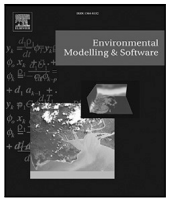
Environmental Modelling and Software 158 (2022) 105555

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



Position Paper

Integrating statistical and agent-based modelling for activity-based ambient air pollution exposure assessment

Meng Lu [a,b,*], Oliver Schmitz [b], Kees de Hoogh [c,d], Gerard Hoek [e], Qirui Li [f], Derek Karssenberg [b]

[a] Chair of Geoinformatics - Spatial Big Data, Faculty of Biology, Chemistry & Earth Sciences, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany
[b] Department of Physical Geography, Faculty of Geoscience, Utrecht University, Princetonlaan 8a, 3584 CB, Utrecht, The Netherlands
[c] Swiss Tropical and Public Health Institute, Basel, Switzerland
[d] University of Basel, Basel, Switzerland
[e] Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands
[f] Africa Multiple Cluster of Excellence, University of Bayreuth, 95440 Bayreuth, Germany

ARTICLE INFO

Keywords:
Agent-based modelling
Statistical modelling
Activity-based
Sampling
Exposure
Air pollution

ABSTRACT

Assessment of long-term human exposure to spatiotemporally highly variable air pollution requires accounting for human space–time activity. Individual exposure and space–time track data are not available over large populations and for long periods and a modelling approach is required. However, activity-based exposure models face here challenges in setting up the model and overly-large computations. Aiming for long-term and large-population simulations, we propose an activity model which integrates statistical and agent-based modelling by treating mobility-related variables as random variables. Probability distributions for these variables are estimated or derived from mobility datasets containing observed activities. On top of the activity model, we implemented an exposure model. A case-study of exposure assessment was developed using hourly air pollution maps. The activity model can potentially integrate any mobility data and is thus applicable when limited time activity data is available at the individual level.

1. Introduction

Estimating the effects of ambient air pollution on health (Luo et al., 2016; Chiusolo et al., 2011) requires assessing the air pollution exposure of the population studied. This is a challenge, particularly for air pollutants with considerable spatiotemporal variation at street level, such as NO_2, as where people are and their activities could greatly determine their exposures. For this reason, spatiotemporal mobility of individuals is important in exposure assessment. Several studies have compared the differences between exposures assessed neglecting space–time activities, i.e. using pollutant concentration values at the home location as a proxy, and exposures assessed accounting for human space–time activities (Duan and Mage, 1997; Lu et al., 2019; Park and Kwan, 2017; Mölter et al., 2012; Zenk et al., 2011). However, if we compare these studies, we could observe that different activity modelling methods lead to inconsistent conclusions on the effects of accounting for space–time activity in exposure assessment. Park and Kwan (2017) and Lu et al. (2019) show lower variations in NO_2 when mobility is modelled while Mölter et al. (2012) show the opposite. The inconsistency is mainly caused by using different methods, calling for further studies in activity-based exposure assessment. Also, air pollution maps generated from different air pollution models could contribute to the differences in activity-based exposure assessment, as shown in Yoo et al. (2015, 2021), who investigate the combined effects of spatiotemporal mapping and space–time activities on exposure assessment using respectively simulated and measured activity data.

Activity modelling is needed for exposure assessment as measured activity data on individuals is mostly unavailable in large-scale epidemiological studies. In transportation studies, progress has been made in the development of "activity models" for simulating transportation patterns. For example, ALBATROSS (Arentze et al., 2000) is a transportation-oriented system that simulates activities for the entire population based on activity diary data and dynamic constraints on scheduling decisions. MATSim (W Axhausen et al., 2016) focuses on large-scale, one-day individual activity simulation based on an activity schedule scoring algorithm and detailed road networks. Activity models such as MATSim that target simulating individual behaviours and

* Corresponding author at: Chair of Geoinformatics - Spatial Big Data, Faculty of Biology, Chemistry & Earth Sciences, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany.
E-mail address: meng.lu@uni-bayreuth.de (M. Lu).

https://doi.org/10.1016/j.envsoft.2022.105555
Received 16 July 2022; Received in revised form 3 October 2022; Accepted 10 October 2022
Available online 21 October 2022

the interactions between individuals and the environment fall within the concept of Agent-Based Modelling (ABM, Crooks and Heppenstall, 2012). As ABM takes a bottom-up approach to understand the emerging or aggregated behaviours, it allows the integration of individual and population behaviours. This property of ABM makes it an important tool not only for transportation studies but also in studies dedicated to human mobility simulation (Čertickỳ et al., 2015; Rosés et al., 2018). More examples include Wu et al. (2019), who attempt to integrate mobile data in ABM for activity simulation and Lu et al. (2019), who focus on simulating the destination locations and comparing NO_2 exposures of homemakers and bicycle commuters. Activity models contribute to the understanding of human activity patterns (Miller and Roorda, 2003). Many of the activity models are open-source and highly customisable, which allows for scenario studies. The activity models consider a comprehensive set of mobility-related and socioeconomic variables such as travel modes, work, education, leisure activities, and traffic (W Axhausen et al., 2016). They are commonly parameterised by mobility microcensus data or diary surveys consisting of locations visited and possibly externally estimated schedules. Then, a continuous-time mobility track for each individual is estimated based on general rules of human mobility patterns and space–time accessibility (Nguyen et al., 2011; Gonzalez et al., 2008; Yang et al., 2010; Yu, 2006; Alessandretti et al., 2017; Miller, 1991). Each route simulated can be contingent on, for instance, distance, safety, city infrastructure, and land use (Law et al., 2014). For example, Shekarrizfard et al. (2017) assign the predictions of a travel demand model to a road network to predict individual hourly trajectories. For each person, the model selects a path from all possible paths by comparing the assigned travel time and the survey travel time.

In recent years, data revealing space–time activity patterns and land use types are becoming increasingly available. National mobility microcensus data (W Axhausen et al., 2016), big social media data (Terroso-Saenz et al., 2022), cellular tower data (Müller et al., 2021), as well as tracking campaign data (Yoo et al., 2021) and the derived information, are becoming more prevalent. The question is how could they contribute to large-scale activity simulations. Statistical modelling has been intensively studied and applied to take opportunities brought by big data. The application of statistical modelling also surged in mobility studies, for descriptive, predictive, and prescriptive analytics (Torre-Bastida et al., 2018). A prominent example is the analysis or simulation of mobility patterns using geo-coded Twitter data (Hawelka et al., 2014; Huang et al., 2020; Jurdak et al., 2015). Statistical modelling has also been integrated into other mobility simulation procedures, for example, Kang et al. (2020) design a series of sample analysis to derive statistical features and combine them with the urban context for trajectory generation.

