Solving the challenges of interpolating NO_2 from SPRINT data and modelling population movements in agent-based modelling

Hyesop Shin[1,2], and Eric Silverman[1]

[1] MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, United Kingdom
[2] School of Environment, University of Auckland, New Zealand

Abstract

This study addresses two critical challenges in urban air quality exposure simulation and offers solutions. The first challenge is to generate nitrogen dioxide (NO_2) fields across the city of London from available stations that provide Spatially Poor but Rich In Time (SPRINT) data. We first used Inverse Distance Weighting (IDW) to spatially interpolate NO_2 at each half-a-day step. Each station had a list of hourly NO_2 values for each time step, from which one NO_2 value was selected by a stochastic process to generate the field. We also added weightings of up to a factor of 3 to London's NO_x emissions to account for emissions from sources other than vehicles. We cross-validated the modelled data with the station data and found beta parameter of 1.5 to be the most appropriate 'power' parameter. The second challenge investigated the use of a fractional origin-destination (OD) matrix to see how to overcome errors when assigning destinations to a small set of population. We tested that 'the nested bin strategy' worked well for our 6,078 London resident agents. To enrich the dynamics to represent people's non-work mobility patterns, we included visits to recreational areas during weekends and festive periods. This, in turn, provides a more comprehensive representation of urban mobility. Solutions to each challenge can provide more accurate assessments of pollution exposure, leading to better informed public health interventions.

Keywords

spatially poor but rich in time (SPRINT); agent-based modelling (ABM); population mobility; OD matrix; NO_2

Code & Data availability

All data and codes are stored in our dedicated GitHub repository: https://github.com/dataandcrowd/ABM-for-Data-Scarcity.

1. Motivation

1.1 Challenges in generating air quality fields

Urban air pollution has been directly linked to increased mortality and a wide range of serious health effects, ranging from minor eye irritation to serious conditions such as asthma, as well as pulmonary and cardiovascular disorders (Brook et al., 2004; L. Chen et al., 2007; IARC, 2013). An overview of pollution distribution reveals significant differences in both time and location, which are influenced by factors such as vehicular traffic, building density, and current meteorological conditions (Guarnieri & Balmes, 2014).

Correspondence:
Contact H. Shin at hyesop.shin@auckland.ac.nz

Cite this article as:
Shin, H., & Silverman, E.
Solving the challenges of interpolating NO_2 from SPRINT data and modelling population movements in agent-based modelling
Socio-Environmental Systems Modelling, vol. 6, 18752, 2024, doi:10.18174/sesmo.18752

This work is **licensed** under a Creative Commons Attribution-NonCommercial 4.0 International License.

Socio-Environmental Systems Modelling
An Open-Access Scholarly Journal
http://www.sesmo.org

There are several established methods for measuring exposure to air pollution over time. One common approach is to use spatial interpolation to generate a pollution field using station data to provide an estimate of ground-level air pollution (Deligiorgi & Philippopoulos, 2011; Li et al., 2012; Naughton et al., 2018). While spatial interpolation brings the advantage of producing maps with a mathematically sound computation, it contains certain drawbacks. One notable drawback is its tendency to overlook small scale variation in air quality, which often results in overly smoothed outcomes. For instance, in London, each borough contains one or two background stations and a few roadside stations. Relying on such a limited number of monitoring sites to generate a pollution field can lead to large misestimation of exposure, both temporally and spatially (Dias & Tchepel, 2018; Hwang & Lee, 2018).

Second, previous studies that used interpolated outcomes often provided a temporally aggregated measure (Min et al., 2020; Naughton et al., 2018). While such an aggregate measure provides an annual average and the likelihood of exposure and potential health risks, some studies have suggested that minority ethnic groups and low socioeconomic group are associated with high air quality (Hajat et al., 2015; Knobel et al., 2023; van den Brekel et al., 2024). At a daily scale, Nyhan et al. (2016) explored population exposure to $PM_{2.5}$ in New York City by combining spatial interpolated $PM_{2.5}$ with population movement using mobile phone data. Although this approach provided valuable insights, the temporal scale was restricted to only two weeks due to the constraints in the validation process. Finally, from a technical viewpoint, feeding in interpolated images at each time step in the simulation would substantially increase the model's memory usage, making it computationally inefficient (Shin, 2021).

To overcome these shortcomings, recent studies using agent-based models (ABM) have attempted to measure exposure to spatial and temporal air pollution data combined with movements (Novak et al., 2023; Shin & Bithell, 2019, 2023). Nevertheless, despite their innovative approaches and findings, there is a lack of in-depth examination of the fundamental methodology used to generate the pollution field.

1.2 Challenges when using OD matrices

Studies combining ABM and activity models have increased significantly (Axhausen et al., 2016; Guarnieri & Balmes, 2014; Lu et al., 2022). One of the most effective methods for estimating population movement involves the use of Origin-Destination (OD) matrices (Maiorov & Saprykin, 2020; Saprykin et al., 2021). For example, in the UK, the 2021 Census has released the counts of individuals moving between origins and destinations, at a census block, sub-district, and regional levels (UK Census, 2023). Their previous census, collected in 2011, "Place of Residence by Place of Work, Local Authority" dataset that offers a comprehensive OD matrix detailing the movements of the employed population over 16 years old during the week preceding the census (Office for National Statistics, 2011). The South Korean Transportation database also provides OD matrices based on vehicles types (e.g., cars, trucks, buses) or purposes (e.g., school, work, leisure) (Korea Transport DB, 2020).

With a simple matrix, an ABM can generate a synthetic population of agents and assign origins and destinations (Maiorov & Saprykin, 2020). To enhance the flexibility in population sampling for a lightweight meso-scale ABM, this matrix could be adjusted based on the fraction of the counts (Shin & Bithell, 2019). This way, the ABM can fine-tune the fraction of the population, applying a fractional OD matrix to guide individual agents to their destinations.

However, challenges arise when assigning destinations of each individual while keeping the sample size small enough to run it on typical software such as NetLogo. In particular, issues arise when the *sample size × percentage* is smaller than 1, leading to some agents not being assigned a destination. For instance, in a city like London, with over 6 million residents between the age of 16 and 64, a 0.1% sample size would be some 6,000 agents. Yet, if the sample size multiplied by the percentage results in a value less than one (e.g., 0.3), some agents might not be assigned destinations. Hence, finding an effective solution is crucial to ensure all agents are appropriately allocated.

A follow-up challenge occurs due to the stark differences between people's work routine and out-of-work patterns. Since OD matrices are designed to model work-related flows (Wheeler, 2005), they might not accurately represent movement during weekends or holiday periods. This discrepancy raises the need for an alternative approach that can better account for varying behavioural patterns during weekends and holidays.

1.3 Research questions

Our primary objective of this project is to assess population exposure to nitrogen dioxide (NO_2), taking into account the varying movements of individuals. However, in order to achieve this, this problem-solving paper addresses two key questions:

1) How can we generate long-term NO_2 on a city-wide scale based on daily measurements when existing datasets are Spatially Poor but Rich In Time (SPRINT)?

2) With a sampled population, how do we assign agents' destinations when the total number of agents in some districts is too small? Additionally, what strategies can be employed to simulate population movements that correspond to weekdays, weekends and festive periods?

