# Route Selection in Low-cost Participatory Mobile Sensing of Air Quality

Mohamed Anis Fekih, Walid Bechkit, Hervé Rivano

# Route Selection in Low-cost Participatory Mobile Sensing of Air Quality

Mohamed Anis Fekih
*CITI Laboratory EA 3720*
*Univ Lyon, INSA Lyon, Inria*
Villeurbanne, France
mohamed-anis.fekih@insa-lyon.fr

Walid Bechkit
*CITI Laboratory EA 3720*
*Univ Lyon, INSA Lyon, Inria*
Villeurbanne, France
walid.bechkit@insa-lyon.fr

Hervé Rivano
*CITI Laboratory EA 3720*
*Univ Lyon, INSA Lyon, Inria*
Villeurbanne, France
herve.rivano@insa-lyon.fr

***Abstract*—Mobile crowdsensing is a powerful paradigm that takes advantage of low-cost sensors and population density. It allows for large-scale deployments and collection of extensive data, offering a great advantage in multiple fields such as air pollution monitoring, which is a major concern worldwide. Given the mobile nature of the crowd, mobile crowdsensing platforms need to implement adequate route selection/planning solutions to better guide the crowd through the area of interest and maximize the quality of monitoring. In this paper, we propose two route selection algorithms that take into consideration the low accuracy of low-cost sensors in order to find the most informative routes. The similarity-based route selection algorithm aims to maximize spatial coverage by reducing overlaps between participant routes. The cluster-based route selection takes advantage of hierarchical clustering to build groups of similar points of the map according to explanatory variables. We compare the proposed solutions to baseline route selection algorithms, and the results show that our solutions allow for a better estimation while being efficient in terms of travel distance.**

***Index Terms*—Route selection, participatory sensing, low-cost sensors, air quality monitoring.**

## I. Introduction

Mobile crowdsensing [1] is an emerging and powerful paradigm which has recently received a great deal of attention, due to rapidly emerging sensing platforms and their growing needs in terms of both quantity and quality of data. This paradigm has become an appealing solution to collect data without making large investments, especially with the emergence and widespread use of low-cost wireless sensors. Mobile crowdsensing highly involves citizens in the sensing process and has been adopted in numerous solutions such as environmental monitoring [2], [3], healthcare [4], [5], transportation [6], [7], etc. Participants in mobile crowdsensing use sensors embedded on their smart devices (smartphones, smartwatches, etc.) or autonomous nodes to accomplish sensing tasks.

Given the mobile nature of the crowd, mobile crowdsensing platforms need to implement adequate route planning/selection solutions to better guide the crowd through the area of interest and maximize the quality of monitoring. Route planning and route selection are of great importance in mobile crowdsensing and smart cities applications [8]–[10]. On one side, participants in route planning delegate the construction of their routes to the monitoring platform, which will drive their movements by leading them through specific points of interest. On the other side, in route selection, each participant can have multiple pre-computed routes, and the role of the system is to select the most appropriate one regarding the needs of the task. It is clear that the second method considerably limits the degree of freedom of sensing platforms as they are restricted to choose from already built candidate routes that may inevitably overlap, thus limiting coverage and increasing data redundancy.

One of the applications that greatly benefits from mobile crowdsensing is air quality monitoring, due to its major importance, the complexity of deploying expensive monitoring stations (especially in dense urban areas), the large number of potential participants, etc. In fact, several million deaths are attributable to air pollution each year [11]. Air pollution consists of chemical and particles in the atmosphere mainly caused by human activities, such as energy production and fuel combustion. This human-generated pollution is prevalent in urban areas [12] due to high population and building densities, which result in an increasing risk of exposure to pollutants, leading to serious health problems such as cancer, respiratory distress, etc.

In our previous research work, we have designed a low-cost participatory air quality monitoring system that features small, portable sensing nodes equipped with three low-cost sensing probes for monitoring $NO_2$, $PM_1$, $PM_{2.5}$, $PM_{10}$, temperature, and relative humidity [3]. The sensor nodes have been evaluated against reference sensors and have been involved in multiple mobile sensing campaigns in Lyon city since 2019 [3]. Besides the mapping of air quality, the collected data also served to understand the relationship between the sampling rate, the energy consumption, and the mapping quality. The capacity of mapping approaches was also investigated, and the results revealed some heterogeneity in mapping quality between different areas, due to the fact that some of them were covered by multiple sensors, while others were not [13]. This led to questions related to the movement of the crowd and how it could be organized in order to maximize the

This work has been supported by the "LABEX IMU" (ANR- 10-LABX-0088) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). It has also been supported by the ANR under the project ANR-21-CE25-0003 (DRON-MAP).

overall monitoring quality while taking into consideration the constraints of the participants.

In this paper, we address the route selection problem in participatory sensing, while focusing on the application of air quality monitoring to validate our proposal. We present two route selection algorithms, where the first one computes the similarity between the different possible routes, while the second algorithm takes advantage of the similarity between the points of the map through clustering. We evaluate the performance of the proposed algorithms and compare them to other baseline route selection algorithms. Furthermore, we show through the results that our clustering-based routing improves the estimation quality while being efficient regarding the travel distance. The reminder of this paper is organized as follows, Section II discusses related works. Section III describes the scenario and the objective of this paper. Section IV presents the baseline route selection algorithms and explains the two proposed route selection solutions. The validation use case of our proposed algorithms is introduced in Section V. Section VI discusses the obtained results. Finally, Section VII concludes this paper.

## II. Related Works

Mobile crowdsensing have raised questions regarding the movement of the crowd and how to maximize coverage while taking into consideration the constraints of the participants and the addressed application. This has opened the door for new research studies that harness citizen's mobility to improve the mapping of environmental phenomena.

