DETECTION AND PREDICTION OF CLIMATE CHANGE WITH TEMPERATURE



Submitted by Adnan Ahmed to the University of Exeter as a dissertation for the degree of Master of Science in September 2020

I certify that all material in this dissertation which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

Signature: <insert your signature here @Adnan>

# Introduction

‘Weather’ is a blanket term under which there are many components to consider. Be it the temperature, humidity, rainfall, air pressure, wind speed, wind direction, and a multitude of factors that play a role in the overall atmospheric conditions of a region, taking this into consideration; it is a safe assumption to say that it is improbable that the weather of a particular region is the same two days in a row. This makes it a challenge to predict the conditions one might face when actively taking part in an activity. It could be argued that the weather of a region is one of the essential features of the place’s development in terms of infrastructure and the foundations of society. For example, if a site is prone to extremes in weather conditions that make it inhabitable by most communities, then obviously, the prosperity of that community living there is also severely affected. Weather is something that has an influence on all living beings daily and cannot be avoided or ignored under any circumstances. Thus, the underlying motive of weather forecasting could be considered one of the ways we can ensure the safety, evolution, and growth of the human species.

Weather forecasting is the science of predicting the weather of a particular region with the help of statistical principles and models that are based on significant and historical data about a place. This, and the principles of physics that we know influence the weather combined are the foundational principles of weather forecasting. In most prediction models in weather forecasting, we must leave some room for error as it isn’t feasible to assume an accurate prediction with no flaws, primarily because of the numerous factors that play a role in the weather. The data that we use tends to have significant factors, but there are multiple factors that could be directly correlated to the weather, such as geographical location, wind speed, moisture, air pressure, seasonality, and a wide range of subsequent features. Now there is a spectrum of variables that indirectly affect the weather, such as calamities like earthquakes, landslides, tsunamis, and other rare occurrences or phenomena that happen so rarely that they may be considered outliers and not considered whilst doing our analysis and predictions, which leads to less accurate results. Science and technology have made a substantial improvement in this field, and so what we can expect is nothing short of sufficient in terms of predictions.

Climate change is a phenomenon that is in direct correlation to weather and may be considered as the periodic modification of Earth’s climate brought about as a result of changes in the atmosphere as well as interactions between the atmosphere and various other geologic, chemical, biological, and geographic factors within the Earth system (Jackson, S. T., 2017). Thus, the main aim of this report is to accurately interpret and predict the change in the average temperature of various regions in the world to determine a trend and estimate if the temperature of the planet is increasing or decreasing over time. This is significant and plays a vital role in the well-being of not just the human species but all the living beings on the planet. Ecosystems around the world, both marine as well as land-dwelling creatures as well as plants and microorganisms, live in a delicate balance with each other, and a tiny change in this ecosystem could cause detrimental damage to it. Thus, climate change growing at a rapid pace is dangerous and has already led to the extinction of multiple species that have been recorded. It is hard to say the damage it has caused as we do know the possible undiscovered species and environments that have been damaged or destroyed. One this is for certain, and that is that climate change growing at such a rapid pace does not have any positive impact on our planet, and many studies have shown how climate change can directly lead to natural calamities such as the melting of the polar caps in the north pole which leads to the increase in the water levels of the planet which subsequently has led to many communities being displaced from their homes as the land has been captured by the ocean. This is just one of the infinite use cases that prove that climate change is real, and we use real-life data to come to our own conclusions as well.

We use the average temperature of different cities across the world from the time period 1995 to 2019 and asses to see if there is an increase in the overall temperature or not; we do this using the average temperature of the place. We implement various statistical data modelling techniques, such as time series analysis, long-term short memory, which is an application of artificial intelligence, and a few other models, after which we pass them through a few agnostic models that are specifically designed to test the interpretability of each model and check to see which gives us the best results. The dissertation may be broken down as follows.

Chapter 2 consists of a literature review and all the ways, methods, and techniques that have been used in the past to determine the weather forecasting and climate change calculations, followed by Chapter 3, which includes a description of the data we use as well as the methodologies implemented by us to determine our results and finally in chapter 4 we discuss the results and the conclusion we have come to and the significance it beholds.

# Literature review

As discussed in the introduction we can see that weather forecasting is an essentiality not just for convenience but rather for the wellbeing and safety of a region. Climate change can be accurately measured with respect to change in the temperature of a region over a period. There are various statistical techniques that have been implemented to determine the change in temperature as well as predict it. In this literature review we are going to delve deep into the ways that it has been done in the past. We may also compare to see the most common methods used and their effectiveness. Using these techniques, we are able to compare and see the interpretability of each model as well as the statistical significance and accuracy they tend to produce. Some of the methodologies used are as follows.

## General Analysis and Pre-processing

This is usually the first stage of most project where we analyse the data to search for missing values, outliers and any trends that may occur over a period. In the case where we use variables such as temperature, snowfall in cms, amount of rainfall, which are numerical we usually clean the data by either eliminating the missing values if they are insignificant or we may also use the mean or median of the column to fill in the missing data from the previous and next day. There are various libraries in R and Python which are designed to combat this, such as MICE, tidyr, Pandas and many more. As temperature is primarily the variable that we used to determine this climate change we see that in the analysis of climate change in Switzerland by M. Beniston et al. The evolution of daily minimum temperatures at the four stations from the beginning of the century to the end of 1992. Based on the daily temperature values, mean annual statistics have been established; a five-year running mean has been applied in order to filter out some of the high-frequency modes inherent to the inter annual variability. (Beniston, M., Rebetez, M., Giorgi, F. , 1994). The temperature has a pattern which shows that it has increased by 2K over the year it has been observed. This is along with several other variables individually is what the analysis was conducted on. Similarly, we see that M.S Shekhar et al in the paper Climate-change studies in western Himalaya used seasonal temperature over a significant period of time where they noted that the maximum temperature as well as the minimum of the region has increased by 2.8 degrees Celsius and 1 degree Celsius respectively. This in turn lead them to believe there will be anomalies in the other climatic conditions of the region such as at the snowfall in the region and after careful analysis they were able to see a significantly less amount of snowfall over the same period. It was observed that there was around 280 cm of snowfall less than the previous years. (Shekhar, M. S et al. (2010)). Another important feature of general analysis and preprocessing is to determine if two or more variables are correlated with each other in any way and try to spot any trends if possible. There are different ways that the correlation can be measured such as Pearson’s Correlation coefficient which is used to quantify the linear relationship between two variables that are random in nature. (W. Xie, M. He et al, ‘2020’) uses Pearson’s correlation coefficient as well as Spearman’s Rank coefficient to understand the type of relationship that exists between wildfire and drought severity, we let the random variable X denote the number of forest wildfire, and the random variable Y denote the Palmer Modified Drought Index (PMDI) which can be used to measure the drought severity. Using the formulas of each correlation coefficient they observed that the Pearson’s correlation coefficient is 0.712. The Spearman’s correlation coefficient has a numerical value of -0. 714. These numerical results indicate that these two variables are highly correlated in this case study. The major drawback in this case is that sometimes we may find that the variables are not correlated with each other which is why we need to conduct this analysis multiple times with different variables taken into consideration. The other major setback is that correlation does not necessarily mean causation. This means even if they are very much correlated we cannot assume that they are dependent on each other in any way. Thus, further analysis must be done, and this can be considered as the initial step of an analysis in the climate change forecasting but definitely not the final step of the process.

## Time Series Analysis (ARIMA)

Models Time series analysis from its name is a statistical modelling technique that is used for analysis and prediction over a period of time. From the term climate change we can see that change is a process that occurs over a period of time, thus this technique would seem to the most apt for climate analysis. According to (Kaufmann, R.K. et al, ‘2016’) the evidence of the effect of human activity on the climate is mainly evident from two sources: The experiments run by climate models and also the statistical analysis of historical data. In general there are various ways to analyse and predict data over a period of time but the most common way to go about it is using the principle of Regression, the three most popular models that are used as AR, MA and ARIMA models. These represent Auto-Regression, Moving Average and Autoregressive Integrated Moving Average. These are of different orders which have different kappa values that are used to measure which is the best fitted model for a particular dataset with respect to the timeseries. The auto arima function is used to determine which is a good model to fit after which we can use the appropriate model to generate predictions. (Dmritri, T. Ahmad, S. et al ‘2020’) used the precipitation and the maximum as well as minimum temperature from the years 1901-2000 by monthly means to generate a timeseries dataset and conduct the analysis. They also fit a separate SARIMA model on the precipitation and temperature timeseries. They then did the following steps to make sure that they had the best model to fit on the timeseries data to find an accurate and reliable prediction.

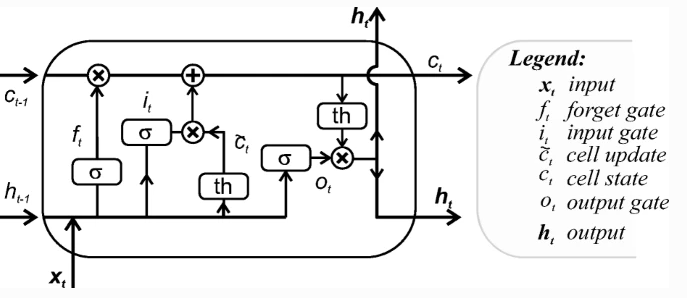
Initially the first and foremost step was to determine which order of the series is best suited to stationarise the series. Different ARIMA models were fitted with different orders but having a constant coefficient. Now that the differenced series exists it is considered to be stationary although it may still have auto-correlated errors.

The next step undertaken was to identify the AR(p) and MR(q) components. This can be done by generating an Auto correlation function (ACF) as well as a Partial Auto Correlation function (PACF) which can show us how well the present value of the series is related to that of the past values. By seeing if there is a sharp cutoff of the differenced series on the PACF graph we can observe that AR needs to be added to the model and if the same occurs on the ACF graph we know that MA needs to be added to the model.

