**Evaluation of a spatiotemporal integration method**

**for individual air pollution exposure calculation**

Ruoying DAI (Earth Surface and Water)

Supervisor: Dr. D. J. Karssenberg and Dr. M. Lu

Table of Contents

List of figures and tables3

Abstract4

1 Introduction5

2 Nitrogen Dioxide Distribution6

2.1 Literature Review6

2.2 Study Area7

2.3 Land Use Regression Model8

2.4 NO2 Temporal Variation in Utrecht8

2.5 NO2 Spatial Variation in Utrecht10

3 Human Trajectory11

3.1 Literature Review11

3.2 Distribution Field12

4 Individual Air Pollution Exposure Model16

4.1 Literature Review16

4.2 Modelling Procedure19

4.3 Reference Maps20

5 Result and Discussion21

5.1 Comparison of Exposure maps calculated from different distribution fields22

5.2 Comparison with Reference Map29

5.3 Limitations33

5.4 Sub-questions & Implications33

6 Conclusion35

Appendix35

1 Script for Distribution Field35

2 Script for Exposure41

3 Script for Evaluation43

References50

List of Tables and Figures

Table 3.1 Profile of Distribution Field13

Table 4.1 Four common genres of air pollution exposure assessment17

Table 5.1 NO2 Exposure result of the 1.5 km-radius fields23

Table 5.2 Rank of the coefficient of correlation for commuters’ pairs23

Table 5.3 Rank of the coefficient of correlation for homemakers’ pairs24

Figure 2.1 Environment zones in the blue area 8

Figure 2.2 Mean nitrogen dioxide workday concentration of four seasons in City Utrecht 9

Figure 2.3 Nitrogen dioxide workday mean concentration of City Utrecht by season and hour10

Figure 2.4 Annual nitrogen dioxide concentration map of Utrecht 10

Figure 3.1 Human travel pattern (Edited)12

Figure 3.2 The relationship between a distribution field and the grid13

Figure 4.1 The calculation of exposure level for one cell19

Figure 4.2 The propagation of calculation of an exposure map19

Figure 4.3 Timetable for commuter and homemaker in Lu's agent-based model21

Figure 4.4 Visiting frequency profile in this study 21

Figure 5.1 Legend of NO2 Exposure 22

Figure 5.2 Scatterplots and correlation coefficient of four distribution fields (R = 0.5 km) 25

Figure 5.3 Scatterplots and correlation coefficient of four distribution fields (R = 1.5 km) 26

Figure 5.4 Scatterplots and correlation coefficient of four distribution fields (R = 2.5 km) 26

Figure 5.5 Difference of the map with roads and the map without roads 127

Figure 5.6 Difference of the map with roads and the map without roads 228

Figure 5.7 Difference of the map with roads and the map without roads 328

Figure 5.8 Scatterplots of three distribution fields of different sizes29

Figure 5.9 Exposure map for commuters from Lu's study30

Figure 5.10 Exposure results with high correlation with commuter's mode reference31

Figure 5.11 Exposure results with lower correlation with commuter's mode reference31

Figure 5.12 Exposure map for Home Makers from Lu's study32

Figure 5.13 Exposure results with high correlation with home maker's mode reference 32

Figure 5.14 Exposure results with lower correlation with home maker's mode reference33

**Abstract**

Both health care providers and policymakers concern chronic individual exposure to air pollution. Estimation from the World Health Organization indicates that the joint effects of household and ambient air pollution contributed to 7 million deaths in 2016 [1]. To ultimately understand the relationship between long-term air pollution exposure with diseases, a global map of exposure level is preliminary. Some current studies have coupled pollution maps and human mobility to calculate exposure, while many adopt static address pollution levels as a proxy. The objective of this study is to ultimately understand the relationship between long-term air pollution exposure to diseases. The method of exposure calculation has two components; one is the annual pollution map based on the Land Use Regression model; the other is the distribution fields used as the mobility input. The Land Use Regression model from Soenario 's study [2] utilizes two variables from road length, one from major road length, and one from traffic load. These four variables change from place to place but not with time, unlike human mobility. From the hourly parameters provided by the model, we made hourly maps of NO2 concentration of weekdays and weekends. As our target is the long-term exposure level instead of a short-term one, an annual NO2 map is generated for use by averaging hourly maps. As to include mobility, we define distribution fields as the probability map of the possible activity range of an individual in the middle inside each distribution field, each cell inside the field has a probability of getting visited by the individual. With the two components above, the long-term exposure of a cell is the weighted sum of its surroundings. The weight equals to the visiting probability in a distribution field, and the range of pollution input equals the range of that specific field. Based on the method and the objective above, this study asks: What is the optimal calculation method for individual air pollution exposure when sparse data is available on the mobility of individuals?

The proposed method is evaluated from an inter-comparison among exposure maps and a comparison between our result and two reference exposure maps. The inter-comparison examines the effect of three field settings: Firstly, field types matter. Center-intense fields can strengthen the input of intense concentration on the roads, as the contour filter in image processing. On the contrary, gradual fields level out concentration input in the neighborhood and will not generate star-shaped centers in any case. Secondly, the exclusion of roads, which are of high concentration than other areas most of the time, can raise exposure in some areas due to the limit of leaving out cells on the road for exposure calculation. Thirdly, the field size is influential to gradual fields, but not to center-intense fields. Because gradual fields take in more diverse values from the expanding range, but the center-intense fields are determined by the few numbers of cells near the center primarily, the trivial change in the border side does not affect much. As for the comparison between results and the reference, both to the commuters and the homemakers, the center-intense fields correlate best with the reference, and gradual fields are less correlated. The difference between the two is that: the homemakers' exposure distribution range is better represented, and there is a constant offset that can be compensated by an intercept after exposure calculation. The optimal distribution fields for both commuters and homemakers are center-intense: three Center Fields and the Exponential Field, which answers the research question.

**1 Introduction**

Chronic exposure to air pollution poses a threat to public health. Estimation from the World Health Organization indicates that the joint effects of household and ambient air pollution contributed to 7 million deaths in 2016 [1]. The same report claimed that even in high-income countries of Europe, the annual death toll related to pollution sums up to 208,000 in 2016. The threat to the living population is equally alarming. From the medical perspective, pollutants disturb the cardiovascular and respiratory systems. A growing number of studies suggest pathogenic effects from a single pollutant or a combination of pollutants like particle matters and nitrogen oxides, with a negative association between concentration degree and health status [3, 4]. A noticeable insight is that for respiratory diseases, the cost of prevention is only a portion of the cost of treatment [5]. Therefore, avoidance of air pollution is preferable.

Although public health concern urges to attenuate the negative effect of air pollution, the exact relationship between air pollution and diseases remains unsolved. As to identify the relationship, the first step is to quantify long-term human exposure to air pollution. The current problem here is that both pollution dispersion and human whereabouts are variable. First, the pollution level can be measured locally, but the spatial density of pollution sensors is insufficient to measure the complete spatial pattern in air pollutants. In other words, we cannot derive the pollution distribution of an entire city just by measurement of several sensors. Second, the complexity of human physical movement in space-time also complicates exposure calculation. This challenge has been empirically addressed by neglecting human activities and using the concentration values at the front door addresses of individuals as a proxy of their exposure. Studies have compared exposure calculation with and without human movement. Without it, the exposure level is either overestimated or underestimated [6]. A more representative exposure model calls for the involvement of movement patterns.

Given the ultimate purpose of understanding the relationship between air pollution and diseases, a global map of individual exposure is desirable. Besides the two inherent challenges above, the coverage of pollution data and human movement data is another concern. Some areas collect a large cohort of replacement or pollution time series for a long time, while others might have a shorter observation in different time resolutions. In this sense, the objective of this study is to map individual exposure with minimal input of mobility assumptions and pollution map. Through the low input, the method applies to future global mapping, and finally enhances studies on the relationship between air pollution and diseases. The low input requirement can realize global mapping over areas with diverse data collections.

As to achieve this objective, this study firstly calculates an annual NO2 map for the study area Utrecht. The map stems from the calibrated Land Use regression model that takes geographic features as input and results in hourly NO2 pollution for a year. The spatial resolution reaches 20 meters for all hours. The annual map takes the mean of all hourly maps. Secondly, we use distribution fields to represent the human mobility pattern. Different distribution fields assumed certain visiting probability for the neighboring areas of one living spot. Inside this area, each place has a probability of getting visited by the citizen. As one citizen is assumed to be somewhere inside the field, the total visiting probability of a distribution field is 1. Our model assumes: citizens of the same movement patterns (i.e., homemakers and commuters) share an identical distribution field. This study compares the final result of the exposure map with two agent-based exposure calculation of homemakers and commuters. Lastly, on the procedure, the individual exposure map is generated from weighing the surrounding pollution value for each living cell with the weight of the distribution field. Apart from the comparison among exposure maps, the primary evaluation method is the comparison with agent-based exposure calculations.

With the objective and the method above, we attempt to answer: What is the optimal individual air pollution exposure calculation method for the situation when sparse data is available on the mobility of individuals? Following sub-questions will break down the research question:

1 If an hourly-averaged pollution map is used for annual exposure assessment, what temporal variation of pollution is neglected?

2 How to include human mobility patterns by distribution fields?

3 What is the effect of three distribution field settings (i.e., field type, field size, the inclusion of roads)?

4 How do the exposure maps match the references of the commuter and home maker's mode separately?

Section 2 studies the existent air pollutionmapping technique and adopts the Land Use Regression method in Utrecht, the Netherlands. In order to answer sub-question 1, we use the Land Use Regression method to map NO2 annual distribution with minimal input. Section 3 dives into the literature of human mobility and propose the use of mobility distribution fields, which answers sub-question 2. Three settings represent different human mobility patterns: field size, field type, and inclusion of roads. With both the pollution input and the mobility input, Section 4 elaborates on the novice method to combine them in individual exposure mapping. Our new exposure calculation method originates in the context of previous individual exposure studies introduced in Section 4. Being the focus of the thesis, Section 5 evaluates the quality of the exposure map and answers sub-question 3 and 4 in Subsection 5.1 and 5.2 separately. Then, we use Statistical comparisons to pinpoint the difference between the reference maps and the new map. The comparison includes both commuter' exposure and homemakers' exposure to see how a particular profile fits into different fields.

**2 Nitrogen Dioxide Distribution**

One major challenge of calculating individual exposure to air pollution is the dynamic of air pollutant concentration. As an indication, our result shows NO2 daily concentration difference at the same spot of more than 20 μg/m3 at Utrecht city center. Besides, seasonality and spatial difference present in our result too. Since the air pollution exerts high variability, mapping air pollution distribution is one of the main challenges we identify in exposure assessment. Due to limited time, this research dives into urban NO2 distribution solely as a start of air pollution mapping. This chapter first discusses several approaches to existent methods for mapping NO2 in Subsection 2.1. Then, Subsection 2.2 continues with an introduction of the Land Use Regression model for mapping NO2. The application of LUR for the city of Utrecht, the Netherlands, is provided in Subsection 2.3.

**2.1 Literature Review**

The following review of NO2 mapping starts with the sources and spatial-temporal variation of NO2 in urban areas. Then it will look into current models for mapping distribution. To start with, to understand the causes of NO2 ‘s emergence in the environment is vital to NO2 mapping. According to ESA, industry production, road traffic, and residential combustion top the NO2 contributors in Europe [7]. In the Netherlands, specifically, road traffic produces most NO2 pollution [8]. Emissions from traffic can be direct and indirect. Atmospheric NO2 is mostly an indirect secondary emission as an oxidation result of emitted NO, while the direct primary emission as NO2 is minor [9]. Given the identified pollution source, studies on the factors for pollution propagation follow. Topography and tropospheric processes both may keep pollutants highly concentrated. For instance, the Po Valley, Italy, was heavily polluted mainly due to the lack of wind and a frequent temperature inversion [10, 11]. As the temperature increases with altitude, the inverse layer serves as a cap for air circulation. Together with low wind, this makes the boundary layer stagnant and keeps NO2 concentrated. Although NO2 particles can be dense temporarily, it is relatively short-lived at the near-surface [12].

Available measurement methods for air pollutants include ground monitors and satellites [13,14,15]. Ground monitors can be routinely run or are project-oriented, which requires extra cost, mainly when other attributes like temperature and wind direction are recorded [16]. Recent studies show a rise in low-cost ground sensors [17, 18, 19], but their study-specific conditions constrain generalization for larger areas. As for satellites, their operation is costly, but the coverage range is more extensive than regional monitor networks [16]. The target is different for ground monitors and satellites: the satellite records the column from the ground to the stratosphere, while the ground monitor takes the near-surface concentration, which is more relevant to human health [16]. This discrepancy may cause different results for exposure assessment given the same study area.

Based on the measurement, conventional air pollution mapping methods are diverse. The first method, named interpolation, derives spatially continuous concentration levels from point records. Typical approaches include Kriging and the Inverse weighted distance method [20, 21]. The second method, the dispersion model, focuses on the atmospheric process of pollutants. It takes into account the dynamics of mass transport, turbulence, and chemical transformation [22]. The third method, called Land Use Regression models is in use widely in epidemiological studies to calculate pollution map for exposure assessment [23]. This approach assumes that the air pollution level relates to geographic attributes. Frequent empirical attributes concern population, traffic, land cover, transportation accessibility [16]. The fourth method, the hybrid model, combines Dispersion Model and the Land Use Regression model (LUR) and is in use as well [24, 25].

To the challenge of NO2 mapping, we picked the LUR model for its brief term and attainable input. Given other data availability and computation environment, alternative models above may serve to map NO2 with certain advantages. From the data availability aspect, the LUR parameter set for the Netherlands is available to our study, so the LUR is most accessible to map NO2 in Utrecht. Another advantage of our LUR over satellite mapping is that its spatial resolution is 20 meters, while the most recent satellite air pollution mapping is sub-10 kilometer from the TROPOMI sensor launched in 2017 [26].

**2.2 Study Area**

City Utrecht, the study area, ranks as the fourth populous city in the Netherlands [27]. The city reports a growing population since 1999, reaching 352,866 in January 2019, according to CBS's survey [28]. Utrecht central railways handle an average of 194,385 travelers per working day last year [29]. Despite heavy traffic that suggests both pollution source and human exposure, city Utrecht (Dutch: Gemeenten Utrecht) has adopted Environment Zone (Dutch: Milieuzone) to reduce air pollution in the city center. Inside the Environment Zone shown in *Figure 2.1*, polluting diesel cars and trucks are forbidden because diesel engines emit harmful substances to health, including soot and nitrogen oxides. Other local policies are useful in tackling air pollution as well. Among them are the introduction of cleaner transport, enhancement of guidance for cars to drive outside of the city, and construction of new buildings for sensitive groups [30].



Figure 2.1 Environment zones in the blue area, Utrecht (Source: Gemeenten Utrecht [30])

**2.3 Land Use Regression Model**

To quantify the air pollution concentration spatial distribution in this study, we adopt a Land Use Regression Model from Soenario's study[2]. Unlike the models above that rely on either big data or computation cost, this model input set is relatively concise and efficient with high spatial resolution. The outcome of LUR in our study is a set of hourly 20 meters \* 20 meters resolution NO2 concentration level map of the city Utrecht, the Netherlands. To reach the outcome, the LUR modeling process from Soenario selects its input variables first by a Lasso regression [31] and then by a best subset regression [32]. After confirming a set of indicative input for pollution concentration level, the LUR model went through model fittings. In the beginning, the model’s candidate input variables contain traffic, infrastructure, and population. Before two regressions, the variables expand into three variables sets within 25, 50, 100, 300, 500, and 1000 meters' buffer zone. This expansion is a common practice in spatial analysis: since it is uncertain within what distance the variable is influential to the center of the buffer, values within the buffer zone of different radius are all listed and checked.

In this study, the LUR model selects the same four derived variables as Soenario does: two variables from road length, one from major road length, and one from traffic load. These four variables change from place to place but not with time, unlike human mobility. The resulting series is hourly maps for weekends and weekdays of each month separately. The concentration difference between weekends and weekdays is significant, so it is necessary to separate them. The model output consists of 24 hours \* 2 modes \* 12 months = 576 maps. One mode is for weekdays, and the other is for weekends. Section 2.3 and 2.4 will look into the temporal and spatial variation of the sample area accordingly.

**2.4 NO2 Temporal Variation in Utrecht**

We want to set the pollution map static for the exposure calculation, so the impact of distribution field settings is clear. The static map is the annual map averaged over hours shown in *Figure 2.4*. By doing this, we neglect the temporal variation of pollution. It is essential to know how much temporal variation is lost due to the hourly average. Although nitrogen dioxide concentration varies spatially, different places share two temporal trends:

1) Seasonality. Most places have their concentration peak in winter and reach the trough in summer. The non-road area in *Figure 2.2* demonstrates this trend.

