IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 21, 2019, accepted March 17, 2019, date of publication April 3, 2019, date of current version April 18, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2909111*

# Optimal Citizen-Centric Sensor Placement for Air Quality Monitoring: A Case Study of City of Cambridge, the United Kingdom

**CHENXI SUN[1], (Student Member, IEEE), VICTOR O. K. LI[1], (Fellow, IEEE), JACQUELINE C. K. LAM[1], (Member, IEEE), AND IAN LESLIE[2]**

[1]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, U.K.

Corresponding author: Jacqueline C. K. Lam (jcklam@eee.hku.hk)

This work was supported in part by the Theme-Based Research Scheme of the Research Grants Council of Hong Kong under Grant T41-709/17-N.

**ABSTRACT** Air quality monitoring plays an increasingly important role in providing accurate air pollution data for assessing the impacts of air pollution on public health. Development of proper sensor networks, by deploying the right air pollution sensors at the right place, in order to meet the needs of different groups in the city and provide the much needed public services, deserves careful attention, especially when smart city development is being considered. However, air quality monitoring can be a costly measure. To tackle such a challenge, air pollution sensor placement can be carefully designed to achieve certain optimal citizen-centric objectives in the absence of field information, which can be formulated as an optimal sensor placement problem. In this paper, we propose three citizen-centric objectives for the optimal sensor placement problem, which does not require the prior deployment of pollution sensors for obtaining any field information. By citizen-centric, we mean that sensor placement puts the citizens' welfare at the center of attention and be able to fulfill the following objectives: 1) better assessing the vulnerable people's exposure to air pollution; 2) maximizing overall satisfaction of obtaining public information on existing air quality; and 3) better monitoring traffic emissions. We formulate the optimization problem for each scenario and propose an effective method to solve the problem accordingly. Last but not least, we conduct a case study in the city of Cambridge to evaluate the feasibility and effectiveness of our proposed methods. Our case study has shown that in order to optimize our citizen-centric objectives, there is a need to re-orient the current sensor placement strategies in the city of Cambridge, U.K.

**INDEX TERMS** Air pollution, citizen-centric, optimization, sensor placement, design methodology, human factors.

## I. INTRODUCTION

According to the World Health Organization (WHO), ambient air pollution presents a major environmental risk to public health, causing around 4.2 million premature deaths worldwide in 2016 [1]. Development of proper sensor networks, by deploying the right air pollution sensors at the right place, in order to meet the needs of different groups in the city, and provide the much needed public information and services, thereby deserves careful attention [2]–[4].

The associate editor coordinating the review of this manuscript and approving it for publication was Saqib Saeed.

Generally, the measurements can be obtained either through low-cost mobile wireless sensors or fixed-location sensors with sophisticated measurement equipment. In this study, we focus on deployment strategies for fixed-location sensors as measurements of such are more accurate and reliable [5]. However, deployment of static sensors can be a costly measure, incurring not only site construction costs but also operational costs. Hence only a limited number of sensors can be deployed given a budget constraint.

There has been much work covering sensor placement in relation to general environmental monitoring. For example, [6] studied the problem of selecting a most informative subset of correlated random variables under size constraints. Reference [7] proposed a near optimal sensor placement strategy for temperature monitoring by maximizing mutual information under size constraint. Reference [8] studied optimal sensor placement for field soil moisture estimation. Reference [9] studied the wind monitoring problem. However, all the studies mentioned above aim to place sensors or stations at the most informative locations defined either in terms of entropy or mutual information. By modeling each location as a random variable, the proposed approach would require the covariance between any two random variables, usually through a strong Gaussian Process assumption on the underlying field and more importantly requires a pilot study to obtain the field information, which is impractical due to complex procedures and high costs.

Some studies focused on the optimal design of air quality monitoring networks using air quality estimation models, such as physical models [10], [11] and learning-based models [12]. Reference [10] studied the optimal redistribution of air quality monitoring network using the commonly adopted atmospheric dispersion model and genetic algorithm. Reference [11] studied the optimization of air quality network design using the chemical transport model and searching algorithm. Reference [12] proposed an entropy minimization model for recommending new station locations based on their proposed semi-supervised air quality inference model. However, the placement quality of these approaches will depend on the accuracy of the air quality model and [10], [12] required existing air quality measurements in the region as inputs to the air quality estimation model.

In this paper, we put forward the following questions regarding for air quality monitoring. Firstly, what should be the key citizen-centric objectives when deploying an air quality monitoring network throughout a city, in the absence of any prior knowledge of the field? By citizen-centric, we mean sensor placement that puts our citizens' welfare at the center of attention. Secondly, given a fixed budget constraint, what is the optimal sensor placement strategy to achieve the objectives? Thirdly, when there are multiple objectives to be considered simultaneously, how should we derive a suitable sensor placement strategy to achieve an integrated goal given the budget constraints?

In our previous study [13], we proposed a citizen-centric objective which aimed to place the sensors to maximize overall citizen satisfaction. They showed that when the individual satisfaction is an exponential decay function of his/her distance to the nearest sensor, the objective function has the nice monotone and submodular property that allows greedy algorithms to solve the problem efficiently with provable approximation guarantee. We generalize the formulation to include many other functions. Furthermore, we formulate two other citizen-centric objectives which aim at: 1. Better protecting the vulnerable people 2. Better monitoring traffic emissions to assess if ambient air quality standards have been met. Our work carries the following significance:

- We formulate three important citizen-centric objectives for our optimal sensor placement problem without the need of obtaining any prior field information. We aim at: 1) Better assessing the vulnerable people's exposure to air pollution 2) Maximizing overall satisfaction in air quality information available to the public 3) Better monitoring traffic emissions in order to assess if air quality objectives or ambient regulatory standards have been met.
- We show that the exact solutions to the first two formulations are NP-hard and propose greedy based approaches to solve the problem efficiently with provable approximation guarantee.
- We propose a simple yet effective approach to solve the multi-objective case by decomposing it into independent single-objective optimizations.
- We conduct a case study in City of Cambridge, the United Kingdom, to assess the feasibility of the formulation and evaluate the effectiveness of the proposed approach. We also compare three optimal sensor placement results with the result of an existing air quality monitoring network (consisting of 20 low-cost pollution sensors) in Cambridge to discuss the individual differences followed by policy suggestions for local environmental decision-makers and urban planners, and sensor placement consultants etc.

The rest of this paper is structured as follows: Section II overviews the related work on sensor placement. Section III describes three placement objectives with a solution method proposed for each formulation. Section IV conducts a case study on City of Cambridge to evaluate the feasibility of three citizen-centric objectives and the effectiveness of our proposed approaches. Finally Section V concludes our study.

