Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

An Algorithm to Coordinate Measurements Using Stochastic Human Mobility Patterns in Large-Scale Participatory Sensing Settings

Alexandros Zenonos, Sebastian Stein, Nicholas R. Jennings
Electronics and Computer Science
University of Southampton, UK
{az2g13, ss2, nrj}@ecs.soton.ac.uk

Abstract

Participatory sensing is a promising new low-cost approach for collecting environmental data. However, current large-scale environmental participatory sensing campaigns typically do not coordinate the measurements of participants, which can lead to gaps or redundancy in the collected data. While some work has considered this problem, it has made several unrealistic assumptions. In particular, it assumes that complete and accurate knowledge about the participants future movements is available and it does not consider constraints on the number of measurements a user is willing to take. To address these shortcomings, we develop a computationally-efficient coordination algorithm (Bestmatch) to suggest to users where and when to take measurements. Our algorithm exploits human mobility patterns, but explicitly considers the inherent uncertainty of these patterns. We empirically evaluate our algorithm on a real-world human mobility and air quality dataset and show that it outperforms the state-of-the-art greedy and pull-based proximity algorithms in dynamic environments.

Introduction

Participatory sensing has gained a lot of attention in the research community in recent years. It has been established as a de facto research methodology, engaging citizens in the collection of information using mobile devices they carry on them. Specifically, this concept has been successfully used in environmental monitoring, as people are able to provide information to help in urban planning and public health. For instance, Noisetube (Maisonneuve, Stevens, and Ochab 2010) and Citisense (Nikzad et al. 2012) are projects that monitor noise and air pollution in cities involving citizens using their GPS-enabled mobile phones.

Monitoring environmental phenomena is crucial because of their potentially detrimental effect on human health. Concretely, noise pollution is associated among others with hearing impairment, and ischematic heart disease (Passchier-Vermeer W 2000). Air pollution is responsible for a range of heart-related and respiratory diseases that lead to millions of annual deaths (Seaton et al. 1995; World Health Organization and others 2014).

However, monitoring environmental phenomena using the participatory sensing paradigm is challenging as users are typically willing to take only a limited number of measurements per day (Chon et al. 2013). Furthermore, people have limited information about the environment and are unaware about how their measurements contribute to the overall campaign. Thus, multiple participants may take measurements at the same locations and times, rather than at those locations that are most informative. Consequently, some areas of interest remain unexplored, which might lead to a false or partial understanding of the environmental phenomenon, or people end up doing duplicate work.

In order to address these challenges, an intelligent system for coordinating the measurements taken by participants in participatory sensing campaigns is needed. Such a system can guide participants by suggesting who should take measurements where and when in order to maximize the information collected and thus improve the understanding of the dynamic phenomenon. In doing so, it can exploit probabilistic information about the participants' future mobility patterns, as these are often highly predictable (McInerney et al. 2013; Baratchi et al. 2014). At the same time, the system must consider the limited number of measurements that individuals are willing to take.

Previous work has addressed these problems only partially. For instance, TRACCS (Chen et al. 2014; 2015) attempted to coordinate participants in a different domain. The aim of that paper is to assign humans to tasks based on their mobility patterns to maximize the payoff of tasks in a given time period. However, there is no limit on how many tasks people can do and the tasks are completely independent from each other. Once executed, they are also no longer available. In environmental monitoring, measurements are dependent on each other and since the phenomenon is dynamic, there is a need to re-visit locations to take more measurements. Other work (Zenonos, Stein, and Jennings 2015), attempted to tackle coordination of measurements for environmental monitoring but they only did so partially. In particular, there was no constraint on the number of measurements an individual can take and it was also assumed that there was complete knowledge of human mobility patterns, which is not true in practice. Also, they showed results for up to 250 individuals, which is somewhat limited for a large-scale participatory sensing application. A large-scale participatory sens-

ing campaign can expect many hundreds or even thousands of people depending on the city and the nature of the phenomenon being monitored. The need for a coordination system is also highlighted in the work of (Zaman et al. 2014; D'Hondt et al. 2014), but they focus on receiving feedback from participants rather than actively coordinating their measurements.

