Devising a Responsible Framework for Air Quality Sensor Placement

Jevon Westcarr
Data Science, Artificial Intelligence and Modelling
University of Hull
Hull, UK
j.a.westcarr-2021@hull.ac.uk

Venkata M. V. Gunturi
School of Computer Science
University of Hull
Hull, UK
v.gunturi@hull.ac.uk

Sheen Mclean Cabaneros
School of Engineering
University of Hull
Hull, UK
s.m.cabaneros@hull.ac.uk

Rameez Raja Kureshi
School of Computer Science
University of Hull
Hull, UK
r.kureshi@hull.ac.uk

Dhavalkumar Thakker
School of Computer Science
University of Hull
Hull, UK
d.thakker@hull.ac.uk

Amanda Porter
Public Health Intelligence
Hull City Council
Hull, UK
mandy.porter@hullcc.gov.uk

***Abstract*—A major challenge faced when developing smart, sustainable urban environments is the reduction of air pollutants that adversely impact citizens' health. The UK has implemented strategies such as clean air zones (CAZs) coupled with the use of sensor technologies to monitor the changes in these pollutant concentrations while simultaneously reducing them over time. Consequently, this poses important concerns: how much consideration has been given to the positioning of these sensors, and are the citizens the focal point of this decision-making process? In this paper, we seek to address these concerns by introducing a framework which responsibly positions air pollution sensors based on three conjoined tenets: 1) sensors should equitably cover areas of varying deprivations faced by citizens, 2) sensors should cover areas that have dense concentrations of specified parameters of interest which directly impact its citizens' health, behavioural practices and other factors, and 3) sensor coverage should be maximised across the expanse of the city. After laying the foundation of our algorithm, we then demonstrate that sensors can be positioned in a responsible manner that upholds the aforementioned objectives using geospatial data for the city of Hull, one of the most deprived areas of England.**

***Keywords*—sensor, smart city, air quality, urban observatory, fairness, geospatial, clean air zone, monitoring**

I. Introduction

Increasingly, local administrations of modern urban cities deploy advanced technologies to enhance the quality of life. Among these technologies, air quality sensors are the most popular tools for real-time monitoring and management of atmospheric pollutants. By providing granular insights into air quality parameters such as particulate matter ($PM_{2.5}$, PM_{10}) and nitrogen dioxide (NO_2), these sensors empower policymakers and urban planners to set up urban observatories [1] to devise targeted interventions (e.g., Clean Air Zones [2]) aimed at curbing pollution levels and safeguarding public health.

However, the determination of "ideal" air quality sensor locations is a non-trivial task for local government officials given the multitude of parameters that need to be considered. These parameters may include but are not limited to population density, the English Indices of Multiple Deprivations (IMD), and the age-standardised rates of pollution-related health conditions such as cardiovascular and respiratory diseases.

This paper delves into designing an algorithmic framework which considers the geographic spread of multiple parameters of interest through which a suitable set of locations for air quality sensor placements is recommended. Our approach achieves fairness by recommending extensive sensor coverage across the varied levels of deprivation. The key aim is to determine the locations such that we "optimise" on both *target parameters of interest* (e.g., assured coverage in hot spots of poor respiratory health) and *fairness* in sensor distribution. We construct our *responsible* framework for air quality monitoring by considering both the "target" and "fairness" parameters.

The current literature in air quality sensor placement puts emphasis on achieving good coverage on either the geographic spread of pollutants (e.g., $PM_{2.5}$) or on the geographic spread of citizens (e.g., optimising on parameters such as population density and key demographics, etc). In the former, sensors are positioned in areas where specific pollutants are more densely concentrated [3], [4], [5], [6]. The algorithmic techniques used consisted of physics-based approaches using proper orthogonal decomposition (POD) [3], clustering based models in conjunction with integer linear programming models [5], and nonlinear programming techniques using the BONUS algorithm [6] when covering air pollutant dense regions. These methods, however, do not emphasise the underlying socioeconomic conditions being faced by the residents of an area of interest.