Integrating activity-based models with predicted air pollution surfaces allows for considering human space–time activities in exposure assessment. This has been reflected in several exposure studies (Shekarrizfard et al., 2017; Deffner et al., 2016; Gulliver and Briggs, 2005; Dons et al., 2011). Beckx et al. (2009) propose the use of the ALBATROSS model to simulate hourly activities and then combine the resulting space–time location of individuals with air pollution dispersion model estimates to assess exposure. However, these activity-based exposure assessment models are not designed to assess long-term air pollution exposures and associated uncertainties for large-scale epidemiological studies. Specifically, *long-term simulation* means that the model is capable of quantifying exposure averaged over a year or multiple years. *Uncertainty* needs to be quantified for each simulated activity, including e.g. time schedules, travel modes, and possible destination locations. A study has focused on long-term exposure assessment of a large population and considering uncertainty in the exposure assessment (Lu et al., 2019), but the activity schedules generated are less realistic activity schedules compared to the exposure models integrating a sophisticated activity model (W Axhausen et al., 2016). The need for quantifying long-term exposures and the uncertainty call for the development of activity simulation models that are dedicated to exposure assessment.

This study aims to incorporate statistical modelling of mobility data in a space–time activity simulation model to enable long-term exposure assessment for large populations, and to illustrate the approach with a case-study assessing exposure in the area of Utrecht, the Netherlands, using the Dutch national mobility microcensus dataset. Our model assumes uncertain mobility-related variables, such as the travel mode and maximum travel distance, as random variables. Statistical modelling is applied to derive, estimate, or empirically specify probability distributions of these variables from mobility microcensus datasets or literature, for each activity and according to the attribute of a person such as age and occupation. This allows a more accurate simulation of activities, as people's activity patterns may share similarities according to their socioeconomic statuses. These mobility-related variables are then used as inputs for the simulation of space–time activities. The Monte Carlo approach is applied to sample from each mobility-related variable for simulating the activities. Lastly, the exposure is calculated by aggregating the air pollution over the activity tracks and over time.

The rest of this paper is organised as follows. We first describe our activity simulation model following the ODD (Overview, Design concepts, and Details) protocol (Railsback and Grimm, 2019, page 37). Then, we show how the activity simulation model is used for exposure assessment and demonstrate the modelling process in a case-study.

2. Design concepts and model structure of the activity simulation model

2.1. Overview

Purpose and scales:

The activity simulation model is developed for large population-scale (e.g. the entire population of the Netherlands), long-term, personal air quality exposure assessment accounting for human space–time activities. The model focuses on having land transportation (e.g. cars, public transport, bikes, on foot) as major commuting means. The activity simulation model is developed with two key features: (1) it can simulate long-term travel behaviours, and (2) the uncertainty of human space–time activity is explicitly quantified. The activity simulation is based on the statistical modelling of mobility-related variables and attributes of a population. The model facilitates integrating mobility-related information from different sources.

Entities, state variables, and an overview of the process:

The activity model treats individuals living at each residential location as entities. The attributes of entities that are related to the activity patterns, such as age, gender, education, occupation, income, working status (e.g. full- or part-time workers), having children or not, having a car or not, and home locations, are considered as static state variables. Mobility-related variables are modelled as random or fixed. The randomness is assigned to variables whose values we are uncertain about, for example travel mode, start time, and duration. They are characterised by a probability distribution. The simulation process includes two steps (Fig. 1), the first is to derive or estimate probability distributions of the random variables from mobility data based on static state variables. In the next step, samples are drawn from these probability distributions to decide sequentially the maximum travel range, the origin and destination locations, and the travel mode. Based on the origin and destination locations and the travel mode, a route is queried from the road network of OpenStreetMap (Boeing, 2017) for the generation of the activity schedules and space–time tracks. The routes may be chosen by different criteria, e.g. based on the shortest distance or the shortest arrival time (for auto-vehicles). The second step is repeated several times (each repetition is called an iteration) for activity prediction and uncertainty quantification, i.e., each time, a different activity schedule and spatial locations are generated. Note the ensembling effect of averaging assessed exposures over several iterations: the precision of the exposure is expected to increase as the variance reduces.

Table 1
An example of a simulated activity schedule. "h2w" means "home to work", and "w2h" means "work to home". The integer part of the start and end times indicates hours, and the digits indicate minutes in percentage, e.g. 9.89 is at around 9:54 am (54 = 89*0.6). The free time activity code is 1, indicating the person is staying at home. As the model assumes the same route and speed going to and back from a destination location, the duration and travel mode are the same for to and back from work.

start_time	end_time	activity	activity_code	travel_mode
0.0	7.06	home	1	–
7.07	7.15	h2w	2	walk
7.16	16.14	work	3	–
16.15	16.23	w2h	2	walk
16.24	17.73	home	1	–
17.74	19.73	free_time	1	–
19.74	23.9	home	1	–

2.2. Design concepts

In exposure assessment, the most important activity variables are the destination locations, the commuting routes largely determined by the travel mode, and the activity schedules. The concept is thus to simulate personal activities based on probability sampling of agents' departure time, travel mode, maximum travel range, free-time activities, and possible destinations. The duration of an activity can be pre-defined or sampled. The probability distributions (i.e. probability density function for continuous variables and probability mass function for discrete variables) can be empirically specified or derived from mobility surveys with different socioeconomic variables or other attributes that relate to mobility (e.g. age) as covariates. The duration of working time is specified for different population groups (e.g. 8 h for full-time workers).

2.3. Model details

2.3.1. Input and output of the model

The input of the activity model consists of (1) home locations of individuals whose exposures are to be assessed, (2) all the possible destination locations for each destination type (e.g. all the work locations, all the school locations), and (3) the mobility data. Alternatively, the mobility data can be replaced by probability distribution functions or probability tables. For example, the probability of each travel mode for different distance ranges and different attributes of the population (e.g. age) can be the input of the model.

The output of the activity model consists of activity schedules and corresponding spatial locations (including geospatial tracks) for each iteration and each individual. The activity schedule consists for each trip of a start time, an end time, a travel mode, and the activity name and a corresponding code (activity_code in Table 1), which separates different free-time activities and links the schedule to the exposure assessment. Table 1 shows an example.

2.3.2. Generating activity schedules and spatial locations

The most important task of the activity model is to generate activity schedules together with the spatial locations (including tracks) for each residential location (Fig. 2). In each iteration, the model sequentially samples or calculates the travel distance, the destination location, the travel mode, the travel duration, a free time activity, and the start and end time of each activity. More specifically, for each residence location, the main steps include:

1. Specify the probability density function of the travel distance, which can be estimated based on the mobility survey data (in this study) or derived from literature.
2. Sample from the density function of the travel distance and use this distance to refine the sampling space of the destination location. Specifically, from the origin location, a buffer is drawn with the travel distance as radius, and a sample is taken out of all the candidate destination locations that lie in the buffer (Fig. 4a). All the potential destination locations, for example, all of the school or sport facility locations, are assumed known and are input to the model. These locations can come from for instance governmental statistics or the OpenStreetMap.
3. If there is no destination location within the travel distance, the nearest destination location is used (Fig. 4b).
4. The probability distribution of the travel mode depends on the distance travelled and is calculated for different population groups (e.g. elderly people, students) and travel purposes (e.g. going to work), based on the mobility survey data.
5. The Euclidean distance between the destination and the origin is used to determine the probability of taking a certain travel mode. Based on this probability, a travel mode is sampled. The model considers three travel modes: walking, cycling, and driving or taking an auto-vehicle. The reason that we use Euclidean distance instead of the distance along the road network is to reduce computational time. This is further discussed in the discussion section.
6. A route is queried from the road network for the sampled travel mode. For example, a walking path is queried if the travel mode is "on foot"; a bicycle path is queried if the travel mode is "by bicycle". By default, the shortest-distance road will be chosen for on foot or by bicycle, and the fastest route will be chosen for auto-vehicles, based on OpenStreetMap.
7. Based on the travel distance and the travel mode, the duration is calculated.
8. Based on the duration, the end time of a trip or the start time of the next activity is calculated.
9. If the start time of the current trip is unknown (e.g. the first trip of the day), the start time is generated with a distribution. The default departure time to work in our model is sampled from a Gaussian distribution with mean 8 (i.e. 8 am) and standard deviation 0.2.
10. A free time activity occurs 1–1.5 h (i.e. a random number is chosen between 1 and 1.5 with equal probability) before a person goes to work in the morning and after a person arrives home from work. It is chosen by randomly sampling from a set of possible free-time activities. Currently, two types of activities are implemented: staying at home and taking a walk from the home location. Taking a walk is implemented with a Gaussian kernel specifying the probability that a person visits a location, the size and variance of the kernel are user-defined.

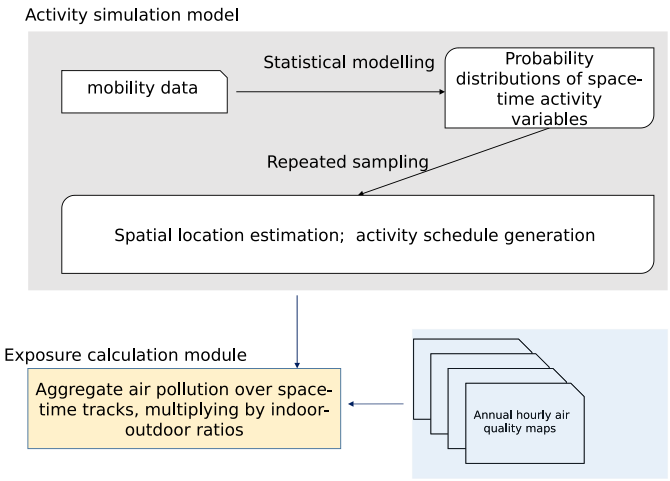


Fig. 1. The structure of the proposed exposure assessment model. The output from the activity simulation model, containing activity schedule and geospatial information, and the temporal air pollution maps, are the inputs to the exposure calculation module.

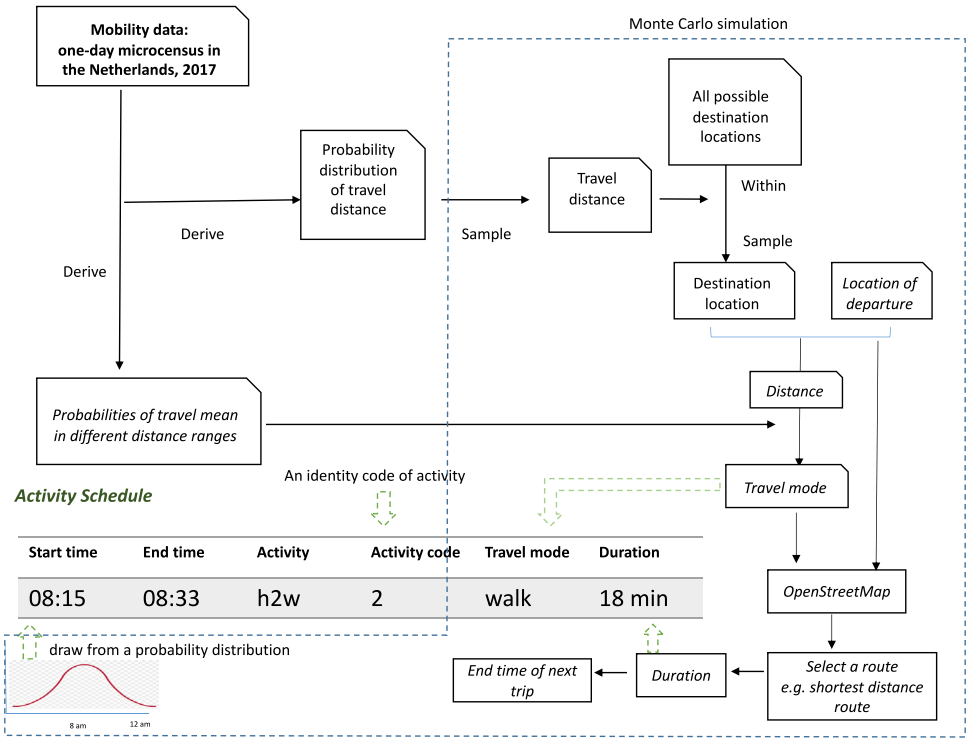


Fig. 2. The procedure of generating activity schedules using the activity model. The activity schedule contains a sequence of activities, one activity (one row) is shown for illustrative purposes. In the activity schedule, "h2w" means home to work. The activity code of the schedule links the schedule to spatial locations.

2.3.3. Routing

Based on the transportation mode, the travel routes are queried from road networks constructed using the Python package OSMnx (Boeing, 2017). The OSMnx processes routes from OpenStreetMap (OpenStreetMap contributors, 2020) into a network and removes the redundant nodes (Boeing, 2017). This reduces the dataset size and accelerates route querying. There are three types of road networks implemented in OSMnx: "auto-vehicles", "bicycle", and "walk" (Boeing, 2017). The road network consists of an attribute travel time. For the travel mode "auto-vehicles", the travel time is calculated in OSMnx, which takes the information of auto-vehicle speed on different roads from the OpenStreetMap for the travel mode "bicycle" and "on foot", the speed is assumed to be constant and can be specified by a user, by default the speed is set to 14 km/hr and 5 km/hr for cycling and walking, respectively. This allows selecting routes either based on the shortest distance or the shortest travel time. By default, we select a route based on the shortest travel time.

3. Exposure assessment

Based on the spatiotemporal tracks of each individual simulated by the activity model and the temporal air pollution maps, the personal exposure is calculated as spatial and temporally-weighted aggregation of air pollution concentration considering the indoor–outdoor ratio (the infiltration of pollution). If the activity model is run N times to simulate different schedules and spatial locations for each person, the exposure is calculated N times, i.e. for each iteration. By default, we use the mean of exposure calculated in the N iterations as the final exposure assessed.