## II. BACKGROUND

### A. MONOTONICITY

Let $\Omega$ be a finite set. A set function $f : 2^{\Omega} \rightarrow R$ defined on subset $V$ is monotone if for any $A \subseteq V$ and $s \in V$, we have

$$f(A \cup \{s\}) - f(A) \geq 0. \tag{1}$$

### B. SUBMODULARITY

A set function $f : 2^{\Omega} \rightarrow R$ defined on subset $V$ is submodular if for all $A \subseteq B \subseteq V$ and any element $s \in V \setminus B$, we have:

$$f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B). \tag{2}$$

Intuitively, it describes the diminishing return property. Equivalently, a set function $f : 2^{\Omega} \rightarrow R$ defined on subset $V$ is submodular if for every $A, B \subseteq V$ we have:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B). \tag{3}$$

Reference [6] showed that monotone submodular functions can be optimized with a greedy algorithm with an approximation guarantee of $1 - 1/e$. Reference [14] further showed that for general case, monotone submodular functions can be optimized with a hybrid greedy approach with an approximation guarantee of $\frac{1}{2}(1-1/e)$.

## III. CITIZEN-CENTRIC PLACEMENT OBJECTIVES

### A. MAXIMIZING OVERALL CITIZEN SATISFACTION

In this formulation, we focus on the utility gain of sensor placement based on citizen satisfaction. Noticing that people naturally rely on observations from the nearest station to obtain environmental information when a number of stations can be found within the city, we follow the assumption in [13] that an individual's satisfaction of a sensor placement scheme $g(d)$ is a function of his/her distance to the nearest sensor $d$. Intuitively, the closer the nearest station is, the more people are satisfied with the resolution of the information. Hence we require the satisfaction function $g(d)$ to satisfy the following:

- $g(d)$ be a monotonically decreasing function, i.e., for any $d_1 \leq d_2$, $g(d_1) \geq g(d_2)$.
- for any $d \geq 0$, $g(d) \geq 0$.
- $g(0) = 1$.

The first requirement matches the intuition while the rest of the requirements ensure that the satisfaction function be a non-negative value between [0, 1]. It is not hard to see that the exponential decay function proposed in [13] satisfies the requirements above. The exponential decay function $g(d)$ is defined as follows:

$$g(d) = \exp\left(-\frac{d}{\theta}\right) \tag{4}$$

where $\theta$ is an exponential decay parameter controlling the decay speed. By default we can set $\theta = 1$. A smaller $\theta$ will cause the satisfaction function to decrease more rapidly as the distance increases.

Following the common practice in environmental monitoring [7], [12], we divide the city into discrete equal-sized square grids, in which sensors are placed, with at most one sensor in each grid. Let $V = \{v_i : i = 1, 2 \dots, n\}$ denote the set of grids in the region of interest, where $n = |V|$ is the total number of grids. Let $p_i$ denote the percentage of population living in the $i$-th grid $v_i$ and $(x_i, y_i)$ denote its location.

We further define the minimum distance between grid $i$ and a set of grids $A$ as follows:

$$d(i, A) = \min_{j \in A} d(i, j). \tag{5}$$

When $A = \emptyset$, we set $d(i, A) = \infty$. We summarize the notations in Table 1.

Then we can define the average satisfaction ratio $f(A)$ of choosing a set of grids $A$ for placing sensors as

$$f(A) = \sum_{i \in V} p_i \cdot g(d(i, A)). \tag{6}$$

Let $c(\cdot)$ be the cost associated with subset $A$. Then our objective is to select a subset of locations $A \subseteq V$ that can maximize the average satisfaction ratio under the total cost constraint $c(A) \leq C$. Formally, our optimal citizen-centric sensor placement problem is as follows:

**TABLE 1.** Notation.

| Notation | Definition |
|---|---|
| $V$ | the set of all grids of interest |
| $n$ | the total number of grids |
| $A$ | the set of selected grids for deploying sensors |
| $p_i$ | the percentage of population living in the i-th grid |
| $d(i,j)$ | the distance between the grid $i$ and the grid $j$ |
| $d(i,A)$ | the minimum distance between the grid $i$ and its closest member of $A$ |

$$\begin{aligned} &\max\ f(A) \\ &\text{s.t. } c(A) \leq C, \quad A \subseteq V \end{aligned} \tag{7}$$

In the simplest case, the cost is uniform for all locations and the cost constraint is reduced to the cardinality constraint:

$$\begin{aligned} &\max\ f(A) \\ &\text{s.t. } |A| = k, A \subseteq V \end{aligned} \tag{8}$$

where $k$ is the total number of sensors that can be placed subject to the total cost constraint $C$.

*Theorem 1:* Finding the optimal solution for the simplest case is NP-hard.

*Proof:* We sketch the proof by a reduction from the p-median on network problem, which is known to be NP-hard [15]. The k-median (of a network) problem is defined as follows. A network $G = (V, E)$ has a nonnegative number $w(v)$ associated with each of its $|V| = n$ vertices and a positive number $l(e)$ associated with each of its $|E|$ edges. Let $X_k = \{x_1, x_2, \dots, x_k\}$ be a set of $k$ points on $G$. The objective is to find $X_k^*$ such that the distance sum of the set $H(X_k) = \sum_{v \in V} w(v) \min_{1 \leq i \leq k}\{d(v, x_i)\}$ is minimized.

The corresponding relationship between these two set systems is as follows: the weight $w(v)$ of $v \in V$ corresponds to the population $p_i$ of grid $i \in V$ and the distance $d(v, x_i)$ between $v$ and $x_i \in X_k$ corresponds to the non-negative satisfaction loss $1 - g(d(i, j))$ of location $i$ by placing sensor at location $j$. Then selection of a set of points $X_k \subseteq V$ on $G$ corresponds to selecting a set of locations $A \subseteq V$ with size $|A| = k$. ■

Fortunately, we prove that the objective function $f$ is non-decreasing and submodular.

*Lemma 1 (Non-Decreasing):* For any $A \subseteq V$ and a sensor $s \in V$, we have

$$f(A \cup \{s\}) - f(A) \geq 0. \tag{9}$$

*Proof:* By definition we have $f(\emptyset) = 0$. For all $i \in V$, we have $d(i, A) = \min_{j \in A} d(i, j) \geq \min_{j \in A \cup \{s\}} d(i, j) = d(i, A \cup \{s\}), \forall A \subseteq V,\ s \in V$. Hence we have $f(A \cup \{s\}) = \sum_{i \in V} p_i \cdot g(d(i, A \cup \{s\})) \geq \sum_{i \in V} p_i \cdot g(d(i, A)) = f(A)$ for any $A \subseteq V, s \in V$. ■

*Lemma 2 (Submodularity):* For all placements $A \subseteq B \subseteq V$ and any sensor $s \in V \setminus B$, we have

$$f(A \cup \{s\}) - f(A) \geq f(B \cup \{s\}) - f(B). \tag{10}$$

*Proof:* By definition of $f$ we need to prove that $\sum_{i\in V} p_i \cdot g(d(i, A\cup\{s\})) - \sum_{i\in V} p_i \cdot g(d(i, A)) \geq \sum_{i\in V} p_i \cdot g(d(i, B\cup\{s\})) - \sum_{i\in V} p_i \cdot g(d(i, B))$. Suppose that $d(i, A) = d(i, a)$ and $d(i, B) = d(i, b)$. Since $A \subseteq B$, $a \in A$, and $b \in B$, we have $d(i, b) \leq d(i, a)$.

Here we have three scenarios for $i$.

