

Chapter 8

Increasing precision using super-sampling

Once we have calibrated our sensors, then we can exploit dense sampling by averaging. Gaussian process regression does this naturally. The diagonal of the estimated covariance function gives the variance of our estimate at each location. We can use the square root of this variance as our estimate of precision at that location. The rest of this section deals with our empirical and simulated validation of this principle, using our test chamber and electrochemical sensors.

8.1 Mobility and Rendezvous

Both our automatic calibration algorithms, and super-sampling depend on users being in close proximity of one another. We know from experience that this, indeed, happens quite often, but in order to quantify how often we will be able to calibrate, and to what extent we can expect users to congregate, we examined records from the Reality Mining database.

The Reality Mining project used software on mobile phones to track a wide variety of user behaviors. The database consists of 100 participants for approximately 9 months in 2004–2005. 75 of the participants were students or faculty at the MIT Media Laboratory, and 25 were at the incoming students at the adjacent MIT Sloan business school. The software collected statistics on incoming and outgoing communication, but more importantly, the Bluetooth and GSM beacons detected by the phone [14].

Parts of this chapter were previously published in “mScience: Sensing, Computing and Dissemination” [7] and “Increasing the precision of mobile sensing systems through supersampling” [29]

Because of the limited range of Bluetooth beacons, we can use this database to infer when the participants were in close proximity with one another. These data provide a larger sample of users, who more closely represent our target population of every-day mobile phone users. Since the users all work in close proximity to one another, they provide a mechanism for studying the effects of increasing the user density in a smaller geographic area.

In Figure 8.1, each square represents the number of times a user rendezvoused with another user in the study, ordered by the total number of rendezvous for a given user. Although some users clearly interact with others more frequently in this cohort, most of the users interact at least slightly with most other users. The most "connected" user rendezvoused 6,256 times during the study, and the most "connected" pair of users rendezvoused 585 times with each other. The median number of rendezvous by a user was 862, and the median number of rendezvous between two users was one.

These data suggest that there will be plenty of opportunities for automatic calibration, as described in Chapter 7. They also raise the question of how many people congregate at one time, and for how long people rendezvous.

In Figure 8.2, for each user in the study, we see the number of other users in the study with whom the user was in close proximity simultaneously, versus time. The maximum number of co-located users was 15. Interestingly, many of the users frequently congregated with many other users.

Figure 8.3 shows the fraction of the time that a given user is in close proximity to a given number of other participants, ranked by time in proximity to at least one other participant. The 5th, 50th and 95th percentile users were at 0.0039, .052 and .16 respectively. This figure shows that, in fact, many participants spent a significant amount of time in proximity to other participants, and even the least "connected" of the participants rendezvous occasionally.

These data suggest that not only will we be able to calibrate users automatically to one another, but that there will be significant opportunity to increase the precision of our system by super-sampling. These data do not however, reveal *where* we will be able to increase our precision.

Figure 8.4 and Figure 8.5 shows that in the Ghana and Oakland, users tend to congregate in "hot-spot" locations, as we would intuitively suspect. The top part of the plot shows the times of rendezvous for each sensors. We can see that the rendezvous were spread throughout the study period, rather than due to a lucky coincidence or abnormal event.

These results are consistent with the findings of Lee et al., who analyzed around 150 GPS traces of 66 users in 5 separate locations. Lee et al. found that people moved between clusters

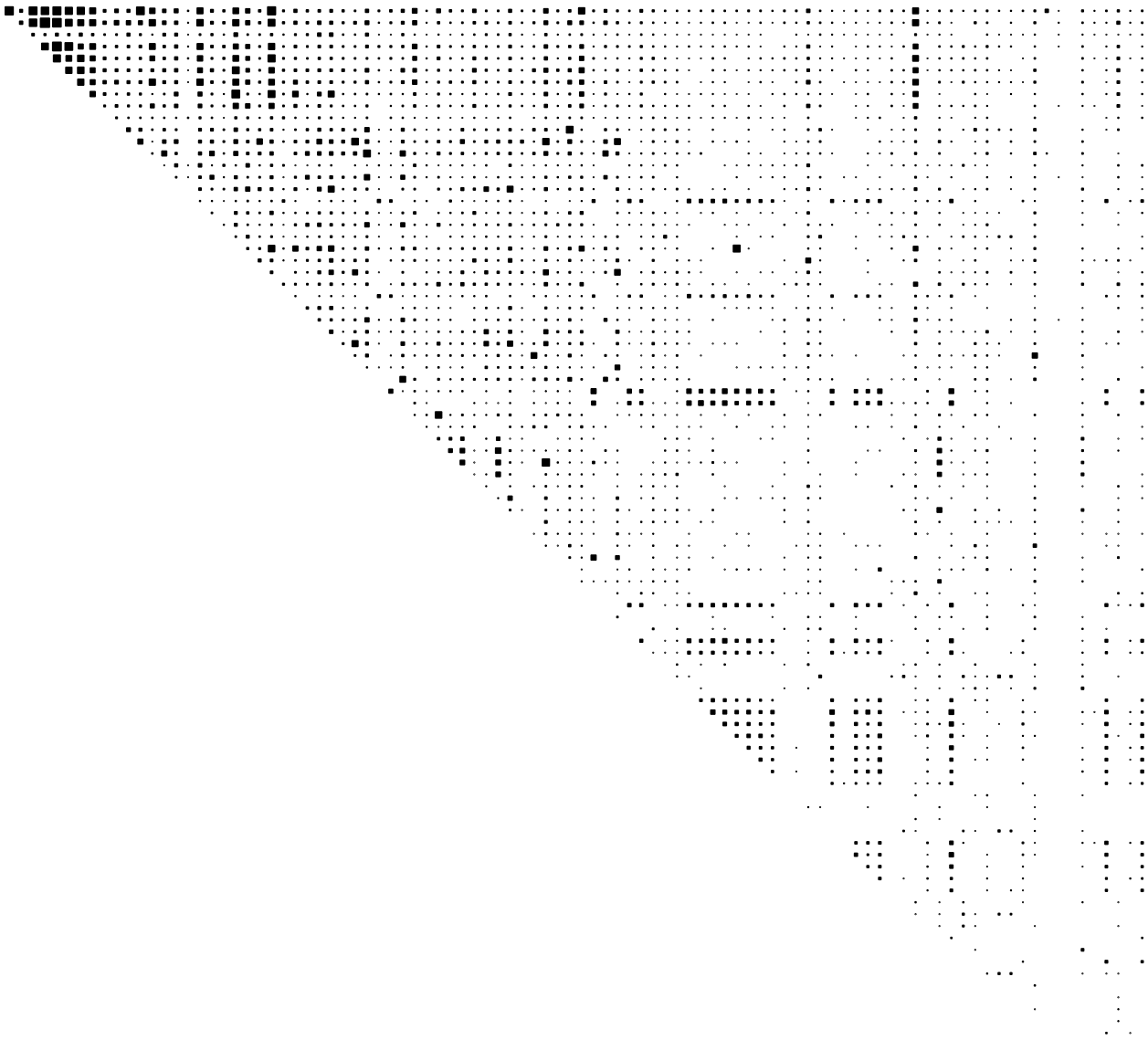


Figure 8.1: Each square in this plot represents the number of times a user rendezvoused with another user in the study, ordered by the total number of rendezvous for a given user. Although some users clearly interact with others more frequently in this cohort, most of the users interact at least slightly with most other users. Only 466 user pairs did not interact with each other at all, or about 11% of the pairs. Furthermore, several of these pairs come from users who did not participate very long in the study.

Figure 8.2: The number of users within proximity of a given user, for each user in the study. The scale of each line can represent between 0 and 15 simultaneous rendezvous.