Synthetic measurements were used in [14] to predict a map of $NO_2$ concentrations in the city of Marseille, France, while considering up to 4500 bike-tracks randomly generated. The simulated concentrations were generated using a numerical model with a spatial resolution of $25 \times 25km^2$. Fictive bike tracks were constructed using a cyclist route planner API. The resulting dataset served to train and compare the performance of three estimation models, namely ordinary kriging, multilayer perceptron neural network, and a generalized additive model (GAM).

In multiple crowdsensing applications, the system owner pays the participants for their collected data. In these use cases, the participants often perform sensing tasks at specific locations rather than continuous sampling. The price is generally related to the number of accomplished tasks, their priority, the quality of the data, etc. In this context, given a budget, the system owner optimizes the mobility of the crowd in order to maximize the number of completed tasks or the coverage of the study area, while satisfying the constraints of the participants.

The work in [15] addresses the coverage problem in vehicular crowdsensing by incentivizing participants to deviate from their original path to improve the overall spatial coverage. The authors take advantage of game theory in which vehicles are the participants and their new routes are their strategies. In the proposed solution, all participants are assumed to start at the same time and the reward for a participant $A$ depends on the number of deviations made to the original path and the diversity of the new route. This diversity metric is computed using the Fréchet distance between the path of $A$ and the routes of all other participants. The goal then is to find the set of deviations to be made to participant routes, such that the coverage is maximized.

A three-phase routing algorithm is proposed in [8] in a task assignment context. A task is characterized by its location and its added value to the system. In the first phase, the algorithm iteratively constructs the route from the origin to the destination, by considering at each step the task that has the largest added value. The task is assigned to the user if his device's energy is enough to travel from the current location to the target task, then to the destination point. The second phase of the algorithm performs the construction backwards (i.e., from the destination point to the starting point of the participant). The final part of the solution is the selection of the route with the highest added value. This solution is appropriate if we consider specific sensing locations.

Instead of suggesting a whole route to a participant, the work in [16] takes a different approach to maximize the coverage quality while avoiding redundant data. The study proposes a reverse greedy algorithm that selects only a subset of segments from the participant's route, based on the cost of the task and the available budget. The selected segments are those along which the participant performs the task. Thus, the participant will be rewarded according to the selected segments instead of the whole journey. To achieve such result, the algorithm eliminates redundant segments. Two segments are considered redundant if the distance separating their respective endpoints is less than a predefined distance. After each round of trimming, if the total cost is greater than the available budget, a new round starts with a larger threshold distance to further reduce the number of selected segments.

Gong et al. [17] addressed the path planning problem to maximize the total task quality in a scenario where users and tasks arrive dynamically. Each user has a limited distance budget and has to register his starting and destination points upon arrival. One of the proposed algorithms in this work selects the tasks that lead to the largest gain-cost ratio, one by one in a greedy manner, as long as they satisfy the travel distance budget of the user. An alternative solution tends to guide users to task-dense areas to maximize the cost-gain ratio. In addition to that, the authors designed an algorithm that takes into account the impact of a candidate task on the travel distance budget, in order to leverage tasks with low impact on the available budget. Given a candidate task, this algorithm evaluates the possibilities of the next step by computing the distance from this task to the others. The task that does not distract the participant from the rest of the tasks is selected. This algorithm yielded better performance compared to the other two solutions.

In [18], the goal is to maximize the sub-profit in each time slot to approximately approach the maximum profit of all slots for a given task. A task is represented as a number of sensing locations during a certain period, the area of interest is divided

into N cells, and the task duration into M time slots with the same length. Each participant in this scenario is guided to move along the shortest path. However, to avoid similar routes and maintain a stable dispersed distribution, the authors use 2D entropy to guide participant distribution. The more decentralized the participants, the higher the entropy value. The solution first randomly selects a number of cells and compute the entropy, then repeats the selection process for multiple iterations. At last, the participant distribution with the largest entropy is selected as an approximately optimal solution.

Although the aforementioned works explore mobile crowdsensing capabilities to accomplish sensing tasks, the majority of them focus more on specific sensing locations rather than sensing the entire zone. This prevents performing continuous sampling which considerably limits the potential of the crowdsensing and significantly reduces the spatial resolution, especially in air quality monitoring applications where there is no sensing range as explained in [19]. Furthermore, most air quality monitoring platforms are based on low-cost sensors, as in [3] [20] [21]. However, the majority do not take into consideration the inaccuracies of these sensors, which is one of the main challenges when it comes to predicting air pollution using low-cost environmental sensors. In addition, in participatory sensing, participants may use embedded sensors on their smartphones, adding a new challenge which is sensor heterogeneity. Therefore, taking into consideration sensing errors during the participant recruitment phase and the route planning process should be of utmost importance.

## III. Problem Statement

In this section, we describe the scenario we focus on as well as the global objective we aim at. Afterwards, we give a mathematical formulation of the problem to solve it.

### A. Scenario Description

In this work, we focus on a scenario in which multiple participants are equipped with heterogeneous low-cost environmental sensors to measure a specific phenomenon in a delimited area. Without loss of generality, we consider air quality mapping as use case. Each participant has a starting point and wants to get a path to reach his destination using a routing service. The participant is also willing to take a path that contributes to the knowledge of the studied phenomenon, without deviating too much from the optimal path.