In this paper, we address these shortcomings by proposing a novel coordination algorithm for large-scale monitoring of dynamic environmental phenomena using the participatory sensing paradigm. The algorithm adaptively selects observations to be taken at each timestep and maps individuals to measurements both in space and time in order to maximize the information learned about the environment over a given time period. The algorithm is able to deal with thousands of participants, incorporates probabilistic knowledge of the mobility patterns of humans and assumes that people have a daily limit/budget on the number of measurements they are willing to take. Our algorithm makes use of clustering techniques, heuristic search and random simulations. In particular, the contributions of this paper are:

- We develop a novel stochastic coordination algorithm that is able to scale up to thousands of participants. The algorithm considers each individual's budget which is realistically determined based on large empirical studies (Chon et al. 2013) and incorporates probabilistic knowledge about human mobility patterns.
- We empirically evaluate our algorithms on real human mobility and air quality sensor data and show that our algorithm significantly outperforms the state-of-the-art algorithms in dynamic scenarios.

Problem Definition

This section formally introduces the problem of coordinating measurements in participatory sensing for environmental monitoring, which is based on (Zenonos, Stein, and Jennings 2015) and extends it to include the limited measurements that users can take per day. An environmental campaign is a campaign initiated by the taskmaster, to collect as much information about a particular phenomenon in an environment as possible. An environment $\mathcal{E}$ is a continuous set of spatio-temporal locations (L, T) that the taskmaster is interested in. This includes the spatial and temporal boundaries of the area of interest and time interval. A set of humans $\mathbf{A}=\{\mathcal{A}_1, \ldots, \mathcal{A}_M\}$ can take a set of discrete measurements[1] within the spatial boundaries of this environment and within the time period of the campaign ($O = L \times T$). The set of observations made before or at time t is denoted as $\mathbf{O}_t \subseteq O$ while the set of observations made at time t is denoted as $O_t \subseteq \mathbf{O}_t$.

A utility function $u : 2^O \rightarrow \mathbb{R}^+$ assigns a utility value to a set of observations. The value assigned by this function is based on the entropy given by the Bayesian D-optimality criterion (Krause, Singh, and Guestrin 2008) and it is further discussed in the next section. Here, it is sufficient to say that the goal is to maximize the sum of utilities over the time period of the environmental campaign. Chon et al. 2013, show that people tend to contribute a specific amount of information in participatory sensing campaigns. Indeed, we cannot assume that people can take an unlimited number of measurements but they rather have a budget. Thus, each individual has a specific budget, i.e., $B_i \in \mathbb{N}$, which is the maximum number of measurements they can take within a day.

Finally the utility can be expressed as:

$$\mathbb{U}(\mathbf{O}_E) = \sum_{t=1}^{E} u(O_t) \tag{1}$$

where $u(O_t)$ considers the observations made by citizens at the right locations and timesteps. The problem is to decide where and when the citizens should make these observations to maximize this function given a probability distribution over people's possible locations at each timestep. Hence, the optimization problem can be formulated as follows: map a set of humans to a set of spatio-temporal coordinates to maximize the utility over the period of the campaign, subject to the individual budget constraints of participants. Formally, $S^* = arg\,max_s\, \mathbb{U}(\mathbf{O}_E)$ where $\mathbf{O}_E$ are the measurements taken according to the mappings $s : \mathbf{A} \times T \rightarrow L$.

Modelling the Phenomenon

This section explains how we model the environmental phenomenon using a non-linear and non-parametric regression model called Gaussian processes (GPs). First, we discretized the environment in a way such that we create a grid of 1000×1000 meters per square and time to 24 timesteps so that each timestep represents an hour. Consequently, we say that locations $\mathcal{L} \subset L$ are the intersections of the grid and $\mathcal{T} \subset T$ are the timesteps. Each location $l \in \mathcal{L}$ and time $t \in \mathcal{T}$ is associated with a random variable $X_{l,t}$, that describes an environmental phenomenon, such as noise or air pollution. We use $X_{l,t} = x_{l,t}$ to refer to the realization of a random variable at a particular spatio-temporal coordinate, which becomes known after an observation is made. In order to describe the phenomenon at time t over the set of locations ($\mathcal{L}$), given that some observations have been made in the past ($\mathbf{O}_{t-1}$), we use $X_{\mathcal{L},t|\mathbf{O}_{t-1}}$. Similarly, we denote by the random variable $X_{\mathcal{L},t|O_t}$, the environmental phenomenon over the set of locations $\mathcal{L}$ at time t given that a set of observations are made at time t. For simplicity in the notation, and unless stated otherwise we use $X_y = X_{\mathcal{L},t|\mathbf{O}_{t-1}}$, $X_A = X_{\mathcal{L},t|O_t}$ and the realization of the measurements over the set of locations $\mathcal{L}$ given a set of observations $X_A = x_A$. Given the nomenclature above, we can now model the phenomenon. As shown in (Krause, Singh, and Guestrin 2008), the measurements of an environmental phenomenon can have a multivariate Gaussian joint distribution over all of their locations $\mathcal{L}$ and timesteps $\mathcal{T}$. It is an effective way to capture the spatio-temporal relationship of different coordinates which enables the use of advanced regression techniques. Since we are interested in monitoring a dynamic environmental phenomenon over some spatial coordinates and a time period we use GPs. GPs can generalize the multivariate Gaussians to an infinite number of random