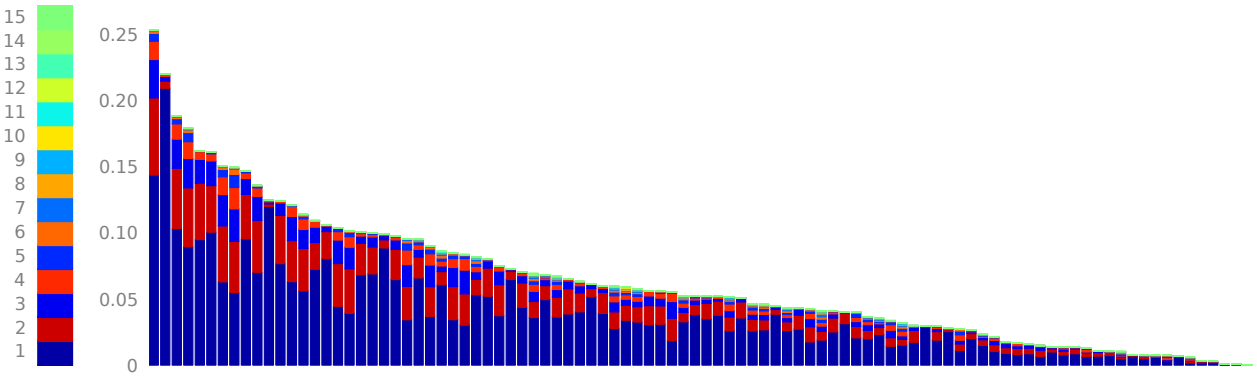


Figure 8.3: The fraction of the time that a given user is in close proximity to a given number of other participants, ranked by time in proximity to at least one other participant. The 5th, 50th and 95th percentile users were at 0.0039, .052 and .16 respectively.

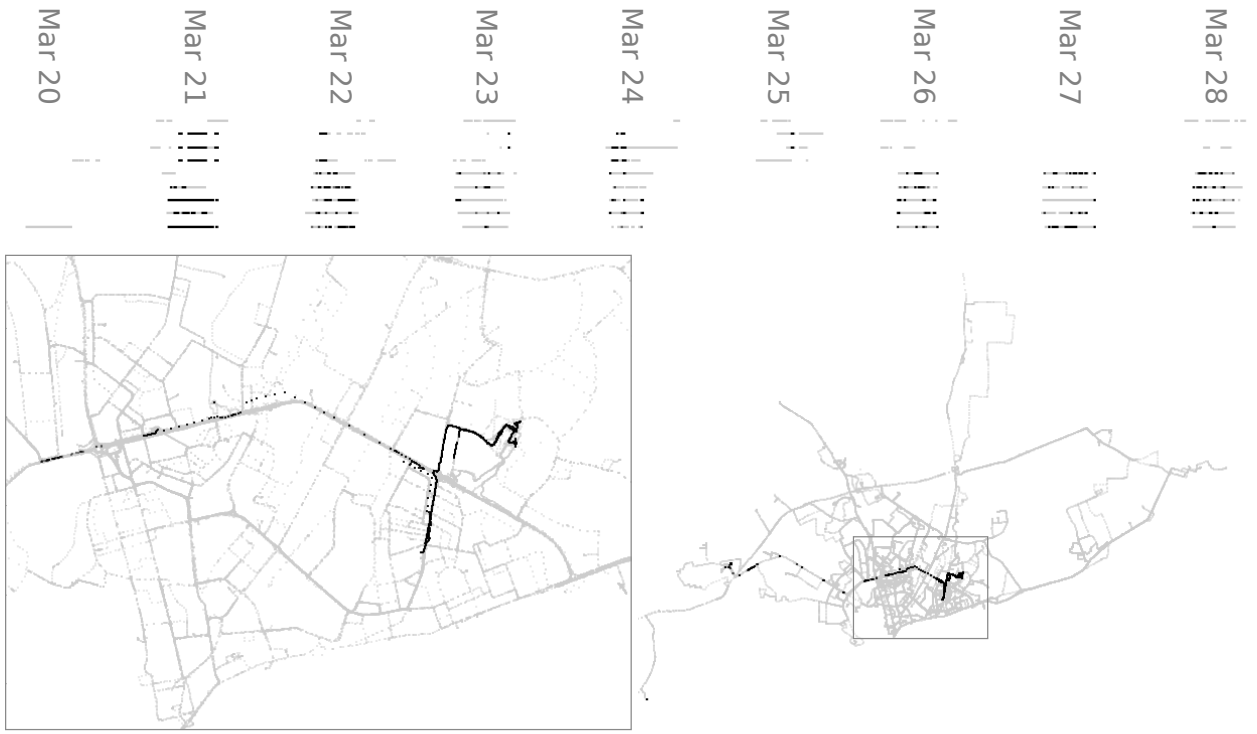


Figure 8.4: Rendezvous points (black) super-imposed on all locations (gray) in the Ghana study. The sample times for each sensor are displayed at the top, again with rendezvous points in black and all sample times in gray. Rendezvous happen in "hot-spot" locations, rather than distributed throughout the map. They are not isolated to a few lucky coincidences, but distributed throughout the study.

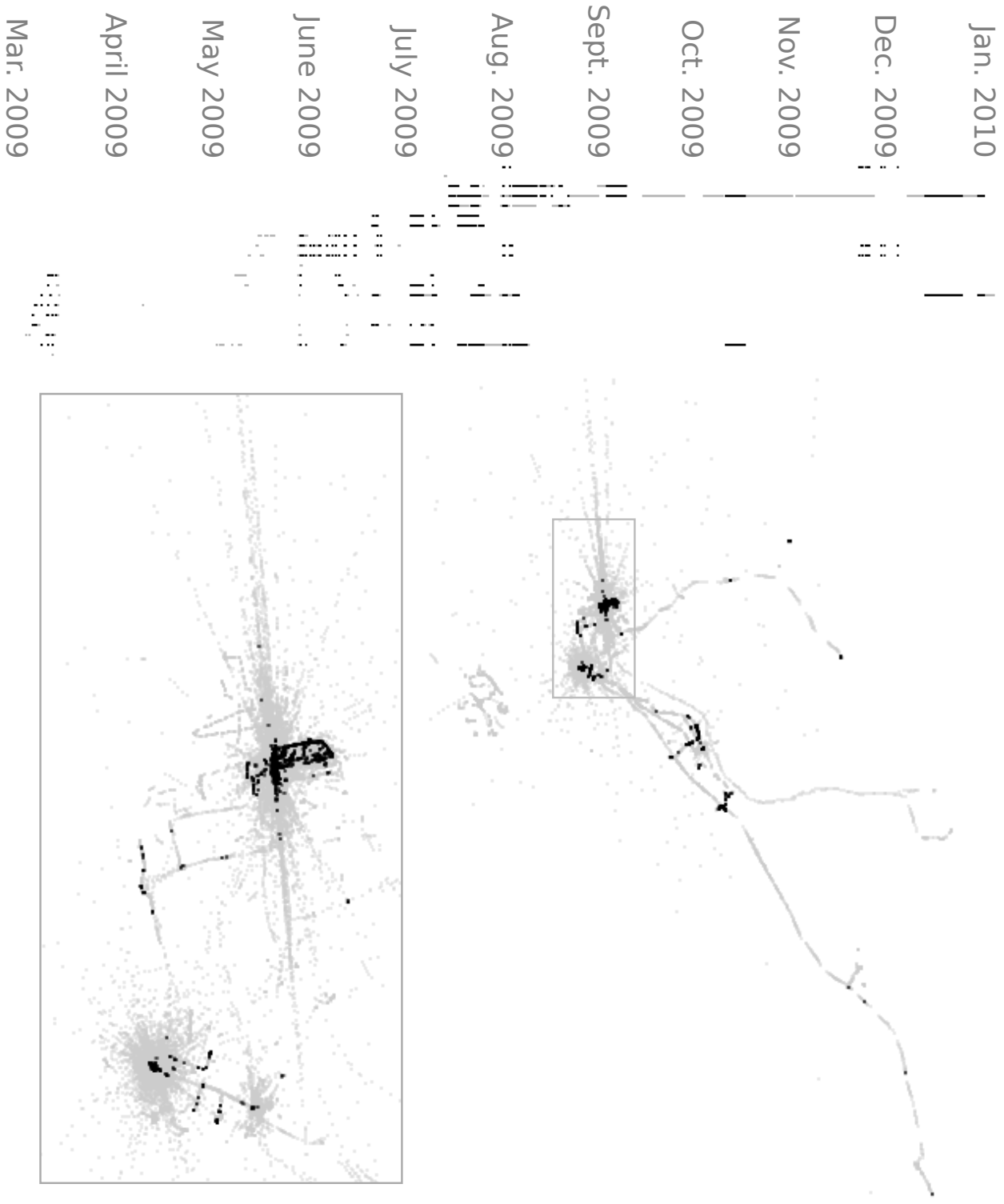


Figure 8.5: Rendezvous points (black) super-imposed on all locations (gray) in the Oakland study. The sample times for each sensor are displayed at the top, again with rendezvous points in black and all sample times in gray. As in the Ghana data, rendezvous happen in "hot-spot" locations, rather than distributed throughout the map. They are not isolated to a few lucky coincidences, but distributed throughout the study.

of locations, and that their movement patterns were well explained by the "least-action" principle: that people plan their routes and visit way-points according to a heuristic that minimizes their total trip distance [36].