The remainder of the paper is organised as follows: Section 2 describes the solution steps for generating NO_2 data across the city over time through statistical imputation and additional road weighting; Section 3 presents solutions for destination allocation for heterogeneous entities that can quantify exposure levels; Section 4 discusses the impact of both solutions; and Section 5 summarises the findings and outlines future directions.

2. Solution for generating an air pollution field with SPRINT data

To develop an ABM to generate air pollution (here NO_2) across the heterogeneous space, we must establish several assumptions (see Figure 1). First, ensuring the availability of data at each time step is crucial (represented as rows). Given that each grid selects one of the 11 work hours (07h-17h) and one of the 13 home hours (18h-06h) throughout the period, stations have to contain data at each time step. However, there may be missing data on a specific day, and the simulation will fail unless we manually correct it. To go a step forward, we use statistical imputation to fill the gaps and to avoid manually tuning the dataset. This section details the steps to 1) import NO_2 and assess the missing data for each station, 2) impute the missing data for each station, 3) generate roadside NO_2 using Inverse Distance Weighting (IDW), 4) add NO_x weightings to add emission sources other than traffic, and 5) cross-validate the model with station-collected data.

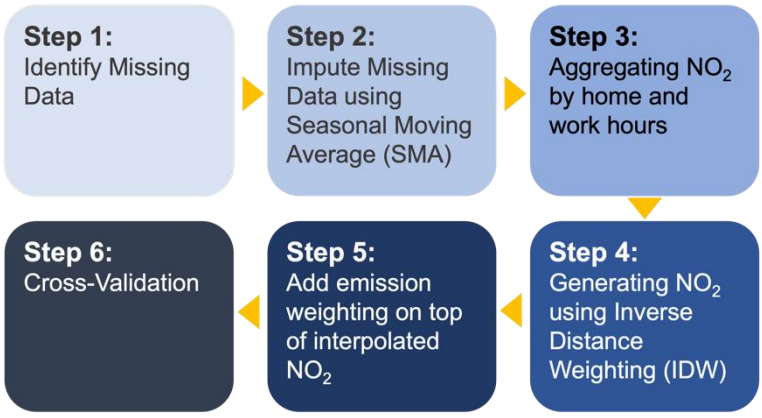

Figure 1: The procedure of generating SPRINT NO_2 for the London ABM model.

2.1 Step 1: Import NO_2 and identify missing data

The data are sourced from monitoring stations across the Greater London Area (GLA). In the R package *openair*, developers have integrated a socket connecting APIs for relevant areas. For instance, users can access data from the UK DEFRA's Automatic Urban and Rural Network (AURN). For London-specific data, we utilised King's College London's recordings, covering the GLA. We collected NO_2 data from four types of stations: Urban Background, Suburban, Roadside, and Kerbside.

2.2 Step 2: Impute missing NO_2 data

We recognise various approaches for managing missing data, such as data being Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR) (Jakobsen et al., 2017). In our case, we determined that the data omission was due to technical, rather than intentional, reasons. Therefore, our situation fits into either a seasonal MAR or MCAR category.

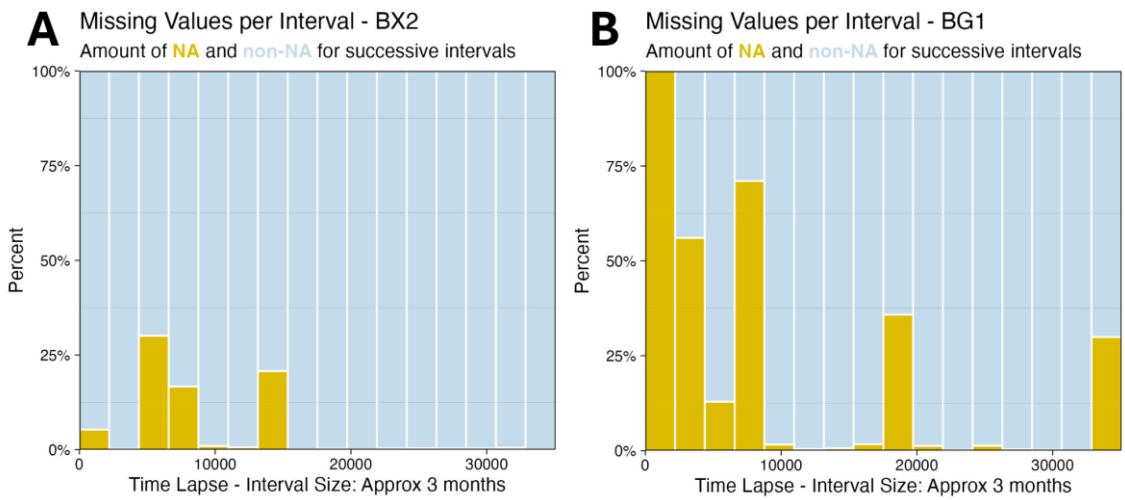

Figure 2: Illustration of the missing values throughout the study period across the stations in London. Among the stations analysed, Barking and Dagenham Station (B) reveals a significant lack of readings in the initial months, as well as sporadic gaps before the 20,000th and after the 30,000th time steps, when compared to Bexley Station that represents a case of good quality.

Focusing on statistical imputation for time series, especially those exhibiting seasonal trends as illustrated in Figure 2B, we opted for a Seasonal Moving Average (SMA) approach. This method imputes NO_2 levels, considering seasonal variations. Utilising the *imputeTS* package in R (Moritz & Bartz-Beielstein, 2017), we initially set four windows for imputation. However, if all data points within a current window were missing, we permitted an exponential extension of the window (see the commands in Figure 3).

```
na_seasplit(timeseriesdataframe,
algorithm="ma", # moving average
find_frequency=FALSE,
maxgap=Inf,
... )
```

Figure 3: Statistical imputation using the moving average method in the "ImputeTS" R Package.

After completing the data imputation process for all four station types, we used NO_2 readings from roadside stations. This was because, firstly, we found that there were not enough background NO_2 readings over London from 2019 to early 2020. This could lead to a large uncertainty in the predicted outcome. In addition, and more practically, NO_2 concentrations generally decrease with distance from roads (Lee et al., 2014; Richmond-Bryant et al., 2017), so the use of roadside NO_2 therefore provides a better representation of both on-road and off-road NO_2 concentrations.

2.3 Step 3: Aggregating NO_2 by home and work hours

After statistically imputing the NO_2 values for each station, we proceeded to extract a representative value for each time step. To match the time step for the origin-destination movement of the agents, we categorised the NO_2 readings into two periods: "home" and "work". The "home" values were gathered between 18:00 and 06:59 next day, assuming that most of the population is at home during those hours. Home hours account for 13 hours. The "work" values were collected from 07:00 to 17:59, a total of 11 hours. This approach is supported by the UK's Department for Transport (2024), which shows that the bus usage (by workers) peaks between 7am and 10am, then falls and stabilises until 4pm, then peaks again until 6pm (see Figure 4). As NO_2 is strongly related to traffic, we have aggregated the readings in this way to reflect typical population movement patterns.