2) Daily variation. *Figure 2.3* shows two peaks in a day for all seasons. One is around 8 a.m., and the other is around 8 p.m. In the Netherlands generally, the concentrations in the morning are relatively high due to the concurrence of the morning traffic peak and favorable boundary layer conditions [8], which may explain the morning peaks. Thus, the temporal variation from the LUR model result is logical.

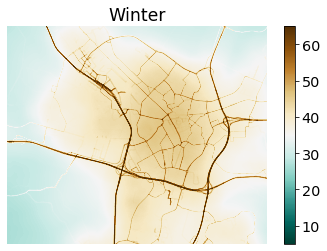
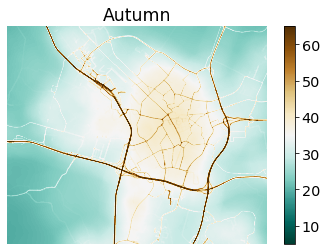
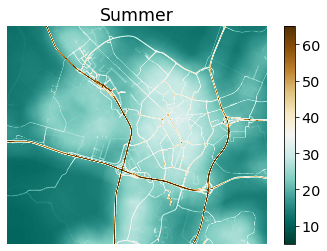
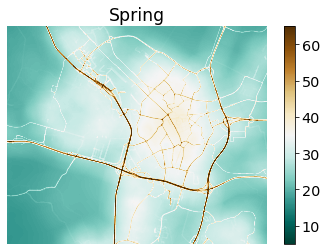


Figure 2.2 Mean nitrogen dioxide workday concentration of four seasons in City Utrecht (unit: μg/m3, LUR model [2]). The four seasons here are March-April-May, June-July-August, September-October-November, December-January-February from spring to winter, respectively. Each season's result is the hourly averaged pollution centration of the corresponding three months. Subsection 2.3 introduces the input for the calculation.

Figure 2.3 Nitrogen dioxide workday mean concentration of City Utrecht by season and hour. (unit: μg/m3, LUR model[2])

**2.5 NO2 Spatial Variation in Utrecht**

*Figure 2.4* shows the annual concentration maps. This map is the direct input to the exposure calculation, and this subsection checks its representative with information on motorways. Note that this map is averaged over each hour’s concentration from the study on five years hourly average introduced above in Section 2.3. The weight ratio of 5:2 applies to weekdays to weekends map. The study area Utrecht has two motorways in the north-south direction: A2 and A27. The west to east direction has motorways: A28 and A12. The area of motorways appears on the map with a high concentration level in dark red color. Between them is the city center that also peaks with dark red color. The peak is likely the result of heavy traffic in the city center.

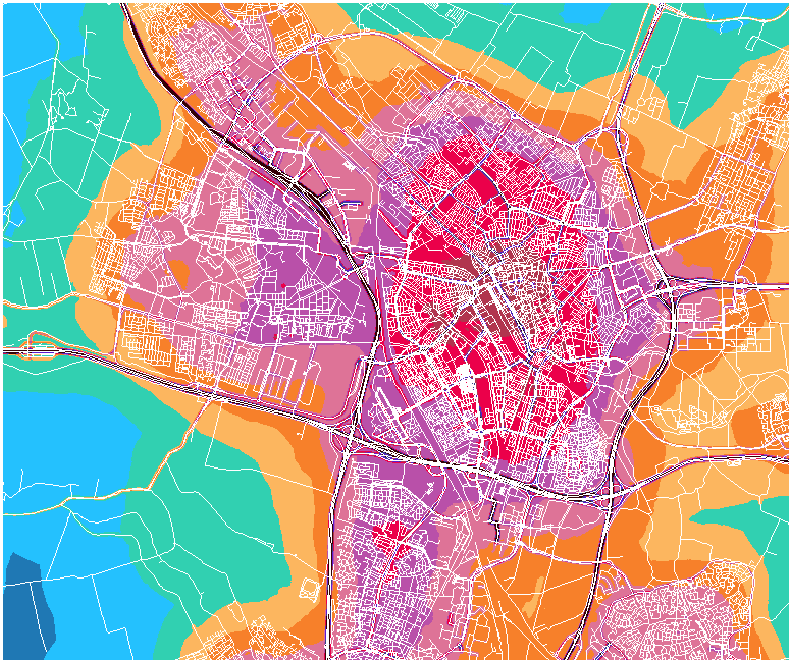
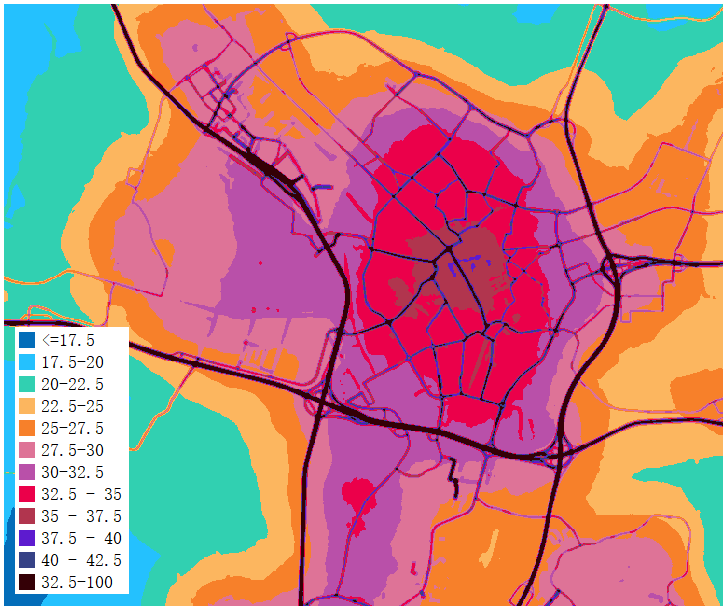


Figure 2.4 Annual nitrogen dioxide concentration map of Utrecht. Left: with roads. Right: without roads (unit: μg/m3, Averaged hourly)

**3 Human Trajectory**

To calculate individual exposure to air pollution, the other major challenge identified in this study is the variability of individual whereabouts. We want to represent long-term human mobility based on theory instead of measurements. The significance of human mobility to exposure measurement is clear: One accesses several places daily. Along the path, one has contact with different levels of pollution concentration, so the total exposure depends on the path mainly. In exposure study, some previous study neglects this mobility factor and uses the residence or work address’ concentration of an individual to calculate all-time exposure [33]. Others do not neglect the trajectory and simulate trajectory based on detailed deterministic input (e.g., GPS, mobile phone signal) [34, 35]. However, to generate a long-term trajectory, it is unlikely to adopt a deterministic model. Probabilistic approaches represent the unknown trajectory instead [36].

The previous study has proposed characteristic patterns of trajectory change: Random Walk theory suggests that short-distance trips are much frequent in personal visit history [37]. This Section focuses on the solution to the human trajectory for exposure study in Subsection 3.2. We design probability fields to simulate human movement range. Probability fields are square grids with individual visiting probability inside each cell. Different probability fields form a set of movement patterns and can be used to represent different individuals. The ideas are inspired by previous human trajectory research widely in Subsection 3.1.

**3.1 Literature Review**

The air pollution exposure calculation relies on the model of individual whereabouts. Existing human trajectory models have their study scale ranging from city-size to intercontinental. The prevalent study topics that apply trajectory model include, from small to large scale: pedestrian movement [38], traffic congestion [39], epidemic propagation [40], migration and emigration [41]. These studies from other fields serve as a valuable reference to trajectory model design. Human trajectory models mostly base on the assumption of mobility motivation or are extracted from real measurements. As for supporting real measurements for trajectory model design, options are GPS data [42], mobile phone records, census, and survey.

Since the range of movement is vast, while our study target is a city-size, here we exclude larger-scale movements like migration. Another focus of our study is on relatively low-velocity traveling because citizens adopt walking, bicycles, and cars mostly, which is the left-down side of the speed-distance spectrum in *Figure 3.1*. The speed constrains the range of their movement inherently.

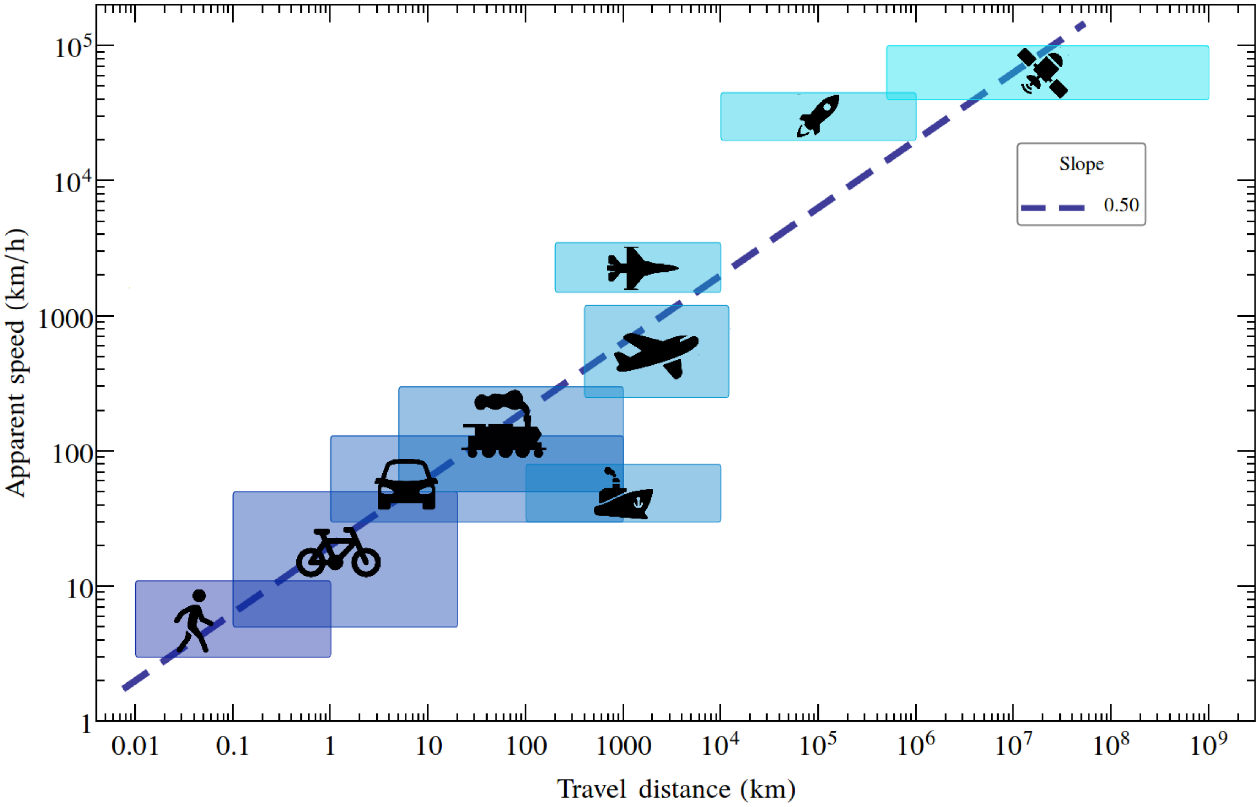


Figure 3.1 Human travel pattern (Edited). Based on the measurement, this figure has shown that the longer the travel distance human reach (x-axis), the faster human travel (y-axis) [11].

Our study stems from a generic individual movement model called Random Walk. In a broader sense, Random Walk is a prototype for human movement. It refers to a trajectory with sequential, random steps. Karl Pearson, in 1905, first coined the term Random Walk. In the beginning, the study object was the distribution of mosquitoes moving in fixed-length each time step [43]. In the same period, Louis Bachelier proposed a similar idea for financial time series [44], and Lord Rayleigh studied sound wave propagation alike [45]. Why are they comparable to each other, and how is human mobility connected to their objects? The link of them is the shared relationship that: the objects relocate themselves based on the previous location. Another shared study interest is the range of one's movement, which is calculated using a scaling factor. Relevant formulas make use of a scaling factor to represent a certain diffusive level. The probability of one object to appear at a spatial-temporal location depends on both the waiting time and length of steps. Here, the waiting-time means the duration of one step, while the length of steps means the spatial distance passed by one step. Each Random Walk application defines two specific distribution functions for waiting-time and length of steps separately.

Levy Flight applies the Random Walk prototype in the sense that it defines the distribution of step-length to be heavy-tailed. The trajectory of Levy Flight would be a set of short-trips clusters linked by long-trips. The trajectory is the direct effect of heavy-tailed step-length distribution. In the application field, Levy Flight is multifunctional. Recent applications aid in wave physics, foraging behavior [46], algorithm optimization [47]. A human mobility study using GPS mining suggests that most people adhere to a simple and reproducible pattern [48]. Our study takes on the idea of Levy Flight in the form of probability fields.

**3.2 Distribution Field**

This study applied Levy Flight with distribution fields. The Levy flight movement patterns that short trip is much more frequent long trips are added with a fixed center. We assume a citizen or an individual dwells at a fixed place, and his or her surrounding area is the movement range of that citizen. Distribution fields refer to the probability map of the possible activity range of an individual in the middle. Shown in *Figure 3.2*, each cell in a distribution field has a probability of getting visited by the citizen. As the individual is assumed to be somewhere inside the field, the total visiting probability of a distribution field is 1. If a cell has a probability of 0.6 for being visited, and each year has 24 \*365 = 8760 hours, then the visiting hour of that cell is 0.6 \* 8760 = 5256 hours. To clarify, the actual distribution fields here are in three sizes, which represent 2.5 km, 1.5 km, and 0.5 km radius of mobility range. Accordingly, the number of cells in use is 251 \* 251, 151\*151, and 51\*51. For instance, a 2.5 km radius covers a range of 5000 meters in diameter and the one cell in the middle. That is 250 cells plus one cell, and 251 cells in one diameter length.

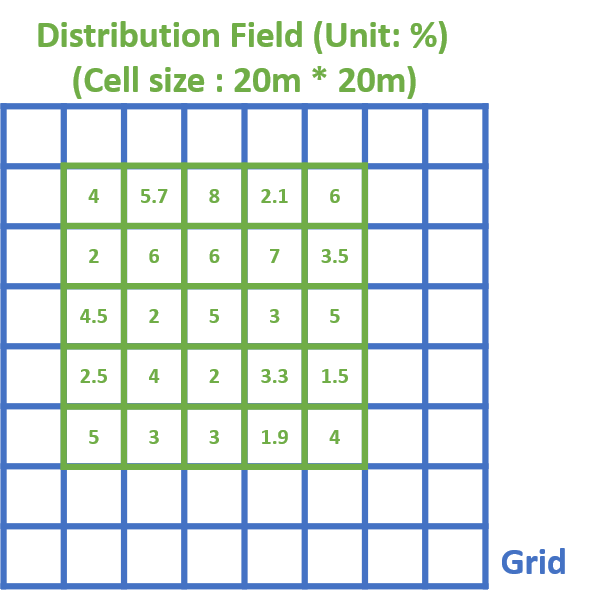


Figure 3.2 The relationship between a distribution field and the grid. Note that they are not in actual size. Here is to state that the distribution field will move on the grid that represents Utrecht.

*Table 3.1* below is the abstract of all distribution fields. The diversity of fields here is a representation of different mobility modes. It is unrealistic to assume that all individuals share an identical distribution field for the simple fact that every habitant exerts his or her movement patterns. This study adopts seven types of fields to cover as many mobility patterns as possible. A distinct difference in the mobility pattern between homemakers and commuters exists. The activity range of homemakers is theoretically smaller since they do not travel between work and home like a commuter. In the result & discussion section, we compare exposure based on each field with commuter's and home maker's exposure map from a previous agent-based study [36]. This comparison will match the optimal distribution field with citizen profiles. For now, we can only define it with probability input and theoretical mobility assumption in the table below. With the reference maps, the further definition will assign a 'commuter profile' or 'homemaker profile' to a specific distribution field.