The exposure assessment module queries and aggregates air pollution concentrations for each individual and over each activity in the activity schedule. We describe the process using the air pollutant NO_2 as an example, in pseudo-code (Algorithm 1) below:

Data: temporal NO_2 maps, for each agent activity schedule and spatial locations associated with each activity in the schedule. The indoor–outdoor ratio.
Result: exposure assessed for each activity and each person.
for *each agent* **do**
  initialization;
  exposure_activity = 0
  for *each activity* **do**
    exposure_activity + = NO_2_of_corresponding_time_over spatial_locations_of_the_agent × activity_duration × indoor_outdoor_ratio
  end
  exposure_agent = exposure_activity /time_of_all_activities
end

Algorithm 1: Exposure calculation, exposure_agent indicates exposure calculated for each agent, exposure_activity indicates exposure calculated for each activity in the schedule for each agent. $a+ = b$ means "$a = a + b$"

The exposure at home and work are calculated as the air pollution concentration at the front door home and work locations multiplied by an indoor infiltration ratio. By default, this ratio is set to 0.7 based on Salonen et al. (2019). Currently implemented free-time activities include (a) staying at home, (b) in the garden or on the terrace, and (c) taking a walk. The free-time activity (b) is implemented as walking randomly within a distance (default 0.002 degree (about 220 m)) close to the home location. The free-time activity (c) is implemented as using a Gaussian kernel of distances away from home to the probability of being a presence at the location (default, mean = 2 degree, about 220 km) and standard deviation = 0.1 degree (about 11 km).

An activity-oriented view of the exposure calculated in a single iteration for two individuals is shown in Fig. 3, which shows the

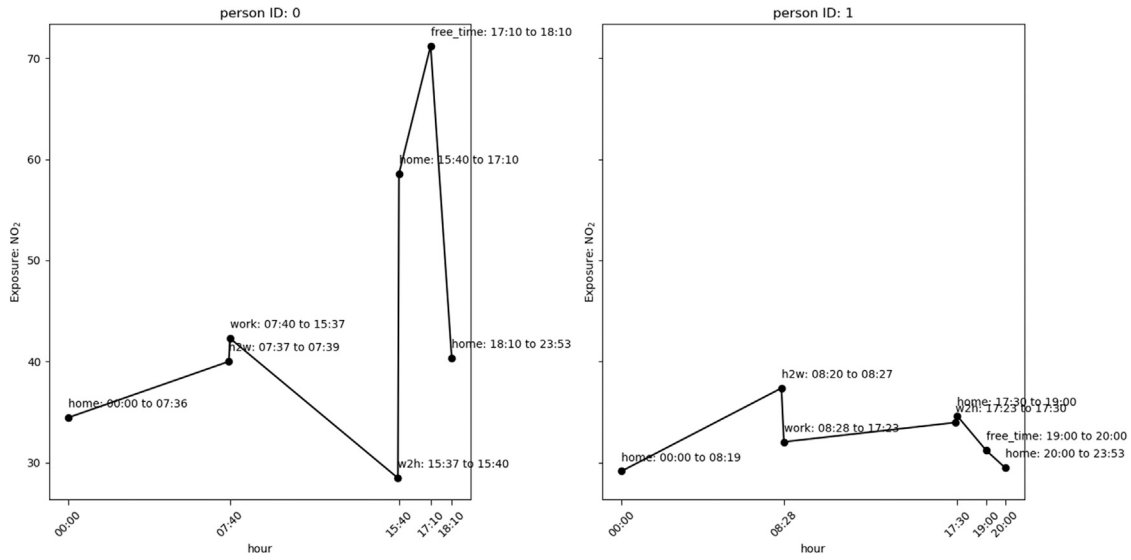


Fig. 3. An illustration of exposure assessed for each activity, plotted at the starting time of the activity, for a single simulation instance. As we are looking at the exposures averaged over each activity, the lines connecting points are simply for visualising the trends, as opposed to showing the exposures continuously over time. The NO_2 exposure is in $\mu g/m^3$.

activities, starting and end time and the average exposure during the time of a certain activity.

4. Case-study: activity modelling using the Dutch national micro-census data

We demonstrate our model by simulating the exposure for the people with occupation "students" and age "18 and older" to represent the population group "university students" in Utrecht. The OViN survey (2010–2017, followed by OVG and MON)) consists of 33 people in this group living in Utrecht at that time.

OViN is collected by Statistics Netherlands (in Dutch: Centraal Bureau voor de Statistiek — CBS) for a one-day trip-based diary. It consists of 0.3% (ca. 52,000) of the Dutch population (17.4 million). We processed the text in the original dataset to extract the range of a certain variable and calculate a mean of it. For example, the variable travel distance is in the form "3 km to 5 km" and it was processed into three variables, "lower limit" (3 km), "upper limit" (5 km), and "mean" (4 km) travel distances.

The students are assumed to go to a university or college in the city during the daytime and do free-time activities (including staying at home) in the evening, a while after returning home. The departure time of going to the university or college is sampled from a Gaussian distribution with a mean of 9 (for 9 am) and a standard deviation of 1 (for 1 h). The Departure time of leaving the university or college is sampled from a Gaussian distribution with a mean of 17 (for 5 pm) and a standard deviation of 1. All the university and college locations queried from OpenStreetMap are used as possible destination locations.

The probability distributions of the travel mode and the travel distance are derived from OViN. For the travel mode, we regrouped the transportation means as they are in the OSMnx, with "auto-vehicles" including taking cars or taxis or all the other land vehicles (e.g. bus, tram). The distance range (less than 1 km, 1–2.5 km, 2.5–3.7 km, . . . ,) is determined empirically. For each distance range, the incidence of each travel mode is divided by the total incidence to obtain the probability (of the travel mode in each travel distance range). For the travel distance, the histogram and log-normal tests through the QQ (Quantile vs Quantile) plot and the Shapiro test indicate the distribution is log-normal.

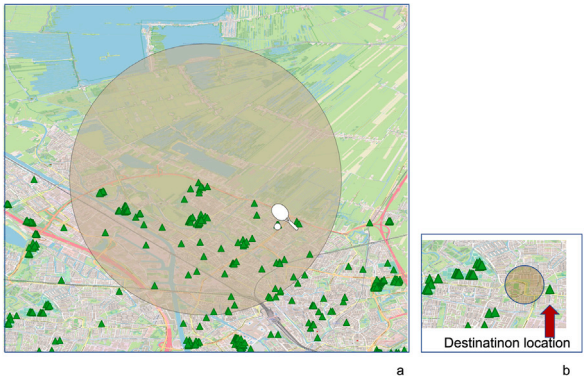


Fig. 4. Illustration of the destination location selection. (a) The origin location (i.e. home location) is at the centre of the buffer. The triangles indicate all the possible destination locations (here: sports facilities in Utrecht, the Netherlands). Only the locations within the maximum travel distances are considered, i.e. the green triangles within the buffer, which is a circle with the estimated maximum travel distance as a radius. Among them, one of the locations is randomly sampled, marked as the *Ping-pong racket and ball*. (b) If there is no destination location within the buffer, the nearest location based on Euclidean distance will be considered.

4.1. Air pollution prediction

Temporal air pollution maps are predicted using statistical modelling. The air pollution measurements are aggregated into hour of the year, e.g. average NO_2 at 12 o'clock of the year. At each time step, the ensemble tree-based algorithm LightGBM (light gradient boosting machine, Ke et al., 2017) is trained using the annually aggregated air pollution measurements of that hour. LightGBM uses histogram-based algorithms to bin the continuous values of each feature. The hyperparameters of the model are tuned using 5-fold cross-validation.