All of these facts suggest that we will have significant opportunities to both calibrate and also increase the precision of our system in "hot-spot" locations.

8.2 Increasing precision in the lab

In this section, we verify that we can indeed increase the precision of our sensing system using the sensors and lab equipment we describe in Chapter 6.

As the density of sensors at a given location increases, we can increase the precision of our system by super-sampling, and averaging. For sensors with Gaussian noise, which our CO sensors exhibit, sampling in the same location, we expect the variance of the signal to be C/n if we average the signals from n sensors with noise variance C. Note that when the noise is not Gaussian, the noise power will still decrease, but at a slower rate.

In Figure 8.6 we see an experiment with six sensors in a chamber in which we can control the concentration of CO (Figure 6.2). In this case, we stepped the concentration of CO by 0.2ppm increments over an hour, and observed the response of the sensors. The light dots show the response of one sensor, and the dark dots show the averaged response of six sensors. Clearly the noise variance has decreased.

Figure 8.7 shows the variance of the signal versus the number of sensors averaged. Not only does this confirms the intuition observed in Figure 8.6, but the empirical results match the theoretical results closely!

8.3 Non-colocated sensors

Using Gaussian process regression (GPR), we can also increase the precision of the system even when samples are not in the same location in space-time (a more realistic situation). The closer the samples are to one another, the greater the increase in the precision.

We should note that GPR is appropriate not only because the sensor noise is Gaussian, but also because the process by which concentrations of gas mix and vary is also often modeled as Gaussian [51]. Modeled this way, we have the sum of two Gaussians, which is itself a Gaussian.

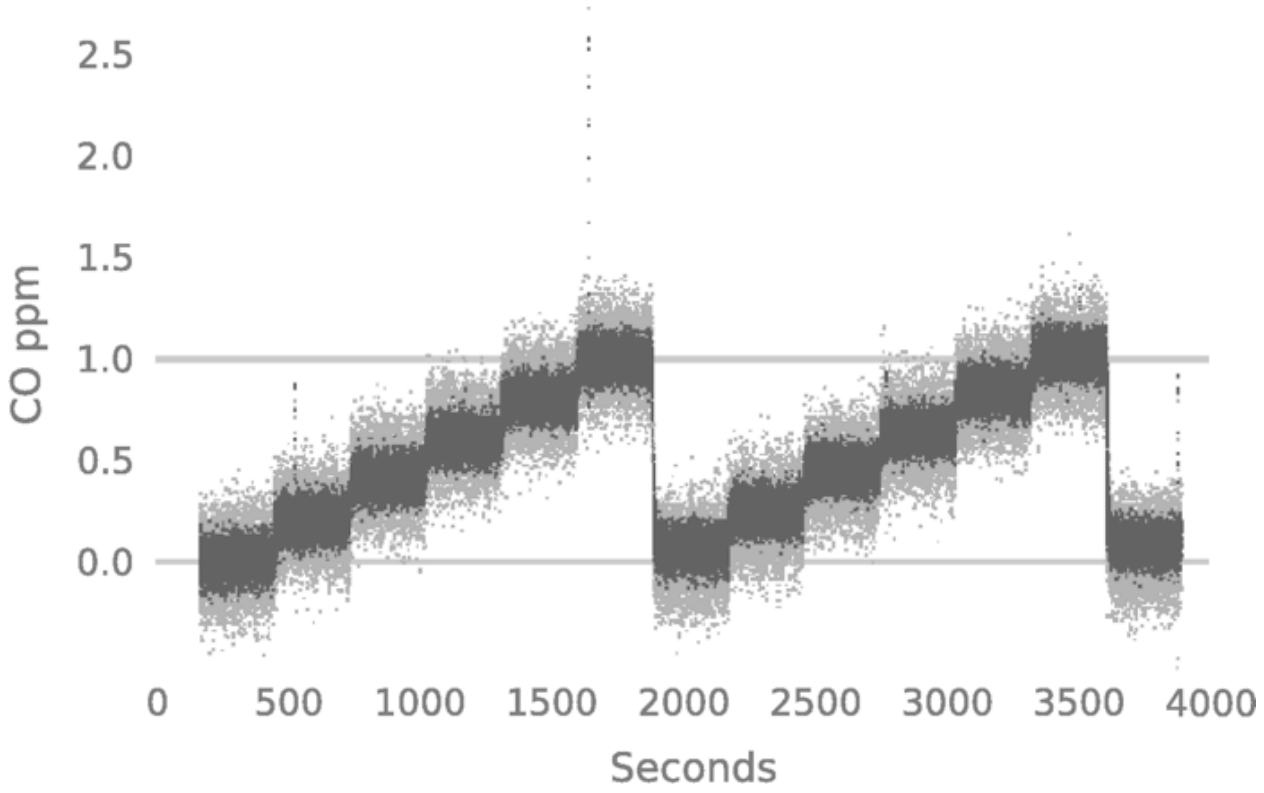


Figure 8.6: The signal from on sensor (light dots) and the average from six sensors (dark dots). Clearly averaging has decreased the noise power.

