DATA EXTRAPOLATION IN SOCIAL SENSING FOR DISASTER RESPONSE

BY

SIYU GU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Tarek Abdelzaher

ABSTRACT

This thesis complements the large body of social sensing literature by developing means for augmenting sensing data with inference results that “fill-in” missing pieces. Unlike trend-extrapolation methods, we focus on prediction in disaster scenarios where disruptive trend changes occur. A set of prediction heuristics (and a standard trend extrapolation algorithm) are compared that use either predominantly-spatial or predominantly-temporal correlations for data extrapolation purposes. The evaluation shows that none of them do well consistently. This is because monitored system state, in the aftermath of disasters, alternates between periods of relative calm and periods of disruptive change (e.g., aftershocks). A good prediction algorithm, therefore, needs to intelligently combine time-based data extrapolation during periods of calm, and spatial data extrapolation during periods of change. The thesis develops such an algorithm. The algorithm is tested using data collected during the New York City crisis in the aftermath of Hurricane Sandy in November 2012. Results show that consistently good predictions are achieved. The work is unique in addressing the bi-modal nature of damage propagation in complex systems subjected to stress, and offers a simple solution to the problem.

To my parents, for their love and support.

ACKNOWLEDGMENTS

This thesis would not be possible without the support of many people.

In particular, I would like to thank my advisor, Tarek Abdelzaher, for his support, patience, and trust. Tarek helps me in various aspects in developing this thesis, including developing the algorithms, analyzing the data, and polishing the write-ups. When I find the experiment results unsatisfactory, he always provides revealing insights where things go wrong. When I find the experiment results as good as expected, he always gives me helpful suggestions on how to strengthen the results.

Hengchang Liu, who was a postdoc student in Tarek's research group, led the first shot to the problem addressed in this thesis. We spent most of my second semester together in University of Illinois, Urbana Champaign collecting datasets, making initial attempts, and writing papers.

Last but not least are my best friends and lab mates here, Chenji Pan, Shaohan Hu, Shen Li, Shiguang Wang, and Su Lu. Having great experience in doing good research, they help me quite a lot in organizing and writing this thesis. And more importantly, they make my stay in Urbana Champaign a remarkable one.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Participatory Sensing System

Participatory Sensing, introduced first by Burke et al. [1] is the concept of communities (or other groups of people) contributing sensory information to form a body of knowledge. It can be used to retrieve information about the environment, weather, congestion as well as any other sensory information that collectively forms knowledge. For example, BeWell [2] developed by Lane et al. is an individual wellbeing tracking system running on smartphones with multiple sensors (e.g., cameras, gyroscope, and accelerometer). P-sense [3] is a participatory sensing system for air pollution monitoring and control. In this system, external sensors are used to collected environmental data and the data is aggregated and propagated via the cell phone network. One other example is ParkNet developed by Mathur et al. [4]. A GPS receiver and a passenger-side-facing ultrasonic rangefinder are installed with each vehicle involved in order to determine parking lot occupancy. And all the data are uploaded to the central server to build the map of parking availability.

The spread of the smartphone brings more chance to participatory sensing system, however, perhaps the most critical issue regarding participatory sensing is privacy. Because if users' or volunteers' privacy is compromised, they are unlikely to contribute to the study. For example, GPS sensor readings can be used to infer private information such as their daily commute, home location, and work location [5]. Many research has been done is this field [6, 7, 8, 9, 10]. Other open issues include, but are not limited to, effective incentives for participation [11, 12], resource limitation [13, 14], and security and data integrity [15, 16].

1.2 Participatory Sensing in Disaster Monitoring and Response

Thanks to the fast development of smartphones and social networks, participatory sensing receives more attention in disaster monitoring and response applications in recent years.

A large body of sensor network literature focused on monitoring and disaster alerts. For example, Werner-Allen et al. deployed three wireless sensor networks on active volcanoes [17]. The initial deployment was a small proof of concept system that monitored acoustic signals from the Tungurahua volcano, in Ecuador. The second deployment was to measure seismic signals at the Reventador volcano, in Ecuador. The third deployment was at Tungurahua in August, featuring a new data collection system. Li et al. deployed a sensor network for monitoring and alerts in a coal mine [18]. Liu et al. present an automatic and reliable sensor network for firefighter applications [19], which allows a firefighter to carry a small dispenser filled with sensor nodes and deploy them one-by-one in a manner that guarantees reliable communication. The SensorFly project [20] develops a sensor cloud, which consists of many low cost and individually limited mobile sensing devices that only when functioning together can produce an intelligent cloud, in disaster situations such as an earthquake and fire.

On the social network side, people share their information about the disaster region to social networks and special-purpose services, to help each other beat the disaster together. For instance, popular social networks such as Facebook [21] and Twitter [22], played an important role after natural disasters such as Japan Tsunami in 2011 [23] and US Hurricane Sandy in 2012 [24]. Many service providers, some notable names including Waze [25] and Gas-

Buddy [26], set up special-purposes services to allow individuals to participate and report the availability of various resources (e.g., gas stations) after Sandy via the web or smartphones. Ushahidi [27] is another notable disaster and crisis management mapping tool. It can be used to collect and visualize data from multiple data streams including text messages, email, twitter and web-forms. However, due to the opportunistic nature of participatory sensing, there are typically "blind points" in the obtained points of interest (POIs) map at any given time point.

1.3 Time Series Forecasting

Usually, a participatory sensing system is deployed to monitor the states of a group of points of interest. The states of points of interest are time series data. So, to fill in "blind points" in the time series data, we use time series forecasting technologies.

Time series forecasting is the use of a model to predict future values based on previously observed values. Given a time series of data, autoregressive moving average (ARMA) [28] is one of the most classic models used to understand the underlying structure. Fed with some training data, ARMA provides a description of the time series data in terms of two polynomials, one for the auto-regression and the second for the moving average. After that, the model can be applied to forecasting future values. ARMA is a good for stationary time series models, while autoregressive integrated moving average (ARIMA) can be utilized when the model is non-stationary, by applying an initial differencing step to the model [29]. These models are widely-used. For example, Van Der Voort et al. [30] use Kohonen self-organizing map and ARIMA model to do the short-term traffic forecasting. Another notable example is Pai et al. [31], in which they try to forecast stock prices. In their work, they first use support vector machines to solve the non-linear regression estimation problem and then apply the ARIMA to capture the patterns.

1.4 Research Contribution of this Thesis

The research topic of this thesis falls in participatory sensing application in disaster settings. More specifically, this thesis explores the question of how to inference missing data in the aftermaths of disasters in a reliable way.

In participatory sensing, sources measure application-related state at locations of interest then usually report it at a later time (e.g., when they encounter a WiFi access point a few hours later). Hence, at any given time, the latest state of some points of interest may be unknown. Incomplete real-time coverage may also arise due to scarcity of sensing resources. For example, volunteers in a disaster-response application may survey and report locations of damage. If there are fewer volunteers than damage locations, the state of some of these locations will not be immediately reported. In such scenarios, one question is: can we infer the missing data? Our thesis is mainly to answer this question in disaster aftermath scenarios. To the best of our knowledge, our approach is the first one addressing this problem effectively.

Disaster aftermath distinguishes from many other scenarios in two aspects, namely, disruptive change and scarcity of training data. Many time-series data extrapolation approaches are based on the assumption that past trends are predictive of future values. These approaches do not do well when disruptive changes occur. For example, a history of no traffic congestion on main highways of some city does not offer a good traffic predictor if a natural disaster causes a mass evacuation. An alternative recourse is to consider only spatial correlations. For example, certain city streets tend to get flooded together after heavy rain (e.g., because they are at the same low elevation), and certain blocks tend to run out of power together after a thunderstorm (e.g., because they share the same power lines). Understanding such correlations can thus help infer state at some locations from state at others when disruptive changes (such as a flood or a power outage) occur. In the following chapters, We show that system state in post-disaster scenarios alternates between periods of calm (when the past is a good predictor of the future) and periods of sudden change, as new parts of the infrastructure are damaged (e.g., due to aftershocks) or repaired. Hence, data extrapolation algorithms that rely predominantly on spatial correlations or predominantly on temporal correlations tend not to work consistently well, as the relative importance weights of temporal versus spatial correlations change significantly between periods of calm and periods of change. Instead, we show that such algorithms must switch intelligently between two extrapolation modes with different emphasis on temporal versus spatial correlations.

Of special interest is the case where correlations needed for extrapolation are themselves not known in advance, but are rather learned on the fly. The need for joint learning and extrapolation distinguishes this thesis from some existing work [32, 33, 34] that predicts missing sensor values assuming a *previously known* correlation structure between sensors, or a known temporal pattern.

We apply the results to an example case study of a New York City crisis in the aftermath of Hurricane Sandy. Many gas stations, pharmacies, and grocery stores around New York City were closed after the hurricane, resulting in severe supply shortage that lasted several days. The outages were correlated, since different stores shared suppliers or power. Our study shows the degree to which extrapolation could infer gas, food, and medical supply availability during the crisis in the absence of complete and fresh information.

To the best of our knowledge, no previous work has been applied to real-world disaster response scenarios where inference algorithms were investigated that (i) specifically address the bimodal nature of damage propagation and that (ii) require very little training data. Our thesis fills in this gap by analyzing the example of New York City gas crisis in the aftermath of Hurricane Sandy via real data traces.

1.5 Organization of this Thesis

The remainder of this theis is organized as follows. We present the general system design and illustrate prediction challenges in Chapter 2. A new algorithm that addresses these challenges via appropriate switching between spatial and temporal extrapolation is presented in Chapter 3. An evaluation is presented in Chapter 4. Chapter 5 demonstrates the working system. Chapter 6 reviews related work. We conclude the thesis in Chapter 7.