[1] Also called observations.

variables and thus generalize over the entire set of locations and timesteps (Rasmussen and Williams 2006). The main advantages of GPs are that they can capture structural correlations of a spatio-temporal phenomenon as well as provide a value of certainty on the predictions, i.e., predictive uncertainty. Crucially, it is sufficient to know the locations of the observations but not the actual value of the measurement, to get the variance over the predictions.

The GP can be fully specified by a mean $m(\mathbf{x})$ and a covariance function (also known as kernel) $k(\mathbf{x},\mathbf{x}')$. It can be interpreted as a distribution over functions, where every random variable represents a value of a function f at a specific point. Formally, $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}'))$. In order to simplify our notation we denote the mean vector of some set of random variables X_A as μ_A. By providing some measurements x_A at some spatio-temporal coordinates, it can predict the value at other spatio-temporal coordinates where no measurement was taken. Most importantly, it provides the corresponding predictive uncertainty which is associated with the values in both the observed and unobserved locations. The distribution $P(X_y|x_A)$, i.e., X_y given these observations x_A, is also Gaussian with mean $\mu_{y|A}$ and variance $\Sigma_{y|A}$, which are formally given by:

$$\begin{aligned} \mu_{y|A} &= \mu_y + \Sigma_{yA}\Sigma_{AA}^{-1}(x_A - \mu_A) \\ \Sigma_{y|A} &= \Sigma_{yy} - \Sigma_{yA}\Sigma_{AA}^{-1}\Sigma_{Ay} \end{aligned} \tag{2}$$

There is a lot of discussion around which kernel to use for each problem. In particular, air pollution could be modelled using a composite non-stationary space-time covariance function that would be able to better capture the change of smoothness of the function depending on the location and time (Garg, Singh, and Ramos 2012). However, in order to preserve time efficiency, a common choice of covariance function is Matérn (Jutzeler, Li, and Faltings 2014; Ouyang et al. 2014), which we adopt in this work. Formally, $k(\mathbf{x},\mathbf{x}') = \sigma_f^2(1+\sqrt{3}r)\exp(-\sqrt{3}r) + \sigma_n^2\delta_{\mathbf{x},\mathbf{x}'}$ where $r = \sqrt{(\mathbf{x}-\mathbf{x}')^T\mathbf{P}^{-1}(\mathbf{x}-\mathbf{x}')}$, $\mathbf{P} = \begin{bmatrix} l_1 & 0 & 0 \\ 0 & l_2 & 0 \\ 0 & 0 & l_3 \end{bmatrix}$ and $\theta = \{l_1, l_2, l_3, \sigma_f^2, \sigma_n^2\}$ are the parameters of the covariance function (also known as hyperparameters) that need to be learned. We are mostly interested in l_1, l_2, l_3 that are crucial to the representation of the dynamism of the phenomenon in both spatial and temporal dimension. We refer to l_1, l_2 as length-scales and l_3 as time-scale.

This formalism allows the GP to update both the spatial, as well as the temporal aspect of the phenomenon. Also, σ_f^2 and σ_n^2 are parameters that control the sensitivity of the kernel to both measurements and noise while $\delta_{\mathbf{x},\mathbf{x}'}$ is the Kronecker delta, which is 1 if $\mathbf{x} = \mathbf{x}'$, and 0 if $\mathbf{x} \neq \mathbf{x}'$.

In this work, hyperparameters are initially unknown. However, we exploit historic fine-grained data provided from a number of static air quality stations in Beijing (Zheng, Liu, and Hsieh 2013) to train the model. To do so, we use a common technique called maximum likelihood estimation (MLE). That is finding the parameters θ that maximize the log marginal likelihood (ML) $\log p(x_A|\theta)$, which is given by $-\frac{1}{2}(x_A-\mu_A)^T\Sigma_{AA}^{-1}(x_A-\mu_A) - \frac{1}{2}\log|\Sigma_{AA}| - \frac{n}{2}\log 2\pi$.