Figure 4: Illustration of the time taken to board buses in the UK (copied from Department for Transport (2024)). Morning peaks around 7-10am and afternoon peaks around 4-6pm.

2.4 Step 4: Generating NO_2 using Inverse Distance Weighting (IDW)

To generate a half-day NO_2 pollution field for GLA from the given roadside stations, we implemented a spatial interpolation technique called Inverse Distance Weighted (IDW). IDW estimates the value of a cell as a weighted average of nearby sample points (Lovelace et al., 2019; Moraga, 2023; Nyhan et al., 2016). A key aspect of IDW is the power parameter, which significantly influences the interpolation results. That is, a higher power gives more weight to nearby points, often resulting in a bullseye effect. In contrast, a lower power results in a smoother output but with less emphasis on proximity (Moraga, 2023; Tripp Corbin, 2015).

The model assumes that agents located on a grid at any given time are directly affected by the air quality there, impacting their exposure. Instead of assigning a uniform pollution value to an entire area by averaging the values at each row, each grid randomly selects a value from the daily pollution field recorded at the nearest local background station (see Figure 5). For example, each grid would randomly choose one of the NO_2 readings recorded at each row. We chose random selection because this approach adds an element of probabilistic exposure to better reflect the fluctuating and uneven distribution of pollutants in space that agents may encounter on any given day (see Figure 6). Furthermore, when comparing the Root Mean Squared Error (RMSE) of randomly selected NO_2 values with the averaged NO_2 per row, we found only a negligible difference (see more in Section 2.6 Cross Validation).

Given that we cover the entire city of London as our study area, we determined that a resolution of 200 metres by 200 metres would be appropriate for generating NO_2 over the year 2019 on a half-day basis.

2.5 Step 5: Incorporating an emission weighting factor on top of the interpolated NO_2

Vehicles are responsible for over 65% of the NO_x at UK roadside locations that leads to NO_2 (DEFRA, 2024). However, it is important to note that the built environment also contributes a considerable amount of NO_x emissions. Buildings emit NO_x due to heating systems, while construction and demolition sites generate NO_x through the operation of industrial trucks and bulldozers. Further, aircrafts are another source of NO_x emissions. We applied a singular weighting factor to the interpolated NO_2 levels based on the London Atmospheric Emissions Inventory (LAEI) 2019 dataset and classified the measures by quintiles (see Figure 7). By adjusting the interpolated NO_2 levels with weightings, we aim to better represent NO_x emissions from sources other than vehicles. In the simulation (also see Table 1), we used weightings of up to a factor of 3, considering emissions from sources other than vehicles. For example, the fourth quintile (weight 1.24) comes from the west of the city where London Heathrow Airport is located. We propose that this method provides a simplified, yet effective way of incorporating traffic-related effects into our NO_2 pollution model.

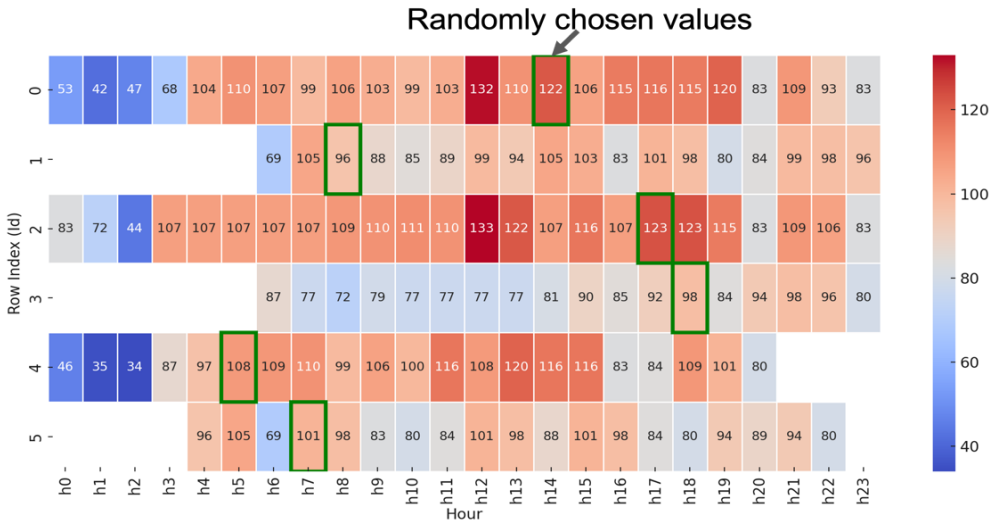

Figure 5: An illustration of NO_2 measurements at various stations across different times of the day (Home hours in even-numbered rows and Work hours in odd-numbered rows). For each time interval (row ID), a randomly selected value from the available measurements is highlighted, offering a simplified yet representative snapshot of the data.

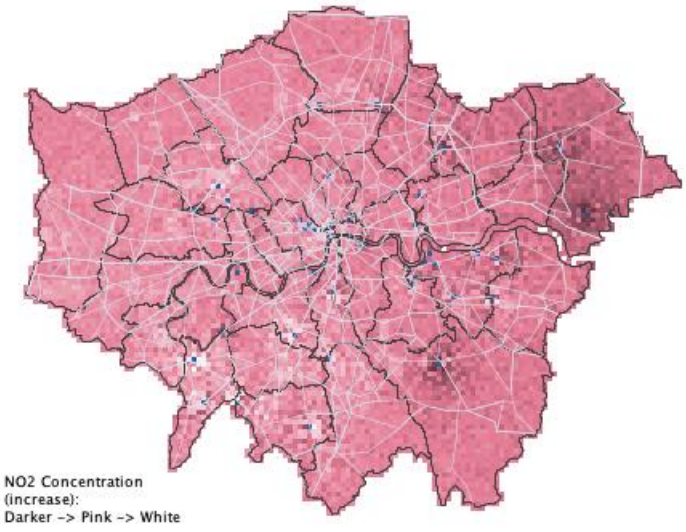

Figure 6: IDW output using a randomly picked value from a list of NO_2.

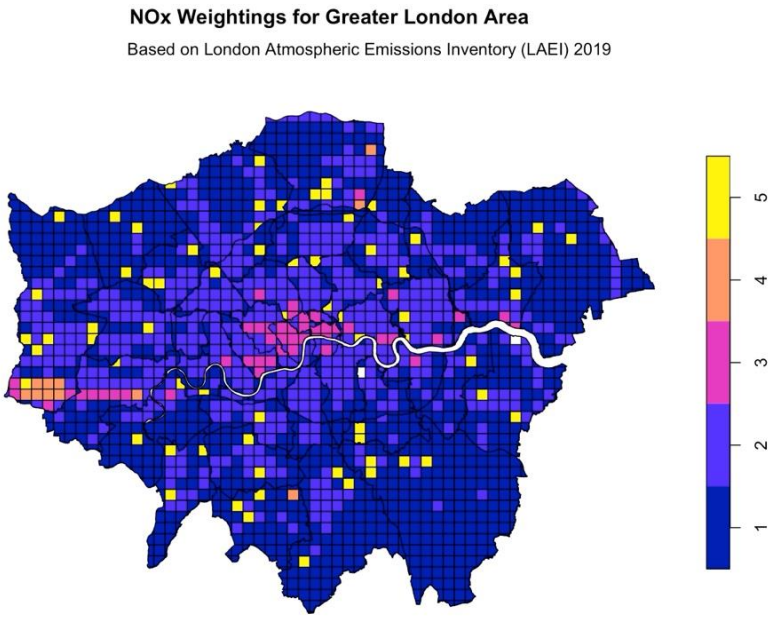

Figure 7: NO_x weightings by quintile classification.