CHAPTER 2

SYSTEM MODEL

We consider a model of participatory sensing applications in which the reported state is binary. It is desired to obtain the state of several points of interest. A central collection node (e.g., the command center) collects the state from participants who make observations and report them later.

The time when participants report their observations may vary. Measurements that are older than some threshold, are deemed stale. Hence, at any given time, there may be "blind points" in the PoI map generated by participants, where fresh information is not available. The challenge is to infer the missing state automatically and accurately.

The main contribution of this work lies in addressing the extrapolation problem in scenarios consistent with disaster response. Two main challenges characterize those scenarios:

- *Disruptive change:* By definition, disasters are unique disruptive events that invalidate normal data trends, making prediction based on historical (time-series) trends largely incorrect.

- *Scarcity of training data:* Since disasters are rare and generally unique, there is very little training data that one can rely on. To understand the worst case, we restrict the prediction algorithm to use only training data available from the current disaster itself. This scarcity of data severely limits the complexity of prediction models that can be used.

We consider applications where today's information matters the most and people prefer undertaking some actions based on best-effort guessing to obtaining exact data at a certain delay. For example, in the case of finding gas stations around New York City that are operational after hurricane Sandy, if one needed to fill up their car now, yesterday's gas availability would be of less use. The challenge is therefore to infer the *current* missing PoI state.

We assume that old (and hence potentially stale) information on PoI state is available. For example, in disaster response scenarios, volunteers might physically report back to the command center daily, which makes yesterday's information available at the center. We call the maximum reporting latency, a *cycle*. Hence, by definition, the backend server knows the state of all PoI sites in previous cycles, but has only partial information in the current cycle. This assumption simplifies our algorithmic treatment. It can easily be relaxed allowing for information gaps in previous cycles as well, since such gaps can always be filled in using the same extrapolation algorithm, applied to past state.

2.1 Problem Statement and Solution Challenges

More formally, our participatory sensing system can be characterized by a weighted graph $G = (V, E)$, $|V| = n$, $|E| = m$, where the node set V represents the n PoIs. We assume that set V is known and remains unchanged. The link set E represents the correlations among PoIs.

One way to compute links E, is to apply the Kendall's Tau statistical method [35] to estimate correlations. More concretely, assume two PoIs, x and y, have data $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$. The Kendall's Tau correlation coefficient, denoted by $KT(x, y)$, can be represented as:

$$KT(x, y) = 1 - \frac{1}{n} \sum_{i=1}^{n} XOR(x_i, y_i) \tag{2.1}$$

Each edge (x, y) between PoI nodes x and y has a weight, $w_{xy} = KT(x, y)$, representing the correlation value. The link set E may be reduced by setting a predefined threshold such that only links with correlations higher than the threshold are retained.

The extrapolation algorithm takes partial state of PoI sites in the current cycle, historical data of PoI sites in previous cycles, and the relationships (i.e., edges) learned so far as inputs. It then infers the current state of missing PoI sites.

As argued above, scarcity of training data renders complex prediction models, such as ARIMA and various data mining models [36], ineffective. For example, on the 4th day of a disaster, we have only 3 past training points, which might be fewer than the number of parameters in some models. This means that our prediction model would have to be very simple. Indeed a contribution of this work lies in arriving at a very simple model that works well with little data, as opposed to beating the current mature state of the art in time-series prediction from large data sets.

We first consider several obvious simple heuristics that can be used for extrapolation. To illustrate the impact of insufficient training data, we also consider ARIMA [36], a standard (and powerful) time series analysis method for non-stationary processes, commonly used in complex forecasting tasks, such as forcasting financial systems [37]. The performance of these solutions will determine whether or not a new extrapolation approach is needed.

- *Random*: It is the most trivial baseline in which the status of missing sites is guessed at random. It shows what happens when no intelligence is used in guessing.

- *BestProxy*: It uses the Kendall's Tau method to find actual pairwise (spatial) correlations between PoIs and predicts missing state based on the state of the best neighbor (i.e., the PoI that has the largest correlation with the one being predicted). It is an example of exploiting local spatial correlations, where state of an individual node is predicted from state of another (well-chosen) *individual node*.

- *Majority*: It computes the majority state of all known PoIs and predicts all missing state to be the same as the majority state. This heuristic is another example of exploiting spatial correlations. It lies at the other end of the spectrum from *BestProxy*, in that it exploits a global notion of spatial correlations, where state of an individual node is predicted from *global state*.
- *LastKnownState*: It explores temporal correlations among PoI sites. Namely, the predicted state today is set equal to the last known state.
- *ARIMA*: This, in principle, is one of the most general forecasting methods for time series data that assumes an underlying non-stationary process [36].

Note that, we include *Random* to understand the baseline performance of a prediction algorithm that has no intelligence. *Best Proxy*, and *Majority* are different versions of algorithms that exploit spatial correlations. *LastKnownState* is a simple way of exploiting temporal correlations. *ARIMA* is a state of the art forecasting method. It is included to illustrate the inefficiency of such methods when training data is minimal. The performance of the above baselines is discussed next.

2.2 New York City Crisis

The dataset used here is the New York City crisis after 2012 US Hurricane Sandy. In November 2012 [38], Hurricane Sandy made landfall in New York City. It was the second-costliest hurricane in United States history (surpassed only by hurricane Katrina) and the deadliest in 2012. The hurricane caused wide-spread shortage of gas, food, and medical supplies as gas stations, pharmacies and (grocery) retail shops were forced to close. The shortage lasted about a month. Recovery efforts were interrupted by subsequent events, hence triggering alternating relapse and recovery patterns.

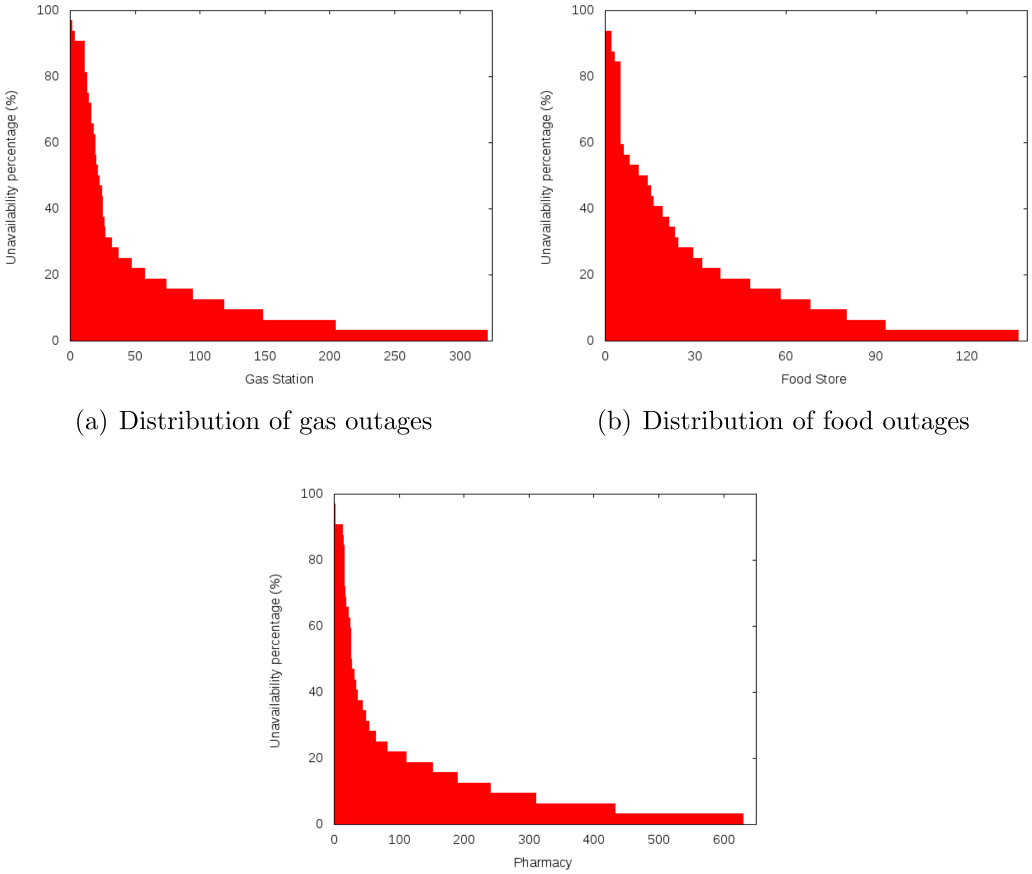
The daily availability of gas, food, and medical supplies was documented by the All Hazard Consortium (AHC) [39], which is a state-sanctioned non-profit organization focused on homeland security, emergency management, and business continuity issues in the mid-Atlantic and northeast regions of the United States. Data traces[1] were collected in order to help identify locations of fuel, food, hotels and pharmacies that may be open in specific geographic areas to support government and/or private sector planning and response activities. The data covered states including West Virginia, Virginia, Pennsylvania, New York, New Jersey, Maryland, and District of Columbia. The information was updated daily (i.e., one observation per day for each gas station, pharmacy, or grocery shop). To give an example of the extent of damage, Figure 2.1(a) shows the distribution of the percentage of time that each of 300+ affected gas stations in the New York area was *unavailable* during the first *month* following the hurricane. We can see that 40 gas stations were not available for more than 1 week and some were out for almost the whole month. Similarly, Figure 2.1(b) shows the distribution of outage for affected food stores and Figure 2.1(c) shows the distribution of outage for affected pharmacies.

Figure 2.2(a) shows the percentage of available gas stations in each cycle. It is clear that there is a disruptive change occurred in the 7th cycle (start from 0). Similar trends are observed for pharmacy and food supply, as shown in Figure 2.2(c) and Figure 2.2(b), respectively.