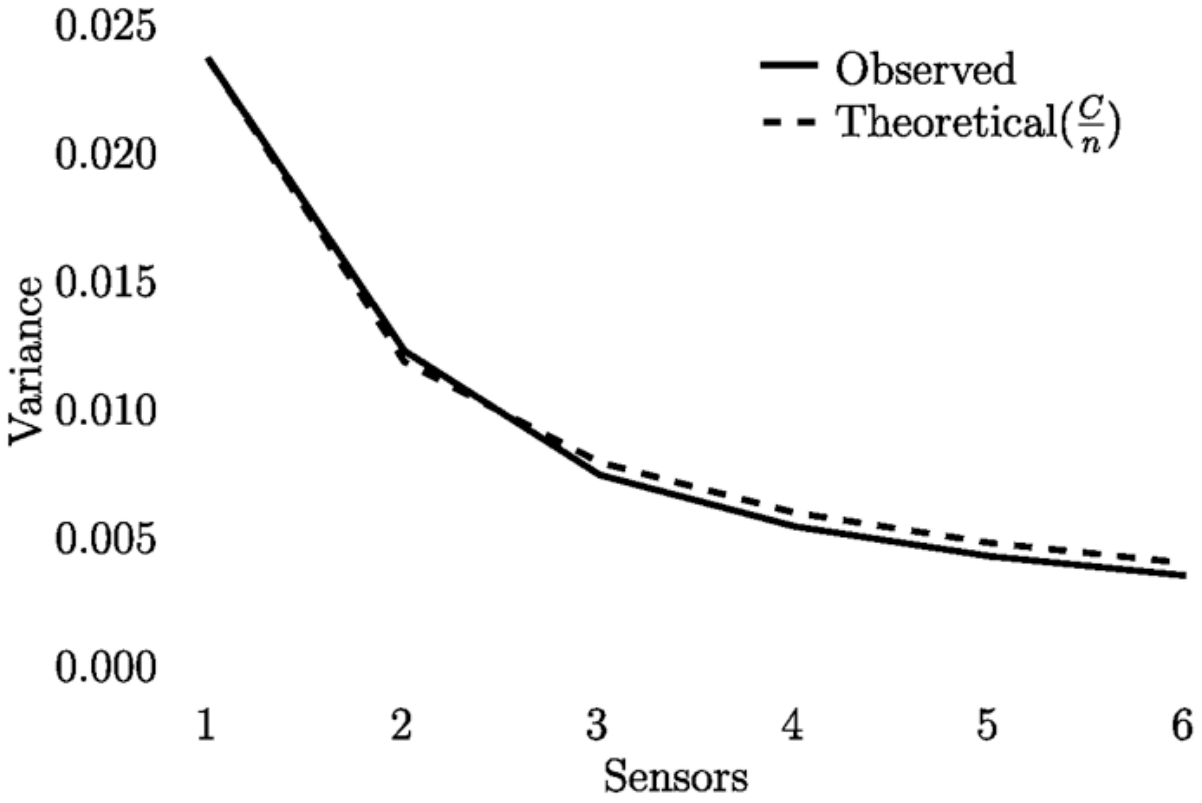


Figure 8.7: The variance of the signal (and thus noise power) decreases as more senors are averaged together, closely matching the theoretical prediction.

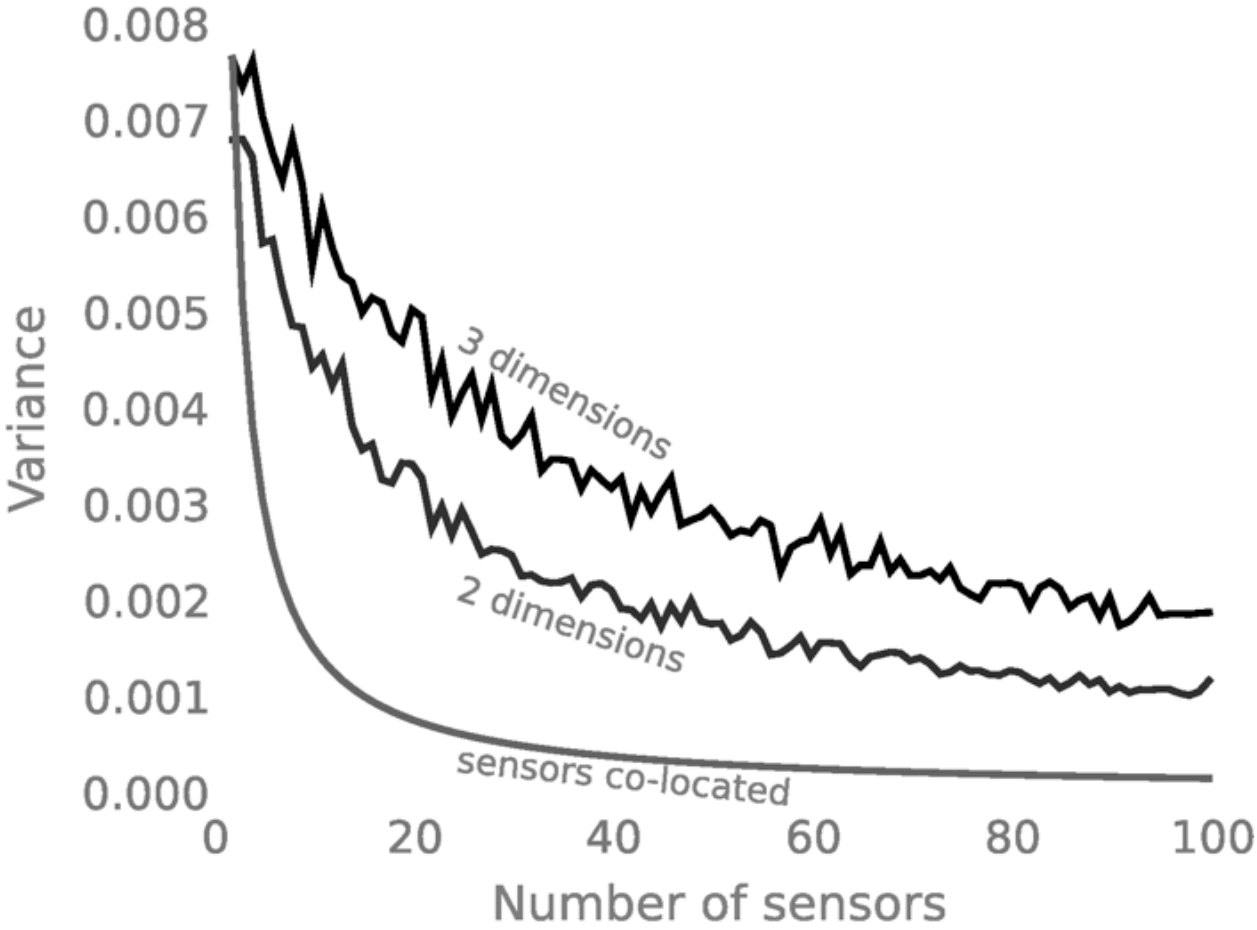


Figure 8.8: Simulated learning curves for two and three dimensional Gaussian processes as the density of sensors in an area increases. Three dimensions could correspond to two spacial dimensions and one temporal dimension. The theoretical C/k variance when the sensors are co-located, as verified in Figure 8.7 is shown for reference.

8.4 Gaussian process learning curves

The amount that the precision of the system increases depends on the density of sampling. As the density of sampling increases, so does the precision.

To quantify this increase in precision for a given algorithm, it is typical to consider the "learning curve" of the algorithm. The learning curve shows the deviation of the true values of samples from the inferred function as the number of training examples increases for a given area. Sollich provides some reasonably tight analytical bounds on the learning curves for GPR [56]. In the future we will present an analysis of the learning curves under various model assumptions.

In Figure 8.8, we see simulation results in which the variance of the signal at a point decreases

when nearby sensor's readings are also taken into account. In this simulation, we randomly choose k points for increasing values of k. These points are all located within a 2x2 square (for two dimensions) or 2x2x2 cube (for three dimensions). We then calculate the estimate variance at the middle of the square or cube. We use a the piecewise polynomial kernel described in (5.20), but other kernels exhibit similar behavior.

The setup of this experiment is conservative in that most of the points will be relatively far away from the point of interest, and will not contribute significantly to reducing the variance. Even under these conservative assumptions, we can see that as the density near the point of interest increases, the variance decreases.

Unfortunately, the data sets we have do not have sufficiently dense rendezvous to study the increase and precision that we can expect from our algorithms. Lee et al. propose a promising model of human mobility [36] that we intend to use to study both the implications of rendezvous for precision and also computational complexity.

Chapter 9

Ongoing work

This chapter briefly enumerates some of the most interesting questions and ideas raised by our research. While it is not exhaustive, it does highlight the main areas we are likely to continue exploring.

9.1 Source inference

One important question that we have not yet discussed is, "Where does the pollution we are sensing originate?" Using a detailed model of how pollution disperses into the atmosphere, we should be able to infer the sources of pollution from our samples of various locations near the sources.

To attack this problem, we have started with a grid-based, steady state model. In this model, we view the pollution we are sensing as emitting at a constant rate from sources over the course of an epic. In that case, the emission in and diffusion out will reach a steady state, as described in (5.3), and we can model the pollution readings as the convolution of the sources of pollution with a kernel function such as (5.3). We then can use Non-negative Matrix Factorization (NMF) [35] to deconvolve the sources from the steady state diffusion (Figure 9.1).

If we observe every point in a grid, then this algorithm does extremely well at determining the source of pollution (Figure 9.1(b)). Even with dense, but incomplete observations, the NMF algorithm seems to do relatively well (Figure 9.1(d). If, however, our observations are sparse, as will be the case with real data, then the algorithm puts sources in odd places (Figure 9.2).