Table 1: Weighting factors according to NO_x quintile provided by the LAEI 2019 measures.

NO_x quintile	Weighting
0 (background)	1.00
1	1.05
2	1.12
3	1.17
4	1.24
5	1.30

The caveat, indeed, is the coarse spatial resolution that may exaggerate possible emission sources. For example, the fifth quintile grids, which are distributed across the city, include junctions and commercial areas in the suburban boroughs, whereas the third quintile grids represent buildings in the city centre and areas near the airport.

2.6 Step 6: Cross-validation

Once the model building was complete, we employed a leave-one-out cross-validation (LOOCV) method to estimate the performance of the interpolated outcome. To check the performance of our proposed IDW method that stochastically selects a random NO_2 value per time step, we also ran averaged NO_2 values per time step (i.e. a traditional approach) and compared the two. We also compared the beta parameter (i.e. power over distance) to identify the minimum and maximum errors.

LOOCV was conducted for the year of 2019. We validated the modelled results against observed data from 17 stations with minimal missing data. Each combination was evaluated using the root mean square error (RMSE). To account for the effects of stochastic variation, our study ran ten iterations for each beta parameter.

The result showed that the mean RMSE was negligible across the beta parameters, but that the median RMSE values decreased as the beta parameter increased (see Table 2). We selected the beta parameter (*ß)* of 1.5 as it resulted in the smallest difference between the mean and median RMSE values. This choice was made because a smaller difference between the mean and the median indicates a more symmetric distribution of errors, thereby reducing the potential impact of extreme outliers.

Table 2: Comparison of RMSE across beta values for randomly chosen NO_2 and averaged NO_2 concentrations.

ß	Model	Mean RMSE	Median RMSE	Absolute difference
ß =1	Average	19.6	20.9	1.3
	Random	19.7	21.1	1.4
ß = 1.5	Average	19.5	18.9	0.6
	Random	19.7	19.1	0.6
ß = 2	Average	19.4	17.2	2.2
	Random	19.5	17.3	2.2
ß = 2.5	Average	19.1	15.7	3.4
	Random	19.3	15.8	3.5

3. Resolving the 'missing destination' issue in agents using existing origin-destination matrices

3.1 Step 1: Obtaining the Origin and Destination dataset

In this paper, we used the 2011 "Place of Residence by Place of Work, Local Authority" dataset that offers a comprehensive OD matrix detailing the movements of the employed population over 16 years old during the week preceding the census (Office for National Statistics, 2011).

While the UK's census is updated every 10 years and the 2021 Census data, including the OD counts, have been published, we chose to use the 2011 OD matrix. This decision was made because, during the survey period for the 2021 Census, a substantial number of people were working remotely from home, often with their designated offices located in different cities or even countries. This abrupt shift in working patterns, looks very different to what it is in 2024, where countries and companies have adopted flexible or hybrid working arrangements, or have returned to more traditional, office-based work. However, for the purposes of this study, the 2011 data provide a more reliable baseline for analysing pre-pandemic commuting patterns and mobility flows. It also matches well with the period of the NO_2 data we used in our analysis, ensuring consistency between datasets.

The OD matrix shows that most people tend to travel within their own boroughs, with the next most common destinations being areas beyond the city boundary, followed by Westminster and Camden (see Figure 8).

3.2 Step 2: Creating the OD matrix on pseudocode

Next, we transformed the OD matrix from each of the 32 boroughs into a fractional table. Note this approach enables the allocation of origins and destinations for individuals in any population sample.

In this study, we simulate population movement with 6,078 agents, which represents approximately a 0.1% sample of London's population between the age of 16 and 64. Taking Greenwich as an example, 200 out of 6,078 is from this borough (see the Supplementary Material for the full matrix). According to the OD matrix 22% of its residents stay within the borough, while only 0.2% travel to Ealing and Enfield. Given small percentages, no agents would be assigned to Ealing or Enfield. This would result in 36 agents being identified as residuals due to rounding errors, and therefore stuck in the model and unable to move.

3.3 Step 3: Demonstrating "nested bin strategy" to overcome the missing agents

To address this, we developed a solution that is straightforward in theory but requires more effort in coding. Our approach, detailed in Figure 9, involves creating a nested matrix. The outer matrix counts agents for each London Borough (denoted as *Num*) and creates corresponding bins (denoted as *totalUsed*). Each agent is assigned an origin name. Agents in a selected borough (*Num*) are then multiplied by the destination percentage in the matrix, and the *totalUsed* is summed up. The algorithm checks if all agents in the selected borough have been assigned to destinations (Lines 15-20). Despite its limitations, this technique reduces inaccuracies in the distribution of agents, a critical advantage when dealing with small populations. In the case of Greenwich, mentioned in the previous step, the 36 agents without specific destinations would be collectively assigned to an "Others" category.

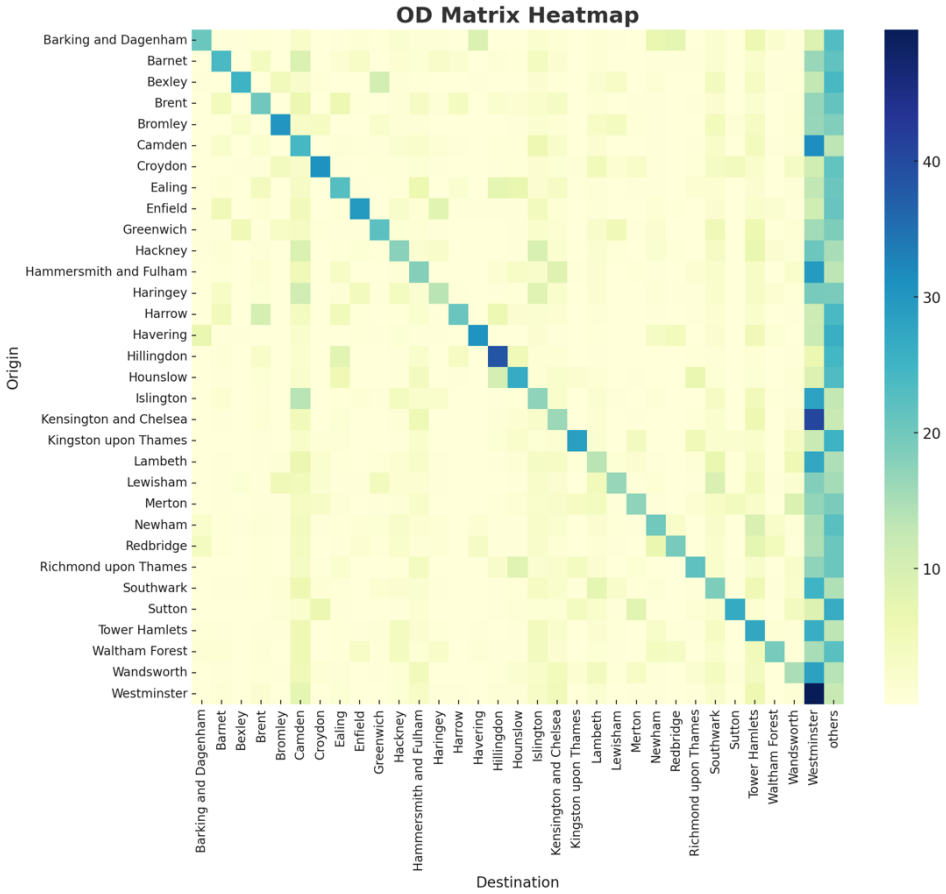

Figure 8: A fractional origin-destination matrix converted amongst London boroughs and out of London areas ("others").