### B. Objective

Our global mission, in this context, is to suggest to each participant a route that might not necessarily be the optimal (shortest) path in terms of distance/duration, but does improve the estimation while being acceptable in terms of journey distance/duration. This implies taking into consideration not only the length of the participant's routes and their relationships with other participant routes, but also the accuracy level of his sensor. In other words, our goal is to find the best combination of routes that allows the system to reduce the overall mapping estimation error while still satisfying participants' constraints in terms of trip distance/duration. The assumptions upon which we build our solution are as follows:

- The duration of the sampling is considerably short (negligible)
- Each route can be divided into smaller segments and routes may overlap along a certain distance (i.e., have segments in common)

### C. Mathematical notation

Let $U = [u_1, u_2, ..., u_n]$ be the set of participants, $S = [s_1, s_2, ..., s_n]$ the set of their respective sensors, $\sigma_k$ the standard error of the *k-th* sensor, and $P_k = [p_{k1}, p_{k2}, ..., p_{kq}]$ the set of possible paths for the *k-th* participant. The goal is to select for each participant $u_k$ a path $p_{kj}$ from his $q$ possible routes, knowing that he is equipped with a sensor $s_k$ which has a standard error $\sigma_k$.

A brute force solution would be to test all the possible combinations (i.e., $q^n$ combinations). However, as the space of solutions grows exponentially with the number of participants, the implementation of such solution in a real life scenario is impractical. To cope with that, we rely on heuristics that are not optimal, but have smaller solution spaces, hence running much faster than the exhaustive search.

## IV. Route selection solutions

Through this contribution, we address route selection in the context of participatory air quality sensing. The objective is to maximize the quality of the prediction where there are no measurements, using a spatial interpolation method and an efficient route selection algorithm. First, we briefly review three traditional solutions that that could be utilized for comparison purposes. We then introduce two algorithms that take into consideration sensing errors and the relationship between participant routes.

### A. Traditional route selection algorithms

Generally, route planning services always offer either the shortest path or $q$ possible paths, while taking into consideration multiple parameters (e.g., traffic condition, waypoints, type of path, etc.). Route selection consists of choosing for each participant a path among the $q$ proposed ones. This is a key element as it determines the geographical zones that will be sampled, which highly impacts the estimation quality. We consider in this part three traditional routing approaches as baseline:

- ***Shortest-path-based routing (SPR):*** This algorithm prioritizes participant comfort by reducing the travel distance for all participants, through suggesting the shortest route among the $q$ paths. In SPR, all route selections are performed independently for each participant.
- ***Longest-path-based routing (LPR) :*** In contrast to SPR, this approach aims at maximizing the set of points to visit. It suggests to all participants the longest route possible among the $q$ suggested routes. It is to be noted that in this contribution, we consider the longest path within

a predefined stretch factor with respect to the shortest path. Suggesting the longest-path for all participants will intuitively expand the set of collected measures, and hence, improve the estimation quality. However, the correlation between paths is not considered, which means that this method does not always maximize the quality of the spatial mapping.

- ***Random routing (RR):*** For each participant, randomly select a route among the $q$ possible ones without taking into consideration its length, the already selected routes for previous participants, or the accuracy level of the sensors.

All three algorithms cited above are easy to implement, but their downside is that they do not take into consideration the correlations between participant routes or the accuracy levels of the sensors. In fact, the routes can pass through similar points or even different points that provide redundant information (points with similar characteristics). These observations feed the need for more sophisticated techniques, that will efficiently attribute routes to the participants that are not necessarily the longest or the shortest ones, but bring as much diversity in the dataset as possible. This can be achieved by introducing metrics and techniques that help to choose the most distinct routes possible and thus bring more information.

### B. First proposition: Similarity-based route selection

The level of accuracy differs from a sensor to another, even between sensors of the same type. Therefore, to offer a good estimation, platforms should have an efficient routing approach that also takes into consideration the heterogeneity of sensing quality. In this regard, given a pool of participants sorted according to their sensors' accuracy, the similarity-based algorithm operates in a greedy manner by considering the $q$ possible routes of the first participant ($q$ routes satisfying a given distance threshold between the shortest and the longest path) and selects the longest one. After that, the algorithm iterates over the next participants in the pool and chooses for each one the route that has the lowest similarity with the already selected routes from the previous participants. Every time a route is selected for a participant, the latter is removed from the pool and the algorithm moves to the participant with the next best sensor, in a greedy manner, until the pool is empty. This process is illustrated by Algorithm 1.

The calculation of the similarity percentage between routes has a major role in finding the most distinct routes possible, and can be computed in multiple ways. Without loss of generality, we have mainly explored a commonly used metric in image segmentation, namely the Sørensen–Dice coefficient (also known as Dice Similarity Coefficient or DSC) [22]. Considering two participant routes $A$ and $B$, the formula for this coefficient is given as follows:

$$DSC_{A,B} = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{1}$$

**Algorithm 1** Similarity-based route selection

**Input:** $U$ {the set of users}
**Output:** $P$ {the set of selected paths}

$P \leftarrow \emptyset$ {initialize the set of selected paths}
$U \leftarrow order(U, descending\_accuracy)$ {order the set of users based on the accuracy of their sensors}
$P \leftarrow longest\_path(u_0)$ {select the longest path for the first user in U}
**for** $u \in U$ with $u \neq u_0$ **do**

$$s\_path \leftarrow \underset{p \in Paths(u)}{\arg\min}\ Similarity(p, P)$$

  $P \leftarrow P \cup s\_path$
**end for**
**return** $P$

Where $|A|$ and $|B|$ are the areas of route A and route B, respectively, and $|A \cap B|$ the area of the intersection between both routes (i.e., area of overlapping).