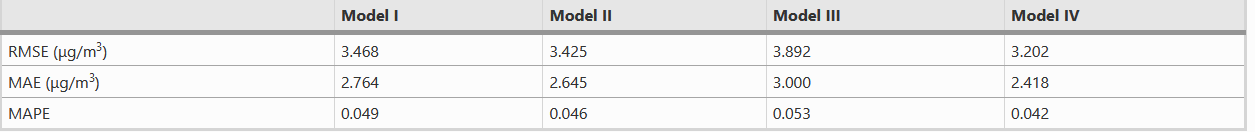
Following this they have made some estimations using appropriate p,d and q values which are fitted on the ARIMA models with appropriate residuals. Then the seasonality is removed for the models and the best SARIMA model needs to be selected. After which the forecasting is done and the results are tested under the Akaike Information Criterion which is an estimate of a constant and the relative distance between the unknown true likelihood and function of the data fitted on the model. Thus we can see that the lower the AIC value the closer it is to the truth. Lower values indicate that the less information the model loses the more higher the quality of the model. The Bayesian Information Criterion (BIC) is also a criterion used with the same principle of lower value the better and is based on the likelihood function. Using this method they were able to observe the maximum and minimum temperature of the regions respectively but also were able to forecast predictions for the precipitation which showed a constant trend in values.

## LSTM

Air Long term short memory is a type of Recurral Neural Network which as a long term dependency which means that it can retain information for a long period of time due to the internal memory it possesses. It is diagrammatically represented in figure 2.2 as shown below.



As we can see the mechanics revolve around gates primarily these gates are comprised of gates, which are structures through which information is added or removed in the cell state; and also has weights Wj , where j signifies the state name. It contains sigmoid neural, σ, net layer and a point wise multiplication operation. The sigmoid layer is called a forget gate denoted by ft . This element decides what information will be thrown away from the cell state. Its decision is made by looking at xt and ht−1. It consists of two property values, one is the hidden state H(t) which is primarily responsible for the long term memory and the forget gate F(t) adjusts the connection of the input with that of the previous hidden state to the cell state, which then allows it to forget when needed. (Pak, U. et all ‘2018) in the prediction of ozone concentration uses LSTM and CNN models to determine the concentration by implementing the following methods. In the project undertaken by Pak, U. et al they combine CNN and LTSM network in different ways to create four model which has a simple combination of the convolution layer and the layer of LSTM in different combinations such that Model I is a simple combination of the convolutional layer with one layer of LSTM. Model II is a combination of the complete convolutional and pooling layers with one layer of LSTM. Model III is a combination of the convolutional layer with two layers of LSTM. Model IV is a combination of the complete convolutional and pooling layers with two layers of LSTM. They then run all the four models with respect to RMSE, MAE and MAPE and we find that Model IV has the lowest values amongst the four and thus we can figure that it is the best fit model for predicting the ozone concentration. As shown in the table below we can see the values of Model 4 is much better than the rest on all accounts.



As we can see the mechanics revolve around gates primarily these gates are comprised of gates, which are structures through which information is added or removed in the cell state; and also has weights Wj , where j signifies the state name. It contains sigmoid neural, σ, net layer and a point wise multiplication operation. The sigmoid layer is called a forget gate denoted by ft . This element decides what information will be thrown away from the cell state. Its decision is made by looking at xt and ht−1. It consists of two property values, one is the hidden state H(t) which is primarily responsible for the long-term memory and the forget gate F(t) adjusts the connection of the input with that of the previous hidden state to the cell state, which then allows it to forget when needed. (Pak, U. et all ‘2018) in the prediction of ozone concentration uses LSTM and CNN models to determine the concentration by implementing the following methods. In the project undertaken by Pak, U. et al they combine CNN and LTSM network in different ways to create four model which has a simple combination of the convolution layer and the layer of LSTM in different combinations such that Model I is a simple combination of the convolutional layer with one layer of LSTM. Model II is a combination of the complete convolutional and pooling layers with one layer of LSTM. Model III is a combination of the convolutional layer with two layers of LSTM. Model IV is a combination of the complete convolutional and pooling layers with two layers of LSTM. They then run all the four models with respect to RMSE, MAE and MAPE and we find that Model IV has the lowest values amongst the four and thus we can figure that it is the best fit model for predicting the ozone concentration. As shown in the table below we can see the values of Model 4 is much better than the rest on all accounts.

## ANN

Agent There are three types of neural networks: recurrent neural networks, multilayer neural networks, and single layer neural networks.

Feed-forward networks with a single layer: The information only moves forward in the network, where the neurons are stacked in layers. The output layers of computational neurons are referred to as the single layers.

Multilayer Feed-forward network : This is different from one layer in that it has one or more hidden layers Due to the fact that these layers don't directly communicate with each other, they are referred to as "hidden" the network's perimeter because their values are not tracked during training.

Recurrent network: Its existence of feedback loops sets it apart from feed-forward neural networks. An example of a recurrent neural network would be neurons that send back their output signal to the input.

Artificial Neural Networks are based on the human brain. In an ANN, the nodes are mostly structured in layers. These layers come in three different types: input, hidden, and output. They are designed to take in a collection of inputs, carry out intricate calculations, and provide an output. The weight of each synapses makes up the ANN's collection of synapses. These weights (wjm) specify how an input will affect a neuron. An adder (additive junction) is a component of the ANN that contributes to adding the weighted signals. A selection criterion may be imposed by the adder based on the architecture. Among these conditions are minimum, maximum, average, and so forth (Kubat, M. , 1999). The basic structural element of ANN is called perceptron, and the transfer function for neuron m is given by yj = ϕ(vj ) = ϕ Xm i=1 wjixj + bj ! with vj = z = Xm i=1 wjixj + bj where {x1, x2, .., xm ∈ R} represents the inputs, w1m, ..., wjm ∈ R are the respective weights of the m neuron. bj ∈ R is the bias that has the effect of increasing or decreasing the net input of the activation function. There exist different kinds of activation functions in machine learning namely sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) (Kubat, M. , 1999). This is diagramtically represented below in figure 2.1.

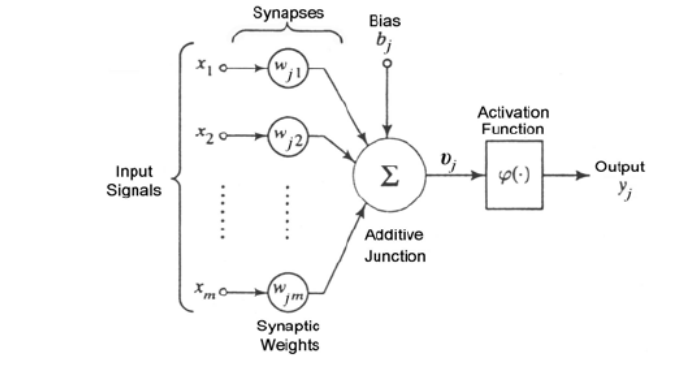


Figure 2.1 Artificial Neural Network Mechanics ((Kubat, M. , 1999)

(Chithra, N.R., Thampi et al. ‘2015’) used ANN-based models were developed for obtaining projections of monthly mean maximum and minimum temperatures at station scale. The capability of the model was assessed by applying it to the Chaliyar river basin in Kerala, India. In the case of prediction of T max, data pertaining to the predictors and the predict and were divided into three seasons, viz. dry period (January–May), wet period (June–November) and the month of December. Although in the case of prediction of T min, they divided it into 2 seasons (wet and dry) and the networks were split and trained separately for each season. They used the correlation coefficient between the predictors and the predicted values. They found that the ANN model is feasible to downscale the climate data and generate appropriate results which in this case was they found both the maximum as well as the minimum value have both increased. The main drawback in this model and approach is that there is a certain level of uncertainty that may be present in the result. This could be combatted using ensemble models.

## RNN

## Wearables and GPS

In recent years, mobile pollutant monitors and Global Positioning Systems (GPS), such as mobile phones, have been utilised to reﬁne personal exposure measurements and activity patterns of individuals. These technologies enable not only the measuring of exposure in

diﬀerent microenvironments, but also (i) how much time an individual spent in the speciﬁc microenvironment, (ii) where the microenvironment is located, and (iii) at what time the individual visited the microenvironment (Steinle et al., 2013; Steinle et al., 2015). How exactly these technologies are deployed varies depending on the study: Sometimes, volunteers wear mobile monitoring systems and GPS so that personal exposure and activity patterns can be tracked in real-time (Liang et al., 2019; Steinle et al., 2015; Sun et al., 2018), other times, GPS, e.g. from mobile phones, is used to record detailed daily activity patterns (Dias and Tchepel, 2014; Breen et al., 2015a) that are then used in combination with modelling techniques, such as LUR, to estimate personal exposure (Su et al., 2015).

These technologies have the advantage that they can account for the heterogeneity of personal exposure in diﬀerent individuals as they deliver activity patterns and personal exposure measurements with a high spatio-temporal resolution (Liang et al., 2019; Dias and Tchepel, 2014). However, mobile monitors are not only labour- and cost-intensive, but they are also an invasive tool for the volunteers because of their size, making them diﬃcult to implement in large scale studies (Su et al, 2015). There are, however, studies addressing the high costs, such as Sun et al. (2018) who promise lower costs and lower technical requirements.

# Methodology

This dissertation follows the underlying framework introduced by Zidek et al. (2003, 2005, 2007) to implement an agent-based model for the prediction of the exposure to PM2.5 in Devon. The reasons for choosing an agent-based model are discussed in chapter 2.4. This chapter gives a high-level overview of this underlying framework. For a more detailed view, the papers mentioned above can be consulted. The general framework for estimating personal exposure is rooted in three main elements: the probabilistic basis that provides the structure for statistical inference, the structural building blocks, such as the data, and the stochastic linkages that connect these building blocks. The following sections describe the three main elements in turn. The details of this chapter are sourced from the papers mentioned above, unless otherwise noted.

## Probabilistic structure

The theoretical probability space (Ω*, A, P* ) forms the basis of the underlying framework where

(i) Ω is the sample space of all information connected to the exposure to the pollutant with individual elements *ω*, (ii) *A* is the collection of subsets of Ω resulting from sampling, and

(iii) *P* is the unknown population distribution of the *ω*s that has to be estimated.