Algorithm 1 Agent Destination Selection

```
for each originName in keys of Matrix iteration = 1, 2, ..., 32 do
    matrix-loop ← 0
    Num ← count people where homeName is originName and age 15-64
    totalUsed ← 0
    number ← 0
    for each percent in Matrix [originName] iteration = 1, 2, ..., 32 do
        newDestination ← destinationNames [matrix-loop]
        if newDestination ≠ "others" then
            number ← round(percent × Num)
            totalUsed ← totalUsed + number
        else
            number ← Num - totalUsed
        end if
        peopleRemaining ← people with homeName originName, destinationName="unidentified", age 15-64
        if count(peopleRemaining) > 0 AND count(peopleRemaining) ≤ number then
            number ← count(peopleRemaining)
        end if
        if number < 0 then
            number ← 0
        end if
        for each of number in peopleRemaining do
            Assign newDestination to destinationName
            Assign one of patches with name newDestination and is-built-area? true to destinationPatch
        end for
        matrix-loop ← matrix-loop + 1
    end for
end for
```

Figure 9: Algorithm of agents selecting destinations.

After assigning destinations to all agents during the setup phase, we were then able to execute the movement functions (see Figure 10). An example of the movement functions on a weekday are visualised in Figure 11.

Algorithm 2 People Movement Simulation

```
function Go
    Move-People
    Tick
    if ticks = 2921 then
        Stop
    end if
end function
function Move-People
    if ticks mod 2 = 0 then
        Move-Out
    else
        Come-Home
    end if
end function
function Move-Out
    Ask(people)
    if patch-here ≠ destinationPatch then
        Move-To(destinationPatch)
        Forward(1)
    end if
end function
function Come-Home
    ask(people)
    if patch-here ≠ homePatch then
        Move-To(homePatch)
        Forward(1)
    end if
end function
```

Figure 10: Pseudocode of agent movement

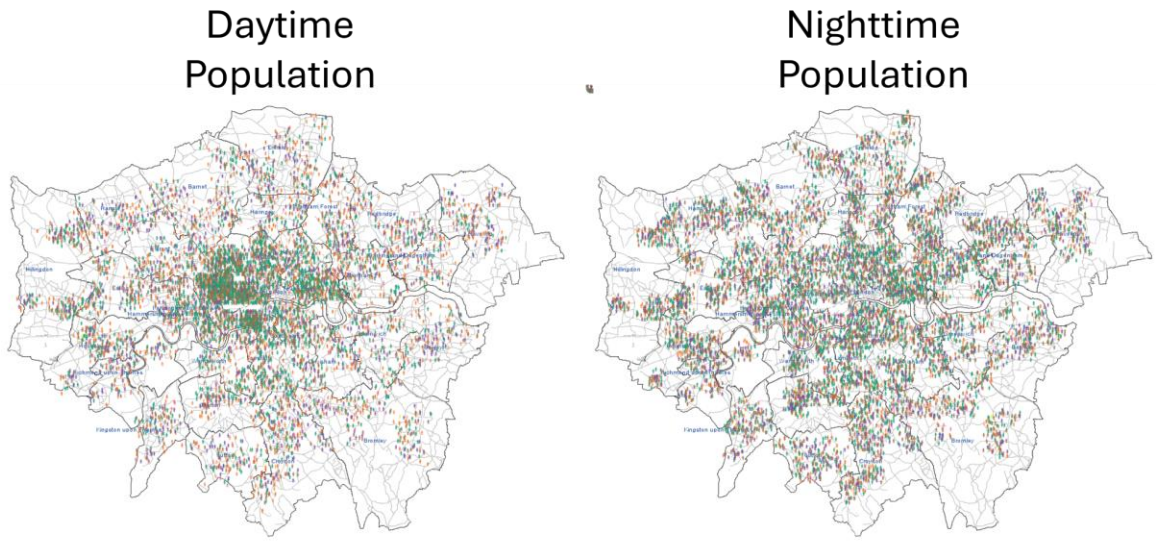

Figure 11: Visual outcome of population mobility of London residents during daytime and nighttime.

3.4 Step 4: Adding dynamic moves over weekends and festive periods

While there is no certain rule way of moving the entities for non-work purposes, we added some dynamics to the behavioural patterns that may happen on weekends and festive periods. Here festive periods include Easter and Christmas.

We used the following steps to implement these dynamic movements.

- Agent classification: agents were classified into different groups based on their likelihood of moving during weekends and festive periods. These groups included 'weekend shoppers' who were more likely to visit the central activity zone and others who would visit recreational areas or stay home.
- Movement Rules: rules were encoded using a probabilistic approach. For 'weekend shoppers,' a random selection algorithm determined 10% of the population residing outside central London would visit the central activity zone during the setup stage. For the remaining population, a stochastic model allocated 75% of agents to recreational areas and 25% to stay at home.
- Spatial Data Integration: spatial data from the GiGL Open Space dataset (https://www.gigl.org.uk/open-spaces/) and the Central Activities Zone (https://data.london.gov.uk/dataset/central_activities_zone) were integrated into the agent-based model to guide agent movements (see Figure 12). Recreational areas were randomly selected for each agent during each weekend or festive period, ensuring that agents did not visit the same place back-to-back.
- Temporal Dynamics: we ask the agents to recognise weekends, Easter break (16-26th April), and Christmas break (22nd-31st December). During these times, the movement rules described above were activated.
- Equal access: since we have assumed a two-point movement between origins and destinations, we also state that all agents have equal access to transport and that the probability of choosing a particular recreational area is the same for all agents.

With these detailed instructions for each individual, we can then allocate the non-work movements in parallel with the existing OD matrix.