### C. Second proposition: Cluster-based route selection

Instead of computing the similarity between participant routes in terms of overlapping, this approach focuses on the similarity between the points of the map. The main idea is to regroup the points of the map, not based on the spatial distance separating them, but on explanatory (also called independent) variables related to surrounding conditions (such as distance to routes, meteorology, land-use data, etc.). Indeed, this design is guided by the fact that air quality mapping often make use of explanatory (independent) variables to predict response (dependent) variable in unmeasured points (pollution concentrations in our case) based on available samples of the same variables in measured points. To achieve our goal, we have opted for the agglomerative hierarchical clustering, which is a widely used technique of hierarchical clustering [23], in order to create groups of similar points that might be spatially far from each other, but present similar properties. First, each point of the map is assigned to an individual cluster, and we calculate the distance between the clusters based on the independent variables. Then, clusters are merged successively while minimizing the sum of squared differences between the clusters being merged. As a result, all points of the map are classified into $c$ clusters. Then, for each route, we calculate the number of clusters it traverses. The main idea of this route selection approach is to choose for each participant the route that passes through the largest number of clusters (see Algorithm 2). It is worth mentioning that the performance of this approach depends on the set of explanatory variables collected, and the preprocessing applied to filter out those with low correlation (with the dependent variable) or redundant information.

## V. Validation

In order to validate our proposal, we followed the methodology presented in Fig 1. We give here an overview of the

**Algorithm 2** Cluster-based route selection

**Input:** $U$ {the set of users}
**Output:** $P$ {the set of selected paths}

$P \leftarrow \emptyset$ {initialize the set of selected paths}
$C \leftarrow \emptyset$ {the set of visited clusters}
$U \leftarrow order(U, descending\_accuracy)$ {order the set of users based on the accuracy of their sensors}
**for** $u \in U$ **do**
  {select the path visiting more new clusters}

$$s\_path \leftarrow \underset{p \in Paths(u)}{\arg\max}\ NbClusters(p, C)$$

  $C \leftarrow C \cup Clusters(s\_path)$
  $P \leftarrow P \cup s\_path$
**end for**
**return** $P$

validation approach while more details are given in next subsections. As shown in Fig 1, we consider simulated pollution concentrations as a reference map of the studied area. We first start with a pool of participants, each with a starting and destination points. Each participant has a pollution sensor with its own standard error $\sigma_k$. Instead of constructing the different routes for each user, we rely on a routing service that provides us with several alternative paths whose length is within a predefined stretch factor of the shortest path. After that comes the route selection phase, during which the algorithm suggests a path for each participant based on the reliability of the sensors and the relationship between the routes. Following that, for each location $l$ visited by a participant's route, a synthetic measurement is generated using a normal distribution of mean $y_l$ and variance $\sigma_k^2$, with $y_l$ being the reference concentration at the location $l$. Finally, the generated synthetic observations are passed to spatial interpolation methods in order to produce a predicted pollution concentrations map. The latter is then compared against the reference map to evaluate the impact and performance of the route selection approaches.

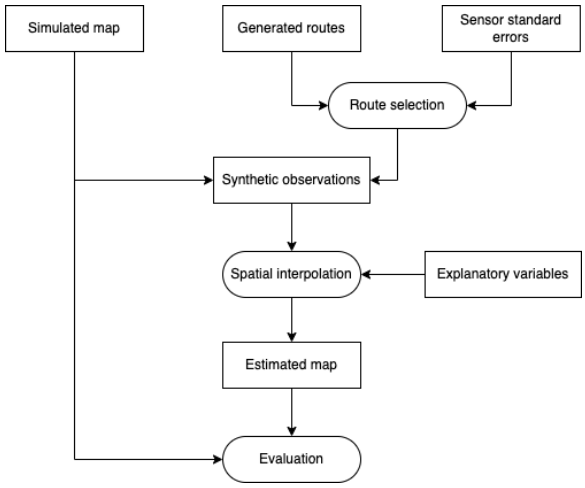


Fig. 1. The general methodology used in validation

### A. Interpolation methods

The use of spatial interpolation methods is a common practice in air quality monitoring field, since there are often fewer sensors to measure every single point of a map, even with a mobile crowd. These techniques give an estimation at unmeasured points based on available samples of the same studied variable [24]. In addition, to better explain the phenomenon, estimation models often need extra information about the surrounding environment. These characteristics may include, but are not limited to, meteorological parameters (e.g., temperature, wind speed, etc.), land-use data (e.g., number of buildings, etc.). Without loss of generality, we have selected three models: Multiple Linear Regression (MLR), K-Nearest Neighbors (KNN) Regression, and eXtreme Gradient Boosting (Xgboost).

Multiple linear regression is an extended version of the simple linear regression that uses just one explanatory variable. MLR aims to model a linear relationship between the studied phenomenon and a number of explanatory variables to predict the concentration values at points where no data was collected [24], [25].

K-Nearest Neighbors is a popular algorithm usually used for classification. The general idea behind it consists of considering the $K$ nearest samples or observations to the point to be predicted based on a distance metric. In regression, the algorithm uses "feature similarity" to identify the $k$ closest points and then assigns the average of their observations to the point of interest. The similarity between the target point and its neighbors is obtained by applying a distance metric to the explanatory variables. The size of the neighborhood (i.e., the value of $k$) is an important part of this algorithm. On one hand, a small neighborhood reduces the number of points used in the regression and leads to overfitting problems. On the other hand, a large value of $k$ means including more points in the estimation causing high sensibility to noise [26], [27].

Xgboost is a fast and powerful ensemble learning method that assumes taking lots of small steps in the right direction results in a better prediction [26]. It implements the gradient boosting decision tree algorithm in which new decision trees are iteratively built to correct the errors of the previous ones, indeed each new tree boosts the attributes that led to estimation errors of the previous trees of the model [28]. Xgboost has shown a good performance in estimating air pollution concentrations in [27].

### B. Study area and reference map

In this paper, we consider the agglomeration of Lyon, which is located in the region of "Auvergne-Rhône-Alpes" in the southeast of France. Our work mainly focuses on a $5 \times 5$ $km^2$ area, corresponding to the center of the Lyon city and its immediate vicinity (see Fig 2).