Zidek et al. show that the population distribution for the exposure *X*, *PX*, can be estimated by the number of person-times *ν* which is the number of times when the exposure *X* = (*X*(*t*1)*, . . . , X*(*tn*)) at times *t*1*, ...., tn* surpasses a speciﬁed exposure threshold *x*0 for all *N*

individuals on all

days of the time frame:

= hΣ*n* Σ*N*

0 i = Σ*n* Σ*N* ,

*n ν E*

*i*=1

*k*=1 *I*{*Xik > x* }

*i*=1

*k*=1 *pik*

where *pik* = *E* [*I*{*Xik > x*0}] which can be simpliﬁed to the expectation of a binomial

distribution *ν* = *nNp* if *pik* = *p* for all *i* and *k*. This number can be estimated through the

unbiased estimator *ν*ˆ = Σ*n*

*i*=1

*I*{*XiK*i

*> x*0}*/πK*i

, where an individual *Ki*

is sampled with the

probability *πK*i = *P* (*Ki* = *k*) for all *k* = 1*, . . . , N* on day *i* = 1*, . . . , n*. When an unbiased sampling design is used *πK* = 1*/N* for all sampled individuals yielding the unbiased estimator *ν*ˆ = *Nn*+, where *n*+ is the number of days on which the sampled individuals exposure exceed *x*0. In this case, a measure of uncertainty can be calculated that includes a term for the variance of the binomial distribution and a term for the correlation of exceedances:

Σ*n*

*N* −2*se*2 =

*ν*ˆ

*p*¯*i*(1 − *p*¯*i*) +

Σ

*p*¯*i*(*p*¯*j*|*i* − *p*¯*j*)*,*

*i*=1 *i j*

where *p*¯*i*

−1 Σ*N*

*k*=1

= *N*

*pik*

is the average probability of exceedance at time *i* and

Σ*N* Σ*N* 0 0

*p*¯*j*|*i*

= *N* −2

*k*=1

*k*𝘫=1 *E* [*I*{*Xik > x* }*I*{*Xjk*𝘫 *> x* }]

*p*¯*i*

is the average probability of exceedance at time *j* conditional on the exceedance at time *i*.

## Structural elements

The purpose of structural elements is to establish a connection between the individual items *ω* of the sample space Ω and the personal exposure X. Structural elements use discrete stratiﬁcation variables to split the sample space Ω into diﬀerent subsets. While

this stratiﬁcation can, in reality, vary in space and time, this framework assumes a static relationship. Stratiﬁcation variables either provide the model with essential qualitative features or split the population into ﬁner sub-populations with regards to the exposure. Possible stratiﬁcation variables include internal factors (I), external factors (E), ambient ﬁelds (AF), and human behaviour (B), such as the microenvironments (ME), the activities (A), and the geographical position (POS).

Internal factors include stratiﬁcation variables that are diﬀerent for all individuals, such as age, sex, and ethnicity, and can be used to stratify the population into a sub-population. External factors are the same for all individuals and are used to stratify circumstances external to the individual, for example, one can imagine that all individuals behave diﬀerently on warm and cold days. An important external factor is the ambient pollution ﬁeld which provides the pollutant concentration on a grid. The concentration varies depending on the location of an individual. Microenvironments stratify the sample space into various smaller environments that are homogeneous and are visited by individuals throughout the day, such as home, work, and outdoors. Because of their homogeneity, they can be generalised through mathematical expressions and modelled later. Activities give information on what an individual does in a microenvironment. These activities inﬂuence the impact that personal exposure has on the individuals. For example, sport increases breathing and, therefore, personal exposure to airborne pollutants through inhalation. The geographical position of each ME is important as the exposure is calculated based on the ambient concentrations that, in turn, depend on the location. Thus, microenvironments located in diﬀerent areas will have diﬀerent pollution concentrations.

## Stochastic linkages

Stochastic linkages connect the structural elements through equations. Oftentimes, the form of these equations is derived or estimated from data but it can also be based on the laws of physics and chemistry. As some uncertainty about the exposure X remains, for example, due to the use of estimates in the linkages, the framework cannot be deterministic but needs a stochastic element. This element is implemented through conditional probabilities and distributions. For example, the conditional distribution (B | I, E) describes the human

behaviour of an individual with certain internal and external factors, or, the conditional distribution (AF | T, AM) gives the ambient pollutant ﬁeld based on monitoring data and temperature. For a closer overview of the conditional probabilities see Zidek et al. (2003; 2005; 2007).

# Data

Before describing the model in detail, it is important to get an overview of what data is used in the model. This chapter presents the primary datasets on which the model is built that are also required as inputs to run the model. In total, there are three datasets providing diﬀerent information: the ﬁrst about the population, the second about the pollutant in question, and the third about temperature. In the following, for each dataset, the source, the contents, the wrangling, and the pre-processing are discussed.

## Population data

The ﬁrst dataset gives information about the population of the study area, in this case, the British county Devon. It is produced from three diﬀerent sources that add general characteristics of individuals living in Devon, health data, and data about time use and employment. In a ﬁrst step, the synthetic population estimation and projection model (SPENSER) developed in tandem by the University of Leeds and The Alan Turing Institute simulates close to 700,000 individuals for the Devon region. It uses spatial microsimulation to combine census data with small scale surveys and datasets to create a synthetic population. The mix of feature-poor census data and feature-rich small scale data results in a feature-rich dataset representing the demography of Devon. This baseline population is then assigned attributes, such as the area they live in (Middle Layer Super Output Area or MSOA), the household they are a part of, sex, age, and ethnicity (The Alan Turing Institute, n.d.). In a second step, information about the health of the synthetic population is added based on the 2017 Health Survey of England. This information includes an indicator of general health, whether the individual is obese, has diabetes, high blood pressure, or smokes. Lastly, the 2014/15 UK Time Use Survey is utilised to add data about the use of time and employment to

the population. This includes the occupation, time spent in diﬀerent environments, such as at home, at work, or in transport, time spent walking and cycling, and time spent in public and private spaces. The use of time follows time-use patterns unique to speciﬁc sub-populations that are then assigned to each individual.

To prepare the data for the model, the proportion of time spent in diﬀerent environments is transformed into hours and rounded up to full hours for each environment. Furthermore, some environments are aggregated so that, ﬁnally, four microenvironments remain: ‘home’, ‘indoor not home’, ‘transport’, and ‘leisure’. As not all time use patterns are complete, individuals with incomplete patterns are assigned a new complete pattern randomly allocated by sampling from the complete time use patterns from the other individuals with the same age, sex, and employment status. This also applies to patterns that do not include at least 8 hours of time spent at ‘home’ (see assumptions made in chapter 5). Lastly, longitude and latitude coordinates of the population-weighted centroid of the MSOA are assigned to each individual.

The overall dataset contains almost 700,000 individuals of which 52.5% are female and 47.5% are male. There are six age groups: younger than 19, 19-29, 30-44, 45-59, 60-74, and over 74

with respective proportions of 20.5%, 9.0%, 18.1%, 23.0%, 20.0% and 9.5%. While only 2.1% of the entire population (not just the workforce) are unemployed, 45.3% are either employed full-time, part-time, or self-employed. The rest of the population is one of the following: under the age of 16, retired, full-time student, homemaker, on maternity leave, long-term sick, or other.

## Pollutant data

The second dataset introduces information about the pollutant in question, in this case ﬁne particulate matter (PM2.5). PM2.5 measurements describe the amount of particulate matter with a diameter of less than 2.5 micrometres per cubic meter air. These particles, such as dust and black carbon, are, for example, emitted by combustion engines from cars or heating (WHO, 2018). The Copernicus Atmosphere Monitoring Service (CAMS) carries out a global reanalysis that provides ambient PM2.5 concentration on a 0.75 by 0.75° grid in three-hour intervals. This reanalysis uses data assimilation to combine estimates from

models and real-life measurements to model the atmosphere of the whole world according to scientiﬁc laws of chemistry and physics. Through this process, biases can be detected, estimated, and excluded, resulting in a uniform dataset for the entire world even for sparsely monitored regions (Copernicus, 2020).

To make this dataset viable for the model proposed in chapter 5, the spatial resolution is increased through linear interpolation from a 0.75 by 0.75° grid to a grid that roughly represents diameters of 0.1 by 0.1°. Similarly, the resolution of the time steps is increased to hourly increments. For illustration, Figure 1 presents the results for Devon on 01 May 2019 at noon with the dots representing the centroids of each grid square. The legend indicates the thresholds of the Daily Air Quality Index (DAQI) used in the UK where green signiﬁes a low level of pollution, yellow a medium level, and red a high level (Defra, n.d.). For the full dataset of 2019, the estimates range from 0.0 to 164.9µg/mS with a mean of 9.6µg/mS and a median of 7.9µg/mS. The disproportionately big spikes in PM2.5 concentration are around 23 April 2019. At this time, the estimates range from 4.5 to 164.9µg/mS with a mean of 51.6µg/mS and a median of 42.7µg/mS. Apart from this time, the highest PM2.5 concentration is 80.7µg/mS on 08 December 2019. The cause of this big spike in April calculated by the CAMS renalysis could not be identiﬁed.

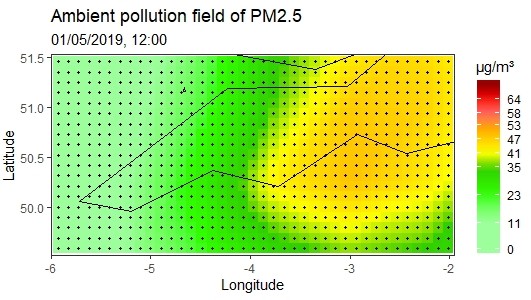


Figure 1: Ambient pollution ﬁeld of PM2.5 for South-East England on 01 May 2019 at noon.