Table 3.1 Profile of Distribution Field

|  |  |
| --- | --- |
| Distribution Field | Description |
|  | The Linear Distribution Field assumes that the visiting probability descends linearly from the central pixel to the border of the field. |
|  | The exponential field assumes an active visiting frequency near the center pixel. Its distribution of distance to the center might be the closest to the Levy flight pattern.  To generate an exponential field, firstly, the line of cells from the center to the border is assigned an exponential of e -i. i is the number of cells from the center. Then the line of values is assigned to all directions. Lastly, all values are divided by the sum of all values to make the sum of probability field 1.  The visiting probability over distance to the central cell is plotted below. Although the distributions are identical for three sizes of fields, since larger the size, more cells included, the accumulated probability for areas away from the center that have more cells than closer areas can make a difference. |
|  | The Center Field assumes that the habitant spent 60% of the time inside the middle pixel, while different ranges from the center receive a specific visit frequency. The plot below shows how they decrease from the center to the border. With the center probability of 0.6, the rest of the cells seem to have the same visiting probability. To clarify their difference, three other plots below leaves out the center cell and only show the remaining cells.    The Center + Even field has 0.6 visiting probability for the center cell and the same probability for the rest of the cells.    One "ring" refers to an area inside which the cell distance to the center is the same or is close. A field of 251 \* 251 cells has 125 narrow rings and five thick rings. The Center + Thick Fields have the same center of 0.6 visiting probability, but the rest of the value is decreasing like stairs from the center to the border. (Plot on the right). Each of them has respectively 1, 3, 5 thick rings around the center cell.  The Center + Narrow Field decreases faster from the center in comparison (Plot on the left). Each ring inside this field shares the same total probability. For example, the smallest narrow ring has eight cells (i.e., 3 \* 3 cells -1 center cell), and they add up to the probability of 0.05. The second smallest ring has eight cells (i.e., 5 \* 5 cells – 3 \* 3 cells), and they also add up to 0.05. The biggest ring can have 1000 cells (i.e., 251 \* 251 cells – 249 \* 249 center cell) and add up to the same value of 0.05. |
|  |
|  |
|  | The Even Field suggests all neighboring area of a center pixel is visited equally. |
|  | The Random field assumes a random frequency of visiting nearby areas. In real life, this implies that there might be hot spots like a community garden, supermarkets, and restaurants that attract individuals often.  The random probability comes from assigning a random integer to each cell and dividing each cell's value by the sum of all integers. The constrain for all fields is the sum = 1 condition. |

**4** **Individual Air Pollution Exposure Model**

This chapter focuses on the exposure model tested in this study. The exposure model is a combination of the human mobility model in *Section 3* and the air pollution distribution model in *Section 2.* Before the introduction of our model, *Subsection 4.1* reviews relevant methods to quantify air pollution exposure, and groups them by the degree of spatial-temporal models’ usage. The degrees of mobility modeling is from using no mobility data to using continuous trajectory models. *Subsection 4.2* and *4.3* continue with the application of our Individual Air Pollution Exposure Model.

**4.1 Literature Review**

In this study, Individual Air Pollution Exposure Model calculates the amount of air pollution concentration exposed to a person or a population group. The route of contact can be inhalation, ingestion, and dermal contact [49]. The exposure is quantified with the unit µg/m3, denoting the averaged concentration value of a period. The terms used by relevant studies that also target the quantification of individual exposure to air pollution include exposure assessment [49, 50] and exposure estimation [51, 52]. They make no meaningful difference in the following narration. Table 1 summarizes four conventional exposure assessment methods. From top to down, their requirement for human mobility input increases. The increase in mobility data input does not necessarily raise research input or result accuracy, but the synthesized or measured movement patterns are increasing. For instance, the agent-based model assigns the workplace and trajectory to the workplace from home, which is a synthesized movement pattern. The exposure study that uses GPS input as mobility information uses measured movement patterns. This study falls into the category of long-term spatial-temporal aggregation.

As one method, exposure assessment with biomarkers gives informative measurement without location information input. Biomarkers are measurements that reflect the interaction between the biological system and a potential hazard [51]. In air pollution epidemiology, for instance, carbon loading in a specific airway cell can be a valid biomarker for particulate matter [53], and contents in urinary excretion are potential for detecting traffic air pollution [53]. In order to capture the effect of air pollution, this method spares the mobility record of an individual and relates effects with biomarkers sample directly. The advantage of it is significant for individual exposure study, while its cost for population study is high. Besides, biomarkers carry other uncertainties for exposure measurements too. Their exposure calculation requires underdeveloped parameters from several domains of study, and the relationship between biomarkers and exposure is unclear in some studies' views [49]. A difficulty in expanding personal biomarkers monitoring to a population level one is noted too [49].

As another method, adopting a single static address to assess exposure in comparison needs a particular location record as input. This method assumes the pollution concentration of one location for a person is representative of his or her exposure calculation. As it did in the case of one global mapping of ambient air pollution, the country-wise population-averaged pollution concentration is taken as national exposure [52]. In epidemiology studies that aim at the link of exposure and diseases, a static address of potential population is often used to represent personal exposure. Improvement for this method is adding more addresses to the study individual. As one study has compared exposure assessment with or without ‘workplace-residence’ pairs: For the same study area, NO2 exposure result is higher when assigning specific working locations [54]. Another concern of this static method is that the concentration it takes on does not distinguish indoor pollution and outdoor pollution [55], that a person at one location faces one concentration level, no matter it is indoor or outdoor. These methods often produce an exposure map of a large area; some even have global converge [52], due to the oversimplified assumption of a single address.

A discrete spatial-temporal aggregation method takes in more input of personal locations than two methods above. This genre of personal exposure assessment combines temporal air pollution concentration with more than one address, which usually is home and workplace [56]. For instance, a study in UK cities found that the annual mean NO2 exposure level with workday locations higher than the population-averaged level calculated without the workplace [54]. This method is classified as "discrete" in one sense that it neglects the trajectory between locations, although it considers more than one address. It can be “discrete” spatially, while in another sense, it can also be “discrete” temporarily, as this method mostly reach mobility input within one year [57]. Like the static method above, in epidemiology studies, this method is also often used to represent personal exposure [58, 59]. However, studies have shown that incorporating activity patterns instead of just several addresses is vital to epidemiology research [56]. Moreover, we argue that: for long-term assessment purposes, short observation of locations change is insufficient.

A continuous spatial-temporal aggregation method for air pollution exposure requires most mobility input. The mobility input can come from real measurements or modeled data. A continuous mobility input from measurement is unlikely because this concerns sensitive data privacy [60]. As the amount of continuous measurement of locations is limited, continuous mobility model is developed [33]. Section 3.1 reviews diverse continuous mobility models. To generalize their ideas: the location of a person comes from assumptions of human movement [61] or stochastic processes [62]. This method can be in the form of agent-based modeling, as well. This study takes a deterministic air pollution map and a hybrid human trajectory model to exposure assessment, aiming to test their effect.

Table 4.1 Four common genres of air pollution exposure assessment

|  |  |  |  |
| --- | --- | --- | --- |
| Exposure Assessment Genre | The requirement for Human Mobility Input | General Definition | Comments |
| Biomarker [53, 63] | No | Measurements to reflect the interaction between the human body and air pollution [51]. | Most functional in individual study, but costly for population study [49]. |
| Static address [52, 64, 65] | No/Yes | Individuals exposure is a function of their home address’s pollutant concentration | Although they usually have broad coverage, they neglect the exposure from several locations and exposure during transportation. |
| Discrete spatial-temporal aggregation [54] | Yes | Its calculation that does not include movement between places or does not include long-term mobility record but short-term one. | They do not conduct representative long-term exposure assessment. |
| Long-term spatial-temporal aggregation | Yes | Accumulate temporal air pollution along the human trajectory | Some have adopted the ideas of agent-based modeling. It makes use of probabilistic methods too. |

Used by exposure assessment methods above, the concept of microenvironment prevails in air pollution exposure study. By definition, one microenvironment is a homogenous temporal area of air pollutant concentration during one's exposure [49]. An individual's location's history can be seen as a set of microenvironments. For example, a doctor's daily location history might be "home-hospital-supermarket-home." Each new place is microenvironment. When the visit duration inside each microenvironment is short enough, this simulation is close to reality. The concept of microenvironment partly inspires the usage of the distribution field: inside one pixel, the NO2 concentration is the same, and visiting one pixel to an individual is analogous to staying in one microenvironment.

**4.2 Modelling Procedure**

The model has three procedures: making the distribution field, mapping the annual NO2 distribution, and calculating the exposure map with the previous two inputs. In the beginning, we define several universal settings for three procedures. Firstly, the work is done on grids of identical square cells with a size of 20 meters 20 meters. A cell here relates to an assumedly homogeneous geographic area. Secondly, each cell has a factor value under a specific procedure. For instance, a cell in a distribution field has a value of the probability of visiting; a cell in the NO2 distribution map has a value of concentration level in μg/m3. Thirdly, the cell size remains the same in the whole procedure, but the number of cells for each procedure varies. Distribution fields are in three sizes, which represent 2.5 km, 1.5 km, and 0.5 km radius of mobility range. The number of cells for the nitrogen dioxide mapping and exposure calculation is the same. They are done on all cells. Based on the universal settings, the following paragraphs explain each procedure.

The first procedure generates distribution fields. It works on 251, 151, and 51 cells one by one. The visiting probability is assigned to each cell. For an exponential field, the visiting probability decreases from the central cell to the border cells. While for an even field, the visiting probability is the same for all cells in the field. The second procedure maps the pollution level distribution. It takes the whole study area of cells. Hourly parameters are combined with four geographic variables for each cell. The hourly set of NO2 maps are averaged later and forms an annual map of pollution. Weekend hours and weekday hours have a weight ratio of 5:2.

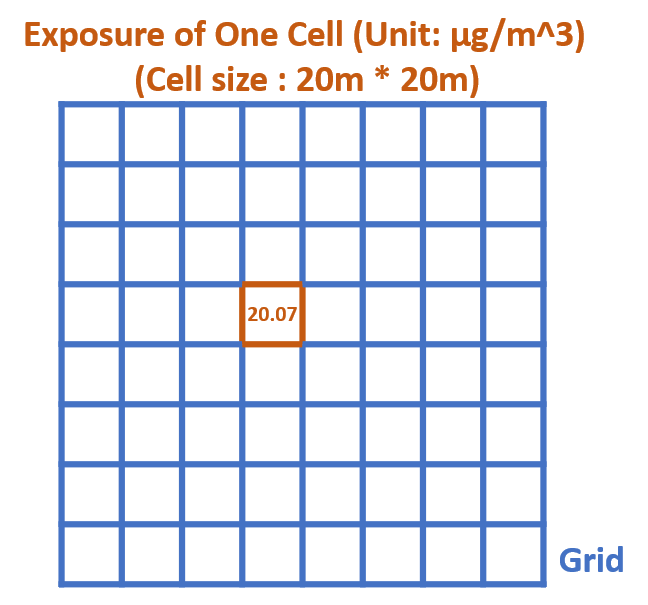
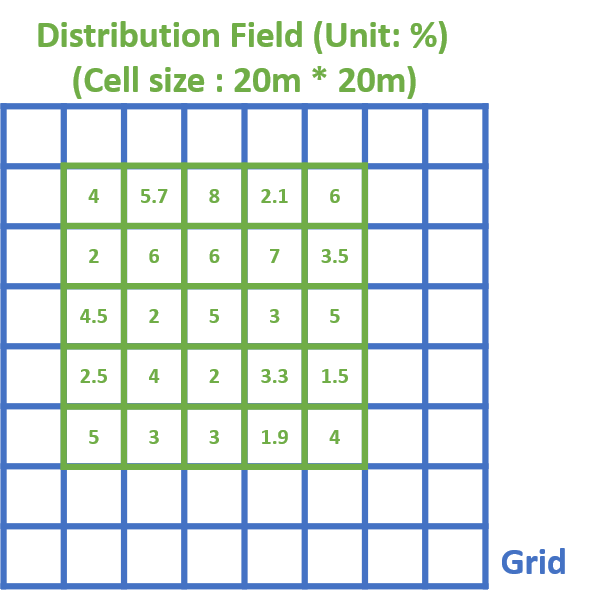
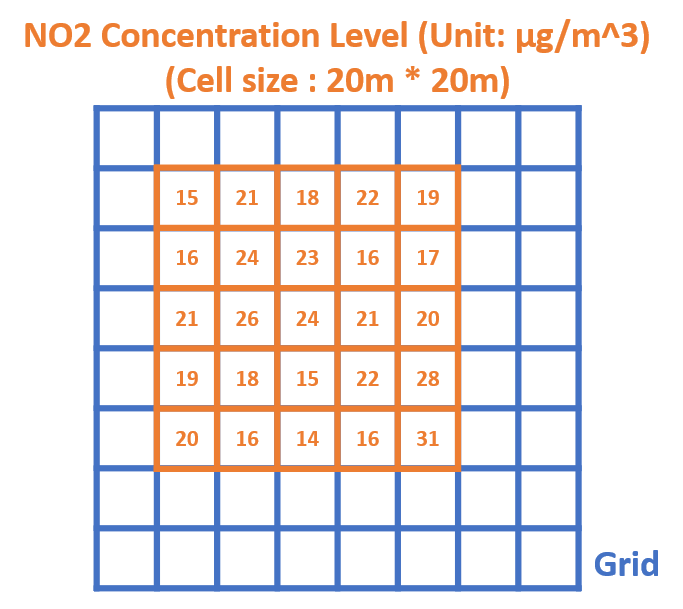


Figure 4.1 The calculation of exposure level for one cell. (Note that both the field and grid are larger in reality. The field and the grid here are for illustration.)

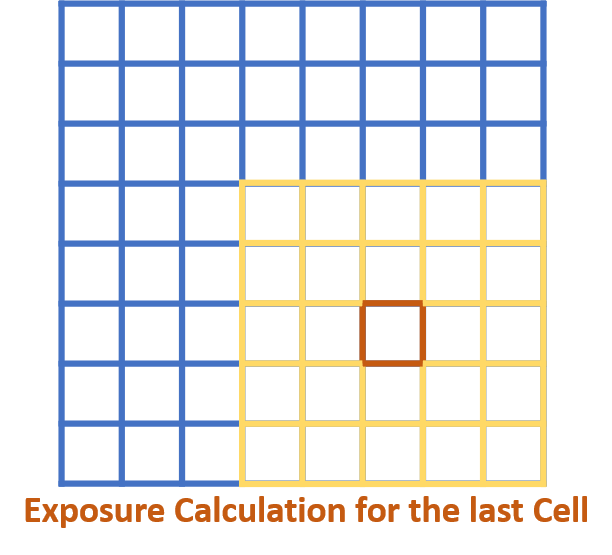
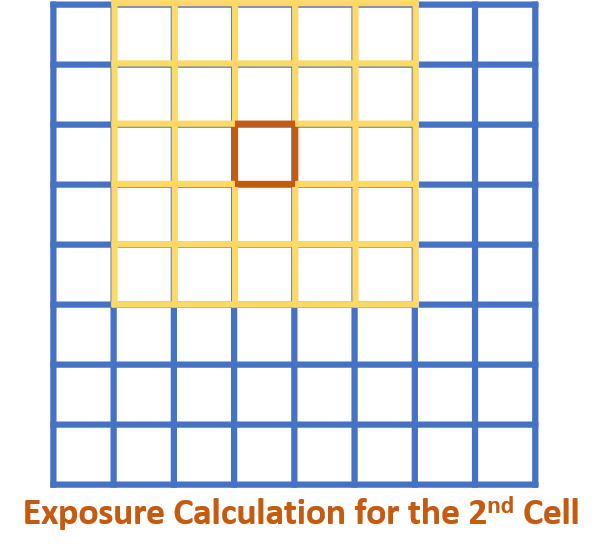
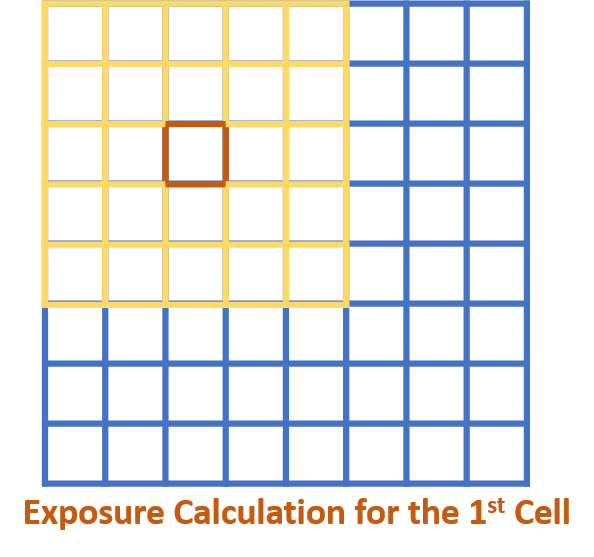


Figure 4.2 The propagation of calculation of an exposure map (Note that both the field and grid are larger in reality. This is for illustration.)