Citizen-based coverage relates to placing sensors in areas where citizens are set as the focal decision point [7], [8], [9]. The use of greedy and genetic algorithms for single and multi-objective optimisation was used for determining citizen based coverage approaches [7]. As well as physics based multiresolution dynamic mode decomposition to position sensors in a way that maximised the coverage of air pollutants while also optimising on socioeconomic factors such as racial minority populations and relative income status by penalising placement in areas that represent the racial majority and wealthy regions [8]. Our work seeks to equitably position sensors across all areas of deprivation without explicitly penalising specific socioeconomic parameters.

While the literature in the area of sensor placement is rich, to the best of our knowledge, previous works so far have developed solutions that only consider a few parameters at a time. Our work is complementary to the existing works in the sense that we propose a generic algorithmic framework that can consider a more expansive combination of "target" and "fairness" geospatial parameters to recommend sensor locations. This paper makes the following contributions:

- We define the novel problem of fairness-based sensor location determination for air quality monitoring which aims to optimise on both *target* and *fairness* parameters.
- We propose a novel algorithm which recommends sensor locations while optimising on both *target* and *fairness* parameters.
- We evaluated our proposed algorithm using real datasets from Hull, UK. More specifically, Lower Layer Super-Output Area (LSOA) data, English IMD, population density, the emergency hospital admissions for all cardiovascular diseases and the emergency hospital admissions for all respiratory diseases. For brevity, the last two are referred to as all cardiovascular diseases and all respiratory diseases throughout the paper.
- Our results on Hull datasets indicate that our proposed algorithm achieves good scores while optimising target parameters (population density, all respiratory and all cardiovascular diseases) and the fairness parameter (IMD). A formal definition of used metrics is detailed later in Section II.

The rest of the paper is organized as follows. In Section II, we present the basic concepts and formally define our problem, then discuss our proposed algorithm in Section IV and present the results of our algorithm on a real-world dataset for the city of Hull in Section V. We then explore some limitations to our approach in Section VI and finally provide our conclusion in Section VII.

II. Basic Concepts

Geospatial Framework: a set, $\mathbf{\Phi}$, of LSOA elements representing geographical areas of interest. LSOAs have been specifically derived for statistical analysis from the UK Censuses by the Office for National Statistics. The boundary files are available in the form of a shapefile which can be loaded into a Geographical Information System (GIS) mapping software (e.g. QGIS).

Geospatial Field Function (F): a function, F, over the geospatial framework that maps Φ to a set of categorical values given by $F : \Phi \rightarrow \{High, Medium, Low\}$.

The function is applied across several geospatial layers corresponding to population density, the emergency admission rates for all respiratory diseases (J00-J99) and all cardiovascular diseases (I00-I99) as defined by the International Classification of Diseases [10], and IMD. Field function values are generally defined over a set of real numbers, but a discretisation procedure was applied in this work to generate the categorical values. The procedure involves tercile divisions, e.g. division of a dataset via the 33rd and 66th percentile of the values or the first and second terciles (T1 and T2), respectively. Thus, all LSOAs with values less than T1 are marked as *Low*, those greater than T2 as *High*, and those in between T1 and T2 (inclusive) as *Medium*.

Target Field Function: a field function that would be used to drive the algorithm in terms of covering high-intensity (on certain parameters of interest) areas. For example, one would want to cover areas that have high-intensity of respiratory diseases. In this paper, we will be using the following Geospatial layers as the target field functions: (i) population density, (ii) all respiratory diseases, and (iii) all cardiovascular diseases.

Fairness Field Function (FF): A field function that would be used to drive the fairness goal of the algorithm. In this paper, we will be using IMD as our fairness geospatial layer.