## Temperature data

The last dataset also stems from the CAMS reanalysis and provides information about the 2m temperature. This temperature describes the air temperature 2 meters above the surface (Copernicus, 2020). Similarly to the pollutant dataset, the spatial resolution is increased from a 0.75 by 0.75° grid to a 0.1 by 0.1° grid through linear interpolation. For the model, only the daily average temperature at each location is of interest. Therefore, the daily average is taken for each spatial point. Figure 2 depicts the ﬁnal results on 03 February 2019 for the Devon area with the dots representing the centroids of each grid square. For the full dataset, the estimates range from -1.0 to 23.9°C with a mean of 12.0°C and a median of 11.7°C. Highest average daily temperatures are reached in July and August while the lowest temperatures occur in January and February.

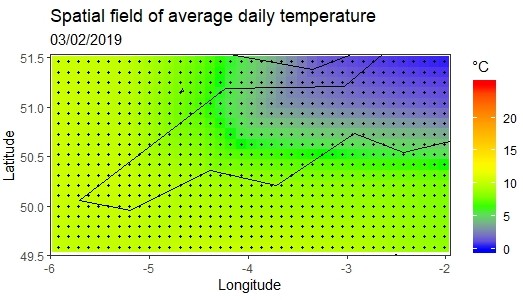


Figure 2: Spatial ﬁeld of average daily temperature for South-East England on 03 February 2019.

# Model

This chapter presents how the theoretical framework described in chapter 3 can be applied in a practical way to build an actual model. After a brief overview of the model, the chapter highlights the individual structural elements and stochastic linkages before discussing its

assumptions and limitations. Finally, the core components of the model, i.e. the activity sampler, the concentration calculator, and the exposure simulator, are explained in detail. While this chapter follows the implementation of a model that estimates the personal exposure to PM2.5, it should be noted that the model can be applied to other pollutants by adjusting the parameters implemented in the individual components and the data provided to the model accordingly.

## Overview of the model

The modelling process can be divided into six steps as Figure 3 suggests: First, the necessary inputs have to be provided. Second, the data needs to be processed and prepared for the subsequent steps. Third, a sample of individuals from the population is taken. Fourth, activity sequences for each day are sampled. Fifth, the exposures for these activity sequences are estimated. Steps four and ﬁve are repeated for each individual of the sample until ﬁnally, sixth, the activity sequence, personal exposures, and the sample of the population is returned as the output of the model run. What follows is a more detailed description of these six steps.

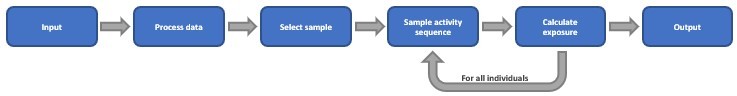


Figure 3: Individual steps of the model.

In the ﬁrst step, the datasets about Devon’s population, ambient particulate matter concentra- tions, and daily average temperature have to be provided to the model. In this step, one also has to deﬁne the general study framework: This includes deﬁning a speciﬁc sub-population or strata, a speciﬁc time frame for which the personal exposure should be calculated, and the number of individuals that should be simulated. The strata is deﬁned by providing the sex, age group, employment status, and smoking status, while the time frame is speciﬁed by giving a start and end date.

The next step ﬁnalises the population dataset by assigning a housing type to each household of the population based on 2011 census data for the South-West of England published by the Oﬃce of National Statistics (ONS). The census data includes many housing types that

are aggregated into four main types for simplicity: detached houses, semi-detached houses, terrace houses, and ﬂats. The probabilities to be assigned a housing type are 30.0%, 27.5%, 23.4% and 19.1% respectively. The housing type is important to estimate the housing volume for the exposure at home in step ﬁve (for more details see chapter 5.3).

Based on the deﬁned criteria of the sub-population in question, the model stratiﬁes the population and then selects a sample of the deﬁned size, i.e. the number of individuals that should be simulated, in step three.

Step four includes the ﬁrst main component of the model, the activity sampler. The objective of the sampler is to create an activity sequence, i.e. a vector of the microenvironments an individual visits each hour throughout the day (Zidek et al., 2007). The sampler uses the information about time spent in diﬀerent microenvironments from the population dataset, the daily average temperatures from the temperature dataset, and information about work patterns, such as working on speciﬁc days of the week and working at night, based on the 2020 Labour Force Survey to create activity sequences tailored to each speciﬁc individual that are unique for each day of the time frame.

The exposure simulator in step ﬁve uses these microenvironments arranged by the activity sampler to estimate the personal exposure for each hour of the day. The simulator diﬀerenti- ates between closed and open microenvironments. The former microenvironment includes ambient as well as non-ambient PM2.5 generating sources. For example, cooking or smoking at ‘home’ could emit PM2.5. In this model, the only environment that is considered to be closed is ‘home’. Open microenvironments do not speciﬁcally model non-ambient sources but use a linear function of the ambient concentration to model the exposure within the microenvironment. Generally, environments that include non-ambient sources that are hard to quantify are modelled as open microenvironments (Zidek et al., 2007). In this model, this is true for the microenvironments ‘indoor not home’, ‘transport’, and ‘outdoor’.

Steps four and ﬁve are repeated for each individual until, ﬁnally, each individual has a unique activity sequence for each day of the time frame and the personal exposure is calculated according to these sequences. The activity sequence, personal exposure, and population data for all individuals of the sample are then returned by the model in step six.

The following sub-chapters present the structural elements and stochastic linkages that are

used in the model in line with the notation used in the underlying framework introduced in chapter 3.

### Structural elements

As described in chapter 3.2, the function of the structural elements is to stratify the whole sample space into smaller sub-spaces from which information is then sampled. The struc- tural elements of this model can be roughly divided into four groups: elements related to internal factors, elements related to external factors, elements related to human behaviour, and elements related to exposure. Internal factors (I) are speciﬁc to each individual and are implemented into the model by the population dataset of Devon that is simulated by SPENSER. External factors (E) are the same for all individuals (Zidek et al., 2007). In this model, they are expressed through the spatio-temporal estimates of the CAMS reanalysis: the PM2.5 pollutant data that provides the background ambient pollution ﬁeld (AF) and the temperature data (T). Human behaviour (B) includes information about the microen- vironments (ME) and the geographical position (POS) of these microenvironments (ibid). The former is sourced from the 2014/2015 Time Use Survey as part of the creation of the population dataset, while the latter is simpliﬁed by using the population-weighted centroids of the individual’s MSOA that are provided by SPENSER. Absorption factors (A) that impact the intensity of the given exposure, such as a higher rate of breathing while running, are not included in this model (ibid.).

### Stochastic linkages

Stochastic linkages connect the structural elements through mathematical equations that contain a stochastic element to reﬂect the uncertainty about the exposure. This is achieved through conditional distributions (see chapter 3.3). In this model, four stochastic linkages connecting the structural elements should be highlighted: The ﬁrst linkage is the distribution of the background ambient pollution ﬁeld conditional on external factors and the ambient monitored ﬁeld (AF | E, AM). In essence, this is a spatial ﬁeld of the ambient pollution that is modelled through external factors, such as temperature, and measurements from monitoring sites. In this model, this is represented by the pollutant data computed by the

CAMS reanalysis model that gives the ambient concentrations in space and time (see chapter 4.2). The second and third linkages are distributions related to geographical locations. The ﬁrst is the distribution of the true geographical position conditional on the microenvironment, the internal factors, the external factors, and the surrogate geographical position1 (POS | ME, I, E, POS\*). The second is the distribution of the surrogate geographical position conditional on the microenvironment, the internal factors, and the external factors (POS\* | ME, I, E). In this model, both linkages are simpliﬁed by using the population-weighted centroids of the area (more precisely the MSOA) that is allocated to each individual during the simulation of the SPENSER model (see chapter 4.1). Lastly, the distribution of the microenvironment conditional on the internal and external factors (ME | I, E) is implemented by the proportion of time spent in each microenvironment on an average day. This is allocated to each individual in the form of diﬀerent time-use patterns based on the 2014/15 Time Use Survey and is part of the population dataset (see chapter 4.1).

## Assumptions and limitations

The model makes some key assumptions that are discussed in the following. However, smaller assumptions, for example, regarding individual parameters or their selection are not listed here but can be found in the individual sections of chapter 5.3..

The ﬁrst key assumption is that there are only four microenvironments: ‘home’, ‘indoor not home’, ‘transport’, and ‘outdoor’. In reality, there are clearly many more microenvironments, however, this model focuses on these four environments for reasons of simplicity and because estimates of the parameters are individual to each microenvironment and the amount of information on PM2.5 for other environments is limited.

Apart from the microenvironments, there are also three assumptions regarding human behaviour and activity. Firstly, individuals are assumed to stay in the microenvironment ‘home’ for at least 8 hours each day during which they sleep. This assumption enables the activity sampler to structure the individuals’ daily activity sequences better which will be demonstrated in the next chapter. Extreme patterns diverging from this assumption are not

1The surrogate geographical position POS\* is an estimation of the location when the real geographical positon POS is not available.

taken into consideration. Secondly, individuals only visit one microenvironment each hour of the day. They cannot change environments half-way through an hour. This assumption makes the generation of activity sequences simpler and computationally less expensive but could be more reﬁned in the future, for example, by decreasing the time increments to 30 minutes or less. However, this change would increase the complexity of the model, aﬀecting the computational requirements that have to be taken into account. Thirdly, individuals work either during the day sometime in the range from 06:00 to 21:00 or during the night in the range from 14:00 to 05:00. This brings more structure into the daily activity patterns and results in a closer mirroring of realistic human behaviour. The other eight hours, i.e. 22:00 to 05:00 and 06:00 to 13:00, are reserved for sleeping.

Moreover, individuals are either smokers or non-smokers. If an individual doesn’t smoke there is no chance that they come in contact with PM2.5 emitted by another individual smoking,

i.e. there is no second-hand smoke. This assumption is based on the data published by the Oﬃce for National Statistics about smoking behaviour of adults living in the UK that shows an overall declining trend (ONS, 2020). It is also assumed that restrictions on smoking in public areas, restaurants, and pubs make exposure to PM2.5 through second-hand smoke less likely.