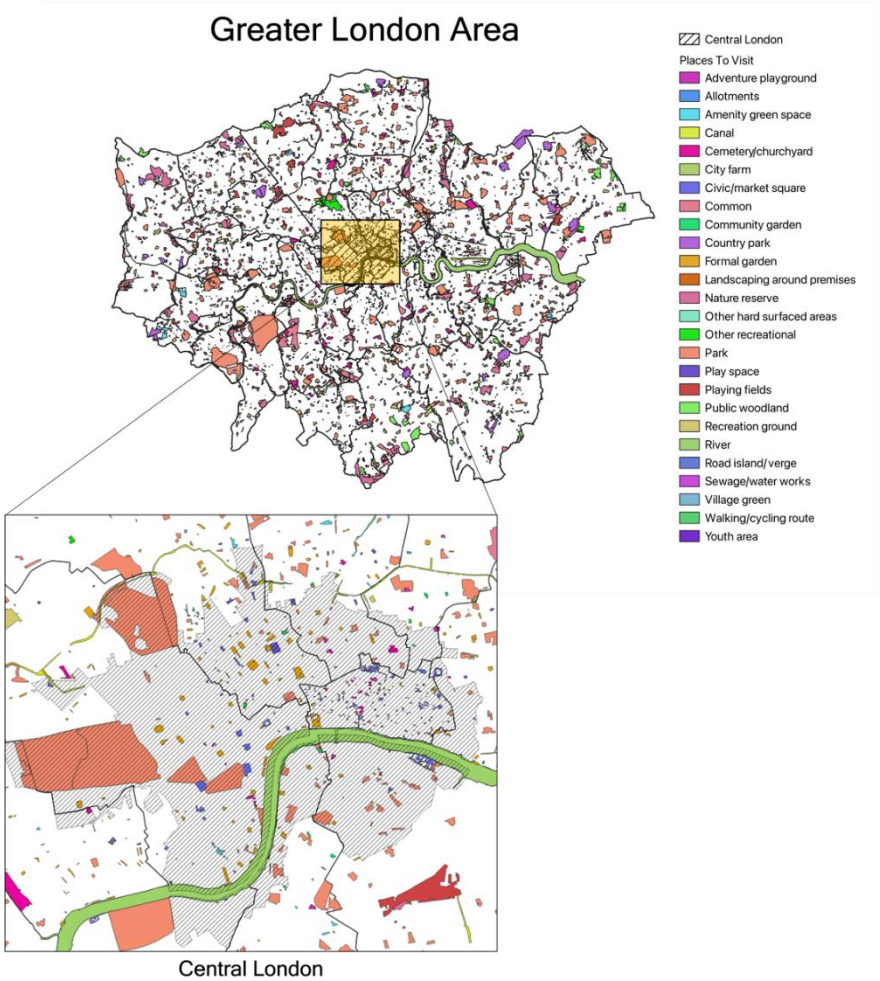

Figure 12: Recreation places to visit during the non-working days.

4. Impact

4.1 Spatially and temporally interpolating NO_2 using SPRINT data

This paper's initial section constructed an air pollution landscape using SPRINT (Spatially Poor but Rich In Time) data, derived from several monitoring locations within the Greater London Area. Through integrating statistical imputation to address gaps in NO_2 measurements and employing cellular automata techniques for spatial value generation, our results suggest that considering roadside NO_2 levels is a substantial advancement.

Unlike the typical IDW interpolation, which is a process that aggregates data and creates a smooth outcome according to spatial autocorrelation (Chen & Lin, 2022; Nyhan et al., 2016), our approach introduces a more dynamic methodology. Instead of averaging, we randomly chose an NO_2 value from the list at every time step. This way, it generates a more rigid outcome to reflect the variability of NO_2 that can occur on a given day.

As with other studies (Chen & Liu, 2012; Risk & James, 2022), IDW needs to go through a rigorous cross-validation to provide the predictions with the lowest error possible. We used a cross-validation and tuned the beta parameters to estimate NO_2 concentrations at half-day intervals. We found that a beta value of 1.5 had the smallest difference between the mean and the median, indicating a well-distributed dataset. By integrating cross-validation with NO_x emission factors, our method moves beyond the limitations of traditional IDW interpolation, which is often prone to local bullseye effects and over-smoothing. This advancement provides a more refined model by considering additional emission sources beyond traffic factors. However, the limited number of monitoring stations and the use of only one year's worth of NO_2 data may present challenges in reducing error margins. Furthermore, it is worth noting that access to NO_x emission data was facilitated by the London-specific context, which might not be as readily available in other locations.

Lastly, our methodology provides a comprehensive perspective on pollution trends and their impact on populations such as long-term population exposure to NO_2 at a city scale. This can be positioned as a viable alternative to both spatial interpolation and personal sensors to some extent. Nonetheless, the use of a mesoscale model (approximately 200m by 200m resolution) with relatively straightforward entity mobility and exposure functions may limit the variability in outcomes concerning personal exposure and health implications.

4.2 OD matrix solution

In this paper, we explored whether an Origin-Destination (OD) matrix of London boroughs can dynamically simulate population movement during the daily commute using a mesoscale agent-based model. Our findings indicate that even with a fractional OD matrix, effective population movement simulation is achievable using a limited data set. In contrast to previous studies primarily focused on vehicular OD matrices (Baek et al., 2010; Schwinger et al., 2022), our approach incorporates population flow using reliable real-world census data. Given the increasing availability of mobile phone data and GPS tracking records, our methodology can serve as an effective proxy, facilitating the generation of population flow simulations that can be calibrated against contemporary datasets (Nyhan et al., 2016).

A significant aspect of using a fractional OD matrix is its flexibility in applying various population samples, such as 1%, 3%, or 10%, depending on the study's requirements. In this study, we used the matrix to simulate city-wide population movement, which required extensive alignment and fractionation tasks. The scalability of this methodology is evident when compared to smaller-scale studies, such as those conducted in a $16km^2$ area (Shin & Bithell, 2023), where traffic movement was based on the OD matrix. Furthermore, the ability to scale the sample size up or down improves computational efficiency (Shin, 2021).

Importantly, employing the OD matrix for population movement assignment enhances reproducibility and consistency in agent-based modelling (ABM) studies (Shin, 2021). This standardised framework allows for the comparison and replication of findings across different scenarios and locations, thereby improving the replicability and generalisability of ABM research. Our approach has been successfully applied in studies involving direct movements between origin points and destination points (Shin & Bithell, 2019), as well as combining shortest-path algorithms (Shin & Bithell, 2023).

Nevertheless, we must acknowledge the inherent uncertainties associated with smaller-scale simulations. While adding an extra bin for allocating left-over agents from rounding errors can mitigate some errors, challenges persist in accurately representing boroughs with minimal samples (i.e. sampling errors). For example, a borough

represented by 0.4 agents in a particular sample might not be represented in that sample but would be represented by 4 agents in a sample ten times larger. This issue of sampling error becomes particularly pertinent when the sample size results in statistically significant differences, or when the presence or absence of even a single agent in a specific area significantly impacts study outcomes (Faber & Fonseca, 2014).

To enrich the OD matrix for work purposes that activate during weekdays, we included recreational activities on weekends and festive periods including Easter and Christmas. Many studies have employed geospatial tools to examine the relationship between the spatial distribution of recreational spaces and the accessibility to them (Both et al., 2022; Olsen et al., 2022; Price et al., 2023). In light of the growing popularity of the 15-minute city concept, these studies have increasingly focused on walkable distances. Additionally, the impact of COVID19 lockdowns has allowed consideration of access to greenspaces (i.e. vegetated land such as parks and playing fields), as they offer engagement with nature that can benefit both physical and mental health (Bustamante et al., 2022; Ha et al., 2022).