2.3 Failure of Individual Baselines

With these PoI sites and input data as ground truth, we evaluate the baselines described. The metrics we use are accuracy of inference and amount of data needed. We break time into cycles as discussed earlier. We set each cycle to

[1] Available at: http://www.ahcusa.org/hurricane-Sandy-assistance.htm

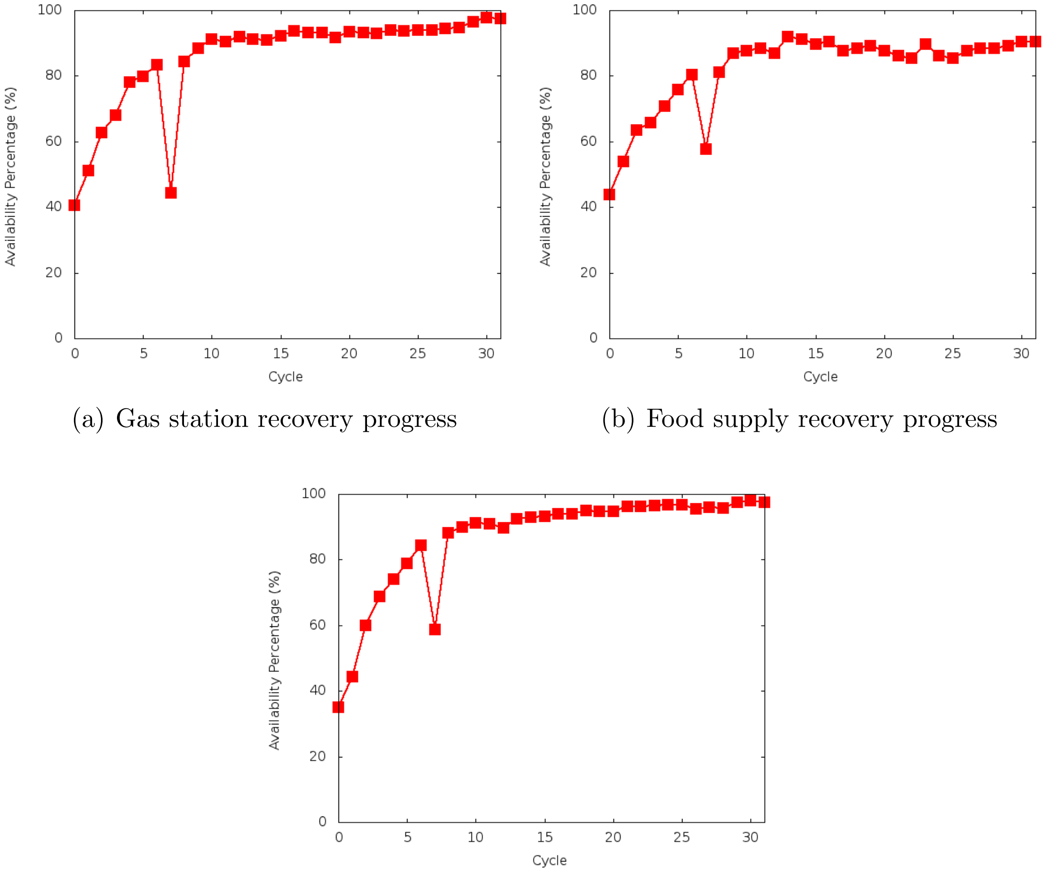


(a) Distribution of gas outages

(b) Distribution of food outages

(c) Distribution of pharmacy outages

Figure 2.1: Distribution of public services outages



(a) Gas station recovery progress

(b) Food supply recovery progress

(c) Pharmacy recovery progress

Figure 2.2: Recovery progress of public services

a day to coincide with the AHC trace. We then plot the performance of the above baselines when a configurable amount of today's data is available (in addition to all historic data since the beginning of the hurricane).

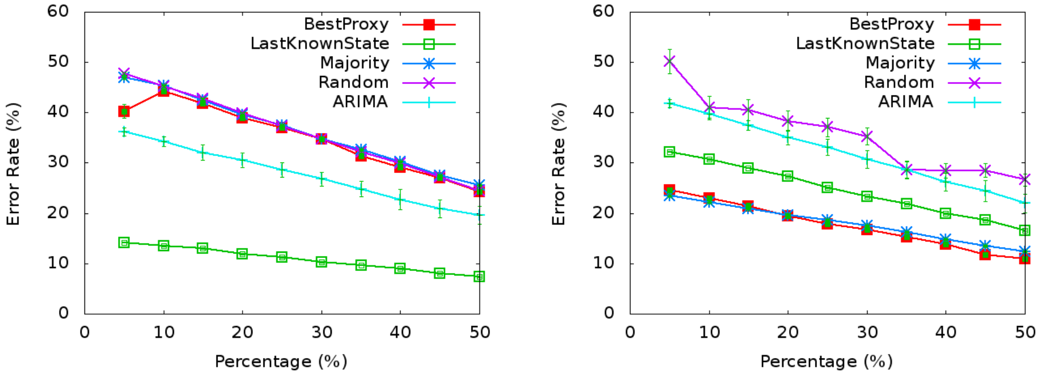
We evaluate the solutions on November 3rd, and November 8th. November 8th corresponds to a period of disruptive change due to a second snow storm that hit after Sandy, causing massive temporary relapse of recovery efforts due to new power outages, followed by a quick state restoration to the previous recovery profile. November 3rd is an example of a period of little change, when damage was incurred but recovery efforts have not yet been effective. The same trend was observed for all datasets we have, namely, gas, pharmacy, and food.

Figure 2.3, Figure 2.4, and 2.5 plot the prediction error with standard deviation shown as error bars in availability of gas stations, food (grocery shops), and pharmacies, respectively. In each figure, sub-figures (a) and (b) refer to November 3rd and November 8th, respectively.

The reader is reminded that we assume that, on a given day, one knows the status of only a fraction of PoIs (where the status refers to whether they are open or closed). The purpose is to extrapolate this data and find out the status of the remaining ones. The horizontal axis in the aforementioned figures varies the percentage of PoIs whose status is known on the indicated day from 5% to 50%. To eliminate bias that may result from knowing the status of specific PoIs, each point (corresponding to a specific percentage of PoIs whose status is known) is an average of 50 different experiments. In each experiment, a different random set of PoIs is selected as known (adding up to the required percentage). The results shown are the average of the 50 experiments.

Consider Figure 2.3-a and Figure 2.3-b, that illustrate the overall prediction error rate for *gas availability* on November 3rd and 8th, respectively, as a function of the percentage of PoIs whose status is known that day. On the vertical axis, the performance of baselines is compared.

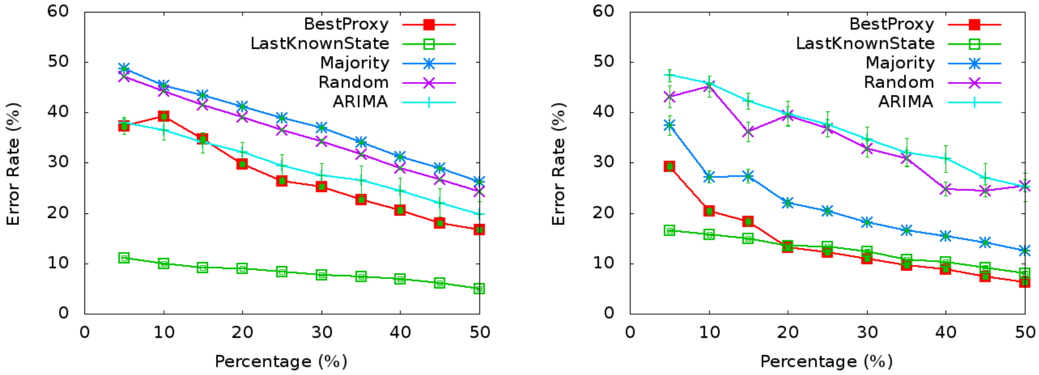
Figure 2.4-a and Figure 2.4-b similarly compare the performance of the



(a) Error rate on November 3rd.

(b) Error rate on November 8th.

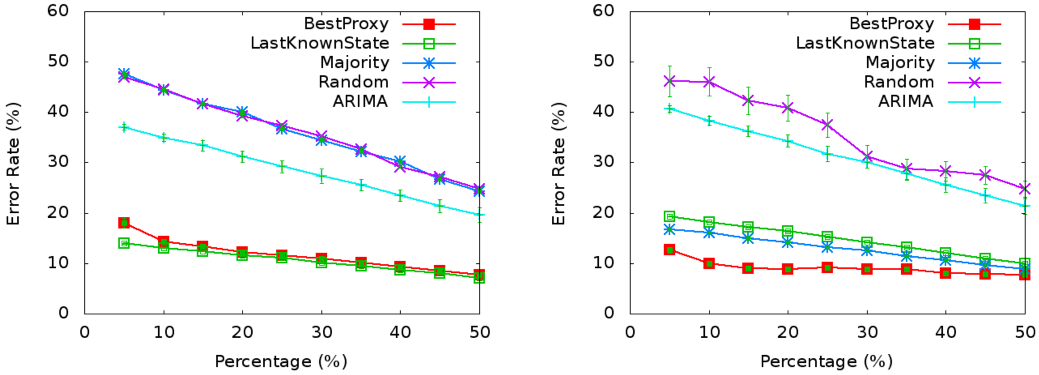
Figure 2.3: Comparing baselines to predict gas availability after Sandy



(a) Error rate on November 3rd

(b) Error rate on November 8th

Figure 2.4: Comparing baselines to predict food availability after Sandy



(a) Error rate on November 3rd

(b) Error rate on November 8th

Figure 2.5: Comparing baselines to predict pharmacy availability after Sandy

baselines in predicting food availability on November 3rd and 8th. Figure 2.5-a and Figure 2.5-b compare the performance of the baselines in predicting pharmacy availability on November 3rd and 8th.