Lastly, it is assumed that individuals and the microenvironments they visit are always at the same geographical location. For each individual, the population-weighted centroid of the MSOA they live in deﬁnes their location for the entire model calculation. Individuals do not travel to diﬀerent cities or go on holiday. As the PM2.5 measurements are roughly on a

0.1 by 0.1° grid this means that individuals spend all day in a roughly 12km by 12km area. This assumption is implemented mainly because of simplicity, however, improvements to this are possible. For example, the geographical location of the workplace could be modelled and used as a second location or the individuals could be randomly distributed around the population-weighted centroid of the MSOA, thus giving individuals of the same MSOA diﬀerent locations.

At this point, it should be noted that some of the parameters used in the estimation of the pollution concentrations in diﬀerent microenvironments are older than 10 to 20 years and are sometimes from other continents, mainly from North America. Where feasible, newer and

more accurate parameters are estimated and sampled, however, this is not possible for all parameters. Ideally, these parameters would be estimated with more recent and precise data for the Devon area yielding more realistic estimates of personal exposure.

## Core components

This chapter describes the three main components of the model in detail: the activity sampler, the calculation of the PM2.5 concentration within the microenvironments, and the personal exposure simulator. It also provides information about the exact parameters used in the model and their sources.

### Activity sampler

The main objective of the activity sampler is to assign a sequence or a vector of microenvi- ronments to each speciﬁc individual on a speciﬁc day. The individuals, then, move through these microenironments included in this sequence as the day progresses. An activity sequence for one day contains 24 elements, one for each hour. Each element speciﬁes which microenvi- ronment an individual spends the given hour in. As discussed in chapter 4.1 and chapter 5.2, it is assumed that there are a total of four possible microenvironments: ‘home’, ‘indoor not home’, ‘transport’, and ‘outdoor’. The activity sampler creates a unique activity sequence for each individual and each day by combining information about time spent in diﬀerent microenvironments, daily average temperatures, and information about work patterns of the population, such as working on speciﬁc days of the week and working at night.

The sampler takes the time an individual spends in diﬀerent microenvironments from the population dataset, transfers this proportion into hours, and rounds these hours up to full hours. For example, consider an individual that spends 50% of their day in the microenviron- ment ‘home’, 35% in the microenvironment ‘indoor not home’, 5% in the microenvironment ‘transport’, and has 10% of leisure time a day (leisure time is allocated to the microenviron- ments in the next step). This individual would spend twelve hours in ‘home’, nine hours in ‘indoor not home’, two hours in ‘transport’, and has three hours of leisure time. Always rounding up is important as, otherwise, the total number of hours could be less than 24 which

would yield an activity sequence shorter than 24 hours.

Next, leisure time has to be allocated to the four microenvironments. In the 2014/15 Time Use Survey leisure activities include going to cinemas, theatres, museums, libraries, restaurants, cafes, or sporting events, and playing sports outdoors. This shows that leisure can be divided into outdoor and indoor activities. Therefore, the sampler splits leisure time and allocates the pieces to ‘outdoor’ or ‘indoor not home’. The sampler assumes that the size of each piece depends on the average temperature of the day. The activity sampler takes the daily average temperature from the temperature dataset described in chapter 4.3 and divides days into cold and warm days assuming a threshold of 14°C. On warm days, 80% of leisure time is allocated to the microenvironment ‘outdoor’, while on cold days only 30% is allocated to ‘outdoor’. These percentages are chosen in a way that the average proportion of time spent indoors, i.e. in ‘home’ or in ‘indoor not home’, over all individuals is between 85 and 90% for the entire year. This is the average amount of time that the European Commission and other sources that analyse time use surveys estimate the average individual spends indoors over a year (European Commission, 2003; ONS, 2017; Klepeis et al., 2001). The threshold of 14°C is taken rather arbitrarily and only has the function of splitting days into cold and warm so that the parameters for time spendt outdoors can be balanced. The sampler then takes the full hours of each microenvironment and aggregates them into a pool from which the activity sequence can be sampled in the next steps. On a warm day, for example, leisure time would be divided into 2.4 hours ‘outdoor’ and 0.6 hours ‘indoor not home’. Adding this to the microenvironments and rounding up the hours results in a pool containing ‘home’ twelve times, ‘indoor not home’ ten times, ‘transport’ two times, and ‘outdoor’ three times.

One assumption noted in chapter 5.2 is that each individual spends 8 hours at home which represents the time used for sleeping. To give the sequences more structure, the sampler allocates home for the ﬁrst six hours and the last two hours of a sequence. Each individual would, therefore, be in the microenvironment ‘home’ from 22:00 to 05:00. Otherwise, every sequence would be completely random giving very unrealistic results, for example, switching between ‘home’ and ‘indoor not home’ for the biggest portion of the day. These eight hours are subtracted from the pool. Then, the remainder of the day, i.e. hours 06:00 to 21:00 are randomly sampled from the ﬁnal pool that includes ‘home’ four times, ‘indoor not home’ ten

times, ‘transport’ two times, and ‘outdoor’ three times. In the example, the sampled activity sequence for one day could look like this:

(home, home, home, home, home, home, home, home, transport, indoor not home, indoor not home, indoor not home, indoor not home, indoor not home, indoor not home, indoor not home, indoor not home, transport, indoor not home, outdoor, outdoor, home, home, home)

However, the above assumes that the individual works every day. To add a more realistic assumption, the sampler uses data from the 2020 Labour Force Survey to decide if an individual works and whether they work during the day or during the night for each speciﬁc day. Whether an individual works on a given day has the following probabilities: Monday (86.9%), Tuesday (89.4%), Wednesday (89.1%), Thursday (88.4%), Friday (84.3%), Saturday (24.2%), and Sunday (17.7%). Additionally, an individual works during the night with a probability of 10.5%. In total, there are three diﬀerent scenarios: (i) the individual works during the day, (ii) the individual works during the night, and (iii) the individual does not work. For (i) the steps described above apply. For (ii) the same steps apply but now the assumption that an individual spends the time from 22:00 to 06:00 at ‘home’ is altered to 06:00 to 13:00 to match the reality of working during the night. This could, for example, yield an activity sequence similar to this:

(indoor not home, indoor not home, indoor not home, indoor not home, indoor not home, transport, home, home, home, home, home, home, home, home, home, home, indoor not home, outdoor, outdoor, home, transport, indoor not home, indoor not home, indoor not home)

For (iii) the same steps as in (i) apply but, now, a new activity pool is needed as the individual no longer works. The proportion of time spent at work is now proportionally allocated to the other microenvironments and then rounded up to full hours. In the example, it is assumed that the individual spends 80% of their time spent in the microenvironment ‘indoor not home’ at work. Hence, on a day the individual does not work, the pool consists of seventeen times ‘home’, four times ‘indoor not home’, three times ‘transport’, and four times ‘outdoor’. A possible activity sequence for a day where the individual does not work could look like this:

(home, home, home, home, home, home, home, home, transport, outdoor, outdoor, indoor not home, indoor not home, transport, home, home, outdoor, indoor not home, transport, home, home, home, home, home)

If an individual does not work in general, e.g. because they are retired, under the age of 16, or unemployed, the sampler automatically uses case (iii) for all days.

### Microenvironments

Zidek et al. (2007) model the personal exposure in any microenvironment, also called the local concentration, as a function Φ of the ambient concentrations of the pollutant ﬁeld and the non-ambient sources within the microenvironment. This chapter follows Zidek et al. (2003; 2005; 2007) unless otherwise noted and deﬁnes the function Φ for each of the four microenvironments.

*local* = Φ(*ambient, source*)

Microenvironments belong to one of two categories: Firstly, there are closed microenvironments that contain ambient and non-ambient sources of PM2.5. In this model, only ‘home’ is modelled as a closed microenvironment. Secondly, there are open microenvironments where no non-ambient source is speciﬁed and the concentration within the microenvironment is modelled solely by a linear transformation of the ambient sources of PM2.5. ‘Indoor not home’, ‘transport’, and ‘outdoor’ are modelled as open microenvironments.

* + - 1. **Open microenvironments**

As there are no non-ambient sources deﬁned in an open microenvironment, these sources are estimated by a linear transformation of the ambient concentrations. In detail, the concentration within microenvironment *m* at the geographical location *d* at time *t* is estimated by an intercept *a* and the ambient concentration at position *d* and time *t* multiplied by a slope *b*.

*locald,m,t* = *am* + *bm* × *ambientd,t*

The intercept and the slope have to be speciﬁed for each of the three open microenvironments

*m*: ‘indoor not home’, ‘transport’, and ‘outdoor’. Estimates for these parameters are sourced from scientiﬁc literature and are presented as distributions to represent the uncertainty in the parameters. Table 1 shows the estimates and their sources. To give the local concentration variability, distributions with standard deviations calculated from measurement data that is also used to estimate the parameters are implemented (Burke et al., 2001). The parameters for ‘indoor not home’ combine Burke et al.’s estimates for work, school, and store. The microenvironment ‘outdoor’ assumes concentrations that are equal to the ambient concentrations. These distributions are truncated at 0 so that sampled values below 0 are not possible.