Intensity Score Metric (ISM): Given a candidate solution set $\mathbf{S} \subset \mathbf{\Phi}$ with k elements and a collection of target field functions (λ), the intensity score metric of $\mathbf{S}$ is calculated as follows:

$$\text{ISM}(\mathbf{S}) = \frac{\omega_H \sum_{F \in \lambda} |F_H| + \omega_M \sum_{F \in \lambda} |F_M| + \omega_L \sum_{F \in \lambda} |F_L|}{\omega_H \sum_{F \in \lambda} (|F_H| + |F_M| + F_L|)} \tag{1}$$

where $|F_H|$, $|F_M|$, and $|F_L|$ represent the number of LSOAs in $\mathbf{S}$ with field function F values of *High*, *Medium* and *Low*, respectively, and the weights ω_H, ω_M, and ω_L are real-valued scalars which satisfy the condition $\omega_H > \omega_M > \omega_L$ to prioritise the selection of elements in Φ that have field function values of *High*. ISM values range from 0 and 1, inclusive, with 1 representing the most favourable outcome. Higher ISM values indicate that most selected LSOAs in $\mathbf{S}$ have values that are mostly *High* and/or *Medium*.

Fairness Score Metric (FSM): Given a candidate solution set $\mathbf{S} \subset \mathbf{\Phi}$ and a fairness field function FF, the fairness score metric of S is calculated as follows:

$$\text{FSM}(\mathbf{S}) = -(p(H)\cdot\log p(H) + p(M)\cdot\log p(M) + p(L)\cdot\log p(L)) \quad (2)$$

such that

$$p(H) = \frac{|FF_H|}{|FF_H| + |FF_M| + |FF_L|} \quad (2a)$$

$$p(M) = \frac{|FF_M|}{|FF_H| + |FF_M| + |FF_L|} \quad (2b)$$

$$p(L) = \frac{|FF_L|}{|FF_H| + |FF_M| + |FF_L|} \quad (2c)$$

where $|FF_H|$, $|FF_M|$, and $|FF_L|$ represent the number of LSOAs in $\mathbf{S}$ with FF values of *High* (H), *Medium* (M) and *Low* (L), respectively. FSM values closer to 0.5 indicate that a balanced spread of placements across the IMD field function was achieved. We expect the division of values to follow $H = 33\%$, $M = 33\%$ and $L = 33\%$ of FSM values.

III. Problem Definition

Given the complexity of the problem of air pollutant monitoring [8], it is critical to implement a solution that addresses these complexities. Our approach seeks to provide a framework that selects sensor locations through a sequential process that allows each parameter to determine a potential location best suited for a sensor. The model aims to produce a set of locations that 1) maximises the ISM value of solution set $\mathbf{S}$, 2) maximises the FSM value of $\mathbf{S}$ and 3) maximises the distance between the LSOAs selected in $\mathbf{S}$. To achieve this, our model accepts five primary input variables. It uses a single fairness field function and a series of target field functions. It then takes the number of LSOAs where sensors should be placed (k) and the minimum number of LSOAs that form a cluster represented by τ. Finally, it uses the minimum distance allowed between sensors, represented by d_{th}.

A. Illustrating the problem

Fig. 1 illustrates our problem using a synthetic dataset comprising four input field functions, namely three target field functions (e.g., population density, all respiratory diseases, and all cardiovascular diseases) and a fairness field function (e.g., IMD). The field functions are identically shaped grids where each grid space is considered as an LSOA. In this example, we have set the number of sensors $k = 4$. For ease of navigation, the grid columns are labeled with letters ranging from A to D while the rows are labeled with numeric values ranging from 1 to 5. The LSOAs in each of the input field functions have been labeled as either H, M, or L represented by the colours red, yellow, and blue, respectively (see Fig. 1E). The result of the algorithm is shown in Fig. 1F in which we represent our candidate solution as a field function that maintains the same grid structure and displays the selected LSOAs where sensors can be positioned. This grid is known as the results field function.