1) $d(i, s) \geq d(i, a)$.
Then $d(i, A\cup\{s\}) = d(i, A)$ and $d(i, B\cup\{s\}) = d(i, B)$. Hence $g(d(i, A\cup\{s\})) - g(d(i, A)) = g(d(i, B\cup\{s\})) - g(d(i, B)) = 0$.
2) $d(i, b) \leq d(i, s) < d(i, a)$.
Then $d(i, A \cup \{s\}) = d(i, s) < d(i, A)$, $d(i, B \cup \{s\}) = d(i, B)$ and hence $g(d(i, A\cup\{s\})) - g(d(i, A)) > g(d(i, B)) - g(d(i, B\cup\{s\})) = 0$.
3) $d(i, s) < d(i, b)$.
Then $d(i, A \cup \{s\}) = d(i, B \cup \{s\}) = d(i, s)$ and hence $g(d(i, A \cup \{s\}) - g(d(i, A)) = g(d(i, s)) - g(d(i, A)) \geq g(d(i, s)) - g(d(i, B)) = g(d(i, B\cup\{s\})) - g(d(i, B))$.

Therefore, for each $i \in V$, we have $g(d(i, A \cup \{s\})) - g(d(i, A)) \geq g(d(i, B \cup \{s\})) - g(d(i, B))$. Since $p_i \geq 0$ for all $i$, the proof is complete. ■

Hence the problem can be approached via the greedy approach with good approximation guarantee. Specifically, at the $i$-th iteration, the algorithm will select the location $s^*$ with the maximum information gain, i.e.,

$$s^* = \arg \max_{s\in V\setminus A_{i-1}} f(A_{i-1} \cup \{s\}) - f(A_{i-1}) \quad (11)$$

where $A_{i-1}$ denotes the selected subset before the $i$-th iteration.

The sensor placement algorithm given a unit cost constraint is summarized in Algorithm 1.

**Algorithm 1** Station Deployment Algorithm Given a Unit Cost Constraint

**Input:** Sensor number constraint $k$, a set of grids $V$ with associated percentage of population $\{p_i\}_{i=1}^n$, distance function $d$, exponential decay parameter $\theta$
**Output:** A subset of locations $A \subseteq V$
$A = \emptyset$
**while** $|A| \leq k$ **do**
select the location $s^*$ with the largest human satisfaction gain $\sum_{i\in V} p_i(g(d(i, A\cup\{s\})) - g(d(i, A)))$
add $s^*$ to location set $A$
**return** $A$

*Theorem 2:* Algorithm 1 finds a set $A^*$ such that

$$f(A^*) \geq (1 - 1/e) \max_A f(A).$$

*Proof:* Since $f$ is submodular and non-decreasing, it follows directly from [6]. ■

When the cost is non-uniform across all locations, the simple greedy algorithm fails to provide a satisfactory result when placing station at a location is much more expensive but only provides a marginally better utility gain. A modified greedy selection rule that takes cost into account will select the location $s^*$ with the maximum information gain per unit cost at the $i$-th iteration, i.e.,

$$s^* = \arg \max_{s\in V\setminus A_{i-1}} \frac{f(A_{i-1} \cup \{s\}) - f(A_{i-1})}{c(s)} \quad (12)$$

where $A_{i-1}$ denote the subset before the $i$-th iteration. However the result for this condition can also be arbitrarily bad. Let $A_G$ denote the greedy selection result using criteria (11) and $A_{CEG}$ denote the greedy selection result using criteria (12). Fortunately, the following theorem shows that at least one of the solutions can provide $\frac{1}{2}(1 - 1/e)$ approximation guarantee.

*Theorem 3:*

$$\max\{f(A_G), f(A_{CEG})\} \geq \frac{1}{2}(1 - 1/e) \max_{A, c(A)\leq C} f(A).$$

*Proof:* Since $f$ is a non-decreasing submodular set function, it follows directly from [14].

The station deployment algorithm given a general cost constraint is summarized in Algorithm 2.

**Algorithm 2** Station Deployment Algorithm Given a General Cost Constraint

**Input:** Cost constraint $C$, a set of grids $V$ with associated percentage of population $\{p_i\}_{i=1}^n$, distance function $d$, exponential decay parameter $\theta$, a cost function $c$
**Output:** A subset of locations $A \subseteq V$
$A_1 = A_2 = \emptyset, V_1 = V_2 = V,\ C_1 = C_2 = 0$
**while** $V_1 \neq \emptyset$ **do**
**for all** $s \in V_1$ **do**
$\Delta_s = \sum_{i\in V_1} p_i(g(d(i, A_1 \cup \{s\})) - g(d(i, A_1)))$
$s^* = \arg\max_{s\in V_1} \Delta_s$
**if** $c(s^*) + C_1 \leq C$ **then**
$C_1 = C_1 + c(s^*), A_1 = A_1 \cup \{s^*\}$
$V_1 = V_1 \setminus \{s^*\}$
**while** $V_2 \neq \emptyset$ **do**
**for all** $s \in V_2$ **do**
$\Delta_s = \sum_{i\in V_2} p_i(g(d(i, A_2 \cup \{s\})) - g(d(i, A_2)))$
$s^* = \arg\max_{s\in V_2} \frac{\Delta_s}{c(s^*)}$
**if** $c(s^*) + C_2 \leq C$ **then**
$C_2 = C_2 + c(s^*), A_2 = A_2 \cup \{s^*\}$
$V_2 = V_2 \setminus \{s^*\}$
$A = \arg\max_{A\in\{A_1, A_2\}} f(A)$
**return** $A$

### B. BETTER PROTECTING THE VULNERABLE PEOPLE

Air pollution has long been proven to have adverse consequences on human health, especially to the asthma and COPD patients. In addition, the young children, and the elderly are more susceptible to the harmful effects of air pollution [16], [17]. Hence, a proper placement objective is to place the sensors at locations that provide accurate ambient air quality information to the vulnerable people and collect air quality data for longitudinal studies of the health effects of air pollution.

Here we assume that the vulnerable people mainly stay at certain locations. Specifically, the young children may stay in the nursery and primary school in most parts of the day while the elderly and patients mainly stay in the elderly care homes and hospitals, respectively. In this context, $V$ denotes the set of locations of interest and $n$ denotes the number of total locations of interest. Let $A$ denote the set of locations for placing sensors and $k$ denote the total number of sensors available. Suppose that the sensors are insufficient to cover all locations, i.e., $k < n$. Let $d(i, j)$ denote the distance between the point of interest $i \in V$ and the location of the sensor $j \in A$. The objective is to minimize the total distance between the points of interest and their nearest sensors $\sum_{i \in V} \min_{j \in A} d(i, j)$. The notations are summarized in Table 2.

**TABLE 2. Notation.**

| Notation | Definition |
|---|---|
| $V$ | the set of locations of interest |
| $n$ | the number of locations of interest |
| $k$ | the number of sensors we can place |
| $A$ | the set of selected locations for deploying sensors |
| $y_i$ | the indicator variable denoting whether the location $i$ is selected |
| $x_{ij}$ | the indicator variable denoting whether the location $i$ is the nearest to the point of interest $j$ |
| $d(i, j)$ | the distance between the location i and the point of interest $j$ |
| $m$ | the number of candidate locations |

Depending on the assumptions on the search space of sensor locations $A$, there are three possible settings:

1) $A \subseteq \mathbb{R}^2$: In this case, there is no restriction on the location of the sensors.
2) $A \subseteq V'$: In this case, a sensor can be placed at one of the locations in a finite location set $V'$.
3) $A \subseteq V$: In this case, a sensor is restricted to be placed in a location that belonged to the set of locations of interest $V$.