Table 1: Parameters implemented in the estimation of open Microenvironments.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Microenvironment m | Intercept a | Slope b | Variability of local | Source |
| Outdoor | 0 | 1 | - | Zidek et al. (2003) |
| Indoor | N(6.467, 2.1) | N(0.507, 0.11) | N(local, 3.467) | Burke et al. (2001) |
| Transport | N(33, 7.2) | N(0.26, 0.14) | N(local, 12) | Burke et al. (2001) |

* + - 1. **Closed microenvironments**

In a closed microenvironment, there are ambient as well as non-ambient sources that generate PM2.5. Non-ambient sources are all sources that generate PM2.5 within the envrionment. In this model, these are restricted to smoking, cooking, and other. Other includes PM2.5 that cannot be directly linked to a speciﬁc source and is commonly known as the personal cloud. Due to the non-ambient sources, the PM2.5 concentration within the closed microenvironment is diﬀerent to the ambient concentration. However, both are connected through ventilation which will bring the indoor concentration towards the outdoor concentration. This process can be modelled by the mass balance equation:

*d*

*Cin*(*t*) =

*dt*

*S*(*t*)

*V*

+ *vFpCout*(*t*) − (*v* + *Fd*)*Cin*(*t*)

Where *Cin*(*t*) is the indoor concentration at time *t*, *Cout*(*t*) is the outdoor concentration at

time *t*, *S*(*t*) is the pollutant emitted by non-ambient sources at time *t*, *V* is the housing volume, *v* is the air exchange rate that connects indoors and outdoors, *Fp* is the penetration factor, and *Fd* is the deposition rate. The penetration factor and deposition rate deﬁne where the equilibrium of the mass balance equation lies while the air exchange rate deﬁnes how quickly this equilibrium is reached. Zidek et al. (2007) come to the following solution for the mass balance equation:

*Cin*(*t* + *u*) =

*Cadd* +

*v* + *Fd*

*Cin*(*t*) −

*Cadd v* + *Fd*

*exp*[−(*v* + *Fd*)*u*]

where *Cadd* ≡ *S*(*t*) + *vFpCout*(*t*) is assumed to be constant for hours (*t, t* + 1). This solution is implemented in the model with *u* = 1 to estimate the indoor concentrations for each hour of the time frame. Parameters are, again, mainly sourced from scientiﬁc literature but also from oﬃcial statistics. However, no estimates for residential housing volume in Devon or England are available at this time. Thus, these are represented by a triangular distribution of the square footage of each housing type that is estimated by a sample (n = 80) from the property website Zoopla and a uniform distribution of ceiling height. The latter is an assumption from average ceiling heights (OnAverage, n.d.). Table 2 presents these estimates and their sources. Again, uncertainty in the estimates is represented by distributions. For the air exchange rate *v*, Murray and Burmaster (1995) give estimates for four diﬀerent zones. The regions are deﬁned by the annual heating degree days. For this, the diﬀerence between the daily temperatures and 65°F are calculated for all days where the temperature is lower than 65°F. This is done for each day and summed up for the entire year. The annual heating degree

*V*

days with 4,170 for Devon and 4,618 for Exeter lie well in region 3 which ranges from 2,500 to 5,500 annual heating degree days. Therefore, estimates for this region are implemented.

Table 2: Parameters implemented in the estimation of closed Microenvironments.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Value | Source |  |
| S(t): cooking | N(1.7, 0.325) | Özkaynak et al. | (1996) |
| S(t): smoking | N(13.8, 1.775) | Özkaynak et al. | (1996) |
| S(t): other | N(1.1, 0.525) | Özkaynak et al. | (1996) |

|  |  |  |
| --- | --- | --- |
| V: square footage detached | Tri(81, 214, 159) | Zoopla |
| V: square footage semi-detached | Tri(56, 204, 84) | Zoopla |
| V: square footage terrace | Tri(33, 155, 59) | Zoopla |
| V: square footage ﬂat | Tri(34, 106, 41) | Zoopla |
| V: ceiling height | Unif(2.1, 2.6) | OnAverage (n.d.) |
| v: Winter | logN(-0.957, 0.589) | Murray and Burmaster (1995) |
| v: Spring | logN(-0.802, 0.782) | Murray and Burmaster (1995) |
| v: Summer | logN(-0.583, 0.612) | Murray and Burmaster (1995) |
| v: Autumn | logN(-0.787, 0.453) | Murray and Burmaster (1995) |
| Fp | N(1, 0.055) | Özkaynak et al. (1996) |
| Fd | N(0.39, 0.0825) | Özkaynak et al. (1996) |
| Hourly cigarette consumption: female | 16-24: Pois(5/15) | ONS (2020) |
|  | 25-34: Pois(6.9/15) | ONS (2020) |
|  | 35-49: Pois(8.33/15) | ONS (2020) |
|  | 50-59: Pois(10.5/15) | ONS (2020) |
|  | 60+: Pois(10.8/15) | ONS (2020) |
| Hourly cigarette consumption: male | 16-24: Pois(4/15) | ONS (2020) |
|  | 25-34: Pois(4.55/15) | ONS (2020) |
|  | 35-49: Pois(8.97/15) | ONS (2020) |
|  | 50-59: Pois(11.9/15) | ONS (2020) |
|  | 60+: Pois(12.9/15) | ONS (2020) |
| Concentration starting point | N(26, 2) | Wallace et al. (1993) |

The standard deviations for the distributions of the non-ambient sources, the penetration factor, and the deposition rate are calculated following Burke et al. (2001) by taking half of one conﬁdence interval. The median indoor concentration at night, taken from Wallace et al. (1993), is a starting point of the indoor concentration *Cin*(*t*). If data from *t*1 to *tn* is provided to the model, this median is *t*0.

The process of calculating the PM2.5 concentration within the microenvironment ‘home’ for a speciﬁc hour is as follows: In a ﬁrst step, the PM2.5 emitted by non-ambient sources is calculated. Here, the model randomly decides if a person cooks in a given hour and whether

the cooking emits PM2.5. The probability for this is assumed to be 10.0%. This means, if a person is at home for ten hours, they would on average emit PM2.5 through cooking in one hour. As this assumption is not based on research the sensitivity of this parameter is analysed in chapter 5.4. If an individual is emitting PM2.5 while cooking, the emissions are sampled from the distribution of PM2.5 emissions due to cooking. Next, PM2.5 emitted by smoking is added to the non-ambient sources if the individual is a smoker. First, the number of cigarettes is decided based on the age of the individual by sampling from a Poisson distribution with the average hourly cigarette consumption as the parameter. The average hourly cigarette consumption is computed by taking the average total number of cigarettes consumed in a day and dividing it by the number of hours a person theoretically smokes. Following Zidek et al. (2003) the latter is assumed to be 15. This number is then multiplied by sampling from the smoking distribution from Özkaynak et al. (1996). Finally, the PM2.5 emitted by other non-ambient sources is sampled from the distribution published in Özkaynak et al. (1996). The total PM2.5 emitted by non-ambient sources *S*(*t*), therefore, consists of the sum of PM2.5 emitted through cooking, smoking, and other.

In the second step, the additional parameters are sampled and estimated: The housing volume is calculated according to the housing type by multiplying a sample from the square-footage distribution and a sample from the ceiling height distribution. Also, the air exchange rate for the given season, the penetration factor, and the deposition rate are sampled from their respective distributions. Finally, all samples are inserted into the mass balance equation to calculate the indoor concentration for a speciﬁc hour.

### Exposure simulator

The exposure simulator simulates the individual moving through the diﬀerent microenviron- ments throughout the day as well as the pollutant concentrations the individual is exposed to. To estimate the personal exposure for each hour of the time frame, the simulator combines the activity sequence created by the activity sampler and the methods for estimating the concentration in open and closed microenvironments. It creates the local array that includes the microenvironment that is visited and the personal exposure for each individual and each hour of the time frame. While the concentrations for open environments only have to

be calculated for the hours when the individual actually visits the microenvironment, the concentrations at home have to be calculated for the entire time frame as past concentrations are required to estimate current concentrations. The median indoor concentration from Wallace et al. (1993) is taken as a rough starting point for past indoor home concentrations. The simulator starts calculating the home concentration three days before the speciﬁed time frame to allow the mass balance equation to get settled and achieve an equilibrium that represents the current concentration well from the ﬁrst hour of the time frame. Therefore, the PM2.5 measurements used as inputs for the model have to start three days before the time frame in question.

## Sensitivity analysis

A sensitivity analysis investigates the sensitivity of one variable, in this case, exposure to PM2.5, to changes of other variables or parameter values. Sensitivity indicates how much the exposure changes if the values of certain parameters are varied. This is especially useful for parameters that could not be estimated through data or sourced from other scientiﬁc literature, i.e. parameters where assumptions do not have a strong foundation. In this model, there are three parameters that ﬁt this description: (i) ‘Cook’, the probability that a person cooks and emits PM2.5 while in the microenvironment ‘home’, (ii) ‘Cold’, the proportion of leisure time spent in ‘indoor not home’ on a cold day, and (iii) ‘Warm’, the proportion of leisure time spent in ‘indoor not home’ on a warm day. The ﬁrst parameter is used in the calculation of the PM2.5 concentration in the microenvironment ‘home’ while the others are used in the activity sampler.

The process of the sensitivity analysis is as follows: First, the model is run repeatedly for diﬀerent parameter values for one speciﬁc individual. How often the model is run with diﬀerent values depends on the parameters in question but also on the cost of running the model. In this case, these are computational costs, i.e. the time it takes to run the model. To ﬁnd a good balance between the accuracy of the results and the cost, the model is run ﬁve times. A space-ﬁlling design in the form of a Maximin Latin Hypercube is used to ﬁnd value combinations of the parameters in question that cover a wide range. The Maximin Latin Hypercube for three parameters creates a three-dimensional space, divides each dimension

into ﬁve intervals of the same size (as ﬁve runs are available), picks a value so that only one point covers one interval of each dimension, and ﬁnally maximises the minimal distance between the ﬁve points in the three-dimensional space to distribute the points within the intervals so that the best coverage is achieved. In the second step of the sensitivity analysis, a regression model is ﬁt to the data. In the last step, using the model, it is determined whether the three parameters are signiﬁcant predictors of the exposure to PM2.5 at the 5% level. If they are not signiﬁcant predictors, the exposure is not sensitive to changes in these parameters and it can be assumed that the assumptions made in determining the parameter values will not have a signiﬁcant inﬂuence on the estimated exposure.

The model is run for the same female individual aged between 30 and 44 that does not work nor smoke. As the parameters in question only concern the home environment, the individual is selected to not work and, therefore, spend a large amount of time at home. Smoking is excluded as it is not an integral part of the sensitivity analysis. The Maximin Latin Hypercube for ﬁve diﬀerent value combinations is calculated and the model is run ﬁve times for the entire year of 2019. Figure 4 shows the daily average exposure of the ﬁve runs. It reveals that, in general, the exposures for each run are similar and only diverge slightly throughout the whole year. The ﬁgure also shows the large spike around 23 April 2019 that is mentioned in chapter 4.2.