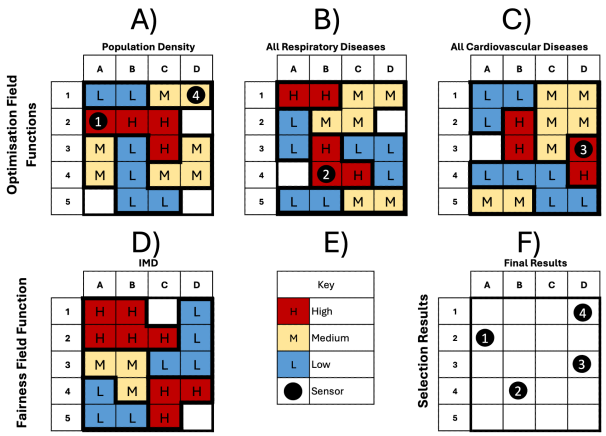

Fig. 1. Values for each grid location represented on the different field functions provided as input to the algorithm. A – C represent the target field functions population density, all respiratory diseases and all cardiovascular diseases. D represents the fairness field function IMD. E defines the colour scheme and F shows the results obtained after processing the algorithm with $k = 4$ sensors selected as input. For the purpose of clustering, we set the minimum number of contiguous locations, $\tau = 2$. The numbered black dots ordinally represents the selection made in each iteration of the algorithm.

Fig. 5A shows the number of selected location for each field function. These values were used to compute the metrics used in the analysis of the algorithm which produced an ISM value of 0.814 and an FSM score of 0.452. The ISM value indicates that most of the selected LSOAs are in areas of high and medium classes while the FSM value underscores the balanced spread of placements across the IMD field function. This indicates that diversity was achieved in the placement of the sensors across the fairness field function.

IV. Proposed Approach

In this section we will provide a detailed description of the operation of the algorithm utilising the dataset introduced in III. Problem Definition section A. An execution trace is provided for each stage that the algorithm performs and results achieved on the dataset.

A. Algorithm Description

1) Removing outlier LSOAs in Field Functions through clustering: As a method of removing outliers from the dataset, we cluster the LSOAs within a field function that share the same tercile category using a contiguity-based approach. The clusters that have less than the minimum number of LSOAs (τ) are considered outliers. In our implementation, $\tau = 2$ which implies that a cluster of only one LSOA is an outlier and is excluded from further consideration by the field function.

We now detail our clustering approach. Consider all the LSOAs of a particular tercile category (say 'H'). Now, we model each LSOA as a node in the graph. We add an edge between two nodes if the corresponding LSOAs are immediate neighbours in the underlying Geospatial framework. Fig. 2 illustrates this process for a tercile category 'H'. Following this, we compute the connected components in this graph. Each connected component in the graph forms a cluster. We

then record the number of LSOAs within each cluster. Note that each LSOA will only belong to one cluster using the approach.

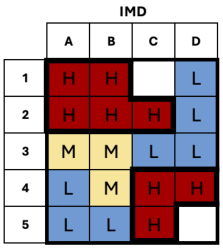

Fig. 2. Performing contiguity based clustering of high (H) locations on the fairness field function (e.g., IMD). White, unlabelled grid locations are classified as outliers.

Let Z_H, Z_M, and Z_L be the list of the clusters obtained (post outlier removal) for the terciles 'H', 'M', and 'L', respectively, in descending order of their size. We then construct a master list Z_μ containing Z_H, Z_M, and Z_L, strictly in that order. Thus, if we were to process the clusters in Z_μ, we would first exhaust those in Z_H, before moving to Z_M, and then Z_L.