For the first setting, if we consider the Euclidean distance in a 2-dimensional space, it aims to solve the following $k$-means clustering problem:

$$\sum_{i \in V} \min_{j \in A} \|i - j\|^2 \tag{13}$$
$$s.t.\ |A| = k \tag{14}$$

However, it is not realistic in general given the actual site requirement for installing sensors. For example, a sensor can not be installed at the center of the road.

The second setting restricts the search space to a finite candidate set and aims to solve the k-median problem. A motivating example is the set of traffic signal pole locations. The Array of Things project [18], for instance, plans to mount the modular sensor nodes on some of the streetlight traffic signal poles to provide real-time measurement on the city's environment, infrastructure, and activity, for research and public use.

The third formulation further restricts the discrete search space to the set of locations of interest. It corresponds to the community-led air quality monitoring scenario when the sensors are operated locally by the communities/organizations. In fact, it is a special case of the second formulation with $V' = V$ and hence in the following we focus on solving the second formulation.

Let $x_{ij}$ be the binary indicator variable denoting whether location $i \in V'$ is the nearest to point of interest $j \in V$ and $y_i$ be the binary indicator variable denoting whether location $i \in V'$ is selected for placing sensors. Let $d(i, j)$ denote the distance between point of interest $i \in V$ and the location of the sensor $j \in V$. Then the problem can be formulated as the following linear integer program:

$$\min \sum_{i \in V'} \sum_{j \in V} d(i, j) \cdot x_{ij} \tag{15}$$
$$s.t. \sum_{i \in V'} y_i = k \tag{16}$$
$$y_i \in \{0, 1\}, \quad \forall i = 1, 2, \ldots, m \tag{17}$$
$$x_{ij} \in \{0, 1\}, \quad \forall i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n \tag{18}$$
$$y_i - x_{ij} \geq 0, \quad \forall i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n \tag{19}$$

The first constraint ensures that at most $k$ locations are selected for placing sensors and the last constraint ensures that each location of interest is covered by at least one sensor. It can be seen that the problem is a variant of the uncapacitated facility location problem with facility cost set to zero and the number of facilities restricted to $k$.

Despite the NP-hardness of the problem, it can be solved by the existing state-of-the-art commercial integer linear programming solvers (e.g. Gurobi [19], CPLEX [20]) exactly or to a near-optimal solution efficiently. Gurobi used Linear Programming (LP) based branch and bound algorithm which consists of a systemic enumeration of candidate solutions by means of state space search: the set of candidate solutions is considered as a rooted tree with the full set at the root. The algorithm explores branches of this tree, which represent subsets of the solution set. Before enumerating the candidate solutions of a branch, the branch is checked against the upper and the lower estimated bounds on the optimal solution, and is discarded if it cannot produce a better solution than the best one obtained so far by the algorithm. The best bound and gap are also provided in the calculation result.

### C. BETTER MONITORING TRAFFIC EMISSIONS

One important consideration for air quality monitoring is to obtain as many measurements as possible at locations of heavy emissions. Generally, the primary sources of air pollution in the urban areas are traffic and power plant emissions. Since there are usually very few power plants in each city and power plant emissions are closely monitored in close proximities [21], we focus on sensor placement strategies for better monitoring traffic emissions. There are many important factors influencing traffic-related air pollution, including road type, car type, driving pattern, traffic congestion level (traffic flow pattern) etc. A higher level of congestion is strongly correlated with a higher level of emission intensity as vehicles spend more time idling, accelerating and decelerating during congestion [22], [23]. Studies have found that traffic congestion degrades ambient air quality [24] and accounts for a significant share of vehicle emissions and air quality impacts as compared to free flow conditions [22].

We assume that the hourly traffic congestion levels of all road segments within City of Cambridge are given and follow a weekly pattern as such data can be easily obtained from Google Map. Let $V$ denote the set of all road segments and $C$ denote the set of all types of traffic conditions. Let $t_i^j$ denote the fraction of time in a week that the traffic condition in road segment $i$ is of type $j$. Let $\lambda_j$ denote the weight parameter for the $j$-th type of traffic condition which can be set according to the relationship between traffic emissions and traffic condition. The notations are summarized in Table 3.

**TABLE 3. Notation.**

| Notation | Definition |
|---|---|
| $V$ | the set of all road segments |
| $n$ | the number of road segments |
| $C$ | the set of all types of traffic conditions |
| $t_i^j$ | the fraction of time in a week that the traffic condition in road segment $i$ is of type j |
| $\lambda_j$ | the weight parameter for the $j$-th type of traffic condition |

We define the importance $I_i$ of road segment $i \in V$ as the weighted sum of the fraction of time under individual traffic conditions:

$$I_i = \sum_{j=1}^{C} \lambda_j t_i^j \tag{20}$$

where the weight is determined by the relationship between traffic emissions and traffic condition. This definition allows us to find the $k$ most important road segments to place the sensors.

When the number of road segments is much larger than the number of sensors available, i.e., $k << n$, (the number of road intersections is larger than $k$,) it is better to find the $k$ most important road intersections for placing the sensors. Let $U$ denote the set of road intersections. For road intersection $u \in U$, let $N(u) \subset V$ denote the set of its neighbor road segments. Then the importance $I_u$ of road intersection $u$ can be defined as:

$$I_u = \sum_{i \in N(u)} I_i. \tag{21}$$

### D. MULTIPLE-OBJECTIVE OPTIMIZATION

Depending on the situations, one or all of the objectives can be taken into account when designing a proper sensor placement scheme. When all objectives are taken into account for placing $k$ sensors, we aim to solve a multi-objective optimization.

A common practice to solve the multi-objective optimization is convert it into a single objective optimization via scalarization, where the new objective is the weighted sum of each objective. However, this approach is not suitable for our case due to the different characterization of each objective. Furthermore this approach may pick locations that are not optimal for any single objective (but optimal for the weighted sum), which is not very helpful in practice due to a lack of clear goal. Notice that our multi-objective optimization problem has the structure of independent optimization coupled with a general knapsack constraint. Hence we propose to solve it by first allocating a certain number of sensors for each objective and then solve the single objective optimization individually. By default the numbers of sensors for each objective are set to be equal. If the location $s_a$ suggested by objective $a$ and the location $s_b$ by objective $b$ are fairly close to each other, i.e., $d(s_a, s_b) < \epsilon$ where $\epsilon$ is a predefined threshold, we can randomly pick one location from $s_a, s_b$, say $s_a$ and select another location $s_b'$ further suggested by the objective $b$. This allows the sensor placement for one objective to benefit another objective, thereby expands the coverage across the city.