In order to build a reference map of pollution concentrations, we consider $NO_2$ pollutant and simulations generated by a numerical model called SIRANE [29] [30] [31]. This model is designed for urban areas and is widely adopted by certified air quality monitoring agencies in France. The simulated data

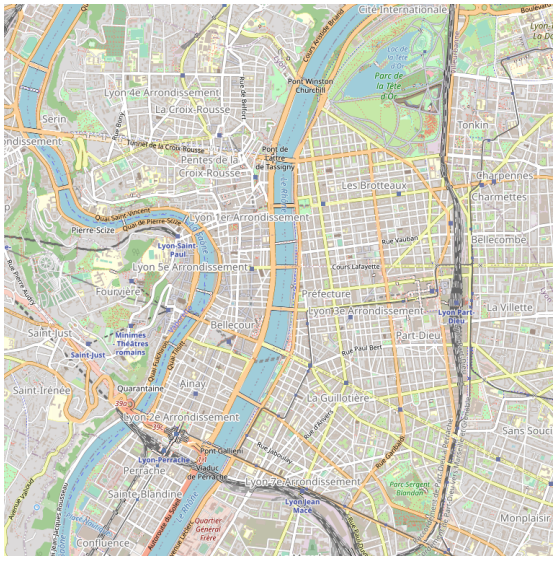


Fig. 2. Map view of the area of interest

correspond to concentrations of $NO_2$ in Lyon city in 2008 (see Fig 3) with a spatial resolution of $20m \times 20m$, resulting in 63.000 points of measure. In addition to that, we use more than 40 explanatory variables for each point on the map, ranging from meteorological data to traffic information and land-use characteristics [27].

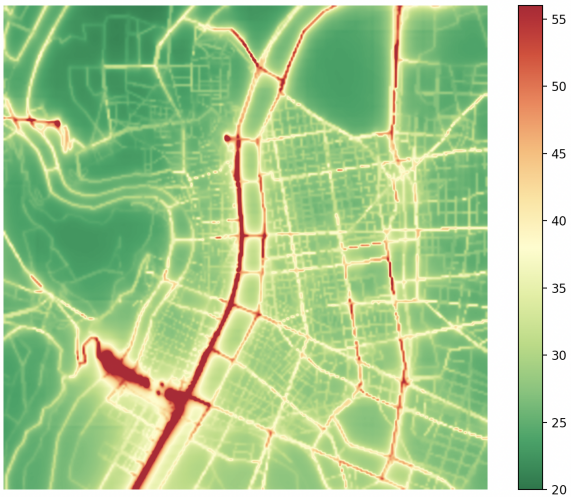


Fig. 3. Reference heatmap of $NO_2$ concentrations (simulated data)

### C. Generation of participant routes and sensor measurements

In our use case, a participant is represented as a triple consisting of his starting point, his destination, and the accuracy level of the sensor he carries. First, we generate 200 random participants with a distance between the starting and destination ranging from $1km$ to $5km$. To match a realistic scenario in which participants have heterogeneous sensors, the accuracy of participant sensors varied between $1\mu g/m^3$ and $20\mu g/m^3$. Then, we use an existing routing service to generate a minimum of 5 alternative routes for each participant to meet a realistic scenario and to have a substantial search space. Furthermore, to avoid ending up with very long alternative routes, we only consider routes that are at most 30% longer than the shortest path.

In the second step, synthetic sensor measurements are generated using the standard errors of the sensors and a normal distribution at each point of the map that a participant's route passes through.

### D. Computing the similarity

In order to implement the similarity-based algorithm, we need to compute the similarity metric between the different routes. This metric should consider segments that do not necessarily overlap but are very close to each other. This may happen when two participants pass by the same road but in opposite directions or parallel sidewalks. For this reason, we build a buffer around each segment of a route, and then compute the similarity between the buffers. The size of the buffer highly influences the similarity metric. Indeed, the larger the buffer is, the higher the similarity value. This size should also be adapted to the spatial resolution of the available data. For the following validation tests, we choose a buffer size of 60 meters around the segment (i.e., 30 meters from each side of the segment).

### E. Clusters construction

The construction of clusters is a crucial phase in the cluster-based route selection approach. It represents the foundation upon which the whole algorithm is built. Therefore, the choice of the optimal number of clusters to construct has a great importance. On one hand, choosing few clusters lowers the similarity threshold. As a result, more points are clustered together without showing much resemblance. On the other hand, a large number of clusters seeks very similar points. Hence, the clusters may gather very few points.

In order to choose an adequate number of clusters in our context, we have evaluated the MAE of the estimation while varying the number of clusters from 250 to 1500 with a step of 250 clusters and 20 iterations for each step. Results depicted in Table I show that the lowest MAE value is reached with 1250 clusters.

TABLE I
MAE VS THE NUMBER OF CLUSTERS WITH 40 PARTICIPANTS

| Number of clusters | 250 | 500 | 750 | 1000 | 1250 | 1500 |
|---|---|---|---|---|---|---|
| MAE ($\mu g/m^3$) | 5.19 | 5.08 | 5.13 | 5.11 | 4.97 | 5.16 |

## VI. Performance evaluation and discussion

To evaluate the performance of our proposal, we first compare multiple spatial interpolation models using the same route selection strategy in order to get insights on which model performs better in estimating $NO_2$ concentrations. The best model is then used in the second part, which compares

TABLE II
AVERAGE EXECUTION TIME OF 1 ITERATION FOR KNN, MULTIPLE LINEAR REGRESSION, AND XGBOOST

| Estimation method | KNN | MLR | XGBoost |
|---|---|---|---|
| Average Execution time (seconds) | 12.95 | 0.31 | 0.73 |

our proposed route selection strategies with the previously presented baseline approaches, in terms of estimation error and travel distance.