Gaussian processes provide the mathematics of the utility function we need to maximize. As mentioned in the previous section, the utility is based on the Bayesian D-optimality criterion, which in terms of our problem measures the reduction of entropy at all locations of the environment (global metric) by making a set of observations. It provides the mutual information between observations made at previous timesteps and observations made at present. Given the knowledge about GPs, it can be seen as proportional to the uncertainty without making any observations minus the uncertainty when observations are made, which is given by: $I(X_y; X_A) = H(X_y) - H(X_y|X_A)$. Using a GP to model the environment, we develop an algorithm to exploit predictive uncertainty and the information metric we designed.

Algorithm Design

Our algorithm is designed to work with thousands of participants with stochastic information about their mobility patterns. As shown in (Krause, Singh, and Guestrin 2008), finding the optimal solution is computationally infeasible. In this work we focus on designing an efficient algorithm that outperforms the state of the art. The challenges are the probabilistic nature of human mobility patterns as well as the large number of participants. Thus, our algorithm must be robust under stochastic information and scalable. In this section, we formally and intuitively explain how our algorithm works. Our approach consists of two main components, the offline component, i.e., the SiScaS algorithm (Algorithm 1), and the online component, i.e., the Best-Match algorithm (Algorithm 3). SiScaS solves the problem of finding the best mapping from observations to agents in space and time in a number of simulations and Best-match deals with finding those mappings in real-time.

Simulations for Scalable Searching (SiScaS)

The Simulations for Scalable Searching (SiScaS) algorithm is a critical component in our work as it is responsible for a number of functions including calling the Stochastic Local Greedy Search (SLGS) algorithm. The algorithm is shown in Algorithm 1. In particular, this algorithm is responsible for sampling from the human mobility patterns distributions (line 4), in order to get possible future locations for each of the participants. It also clusters people in spatially correlated groups for all the timesteps using a well-known clustering technique called DBSCAN (Ester et al. 1996) (line 5). DBSCAN enables the grouping of people based on the distances between each other. Consequently, people close to each other are said to belong to the same cluster and thus can be treated as a single entity, which is crucial in scaling up the number of participants in the campaigns. Since at each timestep people can be in different locations, the algorithm produces a different set of clusters for each timestep. Formally, C is a set of spatio-temporal clusters that include information about each participant's location and budget as well as the centroid of each cluster. Finally, SLGS is called (line 6) and the human mobility patterns as well as the

spatio-temporal clusters are passed to it. For each iteration of the algorithm, SLGS will produce a different mapping of participants to measurements since it will keep sampling from the mobility patterns and forming clusters clusters for a number of times N.

Algorithm 1 Simulations for Scalable Searching (SiScaS) Algorithm

1: **input**: E (timesteps), current (current timestep), $B_{1,...,M}$ (budget)
2: $Simulations = N$
3: **for** $s = 1$ to $Simulations$ **do**
4: $\mathbf{A}, L \leftarrow SAMPLEHMPs$ {Sample from human mobility patterns distribution}
5: $C_s \leftarrow DBSCAN(\mathbf{A}, L, E - current)$
6: $S_s^* \leftarrow SLGS(E, C, current, B_{1,...,M})$
7: **end for**
8: **return**: $S^*_{1,...,N}, C_{1,...,N}$

The Stochastic Local Greedy Search Algorithm (SLGS) algorithm is the core component of SiScaS. It is an anytime stochastic variation of LGS presented in (Zenonos, Stein, and Jennings 2015). The idea is to stochastically evaluate a number policies, according to the utility function defined, and greedily proceed to a neighbouring policy by applying local changes in order to maximize that function. In other words, SLGS is given a set of spatio-temporal clusters, the budget of people and a number of timesteps and finds a mapping between clusters and possible measurements such that the information about the environment is maximized. Each cluster can take a single measurement at a time which is assumed to be taken from its centroid. The reason is to avoid using individuals' locations to make our algorithm more efficient. Since the number of spatio-temporal clusters can be large (up to a maximum of the number of participants times the number of timesteps, i.e., $M \cdot E$) we sample again through space and time. Consequently, we are left with a smaller number of spatio-temporal clusters. We greedily select measurements that maximize the total information. However, in order to save computation time, we stop the process when the increase of information, when taking a specific measurement is below a predefined threshold.

Now, the SLGS algorithm, shown in Algorithm 2, is described in more detail. The algorithm accepts the locations of people spatially clustered per timestep, as well as the budget of each individual, the total number of timesteps and the timestep the campaign is currently up to (line 1). Given that there is sufficient budget left for at least one person in the cluster, it randomly selects a cluster per timestep (line 7). It then checks what the utility would be when adding a measurement from the centroid of each cluster and forwarding the campaign in time to check what the final utility would be (line 9-14). This enables the simulations to run fast since not every single position in the cluster is considered by the Gaussian Process.