One approach to addressing this problem would be to only make inferences about sources at a resolution for which we have data. By partitioning the data using a quadtree, we restrict our

Data: K is the kernel, O is our observations, W is a weight matrix, related to the number of observations at a location

Result: S holds the deconvolved matrix, e.g. the sources

begin

$S_{i,j} = 1$; // initialize the output to ones

$K'_{i,j} = K_{n-i,m-j}$; // K' is the reverse of K

$$O_{W_i} = \begin{cases} W_i \cdot O_i & W_i > 0 \\ \epsilon & W_i = 0 \end{cases}$$

$$W_i = \begin{cases} W_i & W_i > 0 \\ \epsilon & W_i = 0 \end{cases}$$

$N = K' * O_W$; /* N is a the reverse kernel convolved with O */

for $i = 1 : niter$ **do**

$D = K * S$; // Convolve the kernel with the estimate

$D_i = W_i \cdot D_i$; // Apply the weight matrix

$D' = K' * D$; // Convolve with the reverse kernel

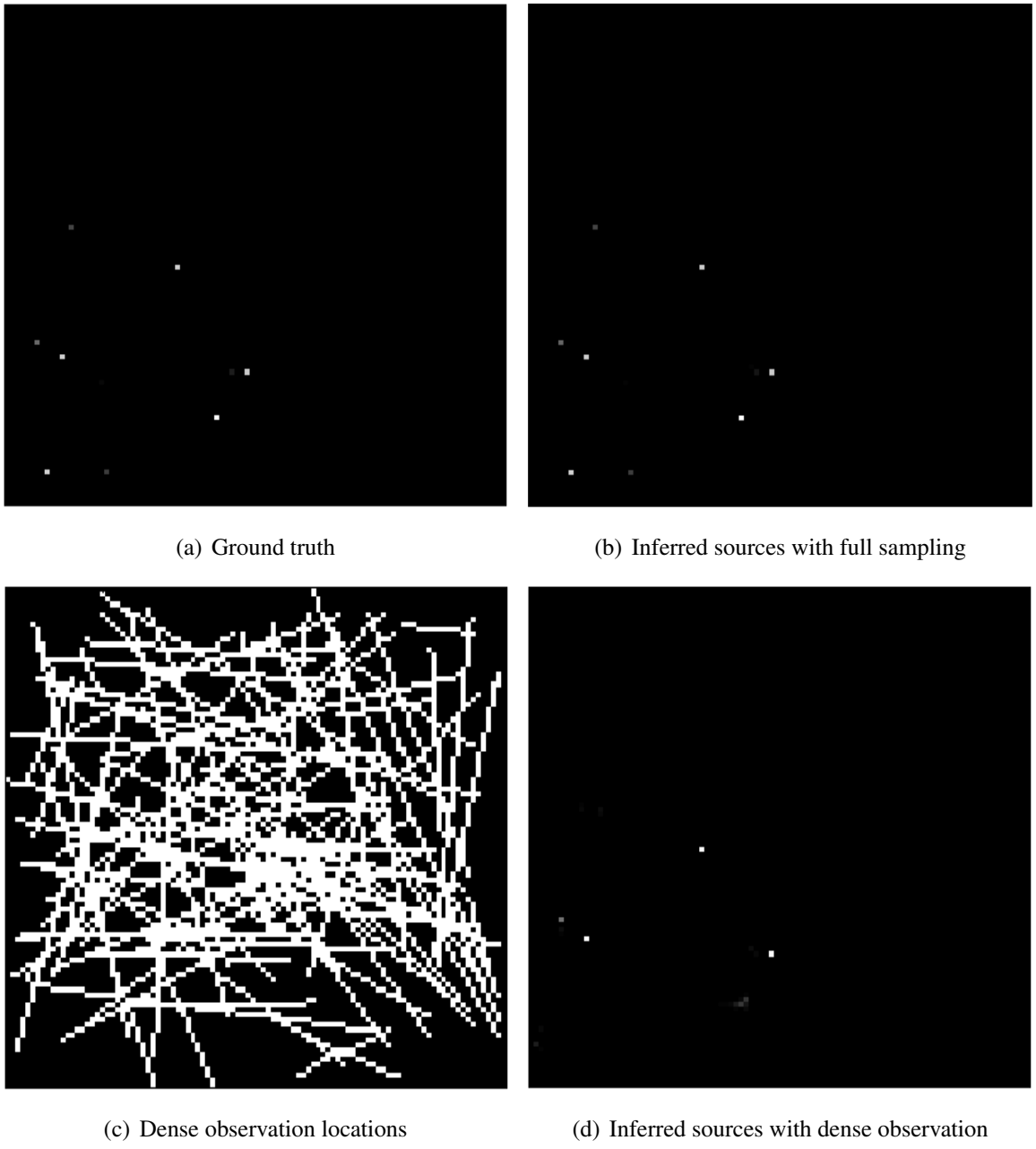
$S_{r_i} = \frac{N_i}{D'}$; /* Compute the ratio between the reversed convolved observations and estimates */

$S_i = S_i \cdot S_{r_i}$; // Apply the ratio to S

end

end

Algorithm 1: Pseudo-code for basic NMF deconvolution of a kernel K from observations O, leaving the sources in S.



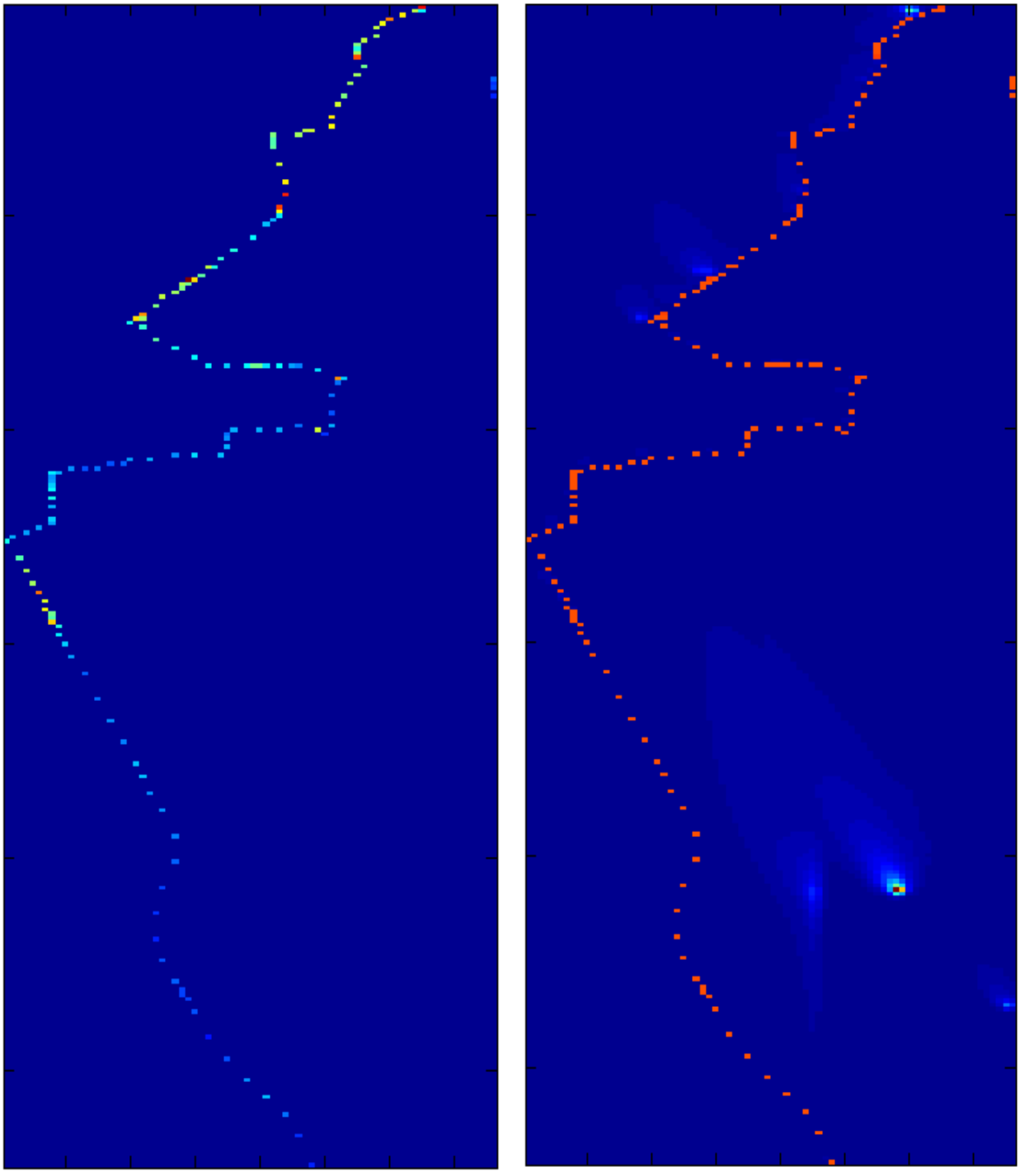
(a) Ground truth

(b) Inferred sources with full sampling

(c) Dense observation locations

(d) Inferred sources with dense observation

Figure 9.1: Source inference using Algorithm 9.1, with both full observation of the field (9.1(b)), and dense observation (9.1(d)).