### A. Comparison of the performance of MLR, KNN, and Xgboost

In order to get an insight on which statistical model gives better results, we conducted multiple simulations with the MLR, KNN, and XGBoost, using the similarity-based route selection approach, while increasing the number of participants from 15 to 40 and randomly generating sensor errors between $1\mu g/m^3$ to $10\mu g/m^3$. The number of selected neighbors did not show a significant impact on the performance of KNN. Therefore, we chose $k = 5$ in our simulations. Fig. 4 depicts the results of the average MAE in function of the number of participants, obtained after 20 iterations. We observe that as the number of participants increases, the prediction error decreases for all three models. XGBoost outperforms KNN by around 30% and MLR by nearly 42% in terms of MAE. Moreover, we present in Table II the average execution time of the three models for one iteration. The table shows that MLR is the fastest model, while KNN is the slowest among the three. The reason that KNN is much slower is that the dataset used to perform the prediction is large in this study, which makes the prediction step computationally intensive for KNN, because it needs to loop over the entire dataset to find the closest points to each unsampled one. The observations from Fig. 4 along with those of Table II give hints about which model is more suitable for a real life scenario. In our case, we chose XGBoost for the upcoming evaluations.

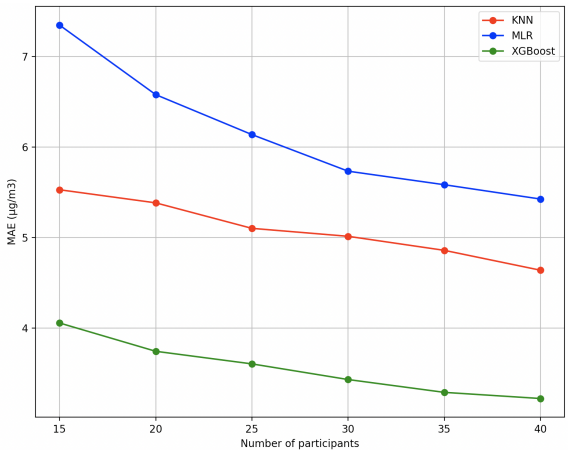


Fig. 4. MAE vs number of participants

### B. Comparison of route selection algorithms

With the aim of evaluating the impact of route selection strategies on the estimation of $NO_2$ concentrations, we carried out multiple tests considering the shortest-path-based routing, the longest-path-based routing, and the two proposed route selection algorithms, i.e., the similarity-based routing, and the cluster-based routing. The two proposed solutions have the same goal, that is, offering the most informative routes possible, while taking into account the correlations between them. However, the difference lies in the evaluation of the relationship between the routes. The similarity-based routing computes the similarity between participant routes and tries to reduce the overlapping between the selected routes, while the cluster-based routing takes advantage of hierarchical clustering to form groups of similar points and then tries to maximize the set of visited groups.

The performance of the four route selection algorithms are evaluated using XGBoost, based on the observations from the previous test. Fig. 5 depicts the MAE of estimation in function of the number of participants. The experiments are performed with a number of participants increasing from 15 to 40 and repeated 20 times. At each iteration, we randomly select a new set of participants with sensor errors between $1\mu g/m^3$ to $20\mu g/m^3$. Results clearly show that increasing the number of users helps decreasing the estimation error due to the presence of more routes and hence a larger area being covered. An interesting observation is that while the shortest-path-based routing has the worst performance in terms of MAE of the estimation, the similarity-based and the cluster-based solutions outperforms the longest-path-based routing. This means that choosing the longest path is not always a good decision to improve the estimation. Indeed, the similarity-based route selection performs up to 15.42% better than the shortest-path-based algorithm and 3.49% better than the longest-path-based solution, while the cluster-based solution surpasses the shortest-path-based and the longest-path-based by 16.84% and 5.22%, respectively.

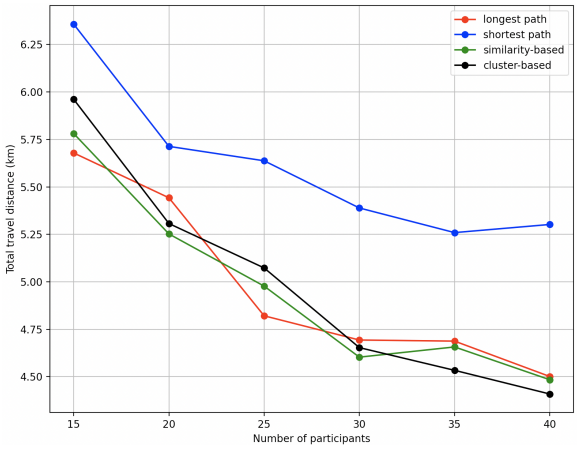


Fig. 5. MAE vs number of participants

In addition, we investigate the impact of the different route

selection approaches on the travel distance of participants. For this purpose, we calculate the total traveling distance for each route selection solution. Fig 6 shows the total travel distance accumulated for all participants. As expected, the shortest-path-based algorithm offers the shortest distance by definition and hence outperforms all the other solutions. Nonetheless, results reveal that our proposed solutions are also more efficient than the longest-path selection in terms of traveled distance. Indeed, the similarity-based algorithm improves the travel distance by nearly 15.27% compared to the longest-path-based algorithm and 10.18% compared to the cluster-based solution. These promising results state that our proposed solutions give the lowest error while offering shorter routes compared to the best of baseline algorithms.