Given these factors, adding an intricate preference metric for each individual can be tempting. While adding a detailed model can be great to represent the preferred place at an individual level, it is also important to balance the focus of the study with maintaining the speed of the model (Badham et al., 2018). As a consequence, we made sure that at least three-fourths of the population, regardless of their physical status, visit their nearest recreation space to recover from their repetitive weekday activities. Moreover, we differentiated recreational venues by age groups. For example, individuals under the age of 30 are more likely to visit theme parks and youth places (e.g. basketball hoops), while city farms and community gardens are frequently visited by middle-aged and elderly people. This addition offers a more comprehensive representation of population movement patterns in London, capturing both daily commutes and leisure activities.

In practice, using OD matrices can be combined with a variety of topics, such as air quality and accessibility (Huang & Ma, 2022; Shin & Bithell, 2023; Sonnenschein et al., 2022). Traffic simulators such as SUMO and MATSim use matrices and routing algorithms to realistically model traffic flows and emission levels that requires activity chains and detailed scheduling (Axhausen et al., 2016; Gurram et al., 2019; Maiorov & Saprykin, 2020; Saprykin et al., 2021). However, traffic simulation, which might be for a full 24-hour period, is already computationally expensive. Therefore, it is crucial to carefully consider the size and scope of the model to balance computational efficiency with the desired accuracy.

5. Conclusion

In this study, we tackled two significant challenges. First, we implemented an Inverse Distance Weighting (IDW) interpolation method using the SPRINT (Spatially Poor but Rich In Time) data and successfully mapped the spatial distribution of NO_2 in the Greater London Area. Second, we addressed the misalignment of agents caused by the inherent limitations of fractional Origin-Destination (OD) matrices. By adopting a nested bin strategy, we efficiently allocated destinations to all agents, which efficiently allocated destinations and reduced allocation errors. Building on the OD matrix, we incorporated visits to recreational areas during weekends and festive periods.

As a next step, we will combine NO_2 simulation with population movement to estimate the exposure levels of London's population. The exposure measures will be estimated based on the exposure metrics tool (US EPA, 2024). Once that is measured, we plan to conduct population exposure across four distinct air quality regulation periods: the introduction of the Congestion Charge Scheme (2008), the implementation of the Low Emission Zone (2015), and the expansion of the Ultra Low Emission Zone (2023). By exploring the distinct periods, ABMs can provide insights into how behaviour and air quality regulations can affect population exposure over time.

Acknowledgements

This paper was originally developed from Dr. Hyesop Shin's Ph.D project "Assessing Health Vulnerability to Air Pollution in Seoul Using an Agent-Based Simulation". We would like to thank Dr Mike Bithell for his valuable guidance in the ideation and practical development of the project.

Supplementary Material

The Supplementary Material for this article can be found online at https://sesmo.org/article/view/18752/18244.

References

Axhausen, K., Horni, A., & Nagel, K. (2016). The multi-agent transport simulation MATSim. Ubiquity Press.

Badham, J., Chattoe-Brown, E., Gilbert, N., Chalabi, Z., Kee, F., & Hunter, R. F. (2018). Developing agent-based models of complex health behaviour. Health & Place, 54, 170–177. https://doi.org/10.1016/j.healthplace.2018.08.022

Baek, S., Lim, Y., Rhee, S., & Choi, K. (2010). Method for estimating population OD matrix based on probe vehicles. KSCE Journal of Civil Engineering, 14(2), 231–235. https://doi.org/10.1007/s12205-010-0231-4

Both, A., Gunn, L., Higgs, C., Davern, M., Jafari, A., Boulange, C., & Giles-Corti, B. (2022). Achieving 'Active' 30 Minute Cities: How Feasible Is It to Reach Work within 30 Minutes Using Active Transport Modes? ISPRS International Journal of Geo-Information, 11(1), 58. https://doi.org/10.3390/ijgi11010058

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., & Smith, S. C. (2004). Air pollution and cardiovascular disease A statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association. Circulation, 109(21), 2655–2671.

Bustamante, G., Guzman, V., Kobayashi, L. C., & Finlay, J. (2022). Mental health and well-being in times of COVID-19: A mixed-methods study of the role of neighborhood parks, outdoor spaces, and nature among US older adults. Health & Place, 76, 102813. https://doi.org/10.1016/j.healthplace.2022.102813

Chen, F.-W., & Liu, C.-W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy and Water Environment, 10(3), 209–222. https://doi.org/10.1007/s10333-012-0319-1

Chen, L., Mengersen, K., & Tong, S. (2007). Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia. Science of the Total Environment, 373(1), 57–67.

Chen, P.-C., & Lin, Y.-T. (2022). Exposure assessment of PM2.5 using smart spatial interpolation on regulatory air quality stations with clustering of densely-deployed microsensors. Environmental Pollution, 292, 118401. https://doi.org/10.1016/j.envpol.2021.118401

DEFRA. (2024). Emissions of air pollutants in the UK – Nitrogen oxides (NOx) [Statistics]. Department for Environment Food & Rural Affairs. https://www.gov.uk/government/statistics/emissions-of-air-pollutants/emissions-of-air-pollutants-in-the-uk-nitrogen-oxides-nox

Deligiorgi, D., & Philippopoulos, K. (2011). Spatial Interpolation Methodologies in Urban Air Pollution Modeling: Application for the Greater Area of Metropolitan Athens, Greece. In F. Nejadkoorki (Ed.), Advanced Air Pollution. InTech. https://doi.org/10.5772/17734

Department for Transport. (2024). Annual bus statistics: Year ending March 2023 (revised) (Accredited Official Statistics). https://www.gov.uk/government/statistics/annual-bus-statistics-year-ending-march-2023/annual-bus-statistics-year-ending-march-2023

Dias, D., & Tchepel, O. (2018). Spatial and Temporal Dynamics in Air Pollution Exposure Assessment. International Journal of Environmental Research and Public Health, 15(3), 558. https://doi.org/10.3390/ijerph15030558

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. Dental Press Journal of Orthodontics, 19(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

Guarnieri, M., & Balmes, J. R. (2014). Outdoor air pollution and asthma. Lancet, 383(9928), 1581–1592. https://doi.org/10.1016/S0140-6736(14)60617-6

Gurram, S., Stuart, A. L., & Pinjari, A. R. (2019). Agent-based modeling to estimate exposures to urban air pollution from transportation: Exposure disparities and impacts of high-resolution data. Computers, Environment and Urban Systems, 75, 22–34. https://doi.org/10.1016/j.compenvurbsys.2019.01.002

Ha, J., Kim, H. J., & With, K. A. (2022). Urban green space alone is not enough: A landscape analysis linking the spatial distribution of urban green space to mental health in the city of Chicago. Landscape and Urban Planning, 218, 104309. https://doi.org/10.1016/j.landurbplan.2021.104309

Hajat, A., Hsia, C., & O'Neill, M. S. (2015). Socioeconomic Disparities and Air Pollution Exposure: A Global Review. Current Environmental Health Reports, 2(4), 440–450. https://doi.org/10.1007/s40572-015-0069-5

Huang, H.-F., & Ma, H.-W. (2022). Redesigning a cap-and-trade program for air emissions by agent-based modeling. Sustainable Environment Research, 32(1), 47. https://doi.org/10.1186/s42834-022-00157-4