The last procedure calculates exposure. It uses a moving distribution field with 51, 151 or 251 cells. The distribution field moves across the pollution map one cell by one cell, as shown in *Figure 4.2*. Each replacement will start a calculation within the distribution field. The formula that returns the exposure value of one cell is:

Expr, c = α Conr, c ⮾ Fieldtype

Where: Expr, c stands for the exposure in µg/m3 inside the cell of row r and column c. On the right, Conr, c denotes the concentration level of the area that is centered around the (r, c) cell. The Fieldtype is one of the distribution fields that have a probability for each cell. It is expected that the concentration levels of cells inside one Fieldtype area are different. α is the ratio of indoor concentration level to adjacent outdoor concentration. As shown in *Figure 4.1*, the calculation operates, such as: in the cell of row r and column c, the exposure level is the sum of the product of neighboring concentration level inside the range of that type of field, and a visiting probability each cell assigned. It is a function of location, field type, and a set of concentration values that field cover. Adopting an indoor ratio α is a practical way to quantify the dispersion effect of air pollutants from outdoor to indoor. As the nitrogen dioxide map stands for outdoor concentration, the indoor concentration is yet to be calculated. An indoor ratio defines the portion of an outdoor pollutant that disperse into an indoor environment. Considering the long stay people tend to take in home, office, and school, the indoor concentration matters to our exposure calculation. In this study, α = 0.7 is used. The magnitude of this value is based on sensitivity analysis for indoor ratio from the agent-based model ran in the same study area that used 0.6, 0.7, and 0.8.[5]. The middle one of that range is taken.

We look into the scenario with and without road concentration. The assumption behind this is that in the time span of one year, an individual would not stay on the road for long. The influence of road traffic can be reflected by the concentration level of areas surrounding roads. With this assumption, the concentration map is processed into one without road input. This is done by overlapping the road raster file with the original concentration map and ignoring the value on roads. *Figure 2.4* shows two pollution maps, one with roads on the left, and one without roads on the right. Since this is an untested assumption, the original map is used for exposure calculation for comparison too. Subsection 5.1 will compare the difference of exposure map with and without road inputs. An important defined rule to mention here is: when a distribution field passes by areas without roads, the cells of probability without roads are set to be zero. After the calculation, the exposure value is divided by the sum of all new probability in that field. This is to make sure the probability adds up to 1. However, this seemingly fair rule will make a difference in exposure result, which is discussed in Subsection 5.1.2. Afterward, we refer to this rule as a "reweighted process."

To analyze the result and to understand the distribution fields, two reference maps from the agent-based model are compared with all results in Subsection 5.2. A distinct difference in mobility patterns is expected between the home maker's reference and the commuter's reference. The activity range of homemakers is theoretically smaller since they save the path to commuter between work and home. With r squared and scatterplots, we will compare the field setting of a good fit with high correlation andthe opposite.

**4.3 Reference Maps**

Two reference maps from Lu’s agent-based exposure model [36] are used as a benchmark for our results. One is made for the commuter; the other is for the homemaker. *Figure 4.3* shows the mobility timetable for individuals. Here the hourly pollution map set is coupled with the timetable hour by hour. Each cell means a product of hourly pollution cell value and the hour. On weekdays, the commutes take the path on the first row. They are either inside an indoor environment, which takes an averaged pollution values of a 60 meter \* 60 meter windows, or between home and workplace. The assignment of the workplace is stochastic, and the path between home and office is determined by the same algorithm afterward. The homemakers take the second row of timetables daily on weekdays. On weekends, two groups of individual shares the same timetable. They are assigned to a 1 km \* 1km averaged pollution intake from their home in the morning.

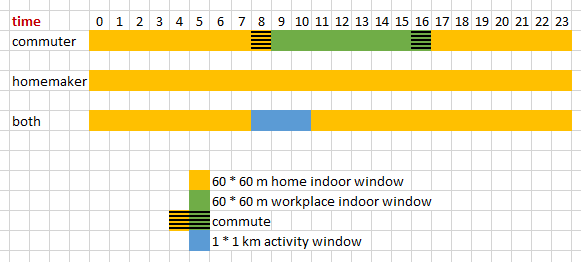
****

Figure 4.3 Timetable for commuter and homemaker in Lu's agent-based model [5].

*Figure 4.4* shows the idea of this study. In comparison to Lu's method, the time dimension is switched to a frequency dimension. The detailed explanation of distribution fields is on Subsection 3.2. We also do not assign either the individual's mobility profile (i.e., homemaker or commuter) nor the movement path ahead of the calculation. This study features a reduce in synthesized mobility. The exact trajectory is replaced with visiting frequency at a different distance to the center. In the following Subsections 5.2, we make use of the two reference maps and try to pick the optimal distribution fields for the two types of individuals.

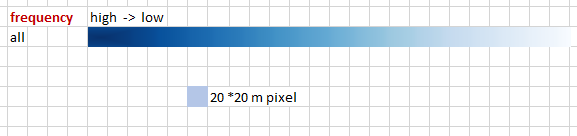
****

Figure 4.4 Visiting frequency profile in this study. "all" here refers to both the commuters and the homemakers during all days. This is to distinguish from Lu's model that separate weekend and weekdays, and commuter and homemakers.

**5 Result and Discussion**

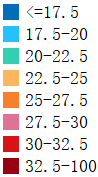
This chapter evaluates the exposure maps generated from the distribution fields and the annual pollution map. Subsection 5.1 starts with a comparison of exposure maps made from different distribution fields. A representative distribution field contributes largely to the final exposure calculation’s accuracy. Here the comparison will tell the impact of three field settings: field size (i.e., 2.5 km, 1.5 km and 0.5 km in radius), field type (i.e., seven distribution fields), and inclusion of roads (i.e., with or without road). If some fields yield similar results, it is desirable to utilize only one of them and leave others out for future study. If not, we try to pinpoint the causes of the differences here. Subsection 5.2 continues with a comparison between the modeled result and its reference. It first introduces the reference of the agent-based modeling exposure method [5]. Then the comparison includes both commuter' exposure and homemakers' exposure to see how different fields fit into a certain mode. Subsection 5.3 discusses the effect of the annual NO2 map. This effect is notable because the annual map is seldom used to map exposure. For exposure study with mobility input, it is common to couple pollution input with higher temporal resolution. Subsection 5.4 addresses the limitations of this method. The temporal variation is neglected, as plotted in the previous chapter. Besides, the grid shape of the distribution field is discussed.

Figure 5.1 Legend of NO2 Exposure (unit: μg/m3)

**5.1 Comparison of Exposure maps calculated from different distribution fields**

This Subsection compares the NO2 exposure maps generated from different distribution fields. It is worth investigating the common traits and differences between them. Essentially, this comparison tells the effect of three field settings, because the distribution fields differ from each other solely in these settings, and no other factors can define the distribution fields. The purpose of this comparison is not only to identify the settings' effect but also to understand the distribution field in Subsection 5.2. The order of analysis on field settings are field types, the inclusion of roads, and the field size. The result is listed in *Table 5.1* for field types and the inclusion of roads.

**5.1.1 Field Types**

To compare the impact of changing the field type, each row in *Table 5.1* shows the exposure results of one distribution field type. The most notable difference is how the shape of the 25-27.5 μg/m3 area changes (color: dark orange). For distribution fields with intensely-visited centers (i.e., exponential field and three Center Fields), the dark orange areas are spread in the shape of a star. On the contrary, for the distribution field with a moderate center (i.e., linear field), the dark orange area is round. This might be explained by the effect of high-frequency centers. Fields under this effect take in the immediate neighboring concentration level. As the immediate neighboring pixels in the city center are usually roads, the star-shaped areas are following the span of roads. In the same fashion, if the distribution fields are even or random, the closest neighbors have less influence to the center pixel, because the probability is more spread to far neighbors and they level down the exposure. Even including the roads in exposure calculation, meaning raising the closest neighboring values, would not bring many changes for even and random fields. Essentially, one pixel ‘s exposure largely depends on the strongest contributors. Different field types decide how much all the pixels in the field contribute to the center value. If the close neighbors contribute most, the city center will have a star-shaped area with the inclusion of road concentration. As a counterpart, if the close neighbors contribute less, the city center will have a round area.

From the image processing point of view, the setting of field types are variations between a mean filter and a contour filter. This link between distribution fields and image filter helps us to understand: Firstly, the specific function of center-intense fields with two types of pollution maps. Secondly, the general function of even fields with all pollution maps. Here, two types of pollution maps refer to maps with and without roads. Roads are substantially higher in pollution value than their surrounding areas, which are seen as contours in image processing. A field with an intense center (i.e., exponential fields and all three center fields) is similar to a contour filter. It can visually strengthen the areas of large differences to their neighbors. In our case, the pollution map with roads assembles an unprocessed image with contours and is expected to be strengthened by a contour filter. On the contrary, the pollution map without road assembles an unprocessed image without contours. So, after the calculation, there are no star-liked areas in the city center. The second function found here is that an even field equals to a mean filter. The function of a mean filter is to replace each cell’s value with the average value of the cell’s neighboring cell’s value. With this function, unrepresentative cells of their surroundings are replaced. The star-liked city center part is missing in with or without roads input. That is the effect of the mean filter that dampens the high value from roads and results in one bulk of area with similar value at the city center.

Table 5.1 NO2 Exposure result of the 1.5 km-radius fields (unit: μg/m3, they share the legend in *Figure 5.1*)

|  |  |  |
| --- | --- | --- |
| Field Type | Exposure without roads’ NO2 concentration | Exposure with roads’ NO2 concentration |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

The scatterplots and correlation index below attempt to tell the similarity between distribution fields. Here random 2000 cells, instead of all cells, are chosen to make the plots clear. The plots here are with the inclusion of roads because the exposure map from them is closer to the reference map in Subsection 5.2. Besides, since the center + even field, center + narrow rings field and center + thick rings field give similar results seen from *Table 5.1*, and random field and even field also give a similar result, four types of fields are chosen to represent all instead.

In *Figure 5,2* to *Figure 5.4*, both the scatterplots and the correlation index tell one major similarity: for a field size of all (i.e., R= 0.5/1.5/2.5 km), linear fields and even fields are the closest. Their scatter plots are focused, and the slope is close to 1. Also, this pair has the highest correlation index for all sizes of comparison. The other pairs suggest high similarity too from the correlation index, but the scatterplots tell their discrepancy with more spread data points.

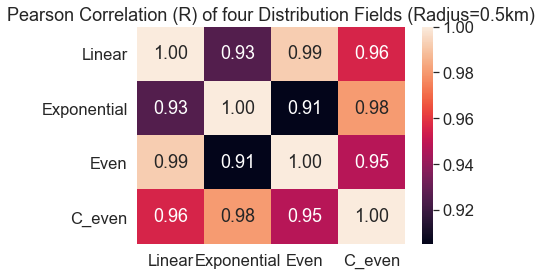
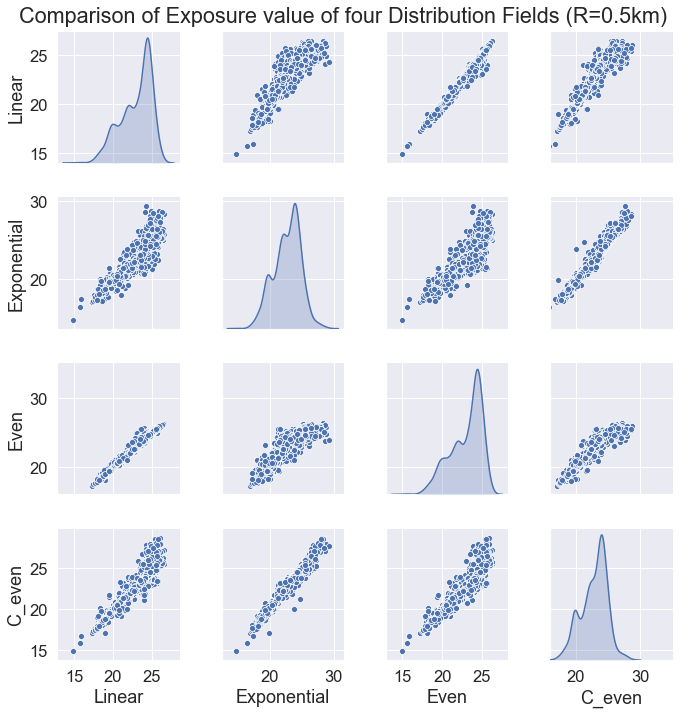


Figure 5.2 Scatterplots(left) and correlation coefficient(right) of four distribution fields (R = 0.5 km)

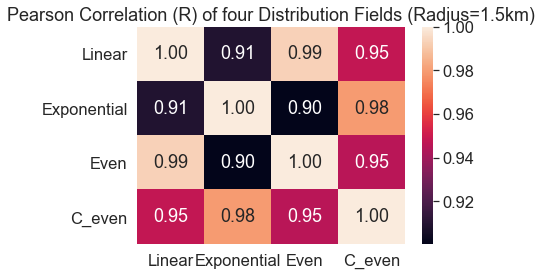
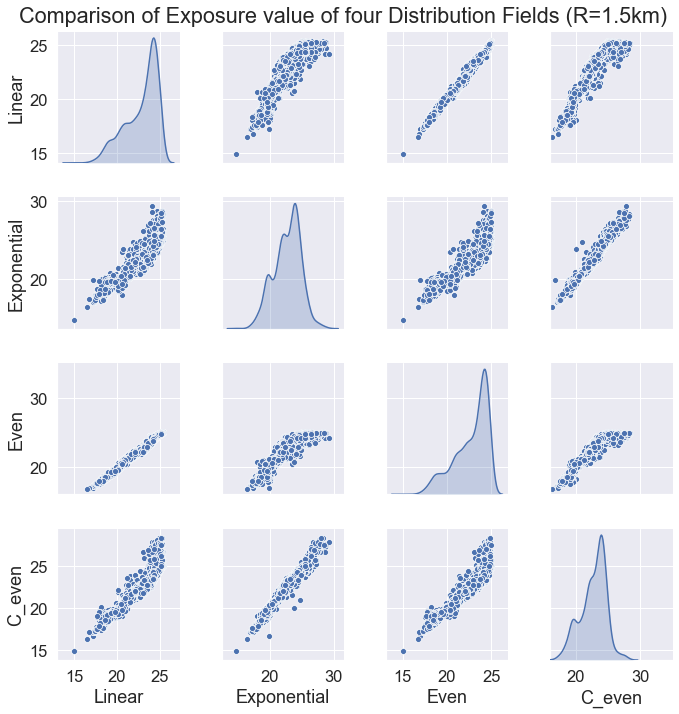


Figure 5.3 Scatterplots(left) and correlation coefficient(right) of four distribution fields (R = 1.5 km)

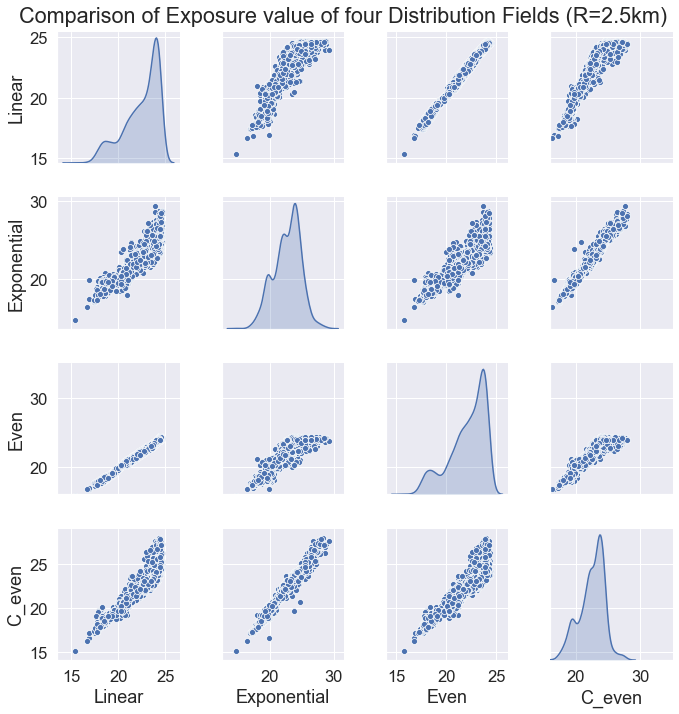
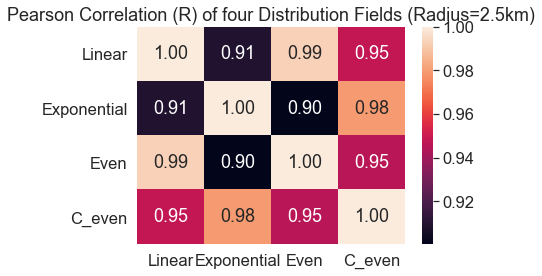
 

Figure 5.4 Scatterplots(left) and correlation coefficient(right) of four distribution fields (R = 2.5 km)

**5.1.1 Inclusion of Roads**

The two columns of exposure maps in *Table 5.1* differ from each other in the inclusion of roads. To see the function of the inclusion of roads, plots on the first column does not include NO2 input on roads, while the second column does. The difference maps between the second column and the first column are below in *Figure 5.5* to *Figure 5.7*. In theory, since the road areas are substantially higher in pollution level than other areas, the exposure map with roads is supposed to have higher value everywhere than the map without roads. Surprisingly, this is not seen in the Exponential Field and the three Center Fields. They have plenty of red areas, meaning these areas have lower exposure level with roads than without roads. This is the joint effect of the reweighted process of distribution fields over areas without roads and the center-intense fields. When the distribution fields overlap areas without roads, the corresponded cells are reset to zero and are neglected in the exposure calculation. The participating cells are weighted again by the remaining sum of probability. This makes high probability areas in the center-intense fields to gain an even higher probability and eventually result in higher exposure levels than the exposure calculated with roads. This is, unfortunately, a limitation of the reweighted process but also a practical way to represent the case with the exclusion of roads.