(a) observations

(b) predicted concentration

Figure 9.2: Source inference on a 2 minute interval of the Ghana data, using Algorithm 9.1, with observation locations in the inferred field shown in red. The algorithm tends to put weight in empty regions, producing strange artifacts upwind and downwind of the observations.

algorithm to only operating on dense observations, implicitly averaging our observations over a larger area in areas with few observations. This approach is partially supported by the fact that the long term average concentrations of pollution seem to be relatively consistent over a wide area (see Figure 1.7).

Another approach we are examining is to give up on pinpointing exact sources, and rather determine the total pollution emitted in a bounded region. If a region is circumscribed by sensor readings, and we know the wind velocity, then we have a good idea of the total flux in and out of the region. We can subtract the flux into the region from the flux out of the region to get an estimate of the amount of pollution produced inside the region.

Of course, this approach depends on a good understanding of the vertical dispersion in the area. If we don't know the vertical dispersion rate, but do, however, know that the vertical dispersion is relatively constant over an area, then we can determine the relative amounts of pollution in different regions.

Research in these promising directions is ongoing.

9.2 Rejecting outlier data

One important consideration, which will certainly arise in any large scale deployment, is how to deal with faulty sensors. When sensors malfunction, they will produce readings that are wildly different from observations by other sensors. Malfunctioning sensors will not only contribute bad data to the data set, but also throw off the automatic calibration algorithms discussed in Chapter 7.

As opposed to a mis-calibrated sensor, a malfunctioning sensor will most likely produce readings that are not only different from other sensors near-by, but have different response characteristics in both the space and time domain. This suggests that we could use the spectrum of the sensor over the space and time domain to identify sensors that are malfunctioning. Using a classifier such as a Support Vector Machine on the spectral data from sensors may allow us to identify those sensors which are malfunctioning, and ignore their data.

Claudel et al. present another promising mechanism for detecting faulty sensors using convex optimization [10]. Their algorithm also relies on similarity between local measurements, much in the way the we use local measurements to automatically calibrate. This similarity is worth exploring.

9.3 Multi-factor calibration

Section 7.7 discusses how to how to automatically calibrate sensors which are cross-sensitive to multiple agents in the atmosphere. This mechanism will be highly dependent on the analytic functions which model the cross-sensitivities of the sensor. Because we don't currently have a clear understanding of the cross-sensitivities the sensors we are using, we have not included an analysis of these techniques. Once we have characterized the cross-sensitivity of our sensors in the lab, we will be able to explore the effectiveness of these algorithms in the "wild."

9.4 Scalability and mobility models

The analysis of mobility presented in Chapter 8 provides some empirical evidence that rendezvous will be frequent enough to calibrate devices automatically, and to super-sample in locations where people congregate. This analysis, however, does not explore the extent of rendezvous in a societal scale system, but merely makes tentative extrapolations from observations of 100 people.

Researchers have recently developed increasingly detailed, accurate and explanatory models of human mobility, (for example, Lee et al. [36]). These models will allow us to analyze the effectiveness and scalability of our algorithms when millions of users contribute data. This will in turn allow us to explore and enhance scalability of our algorithms.

9.5 Context inference

While context inference is a well studied topic (see Section 2.2), we have identified at least one important inference problem unaddressed in prior work, as far as we know: inferring whether the phone is in a pocket, or bag, or exposed to the open air. A related problem is inferring whether the phone is indoors or outdoors. These contexts will have a significant impact on the readings taken by a phone, and must be accounted for in order to accurately measure and label the environment and a user's exposure to pollution.

We believe that these classifications could be performed by either listening to ambient noise and classifying the spectral or cepstral features of the noise, or sending a impulse on the device's speaker, and classifying the transfer function. Some initial results using SVM classification of spectral features are promising, but need more investigation. These algorithms should be evaluated

for effectiveness and intrusiveness in real phones, for a large number of users.

9.6 Scalable computation

This dissertation present efficient algorithms for interpolation, calibration and estimation of sensors and their output. The efficiency of these algorithms largely depends on the algorithms' ability to exploit sparsity in the data. At the core of these algorithms is the solution of sparse linear systems of equations, and sparse matrix inversion. While the computational cost of these fundamental operations are well understood, they depend significantly on the sparsity of the data set, as well as the sparsity structure in the matrices.

The inverse of a sparse matrix is not sparse, in general. The sparsity structure of the inverse depends significantly on the particular structure of a matrix. This has both computational and storage implications, and therefore needs to be explored.

9.7 Signal Strength Map

The algorithms we have developed are not only useful for sensing air pollution. Another interesting characteristic of the environment to sample is the network signal strength in a given location. This information is valuable to both the consumer and the network operators. Network operators, however, apparently do not have a technology or strong incentive with which to compel large numbers of users to give up their privacy in order to improve their network, since they typically use highly instrumented vehicles driving around the major transportation arteries to characterize their network.

A third party which provides a valuable public service, however, does provide the public with an incentive to sacrifice some of their privacy, especially if guarantees can be made with respect to the risk and degree of privacy that a user must sacrifice. A map of network signal strength might therefore provide a reasonable incentive for service providers to subsidize the cost of embedding sensors in devices, and the network traffic that would result. Such a subsidy would also provide good PR and a differentiating factor for a network operator, as well as a phone manufacturer.

Since a signal strength map would be easy to build without any changes to existing devices, it is a quick way to both test the scale of our algorithms, as well as the true value of signal strength data to providers.

Chapter 10

Conclusions

This dissertation has focused on a practical approach to building a large scale, distributed sensing system, using pollutions sensors embedded in mobile phones. Although we have focused on a particular type of air pollution, we believe these results generalize beyond particular sensors, pollution types and even domains to other sensible mediums such as noise levels and signal strength.

We have used the data from several sensing campaigns to validate our approach, and the basic computational models we have proposed. Building on these models, we have demonstrated a scalable mechanism for automatically calibrating devices in the field, thus addressing a major issue with atmospheric sensing. We have also presented evidence that rendezvous is frequent enough to not only support automatic calibration, but also to allow for super-sampling in areas of congregation, thus increasing the precision of the system.

Although we have arguably addressed the two most important and fundamental issues with real world sensing, several important directions of investigation remain, outlined in Chapter 9. These include source inference, rejecting outlier data, multi-factor calibration, more in depth analysis of mobility and its impact on rendezvous frequency and density, some important but unaddressed context inference problems, the exact mechanisms for scalable computation, energy conservation and utilizing the basic machinery we have developed for monitoring other aspects of our environment such as RF signal strength.

In the past few years, smartphones have proliferated wildly, and application developers are racing to take advantage of the computational and communications capabilities that are appearing at the edge of the network. The rise of smart-phones has also meant that GPS and accelerometers are fast becoming standard options in even lower end phones, whereas they were specialty items

only a few years ago.