Then, the algorithm finds the cluster that produced the highest marginal increase (δ) in utility and selects it (line 15). This measurement can no longer be removed or reconsidered in the following iterations. At the same time, the budgets of people in the cluster selected are adjusted accordingly, i.e., the budget of all the people in the cluster is reduced by one since all of them are requested to take a measurement (line 16). The algorithm iterates until the marginal increase is below a percentage threshold or people's budget is depleted (line 19 and line 4 respectively).

Algorithm 2 Stochastic Local Greedy Search (SLGS)

1: **input**: E (timesteps), C (clusters), current (current timestep), $B_{1,...,M}$ (budget)
2: $maxU' = 0,\ S^* \leftarrow null\ matrix(|C|)$
3: **for** $k = 1$ to $|C|$ **do**
4: **if** $max(B_i) == 0$ **then**
5: **return**: S^*
6: **end if**
7: $z \leftarrow RANDOMSAMPLE$ {take a random sample per timestep from clusters where people have some budget left such that $z \subseteq C$}
8: $sz \leftarrow |z|$
9: **for** $l = 1$ to sz **do**
10: **for** $r = current$ to E **do**
11: $\mathbb{U}(O_r) \leftarrow u(O_r)$ {For each l we have a different spatio-temporal cluster}
12: **end for**
13: $s_l \leftarrow$ getMappings($\mathbb{U}(\mathbf{O}_E)$){Get the single best mapping from a user to spatio-temporal location }
14: **end for**
15: Keep maximum $\mathbb{U}(\mathbf{O}_E)$ of s_l in $maxU$ variable
16: Reduce budget from humans in cluster containing s_l
17: Set S^* to be the best configuration in s_l
18: $\delta = (maxU - maxU')/maxU$
19: **if** $\delta <$ threshold **then**
20: **return**: S^*
21: **end if**
22: $maxU' \leftarrow maxU$
23: **end for**
24: **return**: S^*

Best-Match Algorithm

SiScaS will produce a number of mappings (N) of participants to measurements depending on the samples taken from human mobility patterns as well as the clusters that are formed. However, in real-time, participants can actually be in a different location or they may not available at all. The idea of Best-Match algorithm is to decide what measurements to request in real-time, given the output of SiScaS ($S_{1,...,N}$) as well as the state of the world at each timestep.

Concretely, the Best-Match algorithm (Algorithm 3) gets human locations (line 2) in real-time and clusters them using the DBSCAN algorithm (line 3). Then, the algorithm finds the best match between measurements that are most informative, as calculated in advance, and the actual positions of participants in real-time. Specifically, we find the nearest neighbours from the real-time clusters to the clusters produced in SiScaS (line 5) and then the Euclidean distance between them is calculated (line 6). The smaller the distance,

the more similar the clusters are. The index of the clusters with the smallest distance among each other is used to select the simulation that best matches the current clusters (line 8). Given what measurements were selected in the simulations in advanced, the corresponding clusters are said to take those measurements (line 9).

Algorithm 3 Best-match

1: **input**: E (timesteps), current (current timestep), B (budget), $S_{1,\ldots,N}, C_{1,\ldots,N}$
2: $\mathbf{A}, L \leftarrow$ Get human locations{Get GPS coordinates of users}
3: $C'_{current} \leftarrow DBSCAN(\mathbf{A}, L, current)$\{$C'_{current}$ are the clusters formed at the current timestep in real-time\}
4: **for** $s = 1$ to N **do**
5: Find nearest neighbour from $C'_{current}$ to $C^s_{current}$
6: $D_s \leftarrow$Calculate Euclidean distance of $C'_{current}$ nearest neighbours
7: **end for**
8: $ind \leftarrow$ Find minimum D\{Get the index of the minimum distance.\}
9: Select measurements that match S^*_{ind}\{S^*_{ind} is the best match between clusters formed in simulations in advance and real-time clusters\}

In order to speed up our algorithms we reuse some of the results already calculated by partially evaluating policies in SLGS algorithm. In particular, at each iteration of policy evaluation in time, i.e., when forwarding the campaign in time, we store the utility earned from that part of the policy. When this part of the policy appears again, we reuse the utility without the need to re-evaluate it.