## IV. CASE STUDY: CITY OF CAMBRIDGE, THE UNITED KINGDOM

In this section, we conduct a case study in City of Cambridge[1] to evaluate our three proposed placement strategies and compare the placement results with the existing placements of 20 sensors in Cambridge.

### A. EXPERIMENTAL SETUP

#### 1) DATASETS

We collect the population data, the points of interest data, and the traffic patterns for City of Cambridge.

1) **Population data** is obtained from ORNL's LandScan™,[2] one of the most fine grained global population distribution data at approximately 1km ($30'' \times 30''$)[3] spatial resolution.
2) **Points of Interest data** is collected from Google Map.[4] Here we consider three types of points of interest where the vulnerable people spend most of their time in, including: the primary school, the nursery and the hospital.
3) **Traffic pattern data** is obtained from Google Map. Google Map displays real time and typical traffic information represented by the color of the road shown on the Map. There are four colors.
   - Green represents no traffic delays.
   - Orange represents medium road-based traffic volume.

[1] The boundary of City of Cambridge can be found in https://www.cambridge.gov.uk/ward-map

[2] https://https://landscan.ornl.gov/.

[3] Here " marks arcsecond.

[4] https://map.google.com.

- Red represents traffic delays.
- Dark red represents high road-based traffic volume.

The typical traffic patterns are updated every 15 or 20 minutes from 6am to 10pm each day of the week. Since there was no traffic API available, we extracted the road level traffic condition information from the Google Map snapshots via image processing techniques.

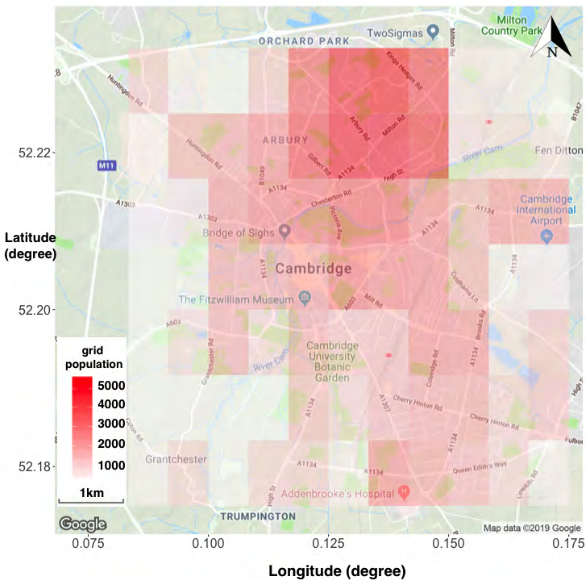


**FIGURE 1.** Grid-level population density of City of Cambridge.

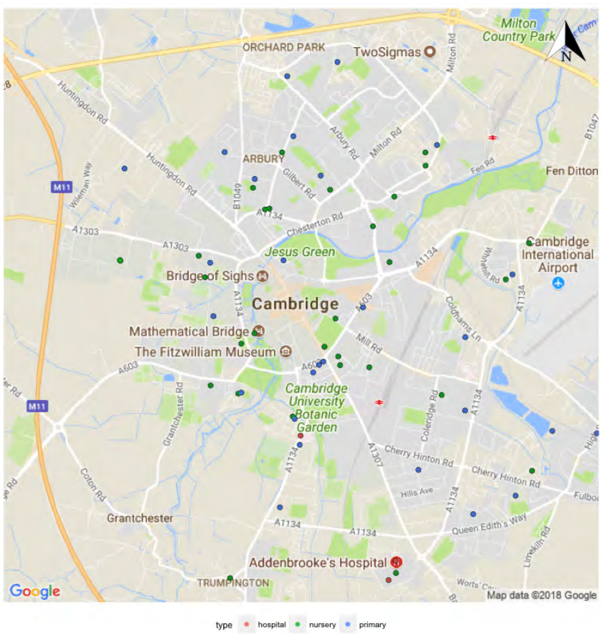


**FIGURE 2.** Points of interest within City of Cambridge.

Figure 1 shows the grid-level population density of City of Cambridge. The redness indicates the population density of each grid. Figure 2 displays the points of interest in City of Cambridge. The red dots represent the hospitals, the green ones represent the nurseries and the blue ones represents the primary schools. As depicted in Figure 2, there are only three hospitals within the region. Furthermore, the primary schools and nurseries are scattered widely across in the region. Figure 3 shows an example of the traffic condition captured from Google Map. As seen, different road segments exhibit different traffic patterns at a certain time.

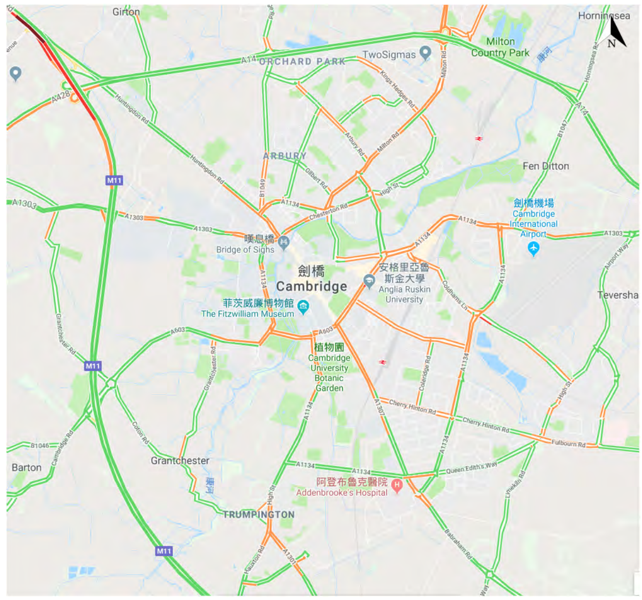


**FIGURE 3.** The traffic condition captured from google map.

Statistics of three datasets are shown in Table 4.

**TABLE 4.** Dataset statistics.

| Dataset | $n = \|V\|$ | Type |
|---|---|---|
| Population data | 84 | real numbers |
| Points of interest data | 50 | 3 classes |
| Traffic data | 100 | 4 classes |

The codes are written in Python 2.7 and the plots are generated with JavaScript.