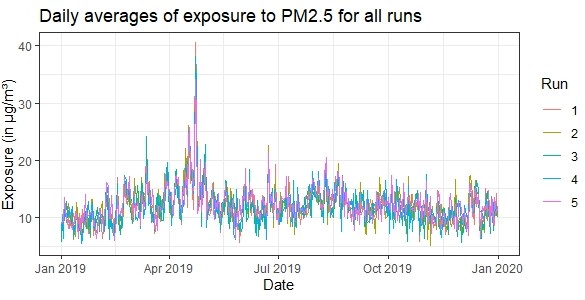


Figure 4: Daily averages of exposure to PM2.5 for the ﬁve model runs of the sensitivity

analysis.

A linear multiple regression model is ﬁt to the data with microenvironments ‘ME’ and the three parameters ‘Cook’, ‘Cold’, and ‘Warm’ as predictors of the response variable exposure to PM2.5. The variable ‘ME’ is a categorical variable where the microenvironment ‘home’ is buried as the baseline in the intercept. ‘MEindoor’, ‘MEoutdoor’, and ‘MEtransport’ describe the inﬂuence of the other microenvironments on the response variable in comparison to ‘MEhome’. All three microenvironments have an impact on the exposure that is signiﬁcantly diﬀerent from the impact of ‘MEhome’. For example, the exposure in the microenvironment ‘transport’ is, generally speaking, increased by 30.275µg/mS compared to the exposure in ‘home’ for a linear model. The linear regression model (1) in Table 3 shows that the three parameters ‘Cook’, ‘Warm’, and ‘Cold’ are all insigniﬁcant on the 5% level. However, the residual plots on the left in Figure 5 question the normality assumption of the linear regression model: The residuals vs ﬁtted values plot shows that the residuals are not randomly scattered around 0. The four clusters show the impact of the microenvironments. In the normal Q-Q plot the tails curl away from the normality line. The upper tail especially diverges greatly from this line. Dropping the insigniﬁcant predictors or adding quadratic terms and interaction terms between the predictors could not change this outcome or the signiﬁcance of the three parameters. Hence, a generalised linear model with a gamma distribution and log link function is ﬁtted to the data. Model (2) in Table 3 shows that the three parameters in question, ‘Cook’, ‘Cold’, and ‘Warm’, are again insigniﬁcant on the 5% level resulting in the model (3) that has the best overall ﬁt. Looking at the residual plots on the right in Figure 5, the residuals look much more normally distributed than for the linear model: the residuals in the residuals vs ﬁtted values plot are scattered around 0 more randomly and the points of the normal Q-Q plot align with the normality line much better. The points only somewhat diverge in the upper tail but it still looks acceptable and is to a much smaller extent than for the linear model. Therefore, the conclusion can be drawn that the exposure is not sensitive to changes in these parameters. This means that even if the parameter values do not match the reality exactly, the eﬀect on the estimated personal exposure to PM2.5 is insigniﬁcant and is unlikely to make a diﬀerence overall. This analysis also shows how important the microenvironments are when predicting personal exposure to a pollutant with a high spatial

heterogeneity, such as PM2.5.

Table 3: Regression table for linear model (1) and gamma regression models (2) and (3).

*Dependent variable:*

Exposure

*OLS glm: Gamma*

*link = log*

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
| Constant | 5.070∗∗∗ | 1.638∗∗∗ | 1.650∗∗∗ |
|  | (0.131) | (0.012) | (0.004) |
| MEindoor | 5.927∗∗∗ | 0.760∗∗∗ | 0.761∗∗∗ |
|  | (0.090) | (0.008) | (0.008) |
| MEoutdoor | 4.126∗∗∗ | 0.582∗∗∗ | 0.581∗∗∗ |
|  | (0.130) | (0.012) | (0.012) |
| MEtransport | 30.275∗∗∗ | 1.919∗∗∗ | 1.919∗∗∗ |
|  | (0.095) | (0.009) | (0.009) |
| Cook | −0.051 | 0.004 |  |
|  | (0.121) | (0.011) |  |
| Warm | 0.207 | 0.012 |  |
|  | (0.109) | (0.010) |  |
| Cold | 0.093 | 0.007 |  |
|  | (0.150) | (0.014) |  |
| Observations | 43,800 | 43,800 | 43,800 |
| R2 | 0.702 |  |  |
| Adjusted R2 | 0.702 |  |  |
| Log Likelihood |  | −128,879.600 | −128,880.700 |
| Akaike Inf. Crit. |  | 257,773.200 | 257,769.400 |
| Residual Std. Error F Statistic | 7.220 (df = 43793)  17,196.010∗∗∗ (df = 6; 43793) |  |  |
| *Note:* | ∗p*<*0.05; ∗∗p*<*0.01; ∗∗∗p*<*0.001 | | |

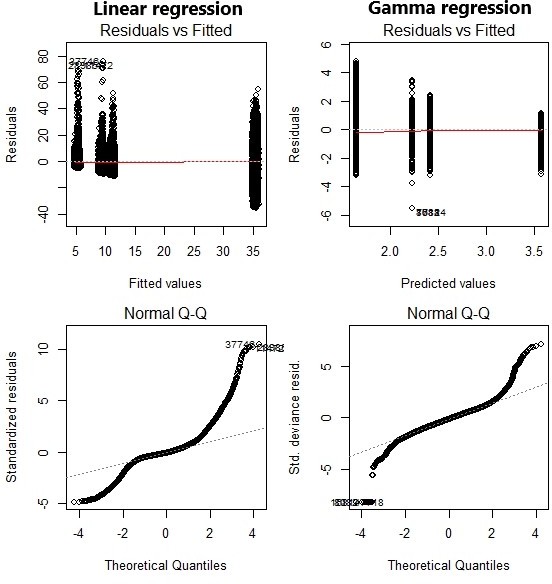


Figure 5: Residual plots of the linear model on the left and the GLM with a gamma distribution on the right.

# Application and results

This chapter shows how the model described in chapter 5 can be applied to a speciﬁc scenario. However, this is only for illustrative purposes as computational restrictions of the hardware do not allow for an extensive application of the model. The scenario investigates the personal exposure for the sub-population of female individuals, aged 19 to 29, who are employed and

do not smoke, for April and July 2019. For both months, 30 individuals from Devon are randomly sampled and simulated.2 The analysis of the results follows Zidek et al. (2007).

Figure 6 shows the predictive distributions that combine the average exposure to PM2.5 of each microenvironment for all 30 individuals in April on the left and July on the right. More precisely, each distribution contains one estimate for each individual. For both months, the median of the predictive distribution is lowest for the microenvironment ‘home’. The medians for ‘indoor not home’ and ‘outdoor’ are close to each other and only slightly higher than the median for ‘home’. However, the median of the predictive distribution for the microenvironment ‘transport’ is much higher than the others. Comparing the variability of the individual predictive distributions, it is noticeable that the distributions for ‘transport’ have a much greater range than the others. The exception is the predictive distribution of the microenvironment ‘outdoor’ for April. The reason for the high variability is that the exposure in the microenvironment ‘outdoor’ is estimated by using the ambient concentrations as a surrogate (see chapter 5.3.2.1). As mentioned in chapter 4.2, there is a very large spike in ambient PM2.5 concentrations in late April. Because of this spike, the daily ambient concentrations range from 6.3 to 65.5µg/mS which translates into a wider range for the daily average exposure in the microenvironment ‘outdoor’. These predictive distributions of personal exposure to PM2.5 show how important the diﬀerentiation between the individual microenvironments is. This is in line with the advantage of the spatial variability discussed in chapter 2.4 in the context of agent-based models. Exposure in diﬀerent microenvironments for the same time and geographical location can diﬀer greatly. For example, consider the case where an individual is in ‘transport’ vs the case the same individual is in the microenvironment ‘home’: The average diﬀerence of these scenarios is more than 25µg/mS. Now, consider the case of taking the ambient concentration from monitoring sites as a surrogate for exposure which is similar to the exposures in the microenvironment ‘outdoor’. The diﬀerence to the other microenvironments can be quite large, especially considering the fact that humans spend most of their days indoors. Using monitoring site measurements would overestimate the exposure for times when individuals are at ‘home’ and underestimate the exposure for

2Ideally more individuals would be simulated, however, due to computational restrictions only 30 are simulated.

times when individuals are in ‘transport’ showing again how important it is to include spatial variability for heterogeneous pollutants, such as PM2.5. Another driving factor of the spatial variability is the individuals’ geographical location that is introduced by the population- weighted averages of the MSOA the individual lives in. As PM2.5 is a heterogeneous pollutant, the variability in diﬀerent MSOAs can diﬀer signiﬁcantly within the same hour. This can be seen in Figure 1 presented in chapter 4.2.

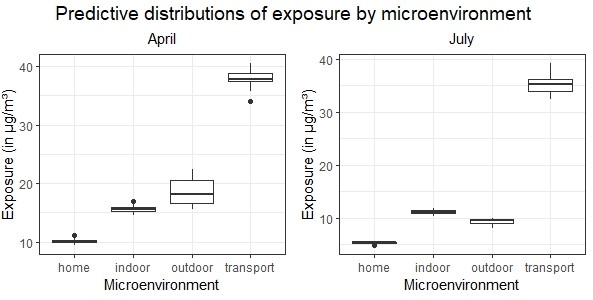


Figure 6: Predictive distribution of personal exposures to PM2.5 by microenvironment for April on the left and July on the right.

Figure 7 shows the predictive distributions of the hourly exposure to PM2.5 averaged over the months for each individual in April on the left and July on the right. Each hourly distribution contains one estimate for each individual. For both months, the day can be divided into two parts. The night from 22:00 to 05:00 forms the ﬁrst part. Here, the exposure and the variability is much smaller than during the second part of the day from 06:00 to 21:00. This mirrors the assumption made in chapter 5 that individuals spend the night at ‘home’, provided they do not work nights. As seen in the previous ﬁgure, the exposure at ‘home’ is very low so that it makes sense that the hours almost all individuals spend at ‘home’ have lower estimated exposure than the rest. During the day, individuals spend more time in the higher exposure environments, such as in ‘transport’, so that for this time the hourly exposure rises as well. These predictive distributions underpin the importance of including

the temporal variability in the model, mentioned in chapter 2.4. There are two components that implement the variability in time: (i) the hourly ambient concentrations and (ii) the activity sequence. While (i) implements the diﬀerent external conditions for each hour that are the same for all individuals, (ii) incorporates the internal variability that is exclusive to the individual. The activity sequence allows the model to deﬁne which hours the individuals spend in diﬀerent microenvironments. As most individuals are at ‘home’ during the night, the personal exposure to PM2.5 is much lower than during the day when individuals spend more time in other environments. Failing to include this variability could paint a wrong picture of the exposure, especially to heterogeneous pollutants.