We combined the official hourly ground station measurements of Germany (416 stations) and the Netherlands (66 stations) to predict annual hourly NO_2 concentrations in Utrecht for 2017. The geospatial predictors include road densities in different buffers (100, 300, 500, 1000, 3000, 5000 m) and of highways, primary roads, local roads from OpenStreetMap (OpenStreetMap contributors, 2020), monthly wind speed and temperature of 2017 from ERA5-Land model re-analysis (Muñoz-Sabater et al., 2021), elevation of 90 m resolution (Dai et al., 2017), radiation (World bank, 2022), and Sentinel 5p L3 product column density

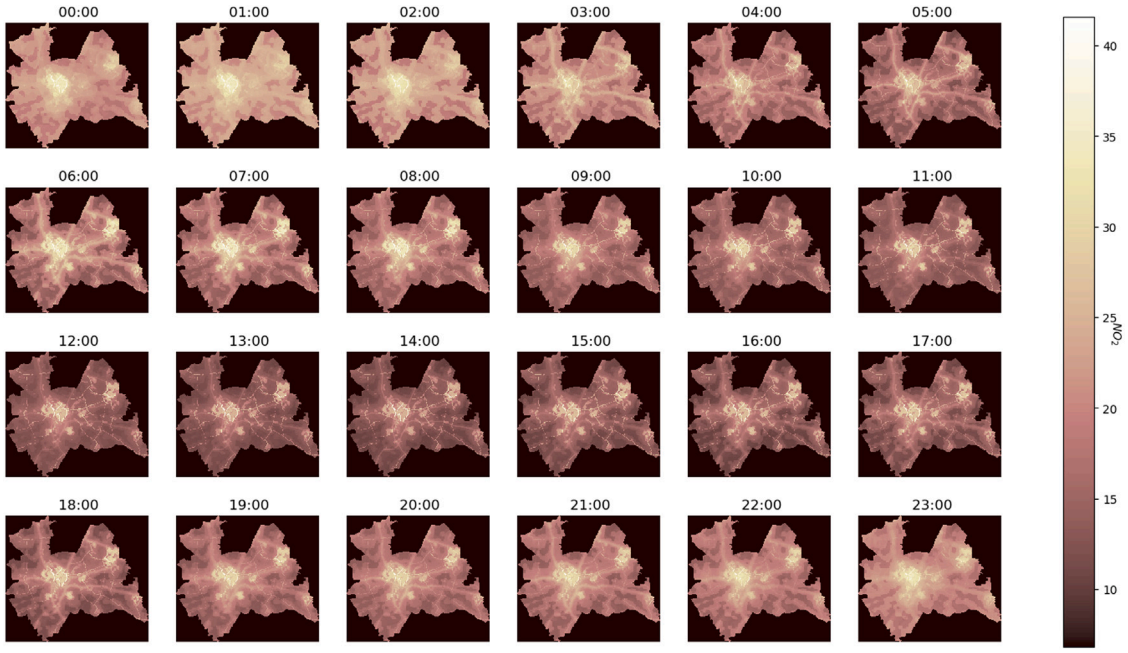


Fig. 5. Annually aggregated hourly NO_2 ($\mu g/m^3$) predicted for Utrecht.

of 2018[1] (10 km resolution, Google Earth Engine, 2019). Population from the Global Human Settlement Population Grid for 2015 (Schiavina et al., 2019) is resampled to 1000, 3000, and 5000 m resolutions. For each pixel of the Earth nightlight satellite data (Elvidge et al., 2017), the mean is calculated in each 450, 900, and 1350 m radius window.

4.2. Results

Annual hourly NO_2 predicted using Light GBM are shown in Fig. 5. The spatiotemporal dynamics of NO_2 are visible. These 24 maps are the input of the exposure calculation component.

The NO_2 exposure is assessed in 11 iterations and the mean of them is shown in Fig. 6. Exposure assessed using real locations (from the survey) in an iteration is shown for comparison. It should be noted that the activity schedules are also simulated in the situation of using true locations. Among all the individuals, 80% of exposure calculated using the real location is within the exposures calculated with simulated locations in multiple iterations. This indicates the uncertainty quantified with exposure assessed in 11 iterations is satisfying in this scenario. In practice, more iterations lead to more accurate uncertainty estimation, as well as exposure prediction (e.g. by taking the mean). The R-squared between exposure assessed using simulated locations and an iteration of exposures assessed using the real locations is 0.35. Besides, different destination locations, departure times, travel modes, and free-time activities likely cause differences in assessed exposure.

The exposure assessed using the proposed simulation model for the university students is shown in Fig. 7. It can be observed that high exposures do not necessarily occur for people with high concentrations at the front door home locations. The exposure assessed is in general lower than the home location concentration. The reason is the use of a relatively low indoor–outdoor ratio (0.7). Note that with university students, the uncertainty in choosing the destination location is relatively low compared to for example full-time workers, as the number of university or colleagues in a city is limited.

5. Discussion

We have described an activity simulation method that uses statistical modelling to parameterise the agent-based model and showed how it can be used together with temporal air prediction maps for exposure assessment.[2] A case-study using the Dutch microcensus data was developed for the assessment of NO_2 exposure by university students in Utrecht. We compared between using the home location concentrations and the exposure assessed using our exposure model with real and simulated destination locations.

The major novelty of our activity modelling concept lies in that it is based on sampling from the probability distributions of the mobility-related variables, and the probabilities could be determined for different attributes of a population such as occupation and age. The activity modelling process is refocused on the statistical modelling of mobility data. Statistical modelling is an important mean to characterise the distribution functions of activity model inputs based on the socioeconomic attributes of a population. This makes the activity modelling not only a simulation problem but also an approach to integrate different sources of information for optimal estimation of the inputs of the agent-based model. The prominent role of data integration and "big data" in gaining us more information and insights about our society is clear. Our model could potentially integrate mobility data from different sources, information from literature, as well as social-economic and environmental data to facilitate characterising the mobility-related variables and reducing the uncertainty. Compared to the proposed model, the approach proposed in Lu et al. (2019) chooses the destination locations of trips without considering people's travel behaviours. Also, there is no randomness in the activity schedule and mobility-related variables.

In the case-study, we singled out a population group (i.e. university students). In practice, we commonly would model for multiple

[1] There is an inconsistency in time but we have evaluated the use of Sentinel-5p of 2018 vs. OMI of 2017, when the NO_2 observations are of 2017 (Lu et al., 2020a) and found that despite the mismatch in time, involving Sentinel-5p L3 product column density of 2018 led to better prediction results. The reason is that the spatial variation is larger than the temporal variation for Sentinel-5p data.