2) Main Implementation: On each iteration of the algorithm, a different target field function is used in the process of selecting sensor locations. To select the first location, the algorithm will choose an LSOA at random from the largest, high-valued tercile cluster in the first target field function. This location is then added to **S** since there are no other locations that have been selected up to this point. After selecting the next target field function in the ordered set, the algorithm then proceeds to determine the distance between each location in **S** and that of each potential location found in the largest high-valued tercile cluster in the target field function. Potential locations that meet the minimum distance between themselves and the locations found in **S** are stored and used in the next stage of the algorithm. If, on the other hand, there are no qualifying points in the largest high-valued cluster, the algorithm will select the 2nd largest high-valued cluster and continue to iterate through the tercile clusters from the largest high-value to the smallest low-value cluster until a potential location is found that satisfies this requirement.

The next stage checks the tercile value of each of the potential locations in the fairness field function and computes the degree of impurity found between each of these locations and those found in S. The potential location that produces the highest level of impurity is then selected as the next location and is added to **S**.

Algorithm 1 Sensor Placement($FF, \lambda, k, \tau, d_{th}$)

Remove outliers for each of the target field functions using the continuity-based clustering of the terciles (using τ)
Create Z_μ for each of the target field functions
Solution set **S** = $\varnothing$
$order\_of\_layers \Leftarrow$ ordering amongst field functions in λ
// (E.g., population density > cardiovascular diseases > respiratory diseases
$current\_layer$ H = next layer in $order\_of\_layers$
$nLSOA \Leftarrow$ arbitrary LSOA from the first cluster in Z_μ list of field function H
Add $nLSOA$ to the solution set **S**
for $i = 2$ *to* k **do**
  $current\_layer$ H = next layer in $order\_of\_layers$
  $solution\_selected = FALSE$
  while $solution\_selected == FALSE$ **do**
    $\Omega \Leftarrow$ Next higher ranked cluster from Z_{μ_H} of layer $H \cup M \cup L$
    $\Omega' \Leftarrow$ Discard LSOAs from Ω which are below distance threshold (d_{th}) of previously selected LSOAs in **S**.
    if Ω' is not $NULL$ **then**
      X = $\alpha \sum_{x_i \in S} dist\,(x, x_i) + \beta \cdot Entropy\,(\mathbf{S} \cup x_i)$
      //The distance values $dist()$ are normalised
      $nLSOA \Leftarrow$ LSOA with the largest value in X
      Add $nLSOA$ to **S**
      $solution\_found = TRUE$
    end if
  end while
end for
return **S**

V. Experimental Analysis

A. Datasets

The data used for this analysis was collated from the Hull City Council (HCC) as well as the Office for National Statistics' Open Geography Portal and includes, the LSOA divisions for the city of Hull, population density, hospital admissions for all respiratory diseases and all cardiovascular diseases.

1) *LSOA data:* Obtained from the Office for National Statistics' Open Geography Portal, this dataset consists of the 168 LSOA geographical boundaries for Hull [11].
2) *English IMD:* Produced by the UK Government in 2019 based on the LSOA geographical boundaries from the 2011 Census and measure relative levels of deprivation across England. The scores have been estimated and adjusted for the 2021 LSOA boundaries by the HCC for this analysis, and the national ranks (out of 33,755) were calculated [12].
3) *Population Density:* Computed by the HCC, it contains the population estimates for each LSOA from the 2021 Census expressed as the number of residents per square kilometre [11].
4) *Emergency Hospital Admissions For All Cardiovascular Diseases and All Respiratory Diseases:* From the NHS Digital's Data Access Environment, the number of emergency hospital admissions over the ten-year period 2013/14 to 2022/23 where the primary diagnosis was cardiovascular disease (ICD I00-I99) or respiratory disease (ICD J00-J99) was used [10]. ONS's mid-year resident population estimates at LSOA level were used