It can be seen that no single baseline does consistently well in all figures. Specifically, LastKnownState does remarkably well on November 3rd, when the change was minimal from the day before. This is especially true for gas and food (grocery) availability prediction, where it beats the next heuristic by a wide margin. However, BestProxy does better on November 8th, when a second snow storm hits and its aftermath causes a lot of perturbation. More specifically, the error rate of BestProxy is around 8% lower than LastKnownState on November 8th. BestProxy clearly outperforms LastKnownState that day for gas and pharmacy availability prediction, and ties for food availability prediction. Majority does poorly on November 3rd and better (but not best) on November 8th. Random does worse. Very interestingly, ARIMA does only marginally better than Random and much worse than the best heuristics on either day. This is attributed to the lack of sufficient training data, and the challenges caused by disruptive changes in the time-series. Also notice that, the standard deviations for all baseline methods are quite small compared to the error rates, which indicates that which PoIs are known does not have a significant effect on the performance. This is quite important for at least two reasons. One is in practice we cannot predict which points of interest we will know at any certain time. The other is that this indicates all these baseline methods can be used as building blocks of some more complex algorithms, for example, the one we will show in the following chapter.

The results confirm that algorithms that do spatial extrapolation (such as BestProxy) are better on days of more change, whereas algorithms that do temporal extrapolation (such as LastKnownState) are better on days of less change. The results also suggest that, due to lack of training data, complex prediction models that normally do well, such as ARIMA, are ineffective. We leverage these observations to guide the design of new algorithms that consistently offers the best performance. The algorithms should appropriately adapt to periods of change versus periods of calm, that is, taking both spatial and temporal factors into account. Attempts to design such algorithms are described in the next chapter.

CHAPTER 3

HYBRID ALGORITHMS

3.1 First Attempt: Spatial-temporal Statistical Model

First, it is natural to think of building a statistical model that takes both spatial and temporal correlation into account.

3.1.1 The Model

The inference component takes partial state of PoI sites in current cycle, historical data of PoI sites in previous cycles, and the inference relationships learned so far as inputs. It then interpolates the current state of missing PoI sites. The relationship between PoI sites and their neighboring sites defines the joint distribution of the state of these sites. Therefore, by feeding the historical data into the joint distribution, this relationship can be estimated and then the missing PoI sites in the current cycle are interpolated using this estimated relationship.

Let X_{ij} be the status of PoI j at cycle i. X_{ij} is binary. Taking the gas station application as an example, $X_{ij} = 1$ indicates that gas station j has gas in cycle i, and $X_{ij} = 0$ means no gas. The conditional distribution of X_{ij} based on its neighbors is modeled by a logistic regression,

$$P(X_{ij} = 1|N_j) = \frac{\exp(z'_{ij}\beta + \rho \sum_{k \in N_j} X_{ik})}{1 + \exp(z'_{ij}\beta + \rho \sum_{k \in N_i} X_{ik})} \tag{3.1}$$

where z_{ij} is a collection of covariates that are related to the status of PoI j at cycle i, and N_j is a collection of sites that are neighbors of PoI j. In this model, β measures the impact of the covariates and ρ measures the between-site dependence.

By the Hammersley-Clifford Theorem [40], the specified conditional distribution well defines the joint distribution

$$\begin{aligned}
& P(X_{ij}, j = 1, ..., N|(\beta, \rho)) && (3.2)\\
= \quad & c_i(\beta, \rho)^{-1} \exp(\sum_{j=1}^{N} X_{ij} z'_{ij}\beta + \frac{1}{2}\sum_{j=1}^{N}\sum_{k\in N_j} X_{ij}X_{ik}) && (3.3)
\end{aligned}$$

where

$$c_i(\beta, \rho) = \sum_{x} \exp(\sum_{j=1}^{N} x_{ij} z'_{ij}\beta + \frac{1}{2}\sum_{j=1}^{N}\sum_{k\in N_j} x_{ij}x_{ik}) \qquad (3.4)$$

The constant $c(\beta, \rho)$ ensures a proper probability distribution. In case that some sites have missing values, for example, at time i, the sites with id in set s_i have missing status. Here, we use X_{i,s_i} to denote the possible status of sites s_i and X_{i,s_i^c} to denote the status of the observed sites. Then we immediately have the joint distribution of X_{i,s_i^c},

$$\begin{aligned}
& P(X_{i,s_i^c}) && (3.5)\\
= \quad & c_i(\beta, \rho)^{-1} \sum_{x_{i,s_i^c}} \exp(\sum_{j=1}^{N} x_{ij} z'_{ij}\beta + \frac{1}{2}\sum_{j=1}^{N}\sum_{k\in N_j} x_{ij}x_{ik}) && (3.6)\\
= \quad & c_i(\beta, \rho)^{-1} b_i(\beta, \rho, X_{i,s_i^c}) && (3.7)
\end{aligned}$$

The parameters β and ρ can be estimated by maximum likelihood estimator (MLE). To simplify the notations, we let $X_i = (X_{i,s_i}, X_{i,s_i^c})'$ be the observed data at cycle i. Therefore we have the log likelihood,

$$\log L(X, \beta, \rho) = \sum_{i} (\log b_i(\beta, \rho, X_{i,s_i^c}) - \log c_i(\beta, \rho)) \qquad (3.8)$$

One challenge introduced by this approach is that, the two constants, $c(\beta,\rho)$ are hard to compute even with a moderately large N and the computation of $b(\beta,\rho,X_{i,s_i^c})$ depends on the number of missing values. Therefore, instead of computing the exact $c(\beta,\rho)$ and $b(\beta,\rho,X_{i,s_i^c})$, we estimate the parameters via the Monte Carlo maximum likelihood estimator (MCMLE) [41], where $c_i(\beta,\rho)$ and $b_i(\beta,\rho,X_{i,s_i^C})$ are approximated by the Monte Carlo Markov Chain (MCMC) method [42]. Note that,

$$\begin{aligned}
& \frac{c_i(\beta_1,\rho_1)}{c_i(\beta,\rho)} && (3.9)\\
= \quad & c_i(\beta,\rho)^{-1}\sum_x \exp(f(x,\beta_,\rho_)) && (3.10)\\
= \quad & E_{x|(\beta,\rho)}\frac{\exp(f(x,\beta_1,\rho_1))}{\exp(f(x,\beta,\rho))} && (3.11)
\end{aligned}$$

where $E_{x|(\beta,\rho)}$ means the expectation is over x based on parameter (β,ρ) and

$$f(x,\beta,\rho)=\sum_{j=1}^{N} x_{ij}z'_{ij}\beta+\frac{1}{2}\sum_{j=1}^{N}\sum_{k\in N_j} x_{ij}x_{ik}$$

Here, (β_1,ρ_1) is the parameter values in the optimization routine from the previous iteration. Then $c(\beta,\rho)^{-1}c(\beta_1,\rho_1)$ is approximated by

$$\frac{1}{M}\sum_{r=1}^{M}\exp(f(x^{(r)},\beta_1,\rho_1)-f(x^{(r)},\beta,\rho)) \tag{3.12}$$

where $x^{(1)},...,x^{(M)}$ are generated from distribution $P(x_i|(\beta,\rho))$. Here (β_1,ρ_1) is used in the approximation to improve the accuracy by importance sampling. Similarly, if a number of sites have missing values,
$b_i(\beta,\rho,X_{i,s_i^c})^{-1}b_i(\beta_1,\rho_1,X_{i,s_i^c})$ is approximated by

$$\frac{1}{M}\sum_{r=1}^{M}\exp(f(x^{(r)},\beta_1,\rho_1)-f(x^{(r)},\beta,\rho)) \tag{3.13}$$

where $x^{(1)}, ..., x^{(M)}$ are generated from conditional distribution $P(x_i|(\beta, \rho))$ given X_{i,s_i^c}.

The MCMC process can be implemented using Gibbs sampler [41]. At each iteration, a Monte Carlo sample is generated by the conditional distribution of each individual PoI site given their neighbors' status from the previous iteration. To estimate $c_i(\beta, \rho)$, the status of all PoI sites is updated at each iteration. To estimate $b_i(\beta, \rho)$, the status of known PoI sites is fixed and only the status of the unknown PoI sites is updated.

When the majority of historical data is available, β and ρ can also be estimated using a pseudo-likelihood (PL) approach [43] to accelerate the inference process. Instead of using the exact likelihood, the parameters (β, ρ) are estimated by maximizing the pseudo likelihood,

$$\prod_i \prod_j \frac{\exp(z'_{ij}\beta + \rho \sum_{k \in N_j} X_{ik})}{1 + \exp(z'_{ij}\beta + \rho \sum_{k \in N_i} X_{ik})}$$

The PL approach bypasses the estimate of $c_i(\beta, \rho)$, and hence is computationally much more efficient.

The parameters β and ρ will be estimated with cumulative data and as the data cumulates, the estimates $\hat{\beta}$ and $\hat{\rho}$ become more robust. If at a new time point, only a subset of full stations can be observed, the status of other stations can be estimated using the Gibbs sampler based $\hat{\alpha}$ and $\hat{\beta}$, similar as estimating $b_i(\beta, \rho)$. Then the estimated probability that the status of j^{th} station is 1 is the mean of the MCMC samples at j^{th} station, denoted by $\hat{p}_j$. Our proposed approach can also provide the variability for interpolated probabilities, which intuitively tells how reliable the results are. The uncertainty of these estimates can be estimated using a parametric bootstrap as follows:

1. generate simulated data using $\hat{\beta}$ and $\hat{\rho}$
2. estimate β and ρ based on the simulated data. Denote these estimators as $\tilde{\beta}^{(i)}$ and $\tilde{\rho}^{(i)}$.