The initial purpose of both including and excluding roads is to examine which one is closer to the reference map and essentially to reality. Both options can be representative to some extent. It is certain that most citizens have visited roads annually, which requires the inclusion of roads. On the other hand, the majority spend less than two hours on the road daily, which questions the importance of roads. Reference maps in *Figure 5.5* to *Figure 5.7 s*hows that the inclusion of roads makes similar exposure results with them. This is especially seen in the star-shaped areas in the city center, which is not reflected in the exposure maps without roads. Therefore, even the time spent on roads is minimal, but the pollution input from roads is important to long-term exposure calculation.

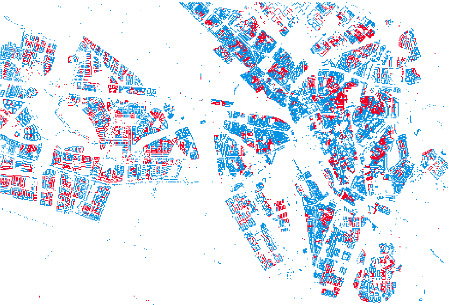


Figure 5.5 Difference of the map with roads and the map without roads (Left to right: Linear Field, Exponential Field, Even Field. Color blue means positive, and color red negative.)

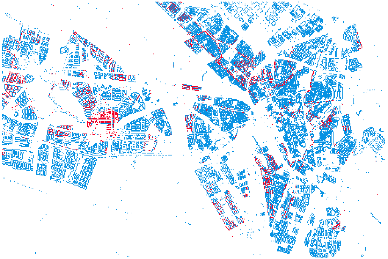
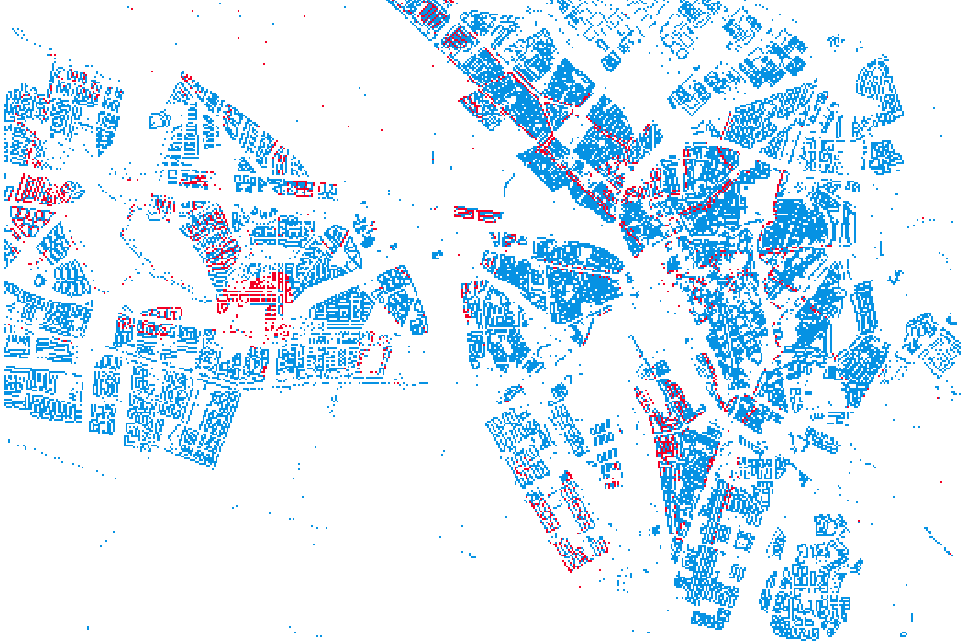
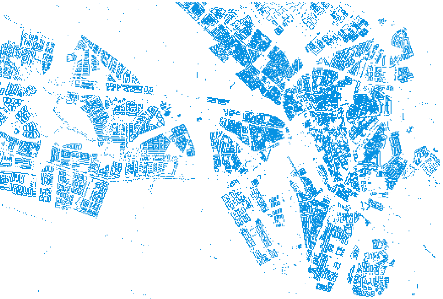


Figure 5.6 Difference of the map with roads and the map without roads (Left to right: Random Field, Center + Even Field, Center + Thick Rings Field. Color blue means positive, and color red negative.)

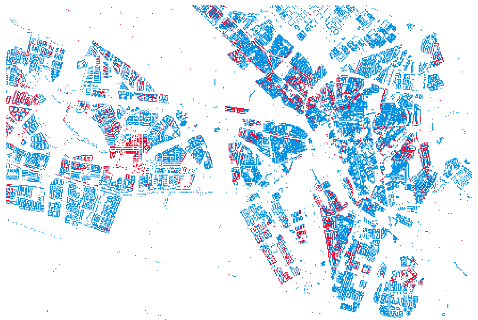
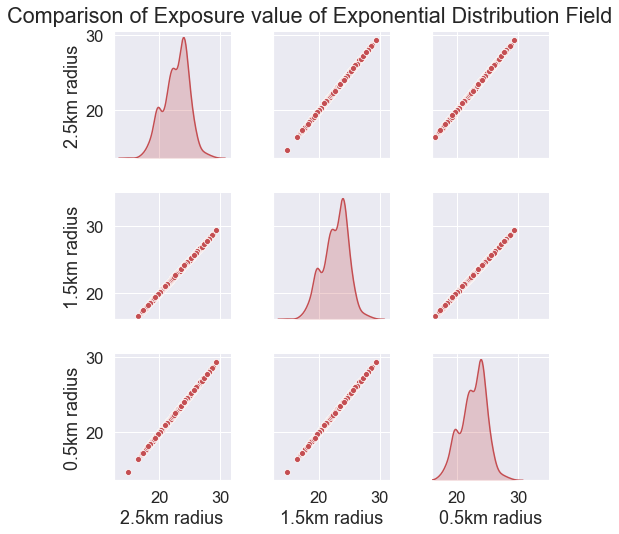
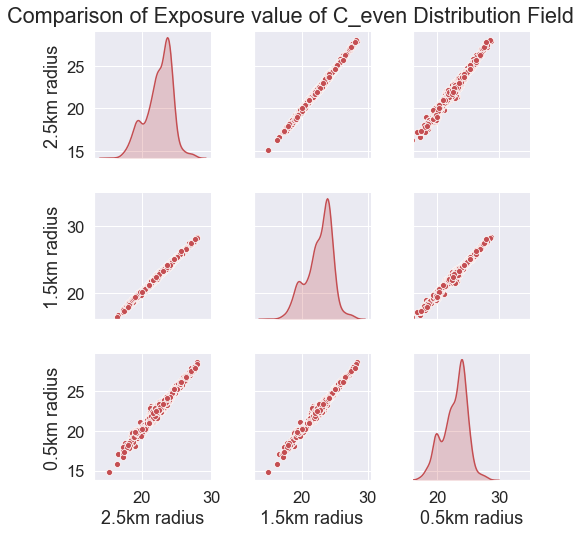


Figure 5.7 Difference of the map with roads and the map without roads (This is from the Center + Narrow Rings Field Color blue means positive, and color red negative.)

**5.1.3 Field Size**

The last field setting to discuss is the function of field size. Theoretically, applying several field sizes equals to assuming different activity range. Larger fields carry the physical meaning of larger individual mobility space, vice versa. The field sizes examined here have a radius of 0.5, 1.5. 2.5 kilometers.

From the scatterplots in *Figure 5.8*, changing field size affects each field type differently. As it did above, here we use Even Field to represent both the Even Field and the Random Field, and one Center Field to represent all due to their similarity. Only 2000 points are chosen to make clearer plots. The Exponential Field is insensitive to change in size. The exposure distribution of the three sizes is highly identical. And the points in scatterplots align with each other, showing the slope of nearly one. This is saying the close distribution of visiting probability over distance to the center of three sizes does not make a difference; even the rings in the far areas have more cells to add probability. The Center Field is more sensitive to the size of fields. The difference between the biggest field and the smallest field can be seen from the deviation of points from the straight line of other points. Both the Linear Field and the Center Field are sensitive to the size of fields. All of their pairs have more distributed points in comparison to both the Exponential Field and the Center Field. A unique feature of these two is that the increase in size will smoothen the exposure distribution. This is due to the fact that larger the input areas, more deviation values will be leveled out.



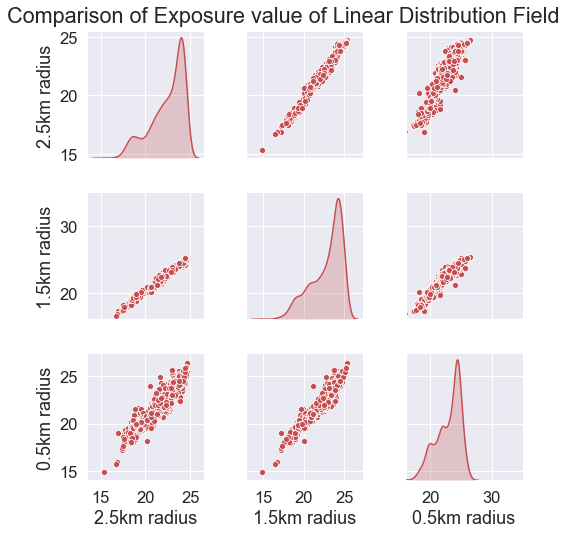
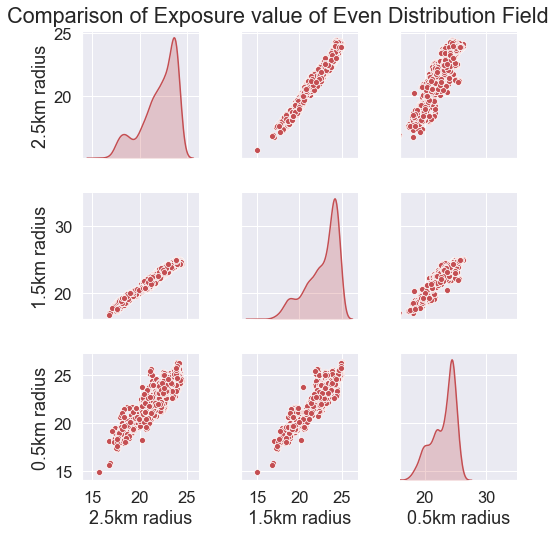
 

Figure 5.8 Scatterplots of three distribution fields of different sizes (From left to right and top to down: Exponential Distribution Field, Center + Even Field, Linear Field, Even Field)

**5.2 Comparison with the reference map**

This section compares our result of exposure maps with two reference maps. One is made for the commuter; the other is for the homemaker. The best and worse fit for both references are discussed. The pollution input is the map with roads’ value.

**5.2.1 Comparison with the commuter’s map**

Here we roughly define the round area with star-shaped orange exposure peak in the reference map as the city center and the rest as the suburb. *Figure 5.9* is the commuter’s reference map. Visual observation between *Table 5.1* and the commuter's map tells that most of the modeled results are close to the reference map in the suburb areas, but the city center area is overestimated mostly. Adequate mapping should show a star-shaped orange area with high exposure values.

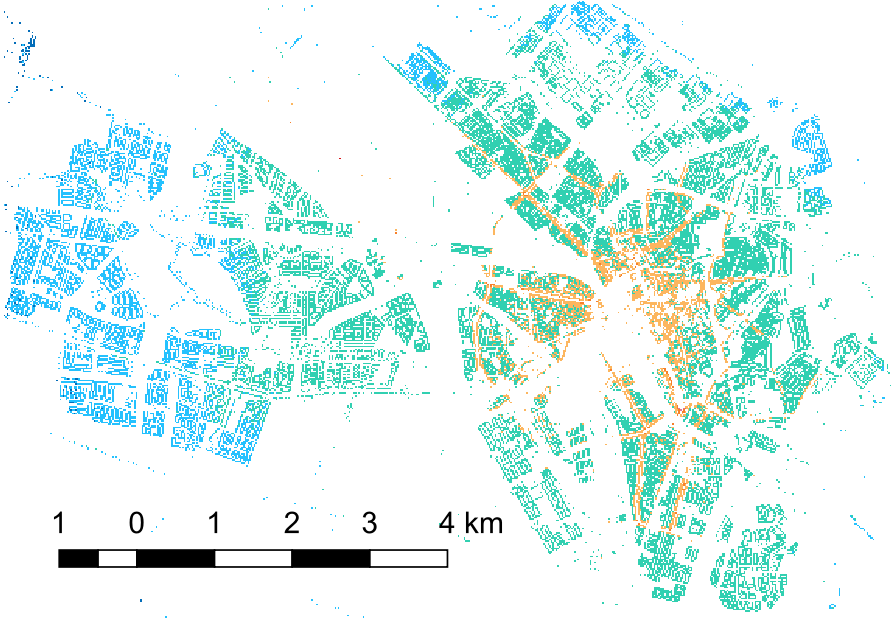
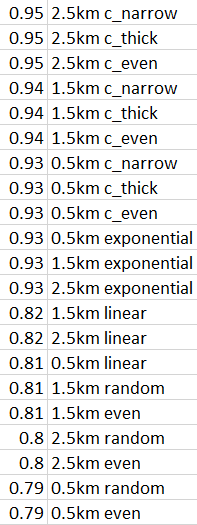


Table 5.2 Rank of the coefficient of correlation for commuters. reference

Figure 5.9 Exposure map for commuters from Lu's study [5]. The color scale is adjusted into Figure 5.1 to compare with our results.

From *Table 5.2*, high correlations are from central-intense fields (i.e., Exponential Field and three Center-Field) with the coefficient higher than 0.93. Gradual fields (i.e., Linear Field, Random Field, Even Field) all give a correlation coefficient lower than 0.85, significantly lower than the central-intense fields. Another trait is that for central-intense fields, larger the field size, the stronger the correlation. This trait is likely to come from a wide range of places commuter visits. Including more areas with a larger radius helps to represent the commuter's trajectory. *Figure 5.10* shows the comparison pair with a high correlation. The agreement comes from the commuter’s feature of large movement outside and a long stay at home, in comparison to the homemaker. However, the range of reference maps and the two peaks are not well-reflected in the three fields' results, which have a higher upper range and three peaks instead. From the result in *Table 5.1*, it is seen that the overestimated value majorly occurs in the city center regions. This is likely to be caused by the discrepancy between actual activity range and the distribution fields’ range. If the exposure of high concentration level areas is overestimated, then the weight of the distribution field should be given to areas with a lower concentration level. This can be achieved by expanding the range of distribution field for city center areas, or high concentration areas in general. This is in line with the increasing fitness when the range gets larger in *Table 5.2*. The neglect of temporal resolution in exposure mapping can also be the cause of the overestimation in the city center. The city center's concentration level fluctuates largely daily, which can be seen in Subsection 2.4. But the pollution input for exposure calculation takes one value for the whole year. This might raise the final exposure value to some extent.