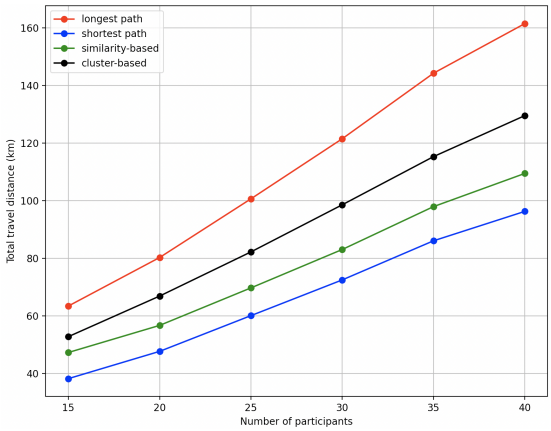


Fig. 6. Total travel distance for the different route selection algorithms

Finally, to give an idea of the final result, we perform simulations to estimate $NO_2$ concentrations with two configurations of sensing errors. For the first simulation, sensing errors are randomly generated between $1\mu g/m^3$ and $5\mu g/m^3$, while in the second one, sensing errors vary between $1\mu g/m^3$ and $20\mu g/m^3$. Both simulations are executed 20 times using 40 participants and the similarity-based route selection algorithm. The results are depicted in Fig 7 and Fig 8, respectively. It is shown that although capturing most of the highly polluted zones, the use of sensors with large errors leads to an overall noisy and poor estimation.

## VII. Conclusion

Air pollution has become a major threat to human health in recent years. Public and government authorities are making considerable efforts to help reduce air pollution around the world. Today, traditional air pollution monitoring stations are clearly not sufficient to assess local exposure to pollutants, which is why mobile crowdsensing, powered by recent developments in sensing probes and communication protocols, has gained a lot of attention. In this work, we address the problem of route selection in participatory sensing with low-cost sensors. We propose two route selection algorithms that take into consideration the characteristics of low-cost sensors in the decision process. The similarity-based routing algorithm

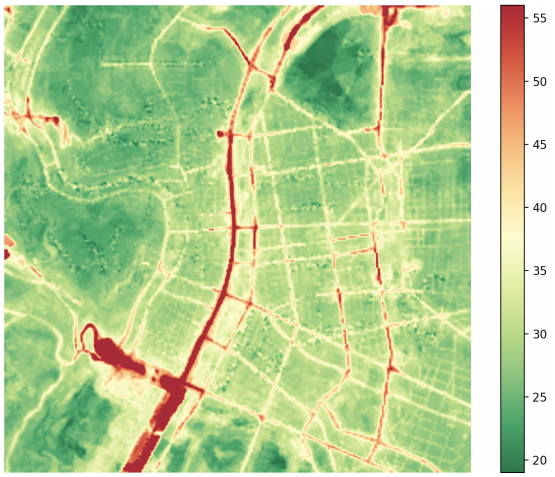


Fig. 7. Heatmap of estimated $NO_2$ concentrations using XGBoost, achieved with 40 participants and sensor errors varying between $1\mu g/m^3$ and $5\mu g/m^3$

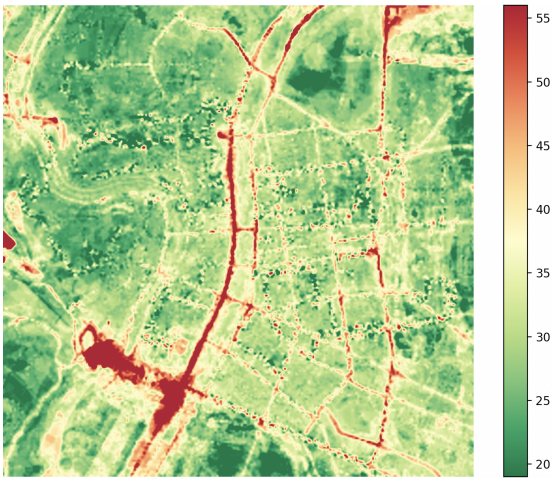


Fig. 8. Heatmap of estimated $NO_2$ concentrations using XGBoost, achieved with 40 participants and sensor errors varying between $1\mu g/m^3$ and $20\mu g/m^3$

aims to give spatially dispersed routes to maximize coverage using the Sørensen-Dice coefficient, a commonly used metric in image segmentation. The cluster-based routing algorithm makes use of hierarchical clustering to build groups of similar points, then tries to visit as many groups as possible to increase the diversity of the collected information. Both algorithms perform in a greedy manner by suggesting at each iteration a route for the user with the most accurate sensor. We compare our solutions to two baseline route selection algorithms, namely the longest-path-based and the shortest-path-based algorithms. We also show through multiple simulations that our route selection approaches can obtain comparable results to the longest-path-based approach or even outperform it, while being efficient regarding the travel distance. Our algorithms are adapted to a scenario in which participants do not necessarily arrive at the same time. Moreover, the idea behind the clustering-based approach can be adapted to other use cases, such as reducing

the individual exposure of citizens to air pollution, by reducing the number of highly polluted clusters to be visited.

## REFERENCES


[1] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[2] M. Zappatore, A. Longo, and M. A. Bochicchio, "Crowd-sensing our smart cities: A platform for noise monitoring and acoustic urban planning," *Journal of Communications Software and Systems*, vol. 13, no. 2, pp. 53–67, 2017.

[3] M. A. Fekih, W. Bechkit, H. Rivano, M. Dahan, F. Renard, L. Alonso, and F. Pineau, "Participatory air quality and urban heat islands monitoring system," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2020.

[4] R. Kraft, W. Schlee, M. Stach, M. Reichert, B. Langguth, H. Baumeister, T. Probst, R. Hannemann, and R. Pryss, "Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain," *Frontiers in neuroscience*, vol. 14, p. 164, 2020.

[5] J. M. Cecilia, J.-C. Cano, E. Hernández-Orallo, C. T. Calafate, and P. Manzoni, "Mobile crowdsensing approaches to address the covid-19 pandemic in spain," *IET Smart Cities*, vol. 2, no. 2, pp. 58–63, 2020.