Empirical Evaluation

In this section, we evaluate the algorithm developed using real human mobility patterns and air quality sensor data. In the first part, we introduce our benchmarks and give a description of the experiments performed. Finally, we discuss our findings.

Benchmarks

The algorithm developed was benchmarked against the state-of-the-art algorithms which are introduced below:

- **Naive Greedy:** This algorithm is based on (Krause, Singh, and Guestrin 2008). It iterates through possible measurements available at each timestep, finding the one that produces the highest utility. It keeps adding measurements until a budget k is met. In our setting k is derived based on the total budget of people available at each timestep. In particular, we divide the total budget that is available with the number of timesteps left.
- **Proximity-driven (Pull-Based):** This algorithm is mostly used in practice to let people execute tasks based on their spatial location. In environmental monitoring this can be interpreted as taking measurements when people are in an area of high uncertainty or when the measurement they take has a high utility. This approach is used by the state-of-the-art mobile crowdsourcing applications such as FieldAgent[2] and GigWalk[3] and it is outlined in (Chen et al. 2014).
- **Random:** This algorithm assumes that measurements are randomly taken by people until no budget is left.
- **Patrol:** This algorithm assumes people take measurements at all timesteps until their budget is depleted.

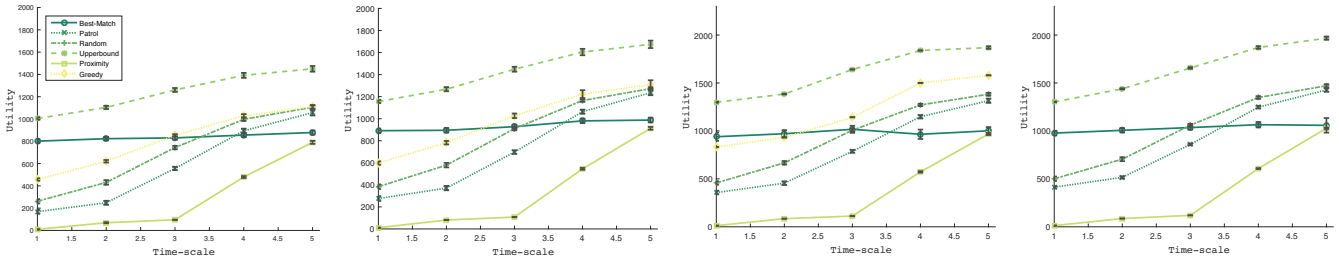
Also, since the optimal algorithm is computationally infeasible we developed an upper bound to the algorithm that can be easily calculated. The upper bound is described below:

- **Upperbound:** We relax the assumption that people have a limited budget. Thus, all participants can take measurements at every timestep, and the total utility of this can be trivially calculated.

Experimental Setup

Our experiments are based on the setup described in (Zenonos, Stein, and Jennings 2015). In particular, we evaluate our algorithm on real air quality (Zheng, Liu, and Hsieh 2013) and human mobility data (Geolife dataset) (Zheng et al. 2009). Air quality data are collected in terms of particulate matter (PM2.5) over a year in Beijing and mobility data in a period of 5 years in the same city. The air quality dataset was used to train our Gaussian process model, i.e., estimate its hyperparameters, as this phenomenon exhibits different behaviour at different spatio-temporal coordinates. In other words, air pollution varies depending on the time of the day and the location. The human mobility patterns dataset contains the trajectories of 182 humans as reported by portable GPS devices. We preprocess the dataset, and take the location of each user every ten minutes. We also take patterns of different weeks or months from the same pool of participants' trajectories. In order to make the system more realistic, we provide a probability distribution of the locations the user could be. This is to simulate the behaviour of a real human mobility prediction system that is able to provide us with probabilities over a number of locations that each user could be at each timestep. In particular, in this work, we assume that the correct locations have a high probability (80%) of being assigned a higher probability than the rest of the locations. Next, we randomly distribute the probability left to a number of historic or future locations. These locations are based on the ground truth provided from the Geolife dataset. Furthermore, people have a limited budget of measurements they are willing to take per day. Specifically, a large-scale empirical mobile crowdsourcing study showed that each individual had on average a contribution of two to eleven measurements per day without any monetary incentives (Chon et al. 2013). So, we are assuming a random budget within that range for each individual. The next section presents the results of our experiments. Our experiment involves comparing the execution time of the algorithms and the performance in terms of utility gained in campaigns with different numbers of partici-

[2] http://www.fieldagent.co.uk/
[3] http://www.gigwalk.com/



(a) (b) (c) (d)

Figure 1: Total utility gained for 24 timesteps when (a) run with 250, (b) 500, (c) 750 and (d) 1000 participants. The error bars indicate the 95% confidence interval.

pants and different time-scale which affects the dynamism of the phenomenon.