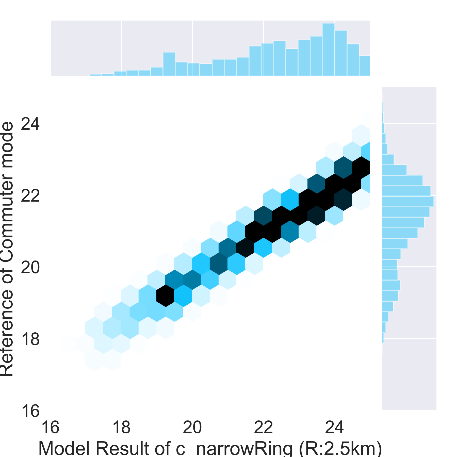
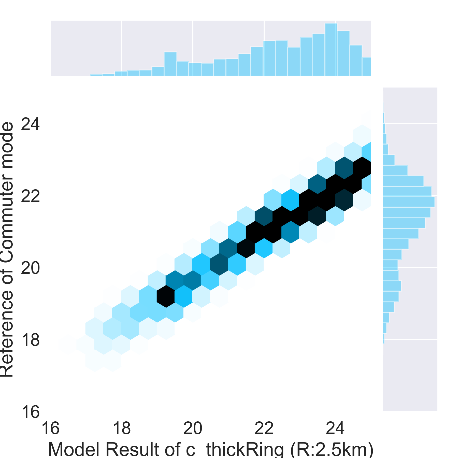
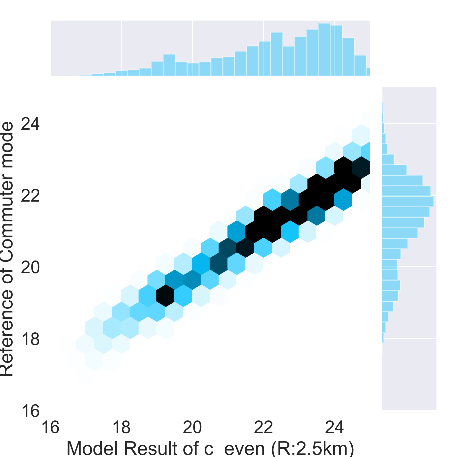


Figure 5.10 Exposure results with high correlation with commuter's mode reference

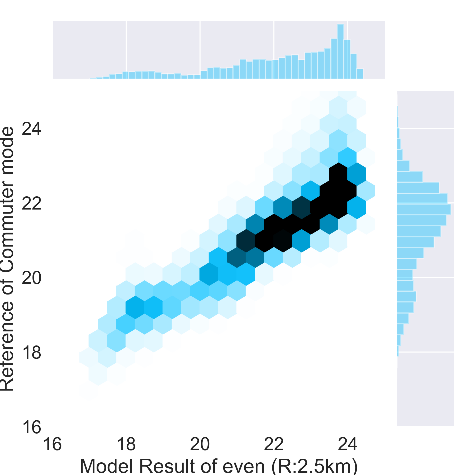
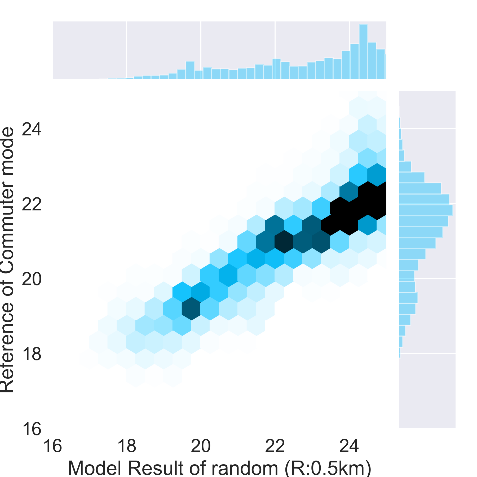
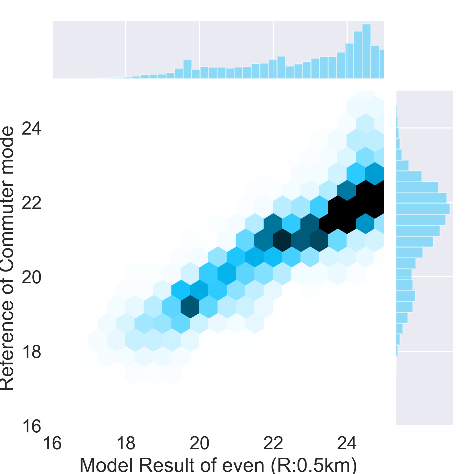


Figure 5.11 Exposure results with lower correlation with commuter's mode reference

**5.2.2 Comparison with the home maker’s map**

From *Table 5.3*, in general, the home maker's reference has a higher correlation with the distribution fields' result than the commuter's, except for the lowest three fields. In *Figure 5.13*, The exposure distribution for homemakers is narrower due to its lower mobility. Higher mobility suggests exposure input from more places and lower mobility the opposite. Although the correlation is high, the modeled values are mostly lower than the reference one with a similar difference, which means adding an intercept after distribution field operation might improve the fitting.

The shared trend with commuters' map is noted too. For homemakers, high correlations are from central-intense with the coefficient higher than 0.97. Gradual fields (i.e., Linear Field, Random Field, Even Field) all give a correlation coefficient lower than 0.83, significantly lower than the central-intense fields. The difference is that the homemakers' comparison pairs do not show an increasing correlation when the size is getting larger as it does in the commuter's comparison pairs. The reason behind this is that the lower mobility of homemakers that makes exposure less sensitive to field size.

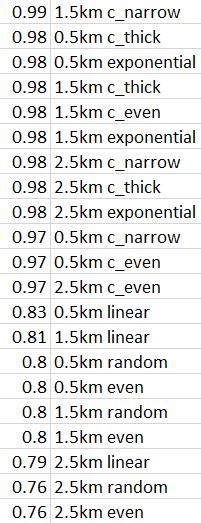
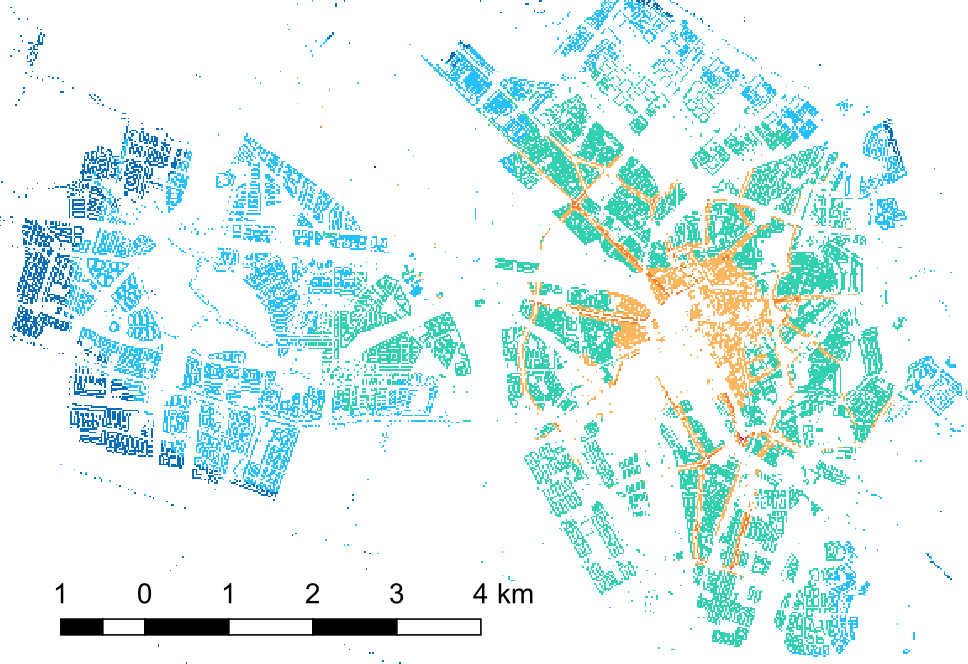
****

Table 5.3 Rank of the coefficient of correlation for homemakers’ pairs

Figure 5.12 Exposure map for Home Makers from Lu's study [5]. The color scale is adjusted into Figure 5.1 to compare with our results.

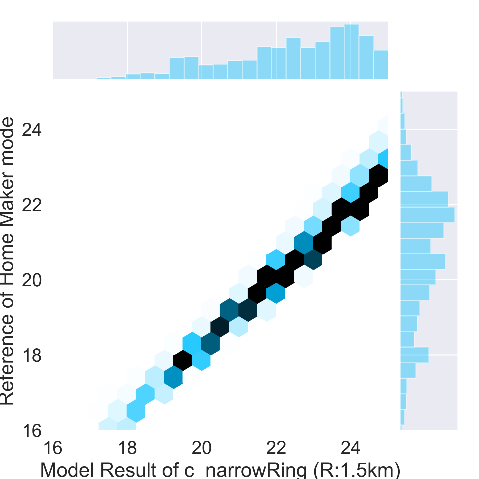
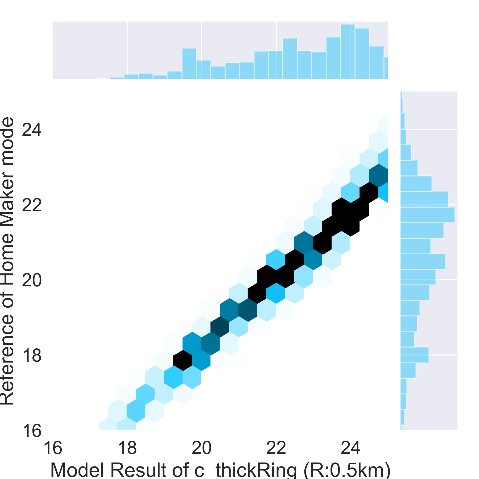
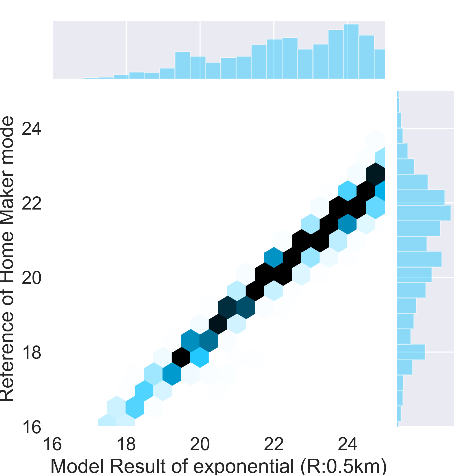


Figure 5.13 Exposure results with high correlation with home maker's mode reference

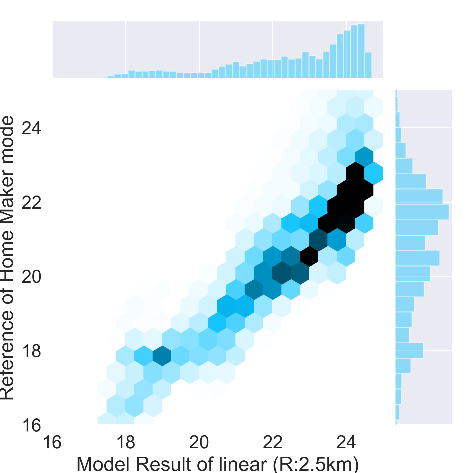
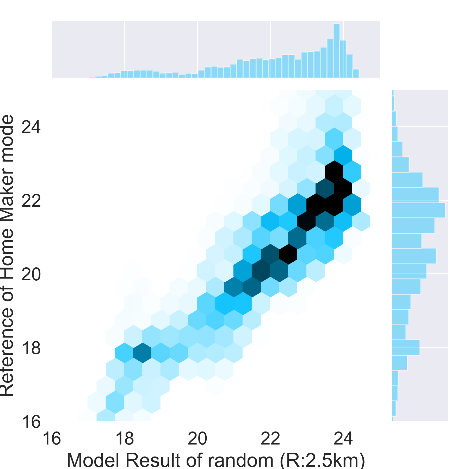
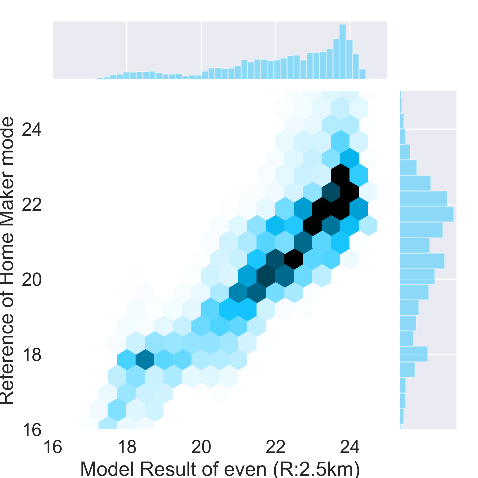


Figure 5.14 Exposure results with lower correlation with home maker's mode reference

**5.3 Limitations**

Three limitations are noted. One limit of the reweighted process is explained in Subsection 5.1.1; one limit concerns the effect of annual pollution map, whose alternatives are hourly, daily, monthly, seasonal maps; the other limit is from the square setting of distribution fields, whose alternative is round distribution fields.

Secondly, the effect of the annual pollution map is notable because it is common to adopt pollution input with the higher temporal resolution, while an annual map is seldom used to map exposure. The frequency one visits a place does not solely determine the exposure. The concentration of one place when the individual is visiting is vital too. A telling example is we will take in more pollutants during peak hour in the city center, but the off-peak hour will not harm us at the same place. Using an annual map is neglecting the variation of pollution concentration over time, and affects the long-term exposure calculation. The temporal variation is studied in Subsection 2.4. If some discrepancies between the exposure calculation and reality are detected, this can be one source. This limitation is mentioned in a study that stresses spatiotemporal pollution variation in exposure assessment, too [66]. However, if we raise the time resolution of pollution input, making several frequency profiles for different pollution time series, the calculation difficulty will increase, and the effect of distribution field settings will get unclear in the joint effects of pollution input and mobility input.

Thirdly, since one individual should not favor any direction for activity from the center of his or her home, the distribution fields are supposed to be round, meaning the radius from center to the border is identical in any direction. The squared fields are chosen for the sake of brevity. In calculation, we first assign a set of probability to a line of cells and then apply it to lines of different directions. The number of cells from the center cell to the diagonal end and the number of cells from the center cell to the vertical or horizontal end are the same. If the round distribution fields are applied, this share of probability profile along a line for all directions is lost and will increase computation cost, especially when multiple field settings are tested.

**5.4 Sub-questions & Implications**

Firstly, in Subsection 2.4, the temporal variation of NO2 concentration in Utrecht is examined with the result from the hourly LUR model calculation. A strong seasonality with a high concentration in winter and low concentration in summer is salient. A great daily difference in the mean city concentration is notable too. Since an annual map of pollution is used for exposure calculation, these temporal variations with a higher frequency are neglected and can pose uncertainty to the result. The overestimation of our exposure results in city center in comparison to the commuter’s map may be caused by this. Till now, the exact impact of the annual pollution map remains unsolved, but the magnitude of the temporal variation is examined. We argue that the optimal exposure calculation has to notice this source of uncertainty.

Secondly, the inclusion of human mobility is important to exposure calculation as previous studies have pointed out, but the mobility input differs largely from GPS, surveys to agent-based models. In this study, distribution fields are designed as a universal mobility input to global exposure mapping. The design of distribution fields mostly stems from the Levy Flight model that notices the higher frequency of short-distance trips and the lower frequency of long-distance trips. It also learns from the concept of windows in Lu's agent-based model that defines homogeneous areas of concentration level as windows. Subsection 3.2 introduces the setting of distribution fields.

Thirdly, to understand distribution fields, the effects of three settings are investigated. Field types can be classified into center-intense fields, including three Center Field and one Exponential Field, and gradual fields, including Even Field, Random Field, and Linear Field. center-intense fields can strengthen the input of strong concentration on roads, as the contour filter in image processing. But when there is no strong concentration identical to contours, the center-intense fields will not result in exposure maps with strong star-shaped centers. Unlike center-intense fields, the gradual fields level out concentration input in the neighborhood and will not generate star-shaped centers in any case. If the application purpose is about the maximum exposure, then the center-intense fields suit better. Apart from field type, the inclusion of roads largely changes the exposure result too. A surprising finding is that the exclusion of roads, which are of high concentration than other areas most of the time, can raise exposure in some areas for center-intense fields. This is due to the limit of reweighting the distribution fields and giving much higher weights to central cells in center-intense fields. As the last setting to investigate, the field size is influential to gradual fields, but not to center-intense fields. Because gradual fields take in more diverse values from the expanding range, but the center-intense fields are determined by the few numbers of cells near the center largely, the trivial change in the border side does not affect much.

Fourthly, both to the commuters and the homemakers, the center-intense fields correlate best with the reference, and gradual fields are less correlated. The difference between the two is that: the homemaker's exposure distribution range is better represented, and there is a constant offset that can be compensated by an intercept after distribution field calculation. The reason for the offset is unsolved. Raising the indoor ratio would not help because this will change the slope of the scatterplot and might create a bigger offset. Besides, the commuters' exposure gets more represented when the field sizes grow, which is in line with the travel profile of a commuter.