[2] the activity model is open-sourced on Github: https://github.com/mengluchu/agentmodel/tree/main/activity_modeling/core

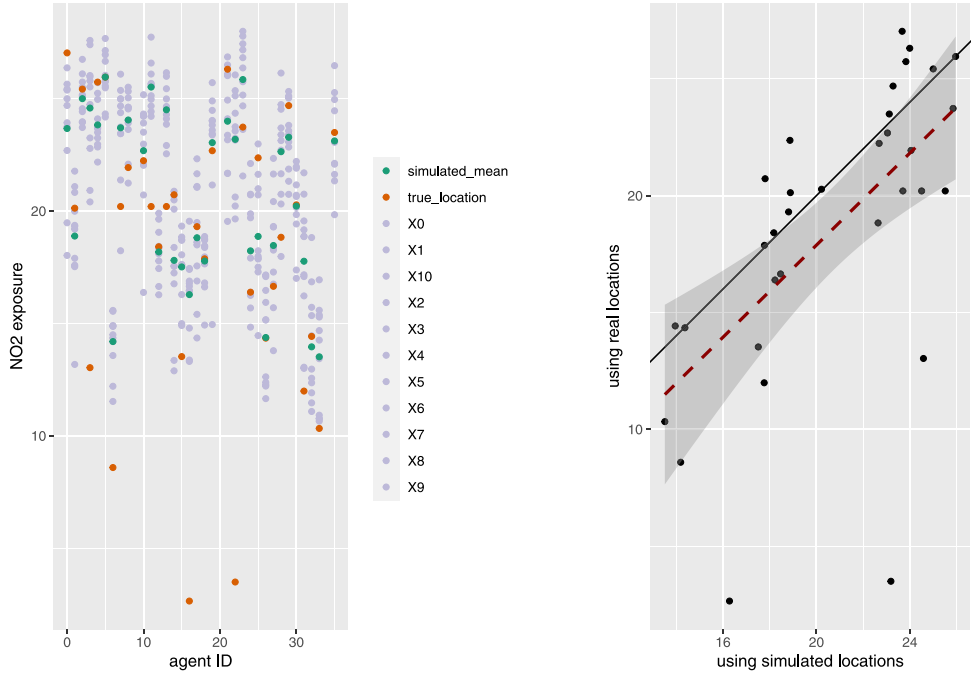


Fig. 6. NO_2 exposure ($\mu g/m^3$) assessed with simulated and real locations based on the microcensus survey. Left: for each individual (agent ID), the red dots indicate exposure assessed using real locations from the microcensus data; the purple dots (X0, . . . , X10) indicate exposures calculated using simulated locations in 11 iterations, with the green dots ("simulated mean") showing the mean of them. Right: The relationships between the exposures calculated using simulated locations (mean of the iterations) and real locations. The black solid line indicates the 1:1 line and the red dashed line the linear regression line. The grey shades indicate 0.95 confidence interval. The R-squared corresponding to the regression line is 0.35.

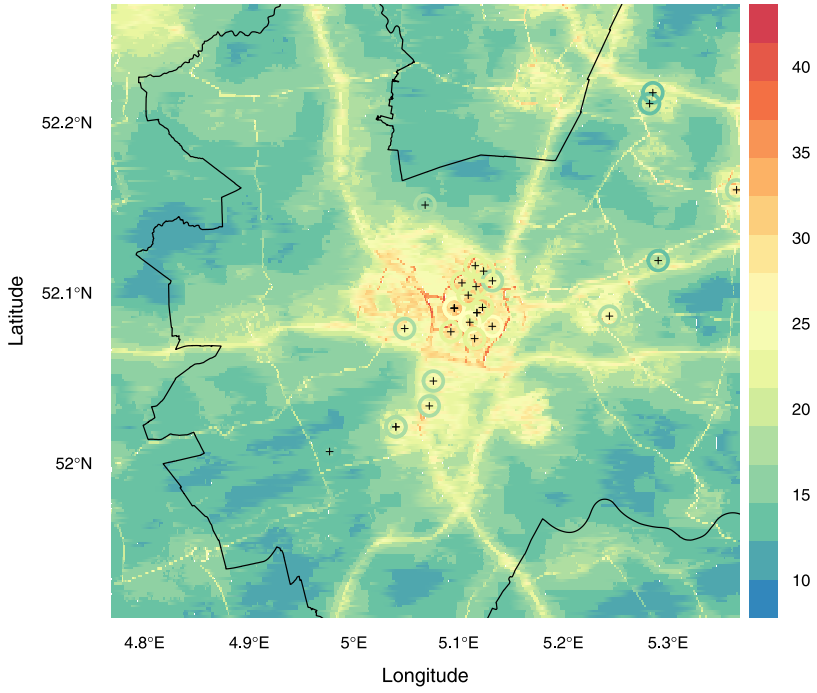


Fig. 7. Each circle shows the exposures assessed using the simulation model for the population group of university students, using the mean of the 11 iterations. To compare the exposure assessed with the annual mean NO_2 concentration for 2017 at the home location, the same legend is used for exposure assessed and NO_2 concentration estimated and the annual mean NO_2 concentration prediction. The circles are enlarged at each location (marked by the crosses) where exposure is assessed to enhance visualisation. The annual mean NO_2 prediction is calculated by taking the average of all the annual hourly NO_2 predictions.

population groups. With our model, two approaches can be conveniently implemented. The first approach is to identify the population groups from the data. More specifically, we could use socioeconomic attributes as independent variables in a statistical model to predict the distribution of mobility-related variables. The second approach is to group the population based on between-group variability of the

activity patterns and then simulate for each group. The first approach allows fitting flexible models (e.g. an ensemble tree model) as many population attributes such as age are in practice numerical. The second approach may include more data in each population group for fitting the mobility-related variables.

The exposure calculation model iterates over each person and then aggregates the air pollution concentrations over the route of each activity of the person, with the indoor–outdoor ratio accounted for. It is highly parallelisable as the exposure calculation is independent for each individual. Note the individuals could still interact with each other and with the environment when simulating the route and the schedules. Another way of calculating the exposure is to iterate over each time step, i.e., for each time step, the exposure of the entire population is calculated. This way can be parallelised for each time step instead of each individual. For a population larger than the number of time steps, this approach may provide an efficient alternative.

This study focuses on the new exposure modelling concept and provides a relatively simple model. Below, we listed the limitations of the current model and our envisioned extensions:

- We currently only generate two probability distributions from the mobility data, the distribution of travel mode and the maximum travel range as a function of the origin distribution locations and the travel mode. The distributions of the departure times and free-time activity are empirical. The model also only chooses one destination location for the entire schedule, i.e. to and back from work. In the future, more commuting activities could be implemented. An effective way would be to include sophisticated transportation modelling in the activity model to simulate secondary and more commuting activities, trips by public transportation, and more detailed road conditions, such as traffic congestion.
- With our case-study, we represented the annually-averaged diurnal variations in exposure, but not variations over a year. This would be an interesting next step, and it might be feasible. In countries where the NO_2 observations are available for each day, it is possible to predict the NO_2 at finer resolutions (Lu et al., 2020b). Since our activity model is flexible in simulating at different temporal resolutions, exposure assessment representing daily variations is possible and the precision could be improved by for example taking into account drivers of seasonal changes in behaviour, conditioned by sufficient computational power and data storage. A more accurate exposure assessment would likely give different exposure maps. How the pattern in Fig. 7 would change with temporally more detailed exposure assessment is beyond the scope of this study but future studies should confirm if the conclusions derived from Fig. 7 stay.
- The proposed activity modelling approach can potentially incorporate environmental, socio-economic, or any variables influencing the travel behaviour. For example, the probability distribution of travel distance could be made dependent on whether the area is urban or rural. In addition to the shortest distance or fastest routes, the greenest route or least polluted routes could be considered.
- The traffic simulated by the activity model could also be used to update air pollution maps (W Axhausen et al., 2016).
- The indoor–outdoor ratio was set to a fixed value of 0.7. It could be better estimated as a function of temperature (Müller et al., 2021). An infiltration ratio may also be applied when taking vehicles for commuting.
- We used Euclidean distance between the departure and arrival locations for selecting the travel mode. As the distance is usually shorter than the actual route, the approach may be more likely to choose a lower-speed traffic mode (e.g. on foot instead of by bicycle). The reason for using Euclidean distance instead of the road distance is to avoid additional route querying, which takes more computation time. Commonly, the Euclidean distance becomes closer to the road distance with increased travel distance. It is, however, possible to calculate the road distance instead of the Euclidean distance if the computational time is not a concern.
- The uncertainties from statistical modelling and air pollution mapping should be quantified and discussed together with the uncertainty from the activity simulation process.
- We used 11 iterations in our case-study as it is shown in Lu et al. (2019) that this number of iterations is sufficient for quantifying the uncertainty. In Lu et al. (2019), though the destination locations are the only random variable, the candidate destination locations are around 500 times more compared to in our case (as it included all potential working locations in Utrecht) and there is no travel distance constraint. For this reason, we believe 11 iterations are also sufficient for the case-study. In practice, the number of iterations could be determined with the method used in Lu et al. (2019).
- The modelling concept can potentially be applied to the exposure assessment of other pollutants such as Ozone and Particulate Matters.

6. Conclusion

An activity simulation model that integrates statistical and agent-based modelling is developed for long-term, large-scale exposure assessment. The key concept of the activity model is to estimate the probability functions of variables that determine the activity patterns (mobility-related variables) from mobility data to parameterise the agent-based model. The mobility-related variables are repeatedly sampled from the corresponding probability function. Correspondingly, personal exposure is repeatedly assessed in multiple iterations to allow uncertainty quantification and improve the prediction. The activity modelling concept could incorporate mobility and environmental information from different sources and makes this process independent of the agent-based simulation component of the approach. The model is easily extensible and applicable geographically. We used the Dutch national microcensus survey to demonstrate our activity simulation model. The simulated activity then combines with hourly NO_2 predicted from the national ground stations of Germany and the Netherlands to demonstrate the exposure assessment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to the code is provided in the manuscript.

Acknowledgements

This study is supported by the Health Effects Institute (No. 4972-RFA19-1/20-6). The authors are grateful to the editors for handling this manuscript and the reviewers for their constructive comments.

References

Alessandretti, Laura, Sapiezynski, Piotr, Lehmann, Sune, Baronchelli, Andrea, 2017. Multi-scale spatio-temporal analysis of human mobility. PLoS One 12 (2), e0171686.

Arentze, Theo, Hofman, Frank, van Mourik, Henk, Timmermans, Harry, 2000. ALBATROSS: Multiagent, rule-based model of activity pattern decisions. Transp. Res. Rec. 1706 (1), 136–144.

Beckx, Carolien, Panis, Luc Int, Arentze, Theo, Janssens, Davy, Torfs, Rudi, Broekx, Steven, Wets, Geert, 2009. A dynamic activity-based population modelling approach to evaluate exposure to air pollution: Methods and application to a Dutch urban area. Environ. Impact Assess. Rev. 29 (3), 179–185.

Boeing, Geoff, 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139.

Čertickỳ, Michal, Drchal, Jan, Cuchỳ, Marek, Jakob, Michal, 2015. Fully agent-based simulation model of multimodal mobility in European cities. In: 2015 International Conference on Models and Technologies for Intelligent Transportation Systems. MT-ITS, IEEE, pp. 229–236.

Chiusolo, Monica, Cadum, Ennio, Stafoggia, Massimo, Galassi, Claudia, Berti, Giovanna, Faustini, Annunziata, Bisanti, Luigi, Vigotti, Maria Angela, Dessì, Maria Patrizia, Cernigliaro, Achille, et al., 2011. Short-term effects of nitrogen dioxide on mortality and susceptibility factors in 10 Italian cities: The EpiAir study. Environ. Health Perspect. 119 (9), 1233–1238.

Crooks, Andrew T., Heppenstall, Alison J., 2012. Introduction to agent-based modelling. In: Agent-Based Models of Geographical Systems. Springer, pp. 85–105.

Dai, Yamazaki, Daiki, Ikeshima, Ryunosuke, Tawatari, Tomohir, Yamaguchi, Fiachra, O'Loughlin, Jeffery, C. Neal, Christopher, C. Sampson, Shinjiro, Kanae, Paul, D. Bates, 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44 (11), 5844–5853.

Deffner, Veronika, Küchenhoff, Helmut, Maier, Verena, Pitz, Mike, Cyrys, Josef, Breitner, Susanne, Schneider, Alexandra, Gu, Jianwei, Geruschkat, Uta, Peters, Annette, 2016. Personal exposure to ultrafine particles: Two-level statistical modeling of background exposure and time-activity patterns during three seasons. J. Exposure Sci. Environ. Epidemiol. 26 (1), 17.

Dons, Evi, Panis, Luc Int, Van Poppel, Martine, Theunis, Jan, Willems, Hanny, Torfs, Rudi, Wets, Geert, 2011. Impact of time–activity patterns on personal exposure to black carbon. Atmos. Environ. 45 (21), 3594–3602.

Duan, Naihua, Mage, David T., 1997. Combination of direct and indirect approaches for exposure assessment. J. Exposure Anal. Environ. Epidemiol. 7 (4), 439–470.

Elvidge, Christopher D, Baugh, Kimberly, Zhizhin, Mikhail, Hsu, Feng Chi, Ghosh, Tilottama, 2017. VIIRS night-time lights. Int. J. Remote Sens. 38 (21), 5860–5879.

Gonzalez, Marta C., Hidalgo, Cesar A., Barabasi, Albert-Laszlo, 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779.

Google Earth Engine, 2019. Sentinel-5P NRTI NO2: Near real-time nitrogen dioxide. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2#description. (Assessed 13 September 2021).

Gulliver, John, Briggs, David J., 2005. Time–space modeling of journey-time exposure to traffic-related air pollution using GIS. Environ. Res. 97.

Hawelka, Bartosz, Sitko, Izabela, Beinat, Euro, Sobolevsky, Stanislav, Kazakopoulos, Pavlos, Ratti, Carlo, 2014. Geo-located Twitter as proxy for global mobility patterns. Cartogr. Geogr. Inf. Sci. 41 (3), 260–271.

Huang, Xiao, Li, Zhenlong, Jiang, Yuqin, Li, Xiaoming, Porter, Dwayne, 2020. Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS One 15 (11), e0241957.

Jurdak, Raja, Zhao, Kun, Liu, Jiajun, AbouJaoude, Maurice, Cameron, Mark, Newth, David, 2015. Understanding human mobility from Twitter. PLoS One 10 (7), e0131469.

Kang, Xu, Liu, Liang, Zhao, Dong, Ma, Huadong, 2020. TraG: A trajectory generation technique for simulating urban crowd mobility. IEEE Trans. Ind. Inf. 17 (2), 820–829.

Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, Liu, Tie-Yan, 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–3154.

Law, Stephen, Sakr, Fernanda Lima, Martinez, Max, 2014. Measuring the changes in aggregate cycling patterns between 2003 and 2012 from a space syntax perspective. Behav. Sci. 4 (3), 278–300.

Lu, Meng, Schmitz, Oliver, de Hoogh, Kees, Kai, Qin, Karssenberg, Derek, 2020a. Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale. Environ. Int. 142, 105856.

Lu, Meng, Schmitz, Oliver, Vaartjes, Ilonca, Karssenberg, Derek, 2019. Activity-based air pollution exposure assessment: Differences between homemakers and cycling commuters. Health Place 60, 102233.

Lu, Meng, Soenario, Ivan, Helbich, Marco, Schmitz, Oliver, Hoek, Gerard, van der Molen, Michiel, Karssenberg, Derek, 2020b. Land use regression models revealing spatiotemporal co-variation in NO2, NO, and O3 in the Netherlands. Atmos. Environ. 223, 117238.

Luo, Kai, Li, Runkui, Li, Wenjing, Wang, Zongshuang, Ma, Xinming, Zhang, Ruiming, Fang, Xin, Wu, Zhenglai, Cao, Yang, Xu, Qun, 2016. Acute effects of nitrogen dioxide on cardiovascular mortality in Beijing: An exploration of spatial heterogeneity and the district-specific predictors. Sci. Rep. 6 (1), 1–13.

Miller, Harvey J., 1991. Modelling accessibility using space-time prism concepts within geographical information systems. Int. J. Geogr. Inf. Syst. 5 (3), 287–301.

Miller, Eric, Roorda, Matthew, 2003. Prototype model of household activity-travel scheduling. Transp. Res. Rec.: J. Transp. Research Board (1831), 114–121.

Mölter, Anna, Lindley, Sarah, de Vocht, Frank, Agius, Raymond, Kerry, Gina, Johnson, Katy, Ashmore, Mike, Terry, Andrew, Dimitroulopoulou, Sani, Simpson, Angela, 2012. Performance of a microenviromental model for estimating personal NO2 exposure in children. Atmos. Environ. 51, 225–233.

Müller, Sebastian A, Balmer, Michael, Charlton, William, Ewert, Ricardo, Neumann, Andreas, Rakow, Christian, Schlenther, Tilmann, Nagel, Kai, 2021. Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. PLoS One 16 (10), e0259037.

Muñoz-Sabater, Joaquín, Dutra, Emanuel, Agustí-Panareda, Anna, Albergel, Clément, Arduini, Gabriele, Balsamo, Gianpaolo, Boussetta, Souhail, Choulga, Margarita, Harrigan, Shaun, Hersbach, Hans, Martens, Brecht, Miralles, Diego G., Piles, María, Rodríguez-Fernández, Nemesio J., Zsoter, Ervin, Buontempo, Carlo, Thépaut, Jean-Noël, 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data 13 (9), 4349–4383.

Nguyen, Anh Dung, Sénac, Patrick, Ramiro, Victor, Diaz, Michel, 2011. STEPS - An approach for human mobility modeling. In: International Conference on Research in Networking. Springer, pp. 254–265.

OpenStreetMap contributors, 2020. Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org. (Assessed 1 January 2020).

Park, Yoo Min, Kwan, Mei-Po, 2017. Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. Health Place 43, 85–94.

Railsback, Steven F., Grimm, Volker, 2019. Agent-Based and Individual-Based Modeling: A Practical Introduction. Princeton University Press.

Rosés, Raquel, Kadar, Cristina, Gerritsen, Charlotte, Rouly, Chris, 2018. Agent-based simulation of offender mobility: Integrating activity nodes from location-based social networks. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. pp. 804–812.

Salonen, Heidi, Salthammer, Tunga, Morawska, Lidia, 2019. Human exposure to NO2 in school and office indoor environments. Environ. Int. 130, 104887.

Schiavina, Marcello, Freire, Sergio, MacManus, Kytt, 2019. GHS-POP R2019A - GHS population grid multitemporal (1975–1990-2000–2015). http://data.europa.eu/89h/42e8be89-54ff-464e-be7b-bf9e64da5218. (Assessed 1 September 2019).

Shekarrizfard, Maryam, Faghih-Imani, Ahmadreza, Tetreault, Louis-Francois, Yasmin, Shamsunnahar, Reynaud, Frederic, Morency, Patrick, Plante, Celine, Drouin, Louis, Smargiassi, Audrey, Eluru, Naveen, 2017. Regional assessment of exposure to traffic-related air pollution: Impacts of individual mobility and transit investment scenarios. Sustainable Cities Soc. 29, 68–76.

Terroso-Saenz, Fernando, Muñoz, Andres, Arcas, Francisco, Curado, Manuel, 2022. An analysis of twitter as a relevant human mobility proxy. GeoInformatica 1–30.

Torre-Bastida, Ana Isabel, Del Ser, Javier, Laña, Ibai, Ilardia, Maitena, Bilbao, Miren Nekane, Campos-Cordobés, Sergio, 2018. Big data for transportation and mobility: Recent advances, trends and challenges. IET Intell. Transp. Syst. 12 (8), 742–755.

W Axhausen, Kay, Horni, Andreas, Nagel, Kai, 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press.

World bank, 2022. Global solar atlas. https://globalsolaratlas.info/download/world. (Assessed 3 March 2022).

Wu, Hao, Liu, Lingbo, Yu, Yang, Zhenghong, Peng, Hongzan, Jiao, Niu, Qiang, 2019. An agent-based model simulation of human mobility based on mobile phone data: How commuting relates to congestion. Int. J. Geo-Inf. 7, 313. http://dx.doi.org/10.3390/ijgi8070313.

Yang, Shusen, Yang, Xinyu, Zhang, Chao, Spyrou, Evangelos, 2010. Using social network theory for modeling human mobility. IEEE Network 24 (5).

Yoo, Eun-hye, Pu, Qiang, Eum, Youngseob, Jiang, Xiangyu, 2021. The impact of individual mobility on long-term exposure to ambient PM2.5: Assessing effect modification by travel patterns and spatial variability of PM2.5. Int. J. Environ. Res. Public Health 18 (4), 2194.

Yoo, EunHye, Rudra, C., Glasgow, M., Mu, L., 2015. Geospatial estimation of individual exposure to air pollutants: Moving from static monitoring to activity-based dynamic exposure assessment. Ann. Assoc. Am. Geograph. 105 (5), 915–926.

Yu, Hongbo, 2006. Spatio-temporal GIS design for exploring interactions of human activities. Cartogr. Geogr. Inf. Sci. 33 (1), 3–19.

Zenk, Shannon N, Schulz, Amy J, Matthews, Stephen A, Odoms-Young, Angela, Wilbur, JoEllen, Wegrzyn, Lani, Gibbs, Kevin, Braunschweig, Carol, Stokes, Carmen, 2011. Activity space environment and dietary and physical activity behaviors: A pilot study. Health Place 17 (5), 1150–1161.