### B. PLACEMENT RESULTS

Figure 4 shows the placement result of 20 sensors for maximizing overall citizen satisfaction. The black polygon is the administrative boundary. Here we assume that the cost is uniform for all location and hence the unit cost constraint is adopted. We further assume $g(d) = \exp(-d)$ since it can capture the fast decay of the satisfaction as the distance increases. As shown in the figure, the sensors are assigned to the densely populated grids in order to provide information for the people living within the region. However it should be noted that the algorithm does not simply pick the top twenty populated grids as people living near by the selected grids also benefit from the sensor placement to some extents. Hence the

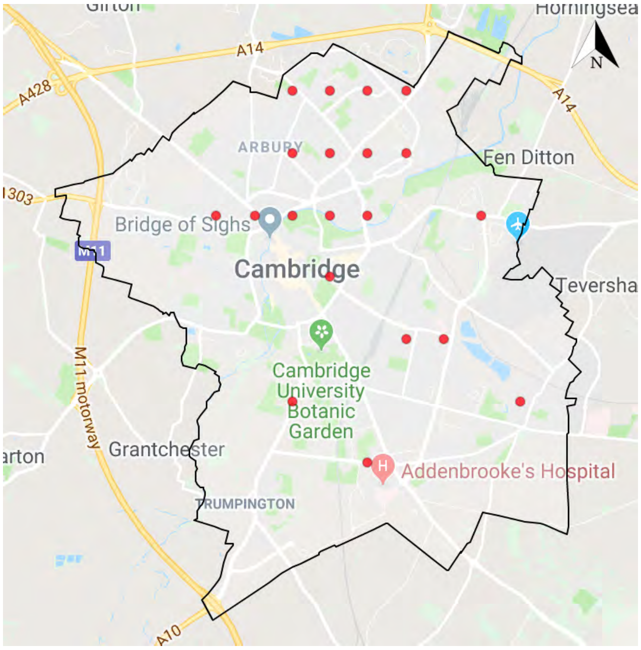


**FIGURE 4.** **Sensor placement for maximizing overall citizen satisfaction.**

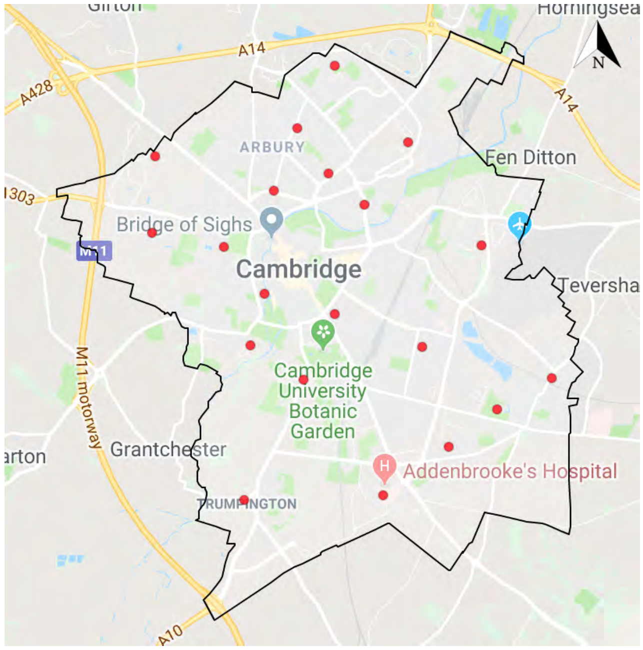


**FIGURE 5.** **Sensor placement for better protecting the vulnerable people.**

optimal placement design to maximize overall satisfaction will also try to spread the sensors around when the number of sensors are limited.

Here the resolution of each grid is 1km $\times$ 1km, implying that any two sensors will be a minimum of 1km apart. If a finer resolution is desired, we can further divide the population grid into equal-size sub-grids (e.g. 4 sub-grids) with the assumption that the population is evenly distributed across the sub-grids and the sensors will be allowed to be placed at a closer distance. This step should be adopted when the number of sensors is comparable to the number of grids.

Figure 5 shows the placement result for better coverage of the areas where the vulnerable people are located. Since the problem size in this case is small, it is possible to obtain the optimal placement result using Gurobi. As shown in the figure, the sensors are widely distributed across City of Cambridge.

Figure 6 further displays the 20 clusters of points of interest (POIs) within City of Cambridge. From Figure 5 and Figure 6 we can see that the centers of each cluster are selected for deploying the sensors. Here we fix the cluster number to be 20 due to the sensor number constraint, while in practice the distribution of the POIs can also suggest a proper number of sensors needed to guarantee the maximum coverage of the vulnerable people based on a certain objective (e.g. the distance between the location of the people covered to their nearest sensor should be no greater than a threshold $d$).

Figure 7 shows the locations of 20 sensors for better monitoring traffic related air pollution. The locations are selected based on the importance of the road segments/intersections in terms of the pollution related factors described above.

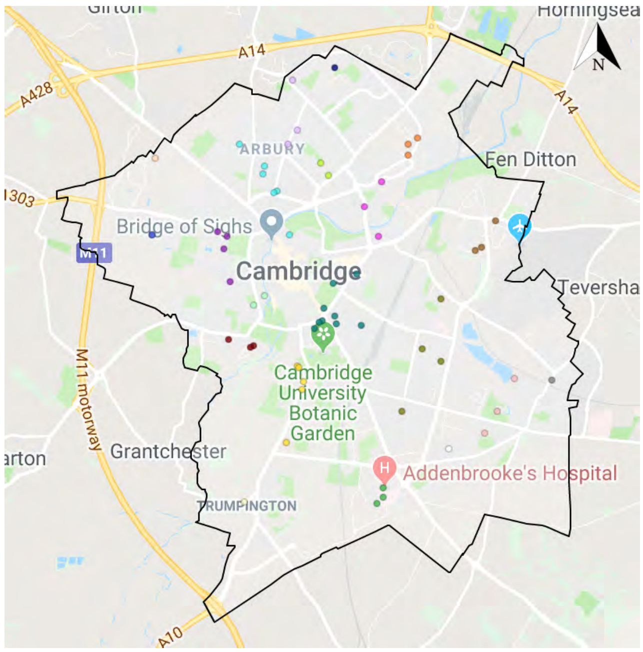


**FIGURE 6.** **Points of interest divided into 20 clusters.**

Compared to Figure 5, we note that some important roads identified are quite close to the points of interest, indicating that the placement results of different objectives may overlap and "redundant" placement should be avoided when three objectives are considered altogether.

The existing locations of 20 sensors deployed in City of Cambridge are shown in Figure 8.[5] It shows that the

[5]The sensor placement data is based on information provided by Prof. Ian Leslie, the University of Cambridge.