There are two strengths of the distribution field in the context of exposure assessment study. The first one is the usage of cells or microenvironments. As pointed out in one study, in order to assess air pollution exposure, the estimation of time spent by individuals in various microenvironments is critical [67]. The distribution fields reflect the visiting frequency for each cell inside a field, which is in line with this finding. The second strength is its low mobility data input. Alternative mobility input can be GPS and mobile phone signal, whose application raises research cost and are constrained by data privacy. Synthesized mobility input is also potential. For instance, the agent-based stochastic trajectory in Lu’s study [36] requires low mobility data input and certain computation, which serves as the reference map in this study.

So far, to our knowledge, similar ideas of using visiting frequency set as mobility input for exposure assessment are lacking. The idea of distribution fields is inspired by the Levy Flight or the Random Walk models [29] that notices the relationship between travel frequency and travel distance. A human mobility study using GPS mining suggests most people adhere to a simple and reproducible pattern [48]. In our study, this is adapted to the setting of "more visited neighboring areas of ones' home" and "less-visited areas far away from one's' home." The idea also borrows from the concept of microenvironments and windows in Lu's study [36].

**6 Conclusion**

To ultimately understand the relationship between pollution and diseases, this study realizes a novice spatiotemporal aggregation method for NO2 exposure calculation with low mobility data input. From the sufficient correlation between our result and the reference, we conclude that this method is representative in mapping individual exposure of NO2 for city Utrecht and is potential for global study. The optimal distribution fields for both commuters and homemakers are center-intense fields: three Center Fields and the Exponential Field, which answers the research question.

This study firstly calculates an annual NO2 map for the study area Utrecht. The map stems from the calibrated Land Use Regression model that takes geographic features as input and results in hourly NO2 pollution for a year. The spatial resolution reaches 20 meters for all hours. The annual map takes the mean of all hourly maps. Secondly, we use distribution fields to represent the human mobility pattern. Different distribution fields assumed certain visiting probability for the neighboring areas of one living spot. Inside this area, each place has a probability of getting visited by the citizen. Lastly, on the procedure, the individual exposure map is generated from weighing the surrounding pollution value for each living cell with the weight of the distribution field.

Further study to improve distribution fields can be done on 1) Raising field size for exposure of commuter's mode since the current map has an overestimated center and require input from areas of lower concentration. 2) Adding an intercept after the exposure calculation with distribution fields for homemakers' mode. This is based on the high correlation and constant offset between model results and the reference. Notable limitations include: 1) The reweighting process is done on center-intense fields that results in lower exposure when the roads are included. 2) The daily and seasonal variation of air pollution concentration covered by an annual map, which may cause unrealistic predictions. 3) The unrepresented concept of round fields is replaced with square fields in order to set up fields faster.

**Appendix**

**1 Script for Distribution Field**

#---------------------------------------------

# PROBABILITY FIELD (3 SIZES AND 7 SHAPES)

#---------------------------------------------

import numpy as np

linesize = [126, 76, 26] # Accordingly, the diameter is 251\*20(5km), 151\*20(3km), 51\*20(1km)

savepath = 'D:/geoscience/MScResearch/Shape/'

# shape 1: linear

shape1 = np.zeros((3,251,251))

line = np.zeros((3, 126))

line[0][:126] = range(126, 0, -1)# Large Field

line[1][:76] = range(76, 0, -1)# Medium Field

line[2][:26] = range(26, 0, -1)# Small Field

for size in range(3):

for row in range(251): #from 0 to 250

for col in range(251):

idx = max(abs(row + 1 - 126), abs(col + 1 - 126))

if idx < linesize[size]:

shape1[size][row][col] = line[size][idx]

shapesum = np.sum(shape1[size])

shape1[size] = shape1[size] / shapesum

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape1.asc',shape1[size])

# shape 2: near-exponential

import math

shape2 = np.zeros((3,251,251))

line = np.zeros((3, 126))

for i in range(0, 126):

line[0][i] = math.e\*\*(-i)# Large Field

for i in range(0, 76):

line[1][i] = math.e\*\*(-i)# Medium Field

for i in range(0, 26):

line[2][i] = math.e\*\*(-i)# Small Field

for size in range(3):

for row in range(251): #from 0 to 250

for col in range(251):

idx = max(abs(row + 1 - 126), abs(col + 1 - 126))

if idx < linesize[size]:

shape2[size][row][col] = line[size][idx]

shapesum = np.sum(shape2[size])

shape2[size] = shape2[size] / shapesum

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape2.asc',shape2[size])

# shape 3: even

shape3 = np.zeros((3,251,251))

shape3[0] = 1/(251\*251)

shape3[1][50:201,50:201] = 1/(76\*76)

shape3[2][100:151,100:151] = 1/51/51

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape3.asc',shape3[size])

# shape 4: no-shape/ random

shape4 = np.zeros((3,251,251))

import random as rd

for row in range(251):

for col in range(251):

shape4[0][row, col] = rd.randint(0, 100) \* 0.01

for row in range(50, 201):

for col in range(50,201):

shape4[1][row, col] = rd.randint(0, 100) \* 0.01

for row in range(100, 151):

for col in range(100, 151):

shape4[2][row, col] = rd.randint(0, 100) \* 0.01

for size in range(3):

shapesum = np.sum(shape4[size])

shape4[size] = shape4[size] / shapesum

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape4.asc',shape4[size])

# shape 5

# The center pixel has a probability of 60%, the rest have the same probability and adds up to 40%.

shape5 = np.zeros((3,251,251))

shape5[0] = 0.4/(251\*251-1)

shape5[1][50:201,50:201] = 0.4/(151\*151-1)

shape5[2][100:151,100:151] = 0.4/(51\*51-1)

shape5[0][125, 125] = 0.6

shape5[1][125, 125] = 0.6

shape5[2][125, 125] = 0.6

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape5.asc',shape5[size])

# shape 6

# The center pixel has a probability of 60%, the rest have the same probability and adds up to 40%.

# This shape differs from shape 5 in the way that each ring has an equal total probability.

# While shape 5 has different total probability for each ring.

shape6 = np.zeros((3,251,251))

ring = [125, 75, 25]

for size in range(3):

ringprob = 0.4/ring[size]

for row in range(125 - ring[size], 126 + ring[size]):

for col in range(125 - ring[size], 126 + ring[size]):

pixels = 8\*max(abs(row - 125), abs(col - 125))

if pixels != 0:

shape6[size][row][col] = ringprob/pixels

shape6[0][125, 125] = 0.6

shape6[1][125, 125] = 0.6

shape6[2][125, 125] = 0.6

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape6.asc',shape6[size])

# shape 7

# The center pixel has a probability of 60%, the rest have the same probability and adds up to 40%.

# This shape has larger rings.

shape7 = np.zeros((3,251,251))

ringprob = 0.4

# large size

ring0 = [[0, 250], [25, 225], [50, 200], [75, 175], [100, 150], [125, 125]]

ring0prob = ringprob/5

# Calculate the probability for each pixel inside each ring

for ringnum in range(5):

pixel = (ring0[ringnum][1] - ring0[ringnum][0] + 1)\*\*2 - (ring0[ringnum + 1][1] - ring0[ringnum + 1][0] + 1)\*\*2

prob = ring0prob / pixel

shape7[0][ring0[ringnum][0]:ring0[ringnum][1], ring0[ringnum][0]:ring0[ringnum][1]] = prob

# Medium size

ring1 = [[0, 150],[25, 125],[50, 100], [75, 75]]

ring1prob = ringprob/5

# Calculate the probability for each pixel inside each ring

for ringnum in range(3):

pixel = (ring1[ringnum][1] - ring1[ringnum][0] + 1)\*\*2 - (ring1[ringnum + 1][1] - ring1[ringnum + 1][0] + 1)\*\*2

prob = ring1prob / pixel

shape7[1][ring1[ringnum][0] + 50 : ring1[ringnum][1] + 50, ring1[ringnum][0] + 50 : ring1[ringnum][1] + 50] = prob

# Small size

ring2 = [[0, 50], [25, 25]]

ring2prob = ringprob

# Calculate the probability for each pixel inside each ring

for ringnum in range(1):

pixel = (ring2[ringnum][1] - ring2[ringnum][0] + 1)\*\*2 - (ring2[ringnum + 1][1] - ring2[ringnum + 1][0] + 1)\*\*2

prob = ring2prob / pixel

shape7[2][ring2[ringnum][0] + 100 : ring2[ringnum][1] +100, ring2[ringnum][0] + 100 :ring2[ringnum][1] + 100] = prob

shape7[0][125, 125] = 0.6

shape7[1][125, 125] = 0.6

shape7[2][125, 125] = 0.6

for size in range(3):

np.savetxt(savepath+ str(size)+ 'shape7.asc',shape7[size])

from mpl\_toolkits import mplot3d

import matplotlib.pyplot as plt

##shape2

zline = np.flip(shape2[0,0:126,125])

xline = [int(i) for i in range(0,126)]

yline = np.ones(126) \*2

ax.plot3D(xline, yline, zline, 'gray')

zline = np.flip(shape2[1,126-76:126,125])

xline = [int(i) for i in range(0,76)]

yline = np.ones(76) \*2

ax.plot3D(xline, yline, zline, 'red')

zline = np.flip(shape2[2,126-26:126,125])

xline = [int(i) for i in range(0,26)]

yline = np.ones(26) \* 2

ax.plot3D(xline, yline, zline, 'blue')

#

#plt.show()

import seaborn as sns; sns.set()

import matplotlib.pyplot as plt

fmri = sns.load\_dataset("fmri")

palette = sns.color\_palette("mako\_r", 6)

#251\*20(5km), 151\*20(3km), 51\*20(1km)

ax = sns.lineplot(x=[int(i) for i in range(25,126)],

y= np.flip(shape2[0,0:126-25,125]),

legend='brief', label='R = 2.5km')

ax = sns.lineplot(x=[int(i) for i in range(25,76)] ,

y=np.flip(shape2[1,126-76:126-25,125]),

legend='brief', label='R = 1.5km')

ax = sns.lineplot(x=[int(i) for i in range(25,26)] ,

y=np.flip(shape2[2,126-26:126-25,125]),

legend='brief', label='R = 0.5km')

ax.set(xlabel='number of cell from the center',

ylabel='probability')

#plt.legend()

plt.show()

**2 Script for Exposure**

import numpy as np

import matplotlib.pyplot as plt

c1 = np.loadtxt('D:/geoscience/MScResearch/Analysis0930/s4mean20.asc',

ndmin = 1, skiprows = 6)

# The exposure result set of c1 has been saved in Analysis0930/exWithRoad

c2 = np.loadtxt('D:/geoscience/MScResearch/Analysis0930/s4mean20noroad.asc')

# The exposure result set of c2 has been saved in Analysis0930/exWithoutRoad

plt.imshow(c2, cmap='Blues')

plt.colorbar()

plt.show()

#

ref = np.loadtxt('D:/geoscience/MScResearch/Analysis0930/reference.asc',

ndmin = 1, skiprows = 6)# The null value is -9999

ex = np.zeros((645, 771))

header = "ncols 771\n"

header += "nrows 645\n"

header += "xllcorner 126432\n"

header += "yllcorner 448710\n"

header += "cellsize 20\n"

header += "NODATA\_value 0"

cc = c1

for shape in range(1, 8):

for size in range(3):

print('shape:{0}, size:{1}'.format(shape, size))

path = 'D:/geoscience/MScResearch/Shape/{0}shape{1}.asc'.format(size, shape)

field = np.loadtxt(path, ndmin = 1)

for row in range(645):

for col in range(771):

if ref[row][col] != -9999:

no2 = cc[row - 125:row + 126, col - 125:col + 126]

if np.isnan(no2).shape[0] != 251 or np.isnan(no2).shape[1] !=251 :

continue

time = np.multiply(no2, field)

kernel = np.sum(time)

prob = np.sum(field[no2 !=0])

ex[row][col] = kernel / prob \*0.7

np.savetxt('D:/geoscience/MScResearch/Analysis0930/exWithRoad/size{0}shape{1}.asc'.format(size, shape),

ex, header = header, comments = '')

print('saved.')

**3 Script for Evaluation**

# plot Difference and calculate RMSE

import os.path, sys

import numpy as np

import pandas as pd

from numpy import array

import matplotlib.pyplot as plt

from scipy import stats

from sklearn.metrics import mean\_squared\_error

import seaborn as sns

header = "ncols 771\n"

header += "nrows 645\n"

header += "xllcorner 126432\n"

header += "yllcorner 448710\n"

header += "cellsize 20\n"

header += "NODATA\_value 0"

field = ['linear (R:2.5km)', 'linear (R:1.5km)', 'linear (R:0.5km)',

'exponential (R:2.5km)', 'exponential (R:1.5km)', 'exponential (R:0.5km)',

'even (R:2.5km)', 'even (R:1.5km)', 'even (R:0.5km)',

'random (R:2.5km)', 'random (R:1.5km)', 'random (R:0.5km)',

'c\_even (R:2.5km)', 'c\_even (R:1.5km)', 'c\_even (R:0.5km)',

'c\_thickRing (R:2.5km)', 'c\_thickRing (R:1.5km)', 'c\_thickRing (R:0.5km)',

'c\_narrowRing (R:2.5km)', 'c\_narrowRing (R:1.5km)', 'c\_narrowRing (R:0.5km)']

path = "D:/geoscience/MScResearch/Analysis0930/exWithRoad"

files = os.listdir(path)

allrmse = []

myresult = np.zeros(( 645, 771))

allx = np.zeros((36739)) # allx: the model result in one row

#refpath = "D:/geoscience/MScResearch/Analysis0930/reference.asc"

refpath = 'D:/geoscience/MScResearch/Data/ref/largeHome5.asc'

refExpo = np.loadtxt(refpath, skiprows =6)

sns.set(font\_scale=1.5)

for file in files:

if file.endswith('asc') ==False:

continue

size = int(file[4])

fieldtype = int(file[10])-1

code = fieldtype\*3 +size

title = field[code]

# resultpath = 'D:/geoscience/MScResearch/Analysis0930/exWithoutRoad/' +file

print(title)

resultpath = 'D:/geoscience/MScResearch/Analysis0930/exWithRoad/' +file

mresult = np.loadtxt(resultpath, skiprows = 6)

for row in range(645):

for col in range(771):

if refExpo[row][col] == -9999:

mresult[row][col] = -9999

if mresult[row][col] == 0:

refExpo[row][col] = 0

myresult = mresult

differ = np.subtract(mresult, refExpo)

y = np.reshape(refExpo, (771\*645))

x = np.reshape(myresult, (771\*645))

x2 = []

# Take out the pixels with no value.

for xitem in x:

if xitem != -9999 and xitem != 0:

x2.append(xitem)

allx = x2

y2 = []

for yitem in y:

if yitem != -9999 and yitem != 0:

y2.append(yitem)

slope, intercept, r\_value, p\_value, std\_err = stats.linregress(x2, y2)

# allr.append([file, r\_value])

print("slope: %f intercept: %f" % (slope, intercept))

print("r-squared: %f" % r\_value\*\*2)

print('this is {0}'.format(file))

mean\_squared\_error(y2, x2)

# plt.figure(figsize=(9,12))

# plt.plot(array(x2), array(y2), 'o', label='original data')

# plt.plot(array(x2), array(x2), '-', label = 'y = x ')

# g = (sns.jointplot(x2, y2, kind="hex"))

# joint\_kws=dict(gridsize=1)

hexplot = sns.jointplot(x2, y2,

color = 'deepskyblue',#gold, 'lightcoral','deepskyblue'

kind="hex",

ylim=(16,25), xlim = (16,25),

joint\_kws={"extent": (0, 25, 0, 25)})

# joint\_kws = joint\_kws)

# hexplot.ax\_joint.set\_title('r^2 = {0}'.format(round(r\_value\*\*2, 2)), pad = 45)

hexplot.ax\_joint.set\_xlabel('Model Result of {0}'.format(title))

hexplot.ax\_joint.set\_ylabel('Reference of Home Maker mode')

hexplot.ax\_joint.set\_xticks([16,18,20,22,24])

hexplot.ax\_joint.set\_yticks([16,18,20,22,24])

# hexplot.ax\_joint.set\_title('{0}'.format(title))

# plt.subplots\_adjust(left=0.2, right=0.8, top=0.8, bottom=0.2)

# plt.subplot(array(x2), array(x2), '-', label = 'y = x')

# shrink fig so cbar is visible

# make new ax object for the cbar

# cbar\_ax = hexplot.fig.add\_axes([.85, .25, .05, .4]) # x, y, width, height

# plt.colorbar(cax=cbar\_ax)

plt.clim(0, 1000)