3. estimate the probability of stations of the unknown stations using $\tilde{\beta}$ and $\tilde{\rho}$, denoted by $\tilde{p}_j^{(i)}$

4. repeat step (1)-(3) B times

Then the variance of $\hat{p}_j$ can be estimated by,

$$var(\hat{p}_j) = \frac{1}{B-1}\sum_{i=1}^{B}(\tilde{p}_j^{(i)} - \hat{p}_j)^2$$

and a $100 \times (1-\alpha)\%$ confidence interval for p_j is $(\tilde{p}_j(\alpha/2), \tilde{p}_j(1-\alpha/2))$, where $\tilde{p}_j(q)$ is the q percentile of $\tilde{p}_j^{(1)}, ..., \tilde{p}_j^{(B)}$.

To summarize, we use a spatial logistic model to characterize the relationship between each PoI site and its neighboring sites. This relationship is estimated using MCMLE based on historical data. A MCMC sampling procedure is used to estimate the normalization parameters while deriving the MCMLE. Then the distribution of each missing PoI site is estimated using these MCMLE. Finally, we also use a bootstrap procedure to derive the uncertainty in our estimated probabilities.

3.1.2 Result

We implement this model in R and apply it to the Sandy Gas dataset we have. However, the parameters used in the model cannot be learned until cycle 10. And similar results are produced in the food and pharmacy datasets. This again proves that due to lack of training data, complex prediction models that normally do well.

That spatial extrapolation heuristics alone and temporal extrapolation heuristics alone do very well in some days, though not in all days, implies that switching between the two might be able to achieve consistently good performance. Note that, we do not aim to outperform any one heuristic at *all* times. Rather, our aim is to match consistently the best performing heuristic at any time, even though that heuristic changes, depending on circumstances. Such an algorithm is described next.

3.2 Second Attempt: A Hybrid Prediction Algorithm

The above study leads to two insights that help develop an algorithm for data extrapolation in disaster response scenarios:

- *Insight #1:* The first insight is that our algorithm should be able to switch between spatial and temporal prediction modes. On days with little change, LastKnownState does really well and should be the default prediction. On days where change is abundant, spatial correlations are more appropriate to use for prediction.

- *Insight #2:* The second insight lies in refining the notion of spatial correlations to be used for prediction. Since our default prediction is LastKnownState (i.e., no change), we need spatial correlations only to predict *change*. Hence, rather than using Kendall's Tau correlation to find a good proxy, we seek a proxy that helps predict change only. In other words, we seek a proxy whose *state changes* (and not overall state) are most correlated with those of the target to be predicted.

The second insight is intuitive in retrospect. Just because two gas stations were out of gas or out of power for a long time, does not mean their state changes are correlated. What's more indicative is whether or not they lost gas or power at the same time. The latter gives a better indication that if gas or power is restored to one, it may also be restored to the other.

More concretely, consider two PoIs, x and y, that have state $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$. Let x_n be unknown (i.e., it has not yet been delivered). Let us define the change time series as $(dx_1, dx_2, ..., dx_n)$ and $(dy_1, dy_2, ..., dy_n)$, where $dx_i = x_i - x_{i-1}$ and $dy_i = y_i - y_{i-1}$ (we assume that $x_0 = 1$ and $y_0 = 1$ (everything was working before the disaster). To predict x_n (or equivalently predict the change dx_n), we would like to find a proxy y, whose current status is known and whose changes are maximally correlated with changes in x. We can then use dy_n to predict dx_n and hence predict x_n. To do so, we compute $P(\mathit{change\ in\ x} | \mathit{same\ change\ in\ y})$ for all gas stations y whose current state is known. This probability can be approximated by:

$$P(\mathit{change\ in\ x} | \mathit{same\ change\ in\ y}) = \frac{count(dx_i = dy_i)}{count(dy_i \neq 0)} \tag{3.14}$$

where $count()$ is a function that counts the number of times the condition in its argument was true for $1 \leq i \leq n-1$. The best proxy for (predicting change in) x becomes the y that maximizes the above probability. Let us call such a y, y_{best}. Let the resulting probability, $P(\mathit{change\ in\ x} | \mathit{same\ change\ in\ y^{best}})$ be denoted P^{best}. Using insight #1 above, the sought algorithm is as follows:

Algorithm 1 Enhanced Best Proxy (x, n)

```
1: IF ( $P^{best} \geq$ threshold T )
2: use **SpatialPrediction**
3: ELSE
4: use LastKnownState (i.e., $x_n = x_{n-1}$)
5:
6: **SpatialPrediction**
7: IF (($dy_n^{best}$ is not zero) AND ($y_{n-1}^{best} = x_{n-1}$))
8: THEN $x_n = y_n^{best}$
9: ELSE use LastKnownState (i.e., $x_n = x_{n-1}$)
```

Lines 1 to 4 indicate that the algorithm alternates between spatial and temporal prediction depending on whether the best found proxy for the target x is sufficiently good (i.e., better than a threshold, T). When spatial prediction is used, we predict that state of x will change (i) if it was the same as the state of the best proxy, and (ii) if the state of that proxy changed. Otherwise, we predict no change. Note that, it is possible that there is no best proxy for a certain PoI. When choosing the best proxy, we require one PoI to have at least a certain number of changes in its own history so far. To see why this is necessary, imagine we are now considering choosing PoI A as B's proxy, however, A has only 1 state change in its history and the change happened in the same cycle as B. In this case, A's P score will be 1, which is always larger than or equal to T and all other proxy candidates. Therefore, A will be selected as B's best proxy, although A is actually not a strong candidate, especially when we have other candidates have scores, for example, 9/10.

It remains to derive the optimal value of the threshold, T. Let M denote the fraction of PoIs that had state = 1 in the last cycle. Hence, $1 - M$ is the fraction of PoIs with state = 0. Furthermore, let F denote the fraction of PoIs (that we are aware of so far) that change state in the current cycle. The optimal value of T is one that minimizes misprediction probability.

The above algorithm mispredicts either (i) when spatial prediction is used and it is wrong, or (ii) when temporal (LastKnownState) prediction is used and it is wrong. Hence, misprediction probability, P_m, is equal to the sum of spatial misprediction probability, P_{sm}, and temporal misprediction probability, P_{tm}. Below, we compute these probabilities.

Spatial Misprediction: From line 7 of Algorithm 1, spatial misprediction occurs when (i) P^{best} exceeds the threshold T and (ii) the best proxy has the same state as x in the last cycle, yet (iii) they have different states in the current cycle. Note that, the first two conditions are what invokes spatial prediction. The third condition causes that prediction to err.

Clearly, the probability of the first condition, $P(P^{best} > T)$, decreases with increasing threshold, T. Let us approximate $P(P^{best} > T) = 1 - T$. The probability of the second condition is simply $1 - 2M(1 - M)$. Since P^{best} is the probability of a correlated change in x (given a change in the proxy), the probability of the third condition (a misprediction) is approximately $1 - P^{best}$. We know that $P^{best} > T$. Assuming that P^{best} could be uniformly anywhere above T, we can replace $1 - P^{best}$ by $(1 - T)/2$. The spatial misprediction probability is then the product of probabilities of the three conditions above, leading to the expression:

$$P_{sm} = (1 - T)[1 - 2M(1 - M)](1 - T)/2 \tag{3.15}$$

Temporal misprediction occurs when the algorithm resorts to temporal prediction and is wrong. According to the algorithm, temporal (LastKnownState) prediction occurs when (i) P^{best} exceeds the threshold T, but (ii) the best proxy does not have the same state as x in the last cycle, or when (iii) P^{best} is less than the threshold T. In either case, a misprediction occurs if the state of x changes (hence contradicting LastKnownState). The latter probability can be approximated by F, the fraction of nodes we know of that changed state today. Hence:

$$\begin{aligned} P_{tm} &= (1 - T)[2M(1 - M)]F \\ &+ [1 - (1 - T)]F \end{aligned} \tag{3.16}$$

Recall that misprediction probability, P_m, is the sum of P_{sm} and P_{tm}. Hence, from Equation (3.15) and Equation (3.16), we get:

$$\begin{aligned} P_m &= (1 - T)[1 - 2M(1 - M)](1 - T)/2 \\ &+ (1 - T)[2M(1 - M)]F \\ &+ [1 - (1 - T)]F \end{aligned} \tag{3.17}$$

The optimal threshold, T, is one that minimizes the above probability. The equation is a quadratic function of T. Because the coefficient of T^2 is $[1 - 2M(1 - M)]$, which is always positive, the optimal threshold can be found by setting the derivative of the above function to zero and enforcing the natural constraints on values of probability (that they are between 0 and 1). In other words:

$$\begin{aligned}\frac{dP_m}{dT} &= -(1-T)[1-2M(1-M)] \\ &- [2M(1-M)]F \\ &+ F = 0\end{aligned} \tag{3.18}$$

subject to the constraint $0 \leq T \leq 1$. After some rearranging and algebraic manipulation, we get:

$$T = 1 - F \tag{3.19}$$

Unfortunately, we do not know the probability of change, F, in advance. In the absence of further knowledge, we can design for $F = 0.5$. In this case, T = 0.5.

Next, we will evaluate this algorithm on the Sandy datasets.