Application developers are also searching for ways to take advantage of the phone's unique status as an ever present electronic companion to people. Augmented reality is one application that has received a lot of interest recently, but it is only the beginning in terms of penetration of computing and sensing into our every-day lives.

Meanwhile, the networks and mobile phone penetration in the developing world continues to expand. While growth rates are not evenly distributed, as a whole, access to telecommunication is expanding quickly, and is largely mimicking the telecommunication expansion in the industrialized world. In fact, because of the phone's relative accessibility to consumers in the developing world with respect to the accessibility of practicality of the PC, there is evidence that users in the developing world are skipping straight to smart-phones.

After a recent increase in concern over climate change and other environmental issues, the prolonged economic problems in the industrialized world has shifted people's focus towards more immediate economic concerns. At the same time, the developing world has continued its rapid economic expansion; particularly in Brazil, Russia, India and China; but many places in Africa and elsewhere as well. Pollution will increasingly become a critical policy issue, and access to good information will be key to making good policy decisions. With the world's economy and environment both highly interconnected, ubiquitous information about pollution and exposure (as opposed to an imbalance of information towards the developed world) will also be important for geopolitical stability and consensus.

As academics, we have a unique opportunity, right now, to influence the direction that the mobile platform takes. Service providers, by their nature, try to control access to information. The smart-phone ecosystem has temporarily wrestled control of the platform away from network providers, but they are seeking for ways to regain control, such as through proprietary access to bandwidth. In this window of opportunity, it is imperative that we establish a platform for gathering information about the environment, and ensuring that the information gathered is fundamentally owned by the public. In this dissertation, we have presented some first steps towards developing such a platform, in service to the goal of providing an important window into the health of all the corners of our global society.

Bibliography

[1] E. Agapie, G. Chen, D. Houston, E. Howard, J. Kim, M. Y. Mun, A. Mondschein, S. Reddy, R. Rosario, J. Ryder, A. Steiner, J. Burke, E. Estrin, M. Hansen, and M. Rahimi. Seeing our signals: combining location traces and web-based models for personal discovery. In *Proceedings of the 9th workshop on Mobile computing systems and applications*, HotMobile '08, pages 6–10, New York, NY, USA, 2008. ACM.

[2] Saurabh Amin, Steve Andrews, Saneesh Apte, Jed Arnold, Jeff Ban, Marika Benko, Re M. Bayen, Benson Chiou, Christian Claudel, Coralie Claudel, Tia Dodson, Osama Elhamshary, Chris Flens-batina, Marco Gruteser, Juan carlos Herrera, Ryan Herring, Baik Hoh, Quinn Jacobson, Toch Iwuchukwu, James Lew, Xavier Litrico, Lori Luddington, Jd Margulici, Ali Mortazavi, Xiaohong Pan, Tarek Rabbani, Tim Racine, Erica Sherlock-thomas, Dave Sutter, and Andrew Tinka. Mobile century: Using gps mobile phones as traffic sensors: A field experiment. In *Proceedings of the 15th World Congress on Intelligent Transportation Systems*, New York, November 2008.

[3] Paul M. Aoki, R.J. Honicky, Alan Mainwaring, Chris Myers, Eric Paulos, Sushmita Subramanian, and Allison Woodruff. Common sense: Mobile environmental sensing platforms to support community action and citizen science. In *Adjunct Proceedings of the Tenth International Conference on Ubiquitous Computing (Ubicomp 2008)*, pages 59–60, September 2008.

[4] Paul M. Aoki, R.J. Honicky, Alan Mainwaring, Chris Myers, Eric Paulos, Sushmita Subramanian, and Allison Woodruff. A vehicle for research: Using street sweepers to explore the landscape of environmental community action. In *Conference on Computer Human Interaction*, pages 375–384, April 2009.

[5] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. *IEEE Pervasive*, 3001:1–17, 2004.

[6] Justin Black, Alex Elium, Richard M. White, Michael G. Apte, Lara A Gundel, and Rossana Cambie. MEMS-enabled miniaturized particulate matter monitor employing 1.6 GHz aluminum nitride thin-film bulk acoustic wave resonator (FBAR) and thermophoretic precipitator. In *Proceedings of the 2007 IEEE Ultrasonics Symposium*, November 2007.

[7] Enrique Canessa and Marco Zennaro, editors. *mScience: Sensing, Computing and Dissemination*. ICTP–The Abdus Salam International Centre for Theoretical Physics, November 2010.

[8] C. Castelluccia, E. Mykletun, and G. Tsudik. Efficient aggregation of encrypted data in wireless sensor networks. In *Mobile and Ubiquitous Systems: Networking and Services, 2005. MobiQuitous 2005. The Second Annual International Conference on*, pages 109 – 117, July 2005.

[9] Dario Catalano, Ronald Cramer, Giovanni Crescenzo, Ivan Darmgrd, David Pointcheval, Tsuyoshi Takagi, Ronald Cramer, and Ivan Damgrd. Multiparty computation, an introduction. In *Contemporary Cryptology*, Advanced Courses in Mathematics - CRM Barcelona, pages 41–87. Birkhuser Basel, 2005.

[10] Christian G. Claaudel, Matthieu Nahoum, and Alexandre M. Bayen. Minimal error certificates for detection of faulty sensors using convex optimization. In *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, Allerton, IL, September 2009.

[11] Sunny Consolvo, David W. McDonald, Tammy Toscos, Mike Y. Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A. Landay. Activity sensing in the wild: a field trial of ubifit garden. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1797–1806, New York, NY, USA, 2008. ACM.

[12] Jason Corburn. *Street Science: community knowledge and environmental health justice*. MIT Press, Cambridge, MA, 2005.

[13] Prabal Dutta, Paul M. Aoki, Neil Kumar, Alan Mainwaring, Chris Myers, Wesley Willett, and Allison Woodruff. Common sense: participatory urban sensing using a network of handheld air quality monitors. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 349–350, New York, NY, USA, 2009. ACM.

[14] Nathan Eagle and Alex Pantland. Reality Mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268.

[15] Shane B. Eisenman. *People-Centric Mobile Sensing Networks*. PhD thesis, Columbia University, October 2008.

[16] Shane B. Eisenman, Hong Lu, and Andrew T. Campbell. Halo: Managing node rendezvous in opportunistic sensor networks. In *Proceedings of the 6th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS '10)*, June 2010.

[17] Shane B. Eisenman, Emiliano Miluzzon, Nicholas D. Lane, Ronald A. Peterson, Gahng-Seop Ahn, and Andrew T. Campbell. BikeNet: A mobile sensing system for cyclist experience mapping. *ACM Transactions on Sensor Networks*, 6(1), December 2009.

[18] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. The pothole patrol: using a mobile sensor network for road surface monitoring. In *Proceeding of the 6th international conference on Mobile systems, applications, and services*, MobiSys '08, pages 29–39, New York, NY, USA, 2008. ACM.

[19] George F. Fine, Leon M. Cavanagh, Ayo Afonja, and Russell Binions. Metal oxide semiconductor gas sensors in environmental monitoring. *Sensors*, (10):5469–5502, 2010.

[20] Kenneth Fishkin, Bernt Schiele, Paddy Nixon, and Aarion Quigley. A practical approach to recognizing physical activities. *IEEE Pervasive*, pages 1–16, 2006.

[21] Garmin. Quest data sheet. https://buy.garmin.com/shop/shop.do?pID=213#specsTab.

[22] Gartner. Competitive landscape: mobile devices, worldwide, 1Q10, May 2010.

[23] B. Gedik and Ling Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *Mobile Computing, IEEE Transactions on*, 7(1):1 –18, January 2008.

[24] Bernd Girod, Rudolf Rabenstein, and Alexander Stenger. *Signals and systems*. Wiley, 2nd edition, 2001.

[25] W. C. Hinds. *Aerosol technology: properties, behavior, and measurement of airborne particles*. Wiley-Interscience, New York, NY, 2nd edition, 1998.