Results

Figure 1 shows results of the performance of the algorithms coordinating a varying number of participants when varying the time-scale which controls the dynamism of the phenomenon. The smaller the time-scale, the more dynamic the phenomenon is. Consequently, as the time-scale approaches zero, each timestep is more independent from the other. In this dynamic environment best-match algorithm is better in terms of total utility gained than the rest of the algorithms. However, when the time-scale is more than 3, simple algorithms perform better and the naive greedy algorithm performs the best. The advantage of the naive greedy algorithm is that it is able to choose individuals who could potentially be in different clusters to take measurements that increase the total utility. However, this comes at a great computational expense as the algorithm needs to consider all the participants one by one until the k best observations are found at each timestep. In particular, we did not calculate the total utility gained for 1000 participants as it was very computationally intense to do so. Also, when the environment is dynamic, more measurements are taken in the beginning of the campaign as it cannot look ahead in time. It is possible that some future measurement is more informative if no measurement was taken at that location in the past. On the other hand, the Best-match algorithm is designed to produce reasonable outcomes in dynamic environments and it is shown to outperform all the benchmarks in these environments. When the phenomenon is not very dynamic, a few measurements, taken either randomly or at the beginning of the campaign (patrol algorithm) can reflect the big picture of the environment without any intelligent algorithm involved. The reason is that a single measurement at the beginning of the campaign can provide the information necessary to understand the phenomenon without the need for more measurements. The Proximity algorithm chooses measurements that are informative, but it does not perform well. The reason is that in some cases the individual measurements are not so informative but if a lot of measurements were taken, the utility would be much greater overall. Moreover, it is difficult to define which measurements are informative as it needs to be empirically determined.

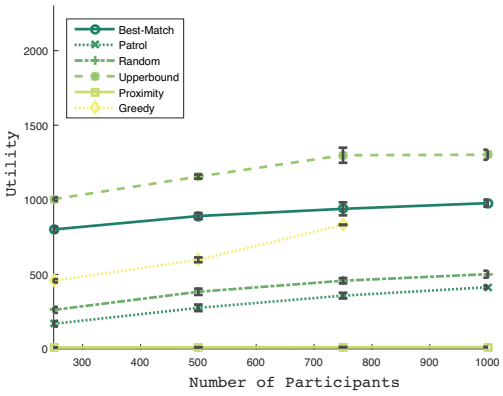


Figure 2: Total utility gained for 24 timesteps and a varying number of participants at a constant time-scale of 1. The error bars indicate the 95% confidence interval.

Figure 2 shows the results of the performance of the algorithms in terms of utility gained when we vary the number of participants M in the campaign. The dynamism in this experiment is fixed at 1, i.e., highly dynamic phenomenon. We can observe that Best-Match is 75.34% better than the second Greedy algorithm for 250 participants, 49.13% better for 500 participants, and 12.71% for 750 participants. This is because Best-match can look ahead in time, and thus make choices that will increase the total utility by the end of the participatory sensing campaign.

Figure 3 shows results of the performance of the algorithms in terms of the total runtime by varying the dynamism of the phenomenon. The results show that the Best-match algorithm is faster than the Greedy algorithm. This is because Best-match stochastically selects measurements taken by groups of people, i.e., people are not considered individually. Also, we can observe the Best-match algorithm has a reasonable runtime performance (approximately 2 hours) with thousands of participants.

Conclusion and Future Work

This paper presents a novel stochastic algorithm for mapping humans to observations based on their daily mobility patterns. We formulate the optimization problem of maximizing an entropy-based objective function over a period of

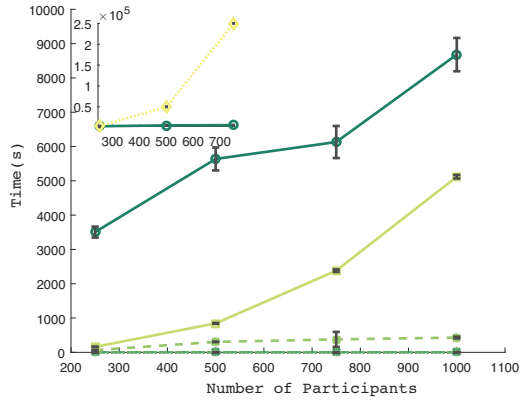


Figure 3: Average runtime for 24 timesteps and a varying number of participants. The error bars indicate the 95% confidence interval.