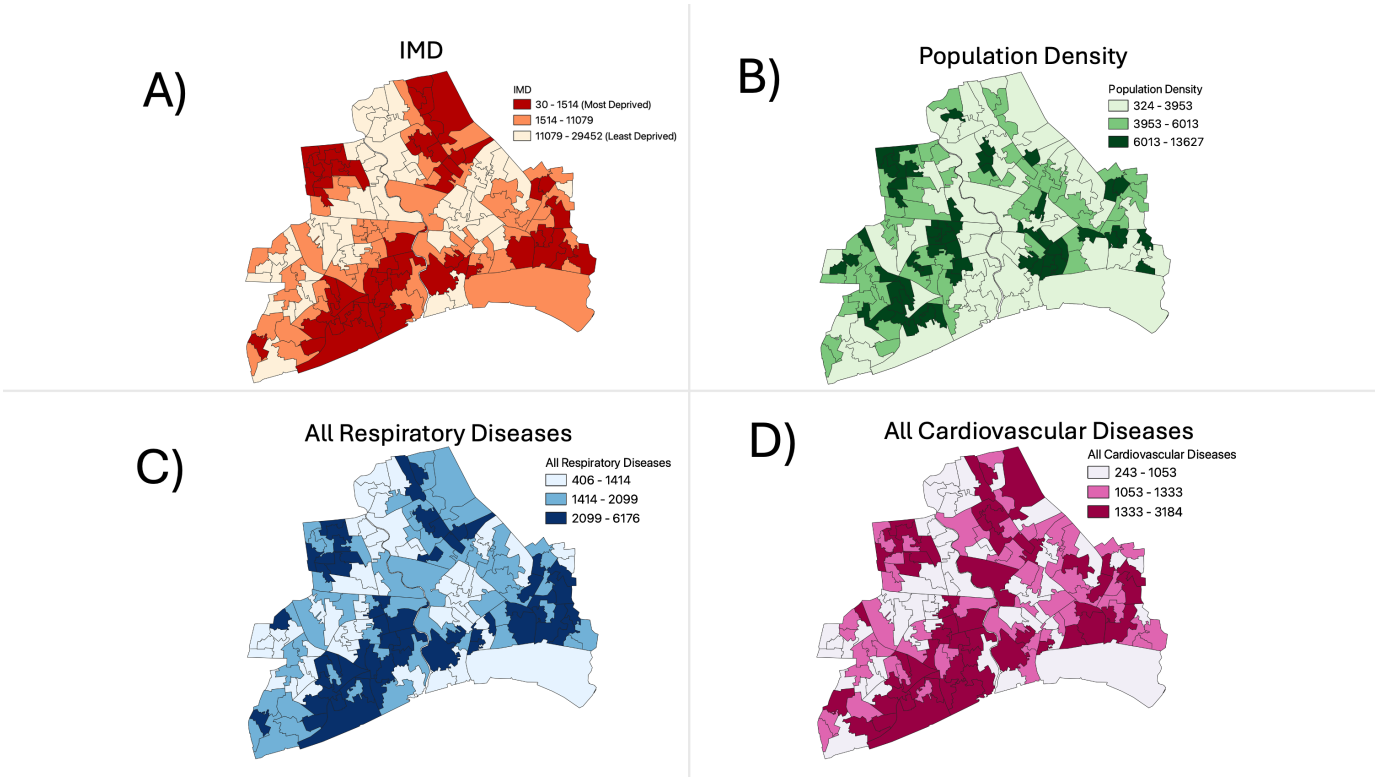

Fig. 3. Tercile representations for the Indices of Multiple Deprivations (A), Population Density (B), All Respiratory Diseases (C), and All Cardiovascular Diseases (D) geospatial field functions.

for 2013 and 2020 (prior to ONS back-revising their estimates for the 2021 Census as these were not available at the time of this analysis) with an adjustment based on the 2021 Census and mid-year 2022 population figures to estimate the population for 2021 and 2022. The directly age standardised emergency admission rate was calculated standardised to the European Standard Population 2013. The rate is expressed as the number of emergency admissions for cardiovascular disease and for respiratory disease per 100,000 residents over the ten-year period 2013/14 to 2022/23. It was necessary to use ten years of data as the number of admissions at LSOA level would have been too small to make the calculation of the directly standardised rate statistically valid otherwise.