Hwang, Y., & Lee, K. (2018). Contribution of microenvironments to personal exposures to PM10 and PM2.5 in summer and winter. Atmospheric Environment, 175, 192–198. https://doi.org/10.1016/j.atmosenv.2017.12.009

IARC. (2013). IARC: Outdoor air pollution a leading environmental cause of cancer deaths. International Agency for Research on Cancer.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology, 17(1), 162. https://doi.org/10.1186/s12874-017-0442-1

Knobel, P., Hwang, I., Castro, E., Sheffield, P., Holaday, L., Shi, L., Amini, H., Schwartz, J., & Yitshak Sade, M. (2023). Socioeconomic and racial disparities in source-apportioned PM2.5 levels across urban areas in the contiguous US, 2010. Atmospheric Environment, 303, 119753. https://doi.org/10.1016/j.atmosenv.2023.119753

Korea Transport DB. (2020). Travel Demand Analysis. https://www.ktdb.go.kr/eng/contents.do?key=249

Lee, J.-H., Wu, C.-F., Hoek, G., De Hoogh, K., Beelen, R., Brunekreef, B., & Chan, C.-C. (2014). Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. Science of The Total Environment, 472, 1163–1171. https://doi.org/10.1016/j.scitotenv.2013.11.064

Li, L., Wu, J., Wilhelm, M., & Ritz, B. (2012). Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California. Atmospheric Environment, 55, 220–228. https://doi.org/10.1016/j.atmosenv.2012.03.035

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. Chapman and Hall/CRC.

Lu, M., Schmitz, O., De Hoogh, K., Hoek, G., Li, Q., & Karssenberg, D. (2022). Integrating statistical and agent-based modelling for activity-based ambient air pollution exposure assessment. Environmental Modelling & Software, 158, 105555. https://doi.org/10.1016/j.envsoft.2022.105555

Maiorov, E., & Saprykin, O. (2020). Intelligent Analysis of City Residents Mobility Data for Transport Simulation. 316–321.

Min, K.-D., Yi, S.-J., Kim, H.-C., Leem, J.-H., Kwon, H.-J., Hong, S., Kim, K. S., & Kim, S.-Y. (2020). Association between exposure to traffic-related air pollution and pediatric allergic diseases based on modeled air pollution concentrations and traffic measures in Seoul, Korea: A comparative analysis. Environmental Health, 19(1), 6. https://doi.org/10.1186/s12940-020-0563-6

Moraga, P. (2023). Spatial statistics for data science: Theory and practice with R. CRC Press.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. R J., 9(1), 207.

Naughton, O., Donnelly, A., Nolan, P., Pilla, F., Misstear, B. D., & Broderick, B. (2018). A land use regression model for explaining spatial variation in air pollution levels using a wind sector based approach. Science of The Total Environment, 630, 1324–1334. https://doi.org/10.1016/j.scitotenv.2018.02.317

Novak, R., Robinson, J. A., Kanduč, T., Sarigiannis, D., & Kocman, D. (2023). Simulating the impact of particulate matter exposure on health-related behaviour: A comparative study of stochastic modelling and personal monitoring data. Health & Place, 83, 103111. https://doi.org/10.1016/j.healthplace.2023.103111

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S. R. H., & Ratti, C. (2016). "Exposure Track"—The Impact of Mobile-Device-Based Mobility Patterns on Quantifying Population Exposure to Air Pollution. Environmental Science & Technology, 50(17), 9671–9681. https://doi.org/10.1021/acs.est.6b02385

Office for National Statistics. (2011). Place of Residence by Place of Work, Local Authority [Dataset]. https://data.london.gov.uk/dataset/place-residence-place-work-local-authority

Olsen, J. R., Thornton, L., Tregonning, G., & Mitchell, R. (2022). Nationwide equity assessment of the 20-min neighbourhood in the scottish context: A socio-spatial proximity analysis of residential locations. Social Science & Medicine, 315, 115502. https://doi.org/10.1016/j.socscimed.2022.115502

Price, A., Langford, M., & Higgs, G. (2023). Quantifying disparities in access to recreational opportunities by alternative modes of transport. Case Studies on Transport Policy, 11, 100949. https://doi.org/10.1016/j.cstp.2023.100949

Richmond-Bryant, J., Chris Owen, R., Graham, S., Snyder, M., McDow, S., Oakes, M., & Kimbrough, S. (2017). Estimation of on-road NO2 concentrations, NO2/NOX ratios, and related roadway gradients from near-road monitoring data. Air Quality, Atmosphere & Health, 10(5), 611–625. https://doi.org/10.1007/s11869-016-0455-7

Risk, C., & James, P. M. A. (2022). Optimal Cross-Validation Strategies for Selection of Spatial Interpolation Models for the Canadian Forest Fire Weather Index System. Earth and Space Science, 9(2), e2021EA002019. https://doi.org/10.1029/2021EA002019

Saprykin, O., Maiorov, E., & Darbinan, M. (2021). Estimating Origin-Destination Matrix with Anonymized Movement Data. 1–4.

Schwinger, F., Forster, L., & Jarke, M. (2022). Population Synthesis by Disaggregating OD Matrices with Time-Progressive Graphs for Agent-based Simulations. Procedia Computer Science, 201, 560–567. https://doi.org/10.1016/j.procs.2022.03.072

Shin, H. (2021). Benefits of open research in social simulation: An early-career researcher's perspective. Review of Artificial Societies and Social Simulation, 23.

Shin, H., & Bithell, M. (2019). An Agent-Based Assessment of Health Vulnerability to Long-Term Particulate Exposure in Seoul Districts. Journal of Artificial Societies and Social Simulation, 22(1), 12. https://doi.org/10.18564/jasss.3940

Shin, H., & Bithell, M. (2023). TRAPSim: An agent-based model to estimate personal exposure to non-exhaust road emissions in central Seoul. Computers, Environment and Urban Systems, 99, 101894. https://doi.org/10.1016/j.compenvurbsys.2022.101894

Sonnenschein, T., Scheider, S., De Wit, G. A., Tonne, C. C., & Vermeulen, R. (2022). Agent-based modeling of urban exposome interventions: Prospects, model architectures, and methodological challenges. Exposome, 2(1), osac009. https://doi.org/10.1093/exposome/osac009

Tripp Corbin, G. (2015). Learning ArcGIS Pro. Packt Publishing Ltd.

UK Census. (2023). Origin-destination (flow) data based on the 2021 UK Census. https://www.ons.gov.uk/census/aboutcensus/censusproducts/origindestinationflowdata

van den Brekel, L., Lenters, V., Mackenbach, J. D., Hoek, G., Wagtendonk, A., Lakerveld, J., Grobbee, D. E., & Vaartjes, I. (2024). Ethnic and socioeconomic inequalities in air pollution exposure: A cross-sectional analysis of nationwide individual-level data from the Netherlands. The Lancet Planetary Health, 8(1), e18–e29. https://doi.org/10.1016/S2542-5196(23)00258-9

Wheeler, J. O. (2005). Geography. In Encyclopedia of Social Measurement (pp. 115–123). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00277-2