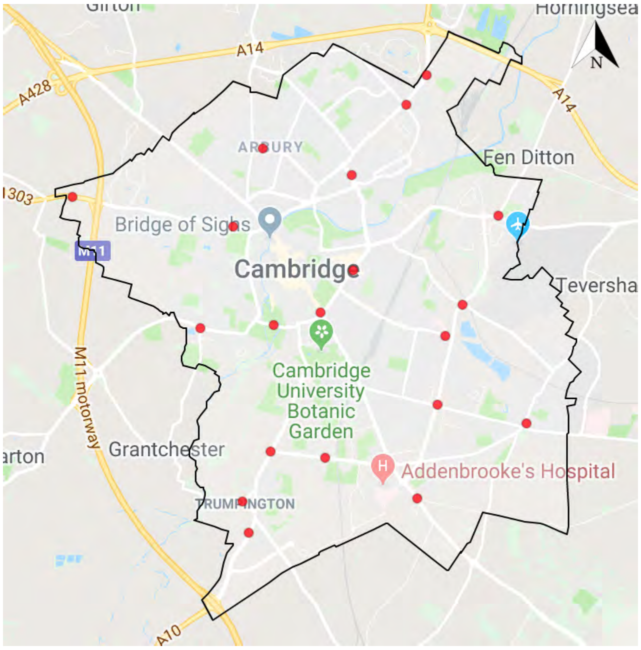


**FIGURE 7.** Sensor placement for better monitoring traffic emissions.

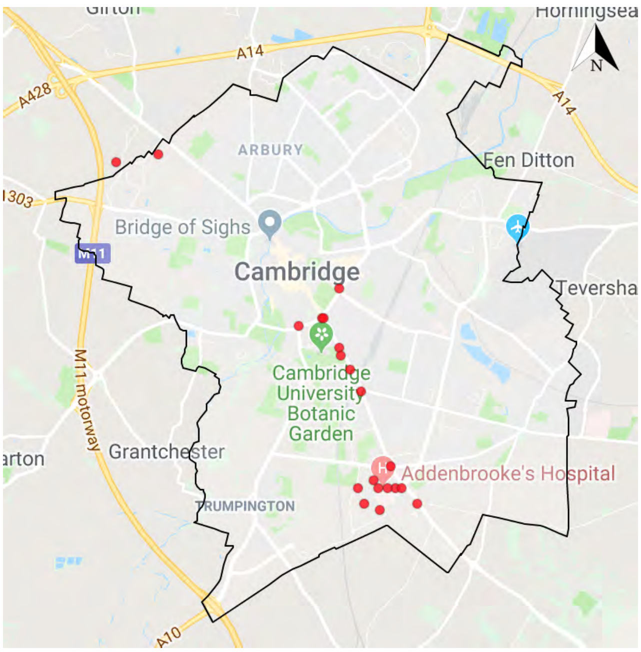


**FIGURE 8.** Original setting of 20 low-cost pollution sensors deployed in City of Cambridge.

sensors are non-uniformly distributed. These sensors are placed mainly near one major road, one hospital and one construction site. However, the rest of the City is yet to be covered by the sensors.

Lastly we show the placement result of 20 sensors that considers all of the objectives above in Figure 9. In order to highlight the difference, we use different colors to represent different objectives: yellow circles for better protecting the vulnerable, blue circles for better monitoring the emissions

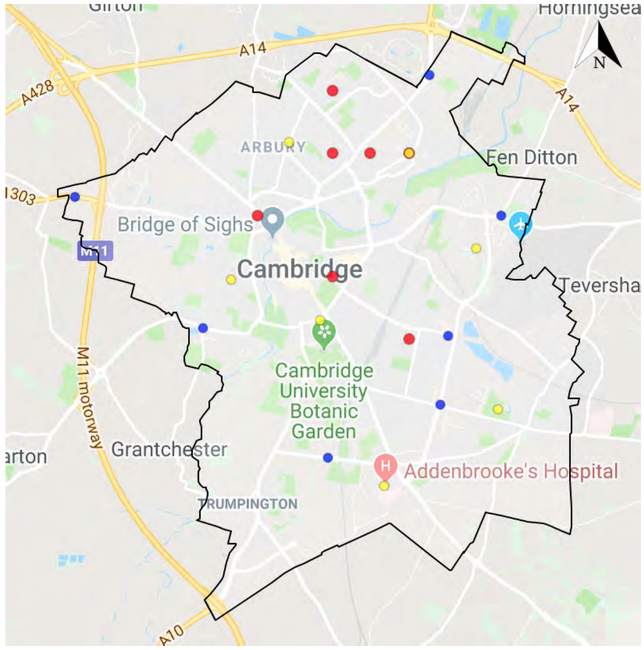


**FIGURE 9.** Multiple-objective placement result of 20 low-cost pollution sensors in City of Cambridge.

**TABLE 5.** Comparison of different objective values of citizen-centric placement.

| Objective | Original placement | Multiple objectives | A only | B only | C only |
|---|---|---|---|---|---|
| A. Citizen satisfaction (%) | 20.95 | **58.02** | **68.18** | 57.23 | 46.08 |
| B. Distance to the vulnerable (km) | 105.08 | **42.93** | 44.57 | **18.86** | 53.49 |
| C. Number of top 20 important roads | 3 | **7** | 4 | 1 | **20** |

and red circle for maximizing overall citizen satisfaction. As shown in the figure, one yellow circle and one red circle are nearly at the same location, indicating that this location will benefit two objectives and one extra location is allowed to further expand the coverage.

The placement results are shown in Table 5, where the better results are in bold. It can be observed that the multi-objective placement result performs much better than the original placement in terms of all three proposed objectives. In addition, although no single optimal result is achieved by the multi-objective approach, it maintains a good balance between different objectives.

## V. DISCUSSION

Our three proposed objectives in air quality monitoring can be easily adopted and applied to other cities or regions of any geographical scale for designing an effective sensor placement scheme as long as urban dynamics data (such as population data, points of interest data and traffic data) are available.

For the first two objectives, it is not always possible to achieve the optimal results due to complexity of the original problems. However, for the first objective, the proposed solution still allows us to obtain a good result with good confidence; for the second objective, the solution sets the bounds for the optimal solution. For the third objective, we have only developed a simple method for collecting the traffic data of City of Cambridge. Since the traffic information provided by Google Map is available on most cities, it can provide the basic traffic data for our optimization model. In case more types of traffic data, for instance, the car type and the traffic speed data are available, it is possible to develop a more refined definition regarding the level of importance of a road intersection/segment, and incorporate such information into our sensor placement optimization.

## VI. CONCLUSION

In this paper, we have introduced three important citizen-centric sensor placement objectives for air quality monitoring. Next, we have formulated three optimal sensor placement problems and proposed efficient methods to obtain the optimal solutions. Finally, we have conducted a case study in City of Cambridge to compare the placement results with the existing sensor placement scheme put forward by Cambridge to verify its effectiveness. Our results have shown that, in order to optimize the citizen-centric objectives, there is a need to re-orient the current sensor placement strategy in Cambridge. For instance, to maximize overall citizen satisfaction, there is a need to spread more of the currently installed low-cost air pollution sensors across the northeastern part of the City, rather than concentrating most of them at the southeastern part. Further, to better monitor traffic emissions, the sensors should be more evenly distributed across the City, rather than being concentrated in the city center and the southeastern part. Hopefully, our results can benefit future government decision-making. Given any budget constraint, we can still maximize the welfare of those citizens residing in Cambridge, by distributing sensors more evenly across different parts of the City.

## REFERENCES

[1] *Ambient(Outdoor) Air Quality and Health*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] X. Chen, S. Shao, Z. Tian, Z. Xie, and P. Yin, "Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample," *J. Cleaner Prod.*, vol. 142, pp. 915–925, Jan. 2017.

[3] J. Lü, L. Liang, Y. Feng, R. Li, and Y. Liu, "Air pollution exposure and physical activity in China: Current knowledge, public health implications, and future research needs," *Int. J. Environ. Res. Public Health*, vol. 12, no. 11, pp. 14887–14897, 2015.

[4] F. J. Kelly and J. C. Fussell, "Air pollution and public health: Emerging hazards and improved understanding of risk," *Environ. Geochem. Health*, vol. 37, no. 4, pp. 631–649, 2015.

[5] N. Castell *et al.*, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environ. Int.*, vol. 99, pp. 293–302, 2017.

[6] C.-W. Ko, J. Lee, and M. Queyranne, "An exact algorithm for maximum entropy sampling," *Oper. Res.*, vol. 43, no. 4, pp. 684–691, 1995.