In Fig. 3 we depict the terciles present in each of our input field functions. Segment A shows the IMD field function with dark red representing areas that are most deprived. Areas represented by cream are less deprived in nature. Population density is represented in Segment B with dark green reflecting the most densely populated areas and light green representing less populated areas. Segments C and D depict all respiratory and cardiovascular diseases respectively with the darkest colours representing areas with the highest rates of emergency hospital admission for the conditions, and the lightest colours showing the areas with the lowest emergency admission rates.

For the final set of inputs, we set the number of locations that will receive a sensor, $k = 40$, and the minimum number of locations that define a cluster, $\tau = 2$. Finally, the minimum distance where sensors can be placed together, $d_{th} = 2$ km. Fig. 4 illustrates the LSOAs that are selected for sensor placement across the city of Hull. The sensors are spread across the city and account for the hotspots found on each of the optimisation layers.

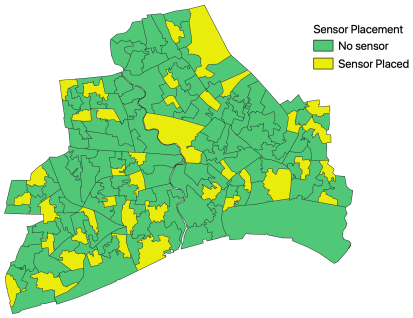

Fig. 4. Result of placing $k = 40$ sensors across the city of Hull.

For evaluation purposes, let $\omega_H = 3$, $\omega_M = 2$, and $\omega_L = 1$. We then compute the ISM and FSM values based on the aggregated data computed illustrated in 5. Of the 40 sensors that were placed by the algorithm, 23 were positioned in the most densely populated areas whereas 9 were placed in areas with the least residents. There were a total of 20 sensors placed in areas with high admission rates for respiratory diseases and

8 in areas where admission rates were relatively low. Similarly, the model allocated sensors for 19 areas with high admission rates for all cardiovascular diseases and 10 sensors in areas with low admission rates. Finally, 19 locations were selected for sensor placement in most deprived areas, 13 in areas with intermediate levels of deprivation, and 8 in areas of lower levels of deprivation.

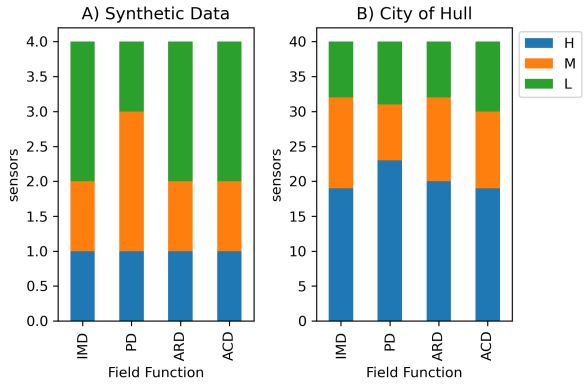

Fig. 5. Number of locations selected in each field function in (A) the synthetic dataset where $k = 4$ and (B) the city of Hull where $k = 40$.

Given the results shown in Fig. 5B, the calculated ISM and FSM values were 0.764 and 0.452, respectively. The ISM value indicates that the majority of sensors were placed in areas where the value of the field function had its densest concentrations. Finally, the FSM value of 0.452 suggests that there was a good balance of the locations selected on the fairness field function. 47.5% of locations selected in areas that are most deprived, 32.5% selected in the medium tercile tier of deprivation, and 20% placed in more deprived areas.