# plt.title('r^2 = {0}'.format(round(r\_value\*\*2, 2)))

# plt.show()

plt.savefig('D:/geoscience/MScResearch/Analysis12/home\_hexagon/{0}.png'.format(str(int(round(r\_value\*\*2, 2) \*100)) + file[:-4]),

dpi = 300)

plt.show()

# .set\_axis\_labels("x", "y"))

# shape = int(file[10])

# size = int(file[4])

# name = field[shape \*3 - 3 + size]

# plt.title('Exposure Comparison (unit: μg/m3)'.format(name), fontsize = 20)

# plt.xlabel('Commuter\'s Exposure', fontsize = 16)

# plt.ylabel('{0}'.format(name), fontsize = 16)

# In statistics, the mean squared error or mean squared deviation of an estimator measures

# the average of the squares of the errors—that is, the average squared difference between

# the estimated values and the actual value.

# allrmse.append([file, mean\_squared\_error(y2, x2)])

# np.savetxt('D:/geoscience/MScResearch/Analysis12/hexagon/{0}'.format(file),

# differ, header = header, comments = '')

np.savetxt('D:/geoscience/MScResearch/Analysis0930/rmse.txt', allrmse)

# library & dataset

import seaborn as sns

dff = sns.load\_dataset('iris')

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

# Basic correlogram

#sns\_plot = sns.pairplot(df)

sns.plt.ylim(16, 25)

sns.plt.xlim(16, 25)

sns.pairplot(df)

plt.ylim(16, 25)

plt.xlim(16, 25)

plt.show()

#ylim=(16,25), xlim = (16,25))

#sns\_plot = sns.pairplot(df,

# kind="scatter",

## kind="reg",

# hue="size",

# markers=["o", "s", "D"],

# palette="Set2")

##sns.pairplot(df3[size == '0'])

#sns\_plot.savefig("IMG/correlogram1.png")

# set of filename of the data to plot.

toplot = ['size0shape1.asc', 'size1shape1.asc', 'size2shape1.asc',

'size0shape2.asc', 'size1shape2.asc', 'size2shape2.asc',

'size0shape3.asc', 'size1shape3.asc', 'size2shape3.asc',

'size0shape5.asc', 'size1shape5.asc', 'size2shape5.asc']

refpath = 'D:/geoscience/MScResearch/Data/ref/largeHome5.asc'

refExpo = np.loadtxt(refpath, skiprows =6)

df0 = np.zeros((3\*34012, 5))# 4 columns + 1 column for label(field size)

import random

list2point5=[]

list1point5=[]

list0point5=[]

for i in range(2500):

# r=random.randint(0,34012)

r=random.randint(0 ,34012)

if r not in list2point5:

list2point5.append(r)

list1point5.append(r+34012)

list0point5.append(r+34012\*2)

for idx, file in enumerate(toplot):

# print(file)

# print(idx)

size = idx%3 # last column

shape = idx//3 # four columns

x = returnOneRow(file)

df0[0 + size\*34012:34012 + size\*34012, shape] = x

df0[0 + size\*34012:34012 + size\*34012, 4] = size

df = pd.DataFrame(df0[list0point5,0:4],

columns=['Linear',

'Exponential',

'Even',

'C\_even'])

def returnOneRow (filename):

size = int(filename[4])

fieldtype = int(filename[10])-1

code = fieldtype\*3 +size

title = field[code]

# resultpath = 'D:/geoscience/MScResearch/Analysis0930/exWithoutRoad/' +file

print(title)

resultpath = 'D:/geoscience/MScResearch/Analysis0930/exWithRoad/' +filename

mresult = np.loadtxt(resultpath, skiprows = 6)

for row in range(645):

for col in range(771):

if refExpo[row][col] == -9999:

mresult[row][col] = -9999

if mresult[row][col] == 0:

refExpo[row][col] = 0

myresult = mresult

x = np.reshape(myresult, (771\*645))

x2 = []

# Take out the pixels with no value.

for xitem in x:

if xitem != -9999 and xitem != 0:

x2.append(xitem)

return np.array(x2)

##########To plot

#Comparison of Exposure value of four Distribution Fields (R=0/1/2.5km

# the index of sampled points are in the folder of Analysis12/scatter

sns.set(font\_scale=1.5)

df25 = pd.DataFrame(df0[list2point5,0:4],

columns=['Linear',

'Exponential',

'Even',

'C\_even'])

g = sns.pairplot(df25, kind="scatter",

diag\_kind="kde",

vars=["Linear", "Exponential", "Even", "C\_even"])

g.fig.suptitle("Comparison of Exposure value of four Distribution Fields (R=2.5km)",y=1)

plt.ylim(16, 35)

plt.xlim(16, 35)

plt.show()

#plt.show()s

# Sharing the data from above called df05, df15, df25

ax = plt.axes()

ax.set\_title('Pearson Correlation (R) of four Distribution Fields (Radius=2.5km)')

sns.heatmap(df15.corr(), annot=True, fmt=".2f")

#g.fig.suptitle("Pearson Correlation (R) of four Distribution Fields (Radius=0.5km)",y=1)

plt.show()

**Reference**

[1] World Health Organization, “Burden of disease from the joint effects of household and ambient Air pollution for 2016,” no. May, 2018.

[2] Soenario Ivan, et al. Land use regression models revealing spatiotemporal co-variation in combustion related air pollutants in the Netherlands (In review).

[3] J. O.Anderson, J. G.Thundiyil, and A.Stolbach, “Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health,” J. Med. Toxicol., vol. 8, no. 2, pp. 166–175, 2012.

[4] T.Bourdrel, M. A.Bind, Y.Béjot, O.Morel, and J. F.Argacha, “Effets cardiovasculaires de la pollution de l’air,” Arch. Cardiovasc. Dis., vol. 110, no. 11, pp. 634–642, 2017.

[5] Forum of International Respiratory Societies, The Global Impact of Respiratory Disease. 2017.

[6] M.Lu, O.Schmitz, I.Soenario, and D.Karssenberg, “Assessment of air pollution exposures with human space-time path : differences between home-based workers and bike commuters,” 2018.

[7] European Space Agency. https://phys.org/news/2019-05-eu-member-states-nitrogen-dioxide.html

[8] Rijksinstituut voor Volksgezondheid en Milieu. https://www.atlasleefomgeving.nl/meer-weten/lucht/stikstofdioxide

[9] Department for Environment, Food and Rural Affairs, UK. Nitrogen Dioxide in the United Kingdom. https://uk-air.defra.gov.uk/library/assets/documents/reports/aqeg/nd-contents-chapter1.pdf

[10] Thunis, P., et al. "Air pollution and emission reductions over the Po-valley: Air Quality Modelling and Integrated Assessment." 18th world IMACS Congress and MODSIM09 International Congress on Modeling and Simulation, Interfacing Modeling and Simulation with Mathematical and Computational Sciences, Cairns, Australia from. 2009.

[11] Caserini, Stefano, et al. "Influence of climate change on the frequency of daytime temperature inversions and stagnation events in the Po Valley: historical trend and future projections." Atmospheric Research 184 (2017): 15-23.

[12] Martin, Randall V. "Satellite remote sensing of surface air quality." Atmospheric environment 42.34 (2008): 7823-7843.

[13] Rivera, Claudia, Wolfgang Stremme, and Michel Grutter. "Nitrogen dioxide DOAS measurements from ground and space: comparison of zenith scattered sunlight ground-based measurements and OMI data in Central Mexico." Atmósfera 26.3 (2013): 401-414.

[14] Paraschiv, Spiru, et al. "OMI and Ground-Based In-Situ Tropospheric Nitrogen Dioxide Observations over Several Important European Cities during 2005–2014." International journal of environmental research and public health 14.11 (2017): 1415.

[15] Kharol, S. K., et al. "Assessment of the magnitude and recent trends in satellite-derived ground-level nitrogen dioxide over North America." Atmospheric Environment 118 (2015): 236-245.

[16] Hoek, Gerard, et al. "A review of land-use regression models to assess spatial variation of outdoor air pollution." Atmospheric environment 42.33 (2008): 7561-7578.

[17] Spinelle, Laurent, et al. "Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide." Sensors and Actuators B: Chemical 215 (2015): 249-257.

[18] Schneider, Philipp, et al. "Mapping urban air quality in near real-time using observations from low-cost sensors and model information." Environment international 106 (2017): 234-247.

[19] Isiugo, Kelechi, et al. "Assessing the accuracy of commercially available gas sensors for the measurement of ambient ozone and nitrogen dioxide." Journal of occupational and environmental hygiene 15.11 (2018): 782-791.

[20] Lloyd, Christopher D., and Peter M. Atkinson. "Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data." International Journal of Applied Earth Observation and Geoinformation 5.4 (2004): 293-305.

[21] Fekri, Reza, Zahra Rostami, and Hadi Tahsini. "Evaluation of the condition of air pollutants in Mashhad city at different stations by using the Inverse Distance Weighting method." Anthropocentric Pollution Journal 2.2 (2018): 26-32.

[22] Leelőssy, Ádám, et al. "Dispersion modeling of air pollutants in the atmosphere: a review." Open Geosciences 6.3 (2014): 257-278.

[23] Mölter, A., et al. "Modelling air pollution for epidemiologic research—Part I: A novel approach combining land use regression and air dispersion." Science of the total environment 408.23 (2010): 5862-5869.

[24] Korek, Michal, et al. "Can dispersion modeling of air pollution be improved by land-use regression? An example from Stockholm, Sweden." Journal of Exposure Science and Environmental Epidemiology 27.6 (2017): 575.

[25] Michanowicz, Drew R., et al. "A hybrid land use regression/line-source dispersion model for predicting intra-urban NO2." Transportation Research Part D: Transport and Environment 43 (2016): 181-191.

[26] Pope, Richard J., et al. "High resolution satellite observations give new view of UK air quality." Weather (2019).

[27] Centraal Bureau voor de Statistiek. https://opendata.cbs.nl/statline/#/CBS/en/dataset/37259eng/map?ts=1568978823529

[28] Centraal Bureau voor de Statistiek. https://opendata.cbs.nl/statline/#/CBS/en/dataset/37259ENG/table?fromstatweb

[29] nltimes. https://nltimes.nl/2019/07/04/dutch-railway-handles-13-million-travelers-per-working-day

[30] Gemeente Utrecht.

https://www.utrecht.nl/wonen-en-leven/gezonde-leefomgeving/luchtkwaliteit/

[31] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996): 267-288.

[32] Hocking, R. R., and R. N. Leslie. "Selection of the best subset in regression analysis." Technometrics 9.4 (1967): 531-540.

[33] Adams, Matthew D., Nikolaos Yiannakoulias, and Pavlos S. Kanaroglou. "Air pollution exposure: An activity pattern approach for active transportation." Atmospheric environment 140 (2016): 52-59.

[34] Breen, Michael S., et al. "GPS-based microenvironment tracker (MicroTrac) model to estimate time–location of individuals for air pollution exposure assessments: Model evaluation in central North Carolina." Journal of Exposure Science and Environmental Epidemiology 24.4 (2014): 412.

[35] Liu, Hai-Ying, Erik Skjetne, and Mike Kobernus. "Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment." Environmental Health 12.1 (2013): 93.[28] Chen, Yuquan, et al. "On optimal tempered Lévy flight foraging." Frontiers in Physics 6 (2018): 111.

[36] Lu, Meng, et al. "Assessment of air pollution exposures across a population: differences between home-based workers and bike commuters." EGU General Assembly Conference Abstracts. Vol. 20. 2018.

[37] Pearson, K. "Thi’Problem of the Random Walk." letter], Nature 72 (1905): 294.

[38] Tang, Tie-Qiao, Yi-Xiao Shao, and Liang Chen. "Modeling pedestrian movement at the hall of high-speed railway station during the check-in process." Physica A: Statistical Mechanics and its Applications 467 (2017): 157-166.

[39] Guo, QiuLei, and Hassan A. Karimi. "A methodology with a distributed algorithm for large-scale trajectory distribution prediction." International Journal of Geographical Information Science 33.4 (2019): 833-854.

[40] Belik, Vitaly, Theo Geisel, and Dirk Brockmann. "Natural human mobility patterns and spatial spread of infectious diseases." Physical Review X 1.1 (2011): 011001.

[41] Massey, Douglas S. "Patterns and processes of international migration in the 21st century." Conference on African Migration in Comparative Perspective, Johannesburg, South Africa. Vol. 4. No. 7. 2003.

[42] Tang, Jinjun, et al. "Uncovering urban human mobility from large scale taxi GPS data." Physica A: Statistical Mechanics and its Applications 438 (2015): 140-153.

[43] Pearson, K. "Thi’Problem of the Random Walk." letter], Nature 72 (1905): 294.

[44] Bachelier, Louis. "Théorie de la spéculation." Annales scientifiques de l'École normale supérieure. Vol. 17. 1900.

[45] Rayleigh, Lord. "On waves propagated along the plane surface of an elastic solid." Proceedings of the London Mathematical Society 1.1 (1885): 4-11.

[46] Chen, Yuquan, et al. "On optimal tempered Lévy flight foraging." Frontiers in Physics 6 (2018): 111.

[47] Ali, Mostafa Z., et al. "A balanced fuzzy cultural algorithm with a modified levy flight search for real parameter optimization." Information Sciences 447 (2018): 12-35.

[48] Lin, Miao, and Wen-Jing Hsu. "Mining GPS data for mobility patterns: A survey." Pervasive and mobile computing 12 (2014): 1-16.

[49] Zou, Bin, et al. "Air pollution exposure assessment methods utilized in epidemiological studies." Journal of Environmental Monitoring 11.3 (2009): 475-490.

[50] Han, Xianglu, and Luke P. Naeher. "A review of traffic-related air pollution exposure assessment studies in the developing world." Environment international 32.1 (2006): 106-120.

[51] World Health Organization. "Biomarkers and risk assessment: concepts and principles." (1993).

[52] Brauer, Michael, et al. "Ambient air pollution exposure estimation for the global burden of disease 2013." Environmental science & technology 50.1 (2015): 79-88.

[53] Bai, Yang, et al. "Carbon loading in airway macrophages as a biomarker for individual exposure to particulate matter air pollution—A critical review." Environment international 74 (2015): 32-41.

[54] Reis, Stefan, et al. "The influence of residential and workday population mobility on exposure to air pollution in the UK." Environment international 121 (2018): 803-813.

[55] Tang, Robert, et al. "Integrating travel behavior with land use regression to estimate dynamic air pollution exposure in Hong Kong." Environment international 113 (2018): 100-108.

[56] Adams, Matthew D., Nikolaos Yiannakoulias, and Pavlos S. Kanaroglou. "Air pollution exposure: An activity pattern approach for active transportation." Atmospheric environment 140 (2016): 52-59.

[57] Dons, Evi, et al. "Impact of time–activity patterns on personal exposure to black carbon." Atmospheric Environment 45.21 (2011): 3594-3602.

[58] Schnell, Izhak, et al. "Human exposure to environmental health concern by types of urban environment: the case of Tel Aviv." Environmental pollution 208 (2016): 58-65.

[59] Liu, Sha, et al. "Association between exposure to ambient particulate matter and chronic obstructive pulmonary disease: results from a cross-sectional study in China." Thorax 72.9 (2017): 788-795.

[60] Siła-Nowicka, Katarzyna, et al. "Analysis of human mobility patterns from GPS trajectories and contextual information." International Journal of Geographical Information Science 30.5 (2016): 881-906.

[61] Yang, Shusen, et al. "Using social network theory for modeling human mobility." IEEE network 24.5 (2010): 6-13.

[62] Pappalardo, Luca, and Filippo Simini. "Data-driven generation of spatio-temporal routines in human mobility." Data Mining and Knowledge Discovery 32.3 (2018): 787-829.

[63] Ciarrocca, Manuela, et al. "Is urinary 1-hydroxypyrene a valid biomarker for exposure to air pollution in outdoor workers? A meta-analysis." Journal of Exposure Science and Environmental Epidemiology 24.1 (2014): 17.

[64] Morello-Frosch, Rachel, et al. "Ambient air pollution exposure and full-term birth weight in California." Environmental Health 9.1 (2010): 44.

[65] Adam, Martin, et al. "Adult lung function and long-term air pollution exposure. ESCAPE: a multicentre cohort study and meta-analysis." European Respiratory Journal 45.1 (2015): 38-50.

[66] Park, Yoo Min, and Mei-Po Kwan. "Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored." Health & place 43 (2017): 85-94.

[67] Breen, Michael S., et al. "GPS-based microenvironment tracker (MicroTrac) model to estimate time–location of individuals for air pollution exposure assessments: Model evaluation in central North Carolina." Journal of Exposure Science and Environmental Epidemiology 24.4 (2014): 412.