[7] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, no. 2, pp. 235–284, 2008.

[8] X. Wu and M. Liu, "*In-situ* soil moisture sensing: Measurement scheduling and estimation using compressive sensing," in *Proc. ACM/IEEE 11th Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2012, pp. 1–11.

[9] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao, "Optimal sensor placement and measurement of wind for water quality studies in urban reservoirs," in *Proc. ACM 13th Int. Conf. Inf. Process. Sensor Netw.*, 2014, pp. 167–178.

[10] Y. Hao and S. Xie, "Optimal redistribution of an urban air quality monitoring network using atmospheric dispersion model and genetic algorithm," *Atmos. Environ.*, vol. 177, pp. 222–233, Mar. 2018.

[11] S. Araki, K. Iwahashi, H. Shimadera, K. Yamamoto, and A. Kondo, "Optimization of air monitoring networks using chemical transport model and search algorithm," *Atmos. Environ.*, vol. 122, pp. 22–30, Dec. 2015.

[12] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 437–446.

[13] C. Sun, V. O. K. Li, and J. C. K. Lam, "Optimal citizen-centric sensor placement for citywide environmental monitoring—A submodular approach," in *Proc. IEEE Int. Conf. Sens., Commun. Netw. (SECON)*, Jun. 2018, pp. 1–4.

[14] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.

[15] O. Kariv and S. L. Hakimi, "An algorithmic approach to network location problems. II: The *p*-medians," *SIAM J. Appl. Math.*, vol. 37, no. 3, pp. 539–560, 1979.

[16] I. Buka, S. Koranteng, and A. R. Osornio-Vargas, "The effects of air pollution on the health of children," *Paediatrics Child Health*, vol. 11, no. 8, pp. 513–516, 2006.

[17] M. Simoni, S. Baldacci, S. Maio, S. Cerrai, G. Sarno, and G. Viegi, "Adverse effects of outdoor pollution in the elderly," *J. Thoracic Disease*, vol. 7, no. 1, pp. 34–45, 2015.

[18] *Array of Things*. Accessed: Sep. 1, 2018. [Online]. Available: https://arrayofthings.github.io/

[19] L. Gurobi Optimization. (2018). *Gurobi Optimizer Reference Manual*. [Online]. Available: http://www.gurobi.com

[20] I. I. Cplex, "V12. 1: User's manual for CPLEX," *Int. Bus. Mach. Corp.*, vol. 46, no. 53, p. 157, 2009.

[21] (2016). *Air Quality in Hong Kong 2016*. [Online]. Available: http://www.aqhi.gov.hk/api_history/english/report/files/AQR2016e_final.pdf

[22] K. Zhang, S. Batterman, and F. Dion, "Vehicle emissions in congestion: Comparison of work zone, rush hour and free-flow conditions," *Atmos. Environ.*, vol. 45, no. 11, pp. 1929–1939, 2011.

[23] R. Smit, A. L. Brown, and Y. C. Chan, "Do air pollution emissions and fuel consumption models for roadways include the effects of congestion in the roadway traffic flow?" *Environ. Model. Softw.*, vol. 23, nos. 10–11, pp. 1262–1270, 2008.

[24] K. Zhang and S. Batterman, "Air pollution and health risks due to vehicle traffic," *Sci. Total Environ.*, vols. 450–451, pp. 307–316, Apr. 2013.

**CHENXI SUN** received the B.E. degree in electrical and electronic engineering from The University of Hong Kong, Hong Kong, in 2015, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering. Her research interests include data analytics, and optimization and decision making in smart cities.

**VICTOR O. K. LI** (S'80–M'81–F'92) received the S.B., S.M., E.E., and Sc.D. degrees in electrical engineering and computer science from MIT. He was the Head of the Department of Electrical and Electronic Engineering (EEE), an Associate Dean (Research) of engineering, and the Managing Director of Versitech Ltd. He co-founded Fano Labs Ltd., an artificial intelligence (AI) company with his Ph.D. student. He was a Professor of electrical engineering with the University of Southern California (USC), Los Angeles, CA, USA, and the Director of the USC Communication Sciences Institute. He is currently the Chair of information engineering and a Cheng Yu-Tung Professor of sustainable development with the Department of EEE, The University of Hong Kong. He is also the Director of the HKU-Cambridge Clean Energy and Environment Research Platform and the HKU-Cambridge AI to Advance Well-being and Society Research Platform, which are interdisciplinary collaborations with Cambridge University. He serves for the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange. His research interests include big data, AI, optimization techniques, and interdisciplinary clean energy and environment studies. He is a Fellow of the Hong Kong Academy of Engineering Sciences, the IAE, and the HKIE. In 2018, he was awarded a U.S. $6.3 million RGC Theme-based Research Project to develop deep learning techniques for personalized and smart air pollution monitoring and health management. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the U.K. Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the HKSAR. Sought by government, industry, and academic organizations, he has lectured and consulted extensively internationally.

**JACQUELINE C. K. LAM** was the Hughes Hall Visiting Fellow. She was the Visiting Senior Research Fellow and an Associate Researcher with the Energy Policy Research Group, Judge Business School, University of Cambridge. She is currently an Associate Professor with the Department of Electrical and Electronic Engineering, The University of Hong Kong, and a Co-Director of the HKU-Cambridge Clean Energy and Environment Research Platform and the HKU-Cambridge AI to Advance Well-being and Society Research Platform. Her research interest includes clean energy and environment using interdisciplinary approaches, with a special focus on China and U.K. Her recent research focuses on the use of big data and machine learning techniques to study personalized air pollution monitoring and health management. Her research is published in the IEEE, *Environment International*, *Applied Energy*, *Environmental Science and Policy*, and *Energy Policy*. She has received three times the research grants awarded by the Research Grants Council, HKSAR Government, from 2011 to 2017. The funded amount totaled U.S. $7.8 million in PI or Co-PI capacity. Her recent research study, in joint collaboration with Y. Han and V. O. K. Li on $PM_{2.5}$ pollution and environmental inequality in Hong Kong, has been published in *Environmental Science and Policy* and widely covered by more than 30 local and overseas newspapers and TVs.

**IAN LESLIE** received the B.E. degree and the master's degree in electrical engineering from the University of Toronto, in 1977, and the Ph.D. degree from the Computer Laboratory, University of Cambridge, in 1983.

From 1999 to 2004, he was the Head of the Computer Laboratory, and from 2004 to 2009, he was the University's Pro-Vice-Chancellor for Research. He was with the Computer Laboratory, where he been involved in networking/operating system research. He was a Co-Founder of Nemesys Research Ltd., which exploited the results of research work done in the department and producing hardware and software to allow cost-effective video connection to ATM networks. Nemesys was acquired by FORE Systems, in 1996. He later co-founded CPlane, Inc., which exploited other work in the department concerned with network control. He is currently the Director of University Information Services and a Senior Adviser to the Vice-Chancellor on information system strategy and environmental sustainability. He has supervised about 25 completed Ph.D.s. His research interests include operating systems, distributed systems, networks, and more recently the use of information systems to reduce energy demand.