[6] S. Hu, L. Su, H. Liu, H. Wang, and T. F. Abdelzaher, "Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification," *ACM Transactions on Sensor Networks (TOSN)*, vol. 11, no. 4, pp. 1–27, 2015.

[7] H. Yan, Q. Hua, D. Zhang, J. Wan, S. Rho, and H. Song, "Cloud-assisted mobile crowd sensing for traffic congestion control," *Mobile Networks and Applications*, vol. 22, no. 6, pp. 1212–1218, 2017.

[8] X. Tao and W. Song, "Efficient path planning and truthful incentive mechanism design for mobile crowdsensing," *Sensors*, vol. 18, no. 12, p. 4408, 2018.

[9] Z. Li, J. Zhang, J. Gan, P. Lu, Z. Gao, and W. Kong, "Large-scale trip planning for bike-sharing systems," *Pervasive and Mobile Computing*, vol. 54, pp. 16–28, 2019.

[10] Y. Zhang, B. Aliya, Y. Zhou, I. You, X. Zhang, G. Pau, and M. Collotta, "Shortest feasible paths with partial charging for battery-powered electric vehicles in smart cities," *Pervasive and Mobile Computing*, vol. 50, pp. 82–93, 2018.

[11] World Health Organization, "Burden of disease from the joint effects of household and ambient air pollution for 2016," 2018.

[12] G. Cannistraro, M. Cannistraro, A. Cannistraro, A. Galvagno, and F. Engineer, "Analysis of air pollution in the urban center of four cities sicilian," *Int. J. Heat Technol*, vol. 34, no. 2, pp. S219–S225, 2016.

[13] M. A. Fekih, W. Bechkit, and H. Rivano, "On the data analysis of participatory air pollution monitoring using low-cost sensors," in *2021 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2021, pp. 1–7.

[14] C. Bertero, J.-F. Léon, G. Trédan, M. Roy, and A. Armengaud, "Urban-scale no2 prediction with sensors aboard bicycles: A comparison of statistical methods using synthetic observations," *Atmosphere*, vol. 11, no. 9, p. 1014, 2020.

[15] H. Chintakunta, X. Wang, and L. G. Jaimes, "Improving sensing coverage in vehicular crowdsensing using location diversity," in *2022 International Conference on Connected Vehicle and Expo (ICCVE)*. IEEE, 2022, pp. 1–6.

[16] Y. Chen, P. Lv, D. Guo, T. Zhou, and M. Xu, "Trajectory segment selection with limited budget in mobile crowd sensing," *Pervasive and Mobile Computing*, vol. 40, pp. 123–138, 2017.

[17] W. Gong, B. Zhang, and C. Li, "Location-based online task assignment and path planning for mobile crowdsensing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1772–1783, 2018.

[18] Y. Chen, D. Guo, and M. Xu, "Prosc+: Profit-driven online participant selection in compressive mobile crowdsensing," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–6.

[19] A. Boubrima, W. Bechkit, and H. Rivano, "On the deployment of wireless sensor networks for air quality mapping: Optimization models and algorithms," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1629–1642, 2019.

[20] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele, "Pushing the spatio-temporal resolution limit of urban air pollution maps," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2014, pp. 69–77.

[21] A. Anjomshoaa, F. Duarte, D. Rennings, T. J. Matarazzo, P. deSouza, and C. Ratti, "City scanner: Building and scheduling a mobile sensing platform for smart city services," *IEEE Internet of things Journal*, vol. 5, no. 6, pp. 4567–4579, 2018.

[22] T. S. Mathai, L. Jin, V. Gorantla, and J. Galeotti, "Fast vessel segmentation and tracking in ultra high-frequency ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 746–754.

[23] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Hierarchical clustering," *Cluster analysis*, vol. 5, pp. 71–110, 2011.

[24] X. Xie, I. Semanjski, S. Gautama, E. Tsiligianni, N. Deligiannis, R. T. Rajan, F. Pasveer, and W. Philips, "A review of urban air pollution monitoring and exposure assessment methods," *ISPRS International Journal of Geo-Information*, vol. 6, no. 12, p. 389, 2017.

[25] A. Larkin, J. A. Geddes, R. V. Martin, Q. Xiao, Y. Liu, J. D. Marshall, M. Brauer, and P. Hystad, "Global land use regression model for nitrogen dioxide air pollution," *Environmental science & technology*, vol. 51, no. 12, pp. 6957–6964, 2017.

[26] X. Ren, Z. Mi, and P. G. Georgopoulos, "Comparison of machine learning and land use regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous united states," *Environment International*, vol. 142, p. 105827, 2020.

[27] M. A. Fekih, I. Mokhtari, W. Bechkit, Y. Belbaki, and H. Rivano, "On the regression and assimilation for air quality mapping using dense low-cost wsn," in *International Conference on Advanced Information Networking and Applications*. Springer, 2020, pp. 566–578.

[28] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[29] L. Soulhac, P. Salizzoni, F.-X. Cierco, and R. Perkins, "The model sirane for atmospheric urban pollutant dispersion; part i, presentation of the model," *Atmospheric environment*, vol. 45, no. 39, pp. 7379–7395, 2011.

[30] L. Soulhac, P. Salizzoni, P. Mejean, D. Didier, and I. Rios, "The model sirane for atmospheric urban pollutant dispersion; part ii, validation of the model on a real case study," *Atmospheric environment*, vol. 49, pp. 320–337, 2012.

[31] L. Soulhac, C. V. Nguyen, P. Volta, and P. Salizzoni, "The model sirane for atmospheric urban pollutant dispersion. part iii: Validation against no2 yearly concentration measurements in a large urban agglomeration," *Atmospheric environment*, vol. 167, pp. 377–388, 2017.