CHAPTER 4

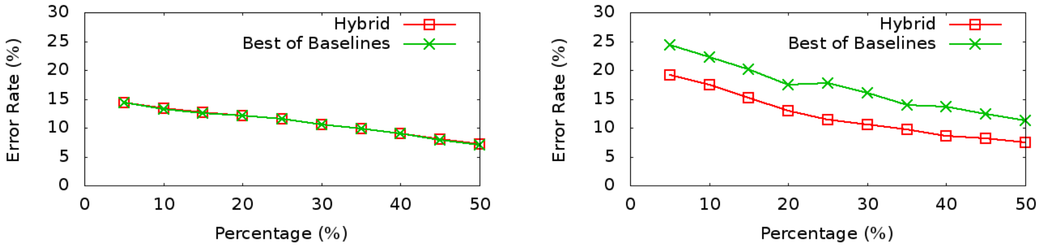
EVALUATION

In this chapter, we evaluate the hybrid approach presented above versus the baselines described earlier in Section 2.1 (i.e., Random, LastKnownState, Best-Proxy, Majority, and ARIMA). For ground truth, we use the same data set, featuring the daily status of gas stations, pharmacies, and food stores in the aftermath of Hurricane Sandy.

4.1 Evaluating Hybrid Algorithm on Period of Calm and Change

First, as before, we opt to predict the status of these PoIs on November 3rd and 8th, as examples of a day of relative calm and a day of significant change. We do so by varying the fraction of PoIs whose state is revealed to the predictor on a given day, and attemtping to predict the rest using each of the compared approaches.

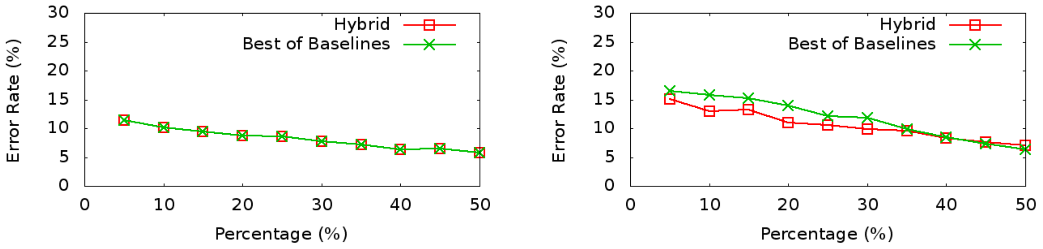
Figures 4.1-a and 4.1-b illustrate the accuracy of prediction of gas availability on November 3rd and 8th, respectively. The horizontal axis shows the percentage of PoIs whose state is known on the given day. As before, each point is the average of 50 experiments featuring different random selections of stations whose status is known. On the vertical axis, two curves are compared. One is the hybrid extrapolation algorithm developed in this thesis. The second is the *best* of the predictions of the five baselines described in Section 2.1. It can be seen that the new algorithm consistently matches or outperforms the best of all others.



(a) Error rate on November 3rd.

(b) Error rate on November 8th.

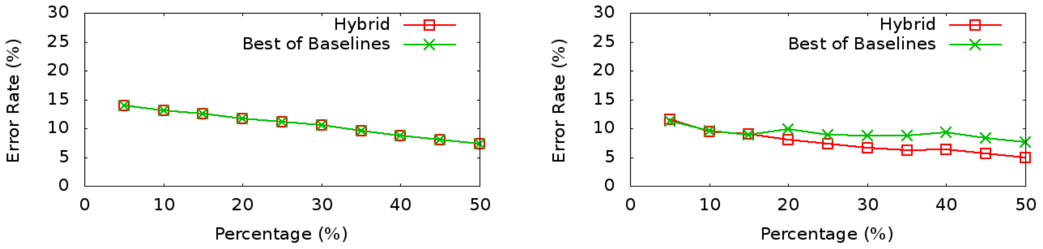
Figure 4.1: Predicting gas availability after Sandy



(a) Error rate on November 3rd

(b) Error rate on November 8th

Figure 4.2: Predicting food availability after Sandy



(a) Error rate on November 3rd

(b) Error rate on November 8th

Figure 4.3: Predicting pharmacy availability after Sandy

Specifically, on November 3rd, the hybrid approach matches the best baseline. This is because it recognizes that change is small, and opts to use LastKnownState, which happens to be the best under the circumstances, as we have seen in Figure 2.3-a). On November 8th, it outperforms the best baseline, which tends to be BestProxy as we have seen in Figure 2.3-b. This is because of the new definition of correlation that it uses, which focuses only on changes, per *Insight #2* discussed earlier.

Figures 4.2-a and 4.2-b repeat the experiment on the food data set. They illustrate the accuracy of prediction of food availability on November 3rd and 8th, respectively. A similar trend is seen, where the hybrid matches the best baseline on November 3rd and outperforms the best baseline on November 8th. Figures 4.3-a and 4.3-b illustrate the same for pharmacies. Further experiments (not shown) demonstrated that the results are largely insensitive to the choice of threshold, T. The superior results presented above can therefore be robustly achieved.

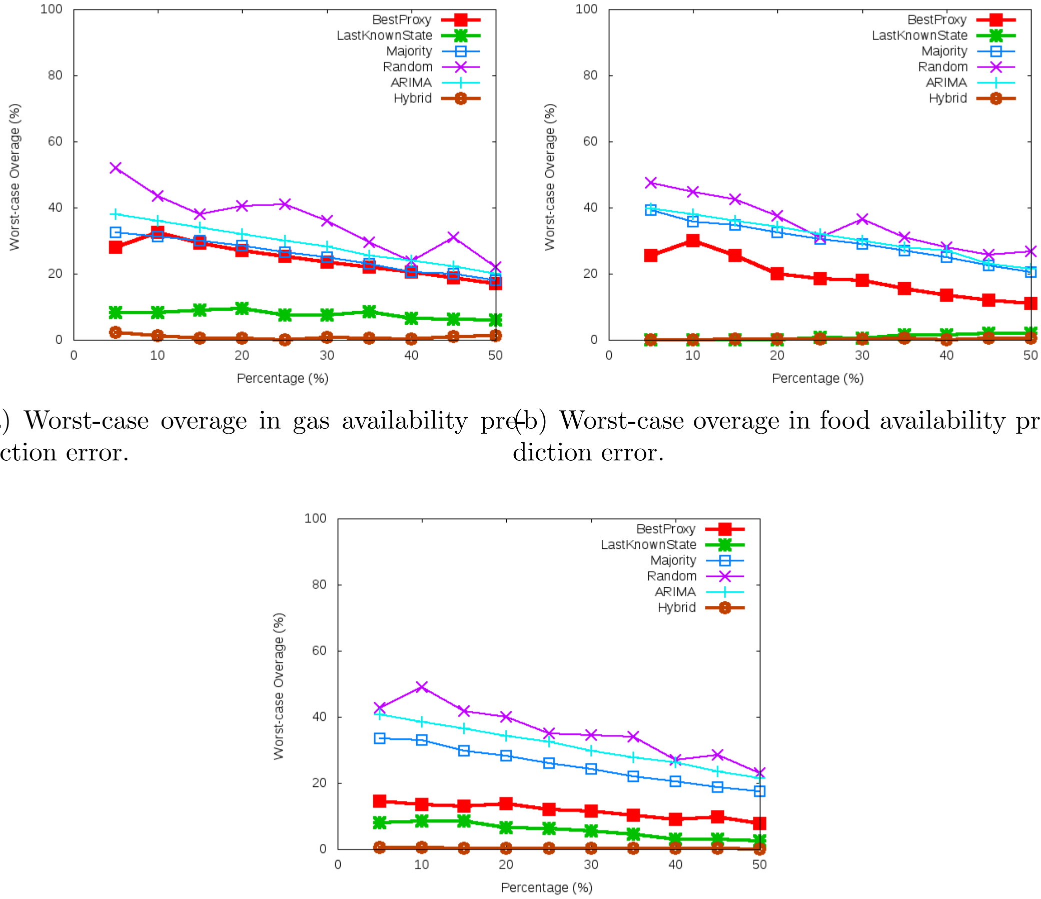
The experimental results presented in this section show that the hybrid approach is as good as or better than the best of all compared algorithms on both November 3rd and November 8th. These two days were selected because of their representative nature, as they exemplified days of calm and days of change, respectively.

4.2 Evaluating Hybrid Algorithm on All Cycles

Next, to show that the above results hold true for other days as well, we compute the *worst case* overage amount by which the prediction error of the hybrid approach, as well as the prediction error of each of the five individual baselines, exceeds the best of the five baselines. Hence, an algorithm that behaves as the best of the baselines under all circumstances will have a worst-case overage of zero. Algorithms that are not consistently the best will have a higher worst-case overage. The results are shown in Figure 4.4, where Figure 4.4-a, Figure 4.4-b, and Figure 4.4-c, are for the case of gas, food, and pharmacy availability prediction, respectively.

In Figure 4.4, the worst-case overage, for each algorithm, is computed by finding the maximum error overage computed over 10 days of the recovery phase (from November 3rd through November 12th). For statistical significance, the performance of each heuristic on each day is first averaged over 50 experiments before the overage is calculated. Consistently with other figures, the horizontal axis shows the percentage of PoIs whose status is known. It is seen that the new Hybrid algorithm has a worst-case overage that is roughly zero. In other words, *it never does worse than the best solution over all days under consideration.*

The figure shows that the overage of other baselines is higher. Their relative prediction (in)accuracy follows roughly the same order in the three data sets. Specifically, LastKnownState is generally the next best algorithm to ours. In the aftermath of disasters, failures take long to fix, so the state changes gradually, making LastKnownState a good predictor most of the time. Errors occur when aftershocks hit or major repairs are made, and are related to the size of such perturbations. BestProxy comes next. Its accuracy depends on how spatially well-correlated the PoI states are. No significant difference is seen between its accuracy in gas and food availability prediction, but pharmacy prediction is better. This can be attributed to the size of the pharmacy data set, shown on the horisontal axis in Figure 2.1(c). Namely, the number of pharmacies is the largest. Hence, the odds of finding a good proxy are better than with the other data sets. Majority comes next after BestProxy. In scenarios where restoration is quicker, PoIs converge to the majority state faster, and the predictor becomes more accurate. Comparing Figure 2.1(a), 2.1(b), and 2.1(c), we can see that pharmacies and gas are restored the fastest, followed by food, which roughly corresponds to how well Majority works in the three cases. Finally, ARIMA and Random consistently do next-to-worst and worst, respectively, showing little variation acorss the data sets. This is be-



(a) Worst-case overage in gas availability prediction error.