time given budget constraints. In particular, we develop the Best-Match algorithm, which is benchmarked against the-state-of-the-art Greedy and Pull-based algorithms. We show that our algorithm outperforms those algorithms in terms of utility for dynamic phenomena while it has a reasonable runtime. There are several future possible avenues for our work. We would like to embed a real human mobility patterns prediction system in our approach and try our algorithm in the field. Also, we could take into account adversarial people that benefit from false or inaccurate measurements as well as provide robustness against sensor failures. In addition, we could evaluate our algorithm in a different domain. Concretely, it can be useful in active learning approaches where a different complex utility function needs to be optimised. For example, in a crowdsourcing classification system, where users are asked to verify objects classified from a machine vision algorithm, the utility could capture how valuable human input is. Our algorithm could be then used to decide which users to ask to increase the overall system's efficiency.

References

Baratchi, M.; Meratnia, N.; Havinga, P. J. M.; Skidmore, A. K.; and Toxopeus, B. A. K. G. 2014. A hierarchical hidden semi-markov model for modeling mobility data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, 401–412. New York, NY, USA: ACM.

Chen, C.; Cheng, S.-F.; Gunawan, A.; Misra, A.; Dasgupta, K.; and Chander, D. 2014. Traccs: Trajectory-aware coordinated urban crowd-sourcing. In *Second AAAI Conference on Human Computation & Crowdsourcing (HCOMP)*, 30–40.

Chen, C.; Cheng, S.-F.; Lau, H. C.; and Misra, A. 2015. Towards city-scale mobile crowdsourcing: Task recommendations under trajectory uncertainties. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, 1113–1119. AAAI Press.

Chon, Y.; Lane, N. D.; Kim, Y.; Zhao, F.; and Cha, H. 2013. Understanding the coverage and scalability of place-centric crowdsensing. *Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'13), Zurich, Switzerland* 3–12.

D'Hondt, E.; Zaman, J.; Philips, E.; Boix, E. G.; and De Meuter, W. 2014. Orchestration support for participatory sensing campaigns. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, 727–738. New York, NY, USA: ACM.

Ester, M.; peter Kriegel, H.; S, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. 226–231. AAAI Press.

Garg, S.; Singh, A.; and Ramos, F. 2012. Learning non-stationary space-time models for environmental monitoring.

Jutzeler, A.; Li, J. J.; and Faltings, B. 2014. A region-based model for estimating urban air pollution. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 424–430.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* 9:235–284.

Maisonneuve, N.; Stevens, M.; and Ochab, B. 2010. Participatory noise pollution monitoring using mobile phones. *Info. Pol.* 15(1,2):51–71.

McInerney, J.; Stein, S.; Rogers, A.; and Jennings, N. R. 2013. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive Mob. Comput.* 9(6):808–822.

Nikzad, N.; Verma, N.; Ziftci, C.; Bales, E.; Quick, N.; Zappi, P.; Patrick, K.; Dasgupta, S.; Krueger, I.; Rosing, T. v.; and Griswold, W. G. 2012. Citisense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system. In *Proceedings of the Conference on Wireless Health*, WH '12, 11:1–11:8. New York, NY, USA: ACM.

Ouyang, R.; Low, K. H.; Chen, J.; and Jaillet, P. 2014. Multi-robot active sensing of non-stationary gaussian process-based environmental phenomena. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, 573–580.

Passchier-Vermeer W, P.-V. W. 2000. Noise exposure and public health. *Environmental Health Perspectives* 123–131.

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

Seaton, A.; Godden, D.; MacNee, W.; and Donaldson, K. 1995. Particulate air pollution and acute health effects. *The Lancet* 345(8943):176 – 178.

World Health Organization, et al. 2014. Burden of disease from household air pollution for 2012. *Summary of Results.*

Zaman, J.; D'Hondt, E.; Boix, E. G.; Philips, E.; Kambona, K.; and De Meuter, W. 2014. Citizen-friendly participatory campaign support. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, 232–235.

Zenonos, A.; Stein, S.; and Jennings, N. R. 2015. Coordinating measurements for air pollution monitoring in participatory sensing settings. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, 493–501.

Zheng, Y.; Zhang, L.; Xie, X.; and Ma, W.-Y. 2009. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 791–800. New York, NY, USA: ACM.

Zheng, Y.; Liu, F.; and Hsieh, H.-P. 2013. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, 1436–1444. New York, NY, USA: ACM.