[26] Baik Hoh, Marco Gruteser, Ryan Herring, Jeff Ban, Daniel Work, Juan-Carlos Herrera, Alexandre M. Bayen, Murali Annavaram, and Quinn Jacobson. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *Proceeding of the 6th international conference on Mobile systems, applications, and services*, MobiSys '08, pages 15–28, New York, NY, USA, 2008. ACM.

[27] J. Michael Hollas. *Modern spectroscopy*. John Wiley and Sons, Hoboken, NJ, 4th edition, 2004.

[28] R.J. Honicky, Eric Brewer, Eric Paulos, and Richard White. N-smarts: Networked suite of mobile atmospheric real-time sensors. In *Networked Systems for Developing Regions, a workshop of SIGCOMM*, 2008.

[29] R.J. Honicky, Eric A. Brewer, John F. Canny, and Ronald C. Cohen. Increasing the precision of mobile sensing systems through supersampling. In *International workshop on urban, community and social applications of networked sensing systems – UrbanSense08*, 2008.

[30] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. Cartel: a distributed mobile sensor computing system. In *Proceedings of the 4th international conference on Embedded networked sensor systems*, SenSys '06, pages 125–138, New York, NY, USA, 2006. ACM.

[31] Apu Kapadia, David Kotz, and Nikos Triandopoulos. Opportunistic sensing: security challenges for the new paradigm. In *Proceedings of the First international conference on COMmunication Systems And NETworks*, COMSNETS'09, pages 127–136, Piscataway, NJ, USA, 2009. IEEE Press.

[32] Apu Kapadia, Nikos Triandopoulos, Cory Cornelius, Daniel Peebles, and David Kotz. Anonysense: Opportunistic and privacy-preserving context collection. In Jadwiga Indulska, Donald Patterson, Tom Rodden, and Max Ott, editors, *Pervasive Computing*, volume 5013 of *Lecture Notes in Computer Science*, pages 280–297. Springer Berlin / Heidelberg, 2008.

[33] Elliot D. Kaplan and Christopher J. Hegarty. *Understanding GPS: principles and applications*. Artech House, Boston, MA, 2006.

[34] Lascar. Usb-el-co data sheet. http://www.lascarelectronics.com/pdf-usb-datalogging/data-logger0333368001254903501.pdf.

[35] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems*, pages 556–562. MIT Press.

[36] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. SLAW: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863, April 2009.

[37] Jonathan Lester, Tanzeem Choudhury, Nicky Kern, Gaetano Borriello, and Blake Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 766–772, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[38] Jing Li. NASA Ames scientist develops cell phone chemical sensor, October 2009. http://www.nasa.gov/centers/ames/news/features/2009/cell_phone_sensors.html.

[39] Emiliano Miluzzo1, Nicholas D. Lane1, Andrew T. Campbell1, and Reza Olfati-Saber. CaliBree: A self-calibration system for mobile sensor networks. *Distributed Computing in Sensor Systems*, Jan 2008.

[40] Mark Moeglein and Norman Krasner. An introduction to snaptrack server-aided gps technology. In *Proceedings of the 11th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1998)*, pages 333–342, Nashville, TN, September 1998.

[41] Min Mun, Sasank Reddy, Katie Shilton, Nathan Yau, Jeff Burke, Deborah Estrin, Mark Hansen, Eric Howard, Ruth West, and Péter Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 55–68, New York, NY, USA, 2009. ACM.

[42] H. Niessner and K. Reichert. On computing the inverse of a sparse matrix. *International Journal for Numerical Methods in Engineering*, 19(10):1513–1526, 1983.

[43] Port of Oakland. Port of oakland facts and figures. http://www.portofoakland.com/maritime/factsfig.asp.

[44] A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design. *Journal of the Royal Statistical Society, Series B.*, 40(1):1–42, 1978.

[45] I. Paprotny, F. Doering, and R. M. White. MEMS particulate matter (PM) monitor for cellular deployment. In *Proceedings of the 9th Annual IEEE Conference on Sensors (IEEE SENSORS 2010)*, Waikoloa, HI, November 2010.

[46] Eric Paulos, RJ Honicky, and Elizabeth Goodman. Sensing atmosphere. In *Proceedings of the 2007 ACM conference on Sensing Systems (SenSys 2007), workshop on Sensing on Everyday Mobile Phones in Support of Participatory Research*, 2007.

[47] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[48] Nishkam Ravi, Nikhil Dandekar, Prreetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. *American Association for Artificial Intelligence*, 2005.

[49] Sasank Reddy, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. Determining transportation mode on mobile phones. In *Proceedings of the 2008 12th IEEE International Symposium on Wearable Computers*, pages 25–28, Washington, DC, USA, 2008. IEEE Computer Society.

[50] Jason Ryder, Brent Longstaff, Sasank Reddy, and Deborah Estrin. Ambulation: A tool for monitoring mobility patterns over time using mobile phones. *Computational Science and Engineering, IEEE International Conference on*, 4:927–931, 2009.

[51] John H. Seinfeld and Spyros Pandis. *Atmospheric Chemistry and Physics – From Air Pollution to Climate Change*, chapter 18, pages 828–861. John Wiley and Sons, 2nd edition, 2006.

[52] Claude E. Shannon. Communication in the presence of noise. In *Proceedings of the IRE*, volume 37, pages 10–21, January 1949.

[53] Katie Shilton. Four billion little brothers?: privacy, mobile phones, and ubiquitous data collection. *Commun. ACM*, 52:48–53, November 2009.

[54] Douglass A. Skoog, Donald M. West, F. James Holler, and Stanley R. Crouch. *Fundamentals of Analytical Chemistry*. Thomson-Brooks/Cole, Belmont, CA, 8th edition, 2004.

[55] Timothy Sohn, Alex Varshavsky, Anthony Lamarca, Mike Y. Chen, Tanzeem Choudhury, Ian Smith, Sunny Consolvo, Jeffrey Hightower, William G. Griswold, and Eyal De Lara. Mobility detection using everyday gsm traces. In *Proceedings of the Eighth International Conference on Ubiquitous Computing (Ubicomp 2006)*, pages 212–224. Springer, 2006.

[56] Peter Sollich and Anason Halees. Learning curves for gaussian process regression: approximations and bounds. 14:1393–1428, 2002.

[57] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002.

[58] K. Takahashi. Formation of a sparse bus impedance matrix and its application to short circuit study. In *Proceedings of the IEEE PICA Conference*, pages 63–69, Minneapolis, June 1973.

[59] BW Technology. Gas alert extreme data sheet. http://www.bwtechnologies.nl/gasalerteng.html.

[60] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, SenSys '09, pages 85–98, New York, NY, USA, 2009. ACM.

[61] International Telecommunications Union. Measuring the information society 2010, 2010.

[62] Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Proceedings of the 13th International IEEE Conference on Intelligent Trasportation Systems*, Washington, DC, USA, 2010. IEEE Computer Society.

[63] Yi Wang, Jialiu Lin, Murali Annavaram, Quinn A. Jacobson, Jason Hong, Bhaskar Krishnamachari, and Norman Sadeh. A framework of energy efficient mobile sensing for

automatic user state recognition. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, MobiSys '09, pages 179–192, New York, NY, USA, 2009. ACM.

[64] Evan Welbourne, Jonathan Lester, Anthony Lamarca, and Gaetano Borriello. Mobile context inference using low-cost sensors. In *In Proceedings of Location and Context Awareness 2005*, pages 254–263. Springer-Verlag, 2005.

[65] Holger Wendland. *Scattered data approximation*. Cambridge University Press, 2005.

[66] Wildtracker. Fastloc snapshot gps. http://www.wildtracker.com/fastloc.htm.

[67] Wesley Willett, Paul Aoki, Neil Kumar, Sushmita Subramanian, and Allison Woodruff. Common sense community: Scaffolding mobile sensing and analysis for novice users. *IEEE Pervasive*, 2010.

[68] R W Wuest, C R adn Werne, B W Colston, and C L Hartmann-Siantar. Applying science and technology to combat WMD terrorism. In *Proceedings of the SPIE Defense and Security Symposium*, May 2006.

[69] www.citytech.com. Microcel cf data sheet.