(b) Worst-case overage in food availability prediction error.

(c) Worst-case overage in pharmacy availability prediction error.

Figure 4.4: Worst-case prediction error overage of individual solutions

cause their worst-case behavior is random (for ARIMA, it occurs in the very early days), and hence not tightly related to the properties of input data.

In conclusion, Figure 4.4 shows that while some prediction algorithms do best under some circumstances, no baseline does consistently well under all circumstances. The contribution of the new approach lies indeed in proposing a method that adapts intelligently between time-based extrapolation and spatial extrapolation, matching or outperforming the best baseline solution at all times.

CHAPTER 5

SYSTEM INTERFACE

We built a working system to demonstrate the functionality presented in the previous chapters. The system architecture is shown in Figure 5.1.

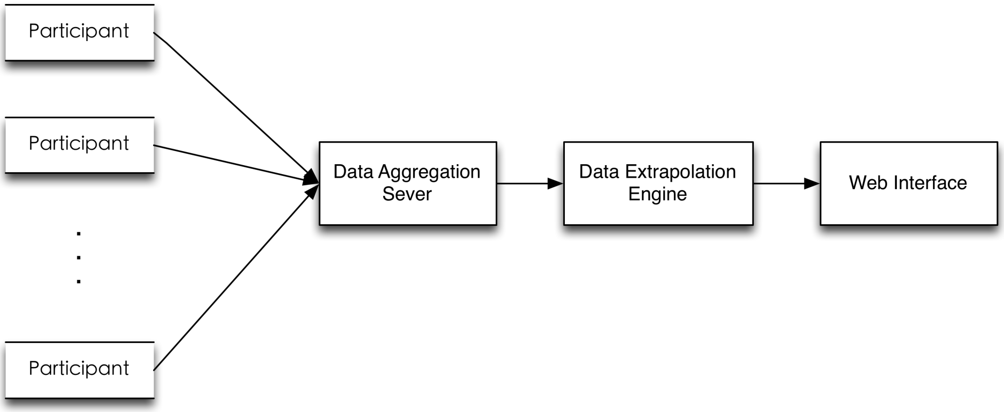


Figure 5.1: System Architecture

There are three major parts, including:

- *Data Aggregation Server*: This is the server that collects data and generates the input for the data extrapolation engine.
- *Data Extrapolation Engine*: This is the core of the system and is described in depth in the previous chapters.
- *Web Interface*: As displayed in Figure 5.2, different colors are used to describe the states of PoIs.

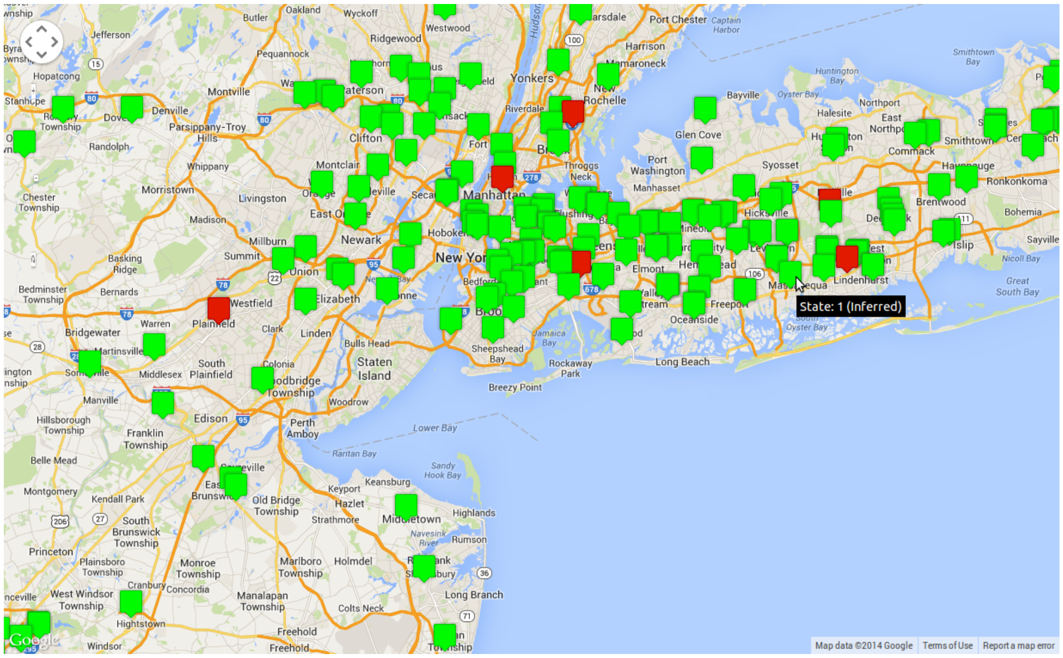


Figure 5.2: Web Interface

CHAPTER 6

RELATED WORK

6.1 Aggregating and Cleaning-up Data

Our work focuses on a new problem in participatory sensing. Namely, the problem of automatically filling in the "blind spots" in reported observations. Past research on participatory sensing describes how to aggregate and clean-up collected data. A survey on analytic challenges in the field recently appeared [44]. For instance, CenWits [45] proposes a participatory sensor network to rescue hikers in emergency situations. The idea is to use the concept of witnesses to convey a subject's movement and location information to the outside world. BikeNet [46] presents a bikers sensor network for sharing cycling related data and mapping the cyclist experience. The Nericell project [47] presents a system that performs rich sensing using smartphones that users carry with them in normal course, to monitor road and traffic conditions. The GreenGPS system [48] provides a service that computes fuel-efficient routes for vehicles between arbitrary end-points, by exploiting vehicular sensor measurements available through the On Board Diagnostic (OBD-II) interface of the car and GPS sensors on smart phones. SignalGuru [49] is a software service that relies solely on a collection of mobile phones to detect and predict the traffic signal schedule, producing a Green Light Optimal Speed Advisory (GLOSA). CarTel is a distributed mobile sensor computing system [50], upon which road sensing application can build. Each CarTel node is comprised of sensing unit and data processing unit and relies on opportunistic wireless connectivity to the Internet or to its peers to communicate with the central data aggregating portal. This thesis complements that past work by looking at the important problem of how to fill in the data gaps. This unique challenge comes from the timeliness constraints in disaster response applications. In the absence of urgency, one can eventually fill in the data gaps by sending (or waiting for) more observers. Hence, there is less need to “guess” them. However, in disaster recovery scenarios, there is no time to wait, so the service provider needs to fill in the gaps immediately as best one can.

6.2 Prediction-based Data Collection in Sensor Networks

Our work is also related to the large body of literature focusing on prediction-based data collection in sensor networks. Le Borgne et al. [51] apply time-series prediction technology to reduce the communication effort while guaranteeing user-specified accuracy requirements on each sensor nodes in wireless sensor networks. Tulone et al. [52] propose a sensor network comprising normal sensor nodes and sink nodes. Sensor nodes transmit their local autoregressive models to sink node, and then sink node uses the models to predict sensor values without communicating with sensors directly. Li et al. [53] presents a similar system, in which sensors check sensed data with predicted values and transmit only deviations from the predication back to the data gathering node. The work by Silberstein et al. [54] points out that one critical weakness of sending changes alone is message failure, which is not negligible in sensor networks. To overcome that, they provide a solution which incorporates the knowledge of the suppression scheme and application-level redundancy in Bayesian inference. Krause et al. [55] develop an algorithm called pSPIEL, which is capable of measuring the predictive quality of sensor locations and then selecting sensor placements at informative and communication-efficient locations. All those researches apply similar prediction technology to ours but focus on improving the communication efficiency while maximizing the quality of collected data.

6.3 Sensor Selection Algorithm Paired with Inference Approach

Our system design is related to state of the art sensor selection algorithms that are paired with inference approaches for missing or incomplete data. For example, Aggarwal et al. formulate the problem of sensor selection, when redundancy relationships between sensors can be expressed through an information network by using external linkage information. They present methods for efficient sensor selection by using regression models to estimate predictability and redundancy [32]. The problem is extended to dynamic sensor selection in data streams [56]. Similarly, PhotoNet [57] provides a picture-collection service for disaster response applications that maximizes situation-awareness. In the aftermaths of disasters, communication infrastructures may not be functional. Under such circumstances, a protocol assigning priorities to images for forwarding and replacement is helpful. Their work designs such a protocol based on the similarity among images. Kobayashi et al. propose a sensor selection method with fuzzy inference for sensor fusion in robot applications [34]. However, this existing work assumes that correlations between data items are known in advance. These correlations are the basis for sensor selection. Also, they assume a stationary process. Biswas et al. proposed a Bayesian inference approach and applied it on a simulated problem of determining whether a friendly agent is surrounded by enemy agents [33]. However, their approach does not work for binary PoI information due to the logistic regression overflow problem. Our work complements these work in that we do not require priori knowledge of the correlations between points of interest. Such knowledge is computed on the fly.

CHAPTER 7

CONCLUSIONS

We presented the design, implementation, and evaluation of an inference-based algorithms for data extrapolation in participatory sensing systems for disaster response applications. It was shown to be capable of accurately predicting the status of PoI sites, when collected data is incomplete. The algorithm exploits correlations among state changes in PoI sites and changes adaptively between temporal and spatial extrapolation. Our experimental results via a real-world disaster response application demonstrate that our algorithm is consistently the best of all compared in terms of prediction accuracy, whereas others may suffer non-trivial degradation. The new algorithm is currently being adapted to more complex prediction tasks (e.g., non-binary variables) and evaluated on new data sets.