VI. Discussion

The algorithm is prone to selecting locations based firstly on their proximity and then uses the qualifying locations to compute the entropy of the candidate solution with the qualifying locations. If, for example, IMD and that particular field function were found to be strongly correlated, the resulting computation may not exhibit the anticipated level of diversity. Another limiting factor in our work centres around the use of terciles as the method of discretising the field function values. Using three classes may inadvertently eliminate some of the distinct spatial variation found across an area of interest. Future iterations of this work will explore the use of more refined divisions such as quintiles or deciles which may lead to sensor placements that more closely align with the aims of the algorithm as detailed in our objective function.

Unlike the target field functions which are capable of utilising 1 or more field functions, the fairness field function is limited to only using a single field function to determine fairness. If more field functions were required to be used to optimise fairness, the algorithm would need to be redesigned to accommodate this change.

VII. Conclusion

Through the analysis of the target and fairness field functions we demonstrated that sensors can be placed in a manner that maximises the diversity of the selection based on the various degrees of deprivation faced in a smart city. This was accomplished while simultaneously maximising the selection of locations that are in densely concentrated areas of the selected target field functions. We were also able to spread the sensors across the area being analysed in a way that maximises their coverage across that area. We proposed metrics (ISM and FSM) that can be used to evaluate the quality of the results that have been achieved by the algorithm and also add a layer of explainability to the locations that have been selected. This approach to placing air quality sensors may positively impact the public's perception since the methods proposed keep fairness as a core tenet of its operation.

Acknowledgment

We would like to thank the EPSRC National Edge AI Hub, Hull City Council and Bradford Metropolitan District Council for their invaluable contributions to this work. Their time, effort and funding has been instrumental to the success of this publication.

References

[1] N. Rusli and et al., "A review on worldwide urban observatory systems' data analytics themes: Lessons learned for Malaysia Urban Observatory (MUO)," *Journal of Urban Management*, vol. 12, no. 3, pp. 231–254, Sep. 2023.

[2] "Clean air zone framework," Oct. 2022. [Online]. Available: https://www.gov.uk/government/publications/air-quality-clean-air-zone-framework-for-england/clean-air-zone-framework

[3] C. Zhou and et al., "Optimal Planning of Air Quality-Monitoring Sites for Better Depiction of PM2.5 Pollution across China," *ACS Environ. Au*, vol. 2, no. 4, pp. 314–323, Jul. 2022, publisher: American Chemical Society.

[4] Z. Mano and et al., "Information Theory Solution Approach to the Air Pollution Sensor Location–Allocation Problem," *Sensors*, vol. 22, no. 10, p. 3808, May 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/10/3808

[5] A. Boubrima and et al., "A new WSN deployment approach for air pollution monitoring," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, Jan. 2017, pp. 455–460, iSSN: 2331-9860.

[6] R. Mukherjee and et al., "Real-time optimal spatiotemporal sensor placement for monitoring air pollutants," *Clean Techn Environ Policy*, vol. 22, no. 10, pp. 2091–2105, Dec. 2020.

[7] C. Robinson, R. S. Franklin, and J. Roberts, "Optimizing for Equity: Sensor Coverage, Networks, and the Responsive City," *Annals of the American Association of Geographers*, vol. 112, no. 8, pp. 2152–2173, Nov. 2022, publisher: Taylor & Francis.

[8] M. M. Kelp and et al., "Data-Driven Placement of PM2.5 Air Quality Sensors in the United States: An Approach to Target Urban Environmental Injustice," *Geohealth*, vol. 7, no. 9, p. e2023GH000834, Sep. 2023.

[9] C. Sun and et al., "Optimal Citizen-Centric Sensor Placement for Air Quality Monitoring: A Case Study of City of Cambridge, the United Kingdom," *IEEE Access*, vol. 7, pp. 47 390–47 400, 2019.

[10] "ICD-10 Version:2019." [Online]. Available: https://icd.who.int/browse10/2019/en

[11] "Census - Office for National Statistics." [Online]. Available: https://www.ons.gov.uk/census

[12] "English indices of deprivation," Sep. 2019. [Online]. Available: https://www.gov.uk/government/collections/english-indices-of-deprivation