REFERENCES

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *In: Workshop on World-Sensor-Web (WSW06): Mobile Device Centric Sensor Networks and Applications*, 2006, pp. 117–134.

[2] N. D. Lane, M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury, and A. Campbell, "Bewell: A smartphone application to monitor, model and promote wellbeing," in *5th International ICST Conference on Pervasive Computing Technologies for Healthcare*, 2011, pp. 23–26.

[3] D. Mendez, A. J. Perez, M. A. Labrador, and J. J. Marron, "P-sense: A participatory sensing system for air pollution monitoring and control," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on.* IEEE, 2011, pp. 344–347.

[4] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe, "Parknet: drive-by sensing of road-side parking statistics," in *Proceedings of the 8th international conference on Mobile systems, applications, and services.* ACM, 2010, pp. 123–136.

[5] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, no. 6, pp. 391–399, 2009.

[6] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[7] R. K. Ganti, N. Pham, Y.-E. Tsai, and T. F. Abdelzaher, "Poolview: stream privacy for grassroots participatory sensing," in *Proceedings of the 6th ACM conference on Embedded network sensor systems.* ACM, 2008, pp. 281–294.

[8] K. L. Huang, S. S. Kanhere, and W. Hu, "Preserving privacy in participatory sensing systems," *Computer Communications*, vol. 33, no. 11, pp. 1266–1280, 2010.

[9] H. Ahmadi, N. Pham, R. Ganti, T. Abdelzaher, S. Nath, and J. Han, "Privacy-aware regression modeling of participatory sensing data," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems.* ACM, 2010, pp. 99–112.

[10] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, and T. Abdelzaher, "Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing," in *Wireless Sensor Networks.* Springer, 2010, pp. 114–130.

[11] J.-S. Lee and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," in *Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on.* IEEE, 2010, pp. 60–68.

[12] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th annual international conference on Mobile computing and networking.* ACM, 2012, pp. 173–184.

[13] J. Paek, J. Kim, and R. Govindan, "Energy-efficient rate-adaptive gps-based positioning for smartphones," in *Proceedings of the 8th international conference on Mobile systems, applications, and services.* ACM, 2010, pp. 299–314.

[14] Y. Chon, E. Talipov, H. Shin, and H. Cha, "Mobility prediction-based smartphone energy optimization for everyday location monitoring," in *Proceedings of the 9th ACM conference on embedded networked sensor systems.* ACM, 2011, pp. 82–95.

[15] V. Lenders, E. Koukoumidis, P. Zhang, and M. Martonosi, "Location-based trust for mobile user-generated content: applications, challenges and implementations," in *Proceedings of the 9th workshop on Mobile computing systems and applications.* ACM, 2008, pp. 60–64.

[16] S. Saroiu and A. Wolman, "Enabling new mobile applications with location proofs," in *Proceedings of the 10th workshop on Mobile Computing Systems and Applications.* ACM, 2009, p. 3.

[17] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh, "Fidelity and yield in a volcano monitoring sensor network," in *Proceedings of the 7th symposium on Operating systems design and implementation.* USENIX Association, 2006, pp. 381–396.

[18] M. Li and Y. Liu, "Underground coal mine monitoring with wireless sensor networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 5, no. 2, p. 10, 2009.

[19] H. Liu, J. Li, Z. Xie, S. Lin, K. Whitehouse, J. A. Stankovic, and D. Siu, "Automatic and robust breadcrumb system deployment for indoor firefighter applications," in *Proceedings of the 8th international conference on Mobile systems, applications, and services.* ACM, 2010, pp. 21–34.

[20] A. Purohit, Z. Sun, F. Mokaya, and P. Zhang, "Sensorfly: Controlled-mobile sensing platform for indoor emergency response applications," in *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on.* IEEE, 2011, pp. 223–234.

[21] "Facebook." http://www.facebook.com/.

[22] "Twitter." http://www.twitter.com/.

[23] "Japan considers using social networks in disaster situations." http://www.engadget.com/2012/08/30/japan-considers-using-social-networks-in-disaster-situations.

[24] "Twitter / search - #hurricanesandy." https://twitter.com/search?q=%23hurricanesandy.

[25] "Waze: Free gps navigation with turn by turn." http://www.waze.com/.

[26] "Gasbuddy: Find low gas prices in the usa and canada." http://www.gasbuddy.com/.

[27] "Ushahidi platform." http://ushahidi.com/products/ushahidi-platform/.

[28] P. Whittle, "The analysis of multiple stationary time series," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 125–139, 1953.

[29] E. McKenzie, "General exponential smoothing and the equivalent arma process," *Journal of Forecasting*, vol. 3, no. 3, pp. 333–344, 1984.

[30] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining kohonen maps with arima time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologies*, vol. 4, no. 5, pp. 307–318, 1996.

[31] P.-F. Pai and C.-S. Lin, "A hybrid arima and support vector machines model in stock price forecasting," *Omega*, vol. 33, no. 6, pp. 497–505, 2005.

[32] C. Aggarwal, A. Bar-Noy, and S. Shamoun, "On sensor selection in linked information networks," in *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, 2011, pp. 1–8.

[33] R. Biswas, L. Guibas, and S. Thrun, "A probabilistic approach to inference with limited information in sensor networks," in *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, 2004.

[34] F. Kobayashi, D. Masumoto, and F. Kojima, "Sensor selection based on fuzzy inference for sensor fusion," in *Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 305–310.

[35] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[36] R. Shumway and D. Stoffer, *Time series analysis and its applications*, 3rd ed. Springer, 2013.

[37] G. Elliott, C. Granger, and A. Timmermann, *Handbook of Economic Forecasting*, 1st ed. North-Holland, 2006, ch. 13.

[38] "Gasoline runs short, adding woes to storm recovery, ny times." http://www.nytimes.com/2012/11/02/nyregion/gasoline-shortages-disrupting-recovery-from-hurricane.html?pagewanted=all&_r=0.

[39] "All hazards consortium." http://www.ahcusa.org/.

[40] N. A. Cressie and N. A. Cassie, *Statistics for spatial data.* Wiley New York, 1993, vol. 900.

[41] J. Zhu, Y. Zheng, A. L. Carroll, and B. H. Aukema, "Autologistic regression analysis of spatial-temporal binary data via monte carlo maximum likelihood," *Journal of agricultural, biological, and environmental statistics*, vol. 13, no. 1, pp. 84–98, 2008.

[42] J. S. Rosenthal, "Minorization conditions and convergence rates for markov chain monte carlo," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 558–566, 1995.

[43] M. Sherman, T. V. Apanasovich, and R. J. Carroll, "On estimation in binary autologistic spatial models," *Journal of Statistical Computation and Simulation*, vol. 76, no. 2, pp. 167–179, 2006.

[44] C. Aggarwal and T. Abdelzaher, *Managing and Mining Sensor Data.* Springer, 2013, ch. Social Sensing.

[45] J.-H. Huang, S. Amjad, and S. Mishra, "Cenwits: a sensor-based loosely coupled search and rescue system using witnesses," in *Proceedings of the 3rd international conference on Embedded networked sensor systems.* ACM, 2005, pp. 180–191.

[46] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "Bikenet: A mobile sensing system for cyclist experience mapping," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 1, p. 6, 2009.

[47] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems.* ACM, 2008, pp. 323–336.

[48] R. K. Ganti, N. Pham, H. Ahmadi, S. Nangia, and T. F. Abdelzaher, "Greengps: A participatory sensing fuel-efficient maps application," in *Proceedings of the 8th international conference on Mobile systems, applications, and services.* ACM, 2010, pp. 151–164.

[49] R. K. Balan, K. X. Nguyen, and L. Jiang, "Real-time trip information service for a large taxi fleet," in *Proceedings of the 9th international conference on Mobile systems, applications, and services.* ACM, 2011, pp. 99–112.

[50] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "Cartel: a distributed mobile sensor computing system," in *Proceedings of the 4th international conference on Embedded networked sensor systems.* ACM, 2006, pp. 125–138.

[51] Y.-A. Le Borgne, S. Santini, and G. Bontempi, "Adaptive model selection for time series prediction in wireless sensor networks," *Signal Processing*, vol. 87, no. 12, pp. 3010–3020, 2007.

[52] D. Tulone and S. Madden, "Paq: Time series forecasting for approximate query answering in sensor networks," in *Wireless Sensor Networks.* Springer, 2006, pp. 21–37.

[53] M. Li, D. Ganesan, and P. Shenoy, "Presto: feedback-driven data management in sensor networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 4, pp. 1256–1269, 2009.

[54] A. Silberstein, G. Puggioni, A. Gelfand, K. Munagala, and J. Yang, "Suppression and failures in sensor networks: A bayesian approach," in *Proceedings of the 33rd international conference on Very large data bases.* VLDB Endowment, 2007, pp. 842–853.

[55] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Near-optimal sensor placements: Maximizing information while minimizing communication cost," in *Proceedings of the 5th international conference on Information processing in sensor networks.* ACM, 2006, pp. 2–10.

[56] C. Aggarwal, Y. Xie, and P. Yu., “On dynamic data driven selection of sensor streams,” in *ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2011.

[57] M. Y. S. Uddin, H. Wang, F. Saremi, G.-J. Qi, T. Abdelzaher, and T. Huang, “Photonet: a similarity-aware picture delivery service for situation awareness,” in *Real-Time Systems Symposium (RTSS), 2011 IEEE 32nd*. IEEE, 2011, pp. 317–326.