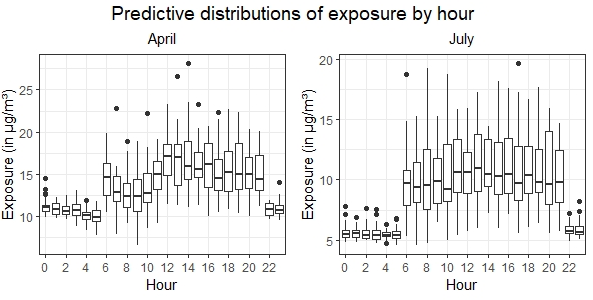


Figure 7: Predictive distribution of personal exposure to PM2.5 by hour for April on the left and July on the right.

As discussed in chapter 2.4, agent-based models use information (such as human behaviour and microenvironments) from many more sources than monitoring sites or the other presented models do. In theory, this should result in more accurate personal exposure estimates. Therefore, Figure 8 compares the predictive distributions of the daily average exposure of all 30 individuals (each daily distribution contains one estimate for each individual) to the ambient concentration of the CAMS reanalysis. In general, the predictive distributions approximately follow the line of daily average ambient concentrations but are lower in many cases. This is especially noticeable from April 15 to April 25: Here, the distributions are much

lower than the extreme spike in ambient concentration, with almost no estimates touching the line. This makes sense as individuals spend most of their days indoors. To an extent, this shields them from the high ambient concentrations outdoors. However, the opposite is also possible: In July, there are no spikes in ambient concentrations and the concentrations are very low in general. For some days, the medians of the predictive distributions exceed the ambient concentrations as the exposure in the microenvironments might be higher due to other sources, such as cooking. This is especially true if people spend a lot of time in the microenvironment ‘transport’ on a given day. The high exposure in ‘transport’ as seen in Figure 6 increases the average daily exposure signiﬁcantly, resulting in higher exposure than the ambient concentration. The exclusive use of monitoring site measurements could, therefore, potentially over- or underestimating personal exposure.

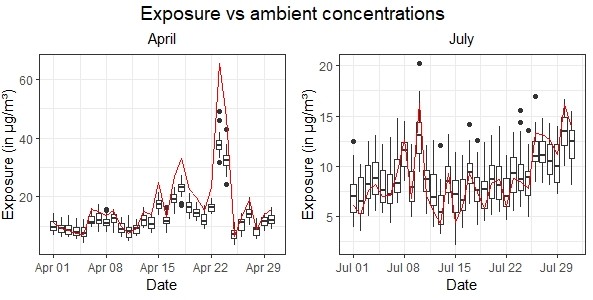


Figure 8: Predictive distributions of personal exposure to PM2.5 by day vs daily average ambient concentrations (red line) in April on the left and July on the right.

The probabilistic framework outlined in chapter 3.1 enables the calculation of the number of days on which the exposure exceeds a certain threshold along with the corresponding measure of uncertainty. In the following, the air quality standards of the Daily Air Quality Index (DAQI) used in the UK are chosen as the thresholds. The DAQI deﬁnes the pollution level of PM2.5 on a scale of 1 (low) to 10 (very high) based on 24 hour means. Indices 1 to 3 (0 to 35µg/mS) are classiﬁed as low, indices 4 to 6 (35 to 53µg/mS) are classiﬁed as moderate,

indices 7 to 9 (54 to 70µg/mS) are classiﬁed as high, and everything above 70µg/mS gets allocated to index 10 which indicates very high pollution (Defra, n.d.). For a random female individual in Devon aged 19 to 29 who is employed and does not smoke 17.2 (3.68) and

2.6 (0.25) days in April exceed the ﬁrst (11µg/mS) and second (23µg/mS) threshold3. Only

0.9 (0.28) days exceed average exposure of 35 µg/mS and 0.2 (0.15) days exceed 41 µg/mS. Therefore, the majority of days is classiﬁed as low pollution. For July, the sub-population is only exposed to low pollution, with only 6.9 (3.15) days exceeding 11µg/mS. These results can be compared to the days on which ambient concentrations exceed these thresholds: In April, 21 days exceed the ﬁrst threshold, 6 the second, 2 the third and fourth, and 1 day exceeds the ﬁfth, sixth, seventh and eighth threshold. In July, only 8 days exceed the ﬁrst threshold. For April, two more days are categorised as moderate to high using the ambient concentrations compared to the model results. For July, about one more day exceeds the ﬁrst threshold. This underpins the previous conclusion that using ambient concentrations as a surrogate for exposure can overestimate individual exposure.

The main advantage of using agent-based models is that the population can be stratiﬁed by diﬀerent criteria and the exposure can be investigated speciﬁc to the sub-population. This makes it simple to explore the eﬀect of possible policy changes or rollbacks. In the following, a policy change decreasing the overall ambient concentration of PM2.5 by 20% is introduced. The results are depicted in Figure 9 which compares the daily average exposures of all 30 individuals for April and July with and without the policy change. Each distribution contains one estimate for each individual. No major eﬀect is noticeable for either of the months. This can have multiple reasons: Firstly, because individuals spend most of their days indoors and ambient concentration is only part of the equation when estimating indoor exposure, it could be that a decrease of 20% in ambient concentration does not have a noticeable impact on personal exposure. Secondly, the ambient exposure in Devon is low in general. Decreasing an already low concentration might ultimately not make a big diﬀerence. In Zidek et al. (2007), the eﬀect of a rollback is much more signiﬁcant. However, Zidek et al. investigate a sub-population located in London, a city with a much higher average ambient concentration than Devon. This analysis could be repeated with a bigger sample and possibly

3Measure of uncertainty displayed in brackets.

even a larger decrease in the ambient concentration. Nevertheless, the result shows how this type of analysis can easily be executed using this model.

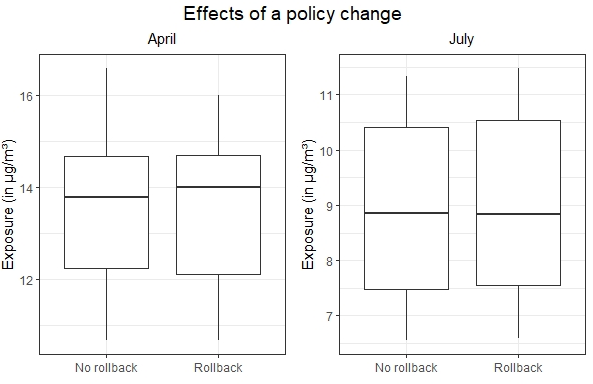


Figure 9: Predictive distributions of personal exposure to PM2.5 with and without a policy change for April on the left and July on the right.

The limited time available to complete this dissertation has led to restrictions on the complexity of the model. While there are many ways to improve the model, the following paragraphs present ﬁve key ideas that could make the model more accurate and realistic. Firstly, the activity sequence that, in the current state, diﬀerentiates between (i) days on which individuals work during the night, (ii) days on which individuals work during the day, and (iii) days on which individuals do not work could be reﬁned. For (ii) and (iii), the individuals are assumed to be at ‘home’ from 22:00 to 05:00 while for (i), the individuals are assumed to be at home from 06:00 to 13:00 (see chapter 5). The microenvironments visited during the rest of the day are sampled from the individual activity pools. This could be improved on by either sampling activities directly from an activity diary, such as the 2014/15 Time Use Survey, or using Markov chains to decide what microenvironment is more

likely to be visited next given the microenvironment the individual is in currently. This would give the daily activity sequences more structure and make them more realistic. A pattern that, for example, switches back and forth between ‘home’ and ‘indoor not home’ is, generally speaking, not realistic but could be sampled with the current approach. With the implementation of these new changes, this would not be possible or at least highly unlikely. Secondly, the model restricts the activity to four microenvironments. Had there been more time, the number of microenvironments could have been increased. For example, the microenvironment ‘indoor not home’ could have been split up in microenvironments for work, school, and shops. However, the ability to introduce more microenvironments is dependent on the data about time-use and the availability of accurate parameters for the calculation of the concentration in each microenvironment.

Thirdly, all microenvironments of all individuals of a speciﬁc area are assumed to be located at the same geographical position: the population-weighted centroids of the MSOA. While this is a good place to start, it would be beneﬁcial to distribute the individuals more realistically in the areas. This could be implemented by, for example, using address data to model where individuals are more or less likely to live. Additionally, diﬀerent geographic locations for diﬀerent microenvironments could be added. For example, the location of work could be sampled from labour force data. Zidek et al. (2003) stress the importance of accurate geographic locations in their papers as, otherwise, measurement errors can occur.

Fourthly, diﬀerent activities that inﬂuence the impact the exposure to PM2.5 has on the individual could be included. Certain activities, such as running and playing sports, increase breathing rates and in turn increase the inhalation (Zidek et al., 2007). Therefore, individuals pursuing these activities would be exposed to more PM2.5 than individuals who, for example, sit still. Adding these activities could have an impact on the overall personal exposure and could improve the estimates.

Lastly, some parameters could be made more precise. For example, taking a larger sample for the square-footage of all housing types and stronger assumptions for the ceiling height for each housing type could improve the accuracy of estimating the housing volume for the microenvironment ‘home’. Also, the scope of the sensitivity analysis could be widened to investigate if changes of these parameters have an impact on the ﬁnal personal exposure

estimate.

As mentioned before, the application part of this dissertation would beneﬁt immensely from a model run with a higher number of simulated individuals. A larger time frame for this dissertation and access to more powerful computational hardware could have allowed for such an analysis.

# Conclusion

This dissertation presents how an agent-based model for predicting the personal exposure to pollutants, in this case, particulate matter with a diameter smaller than 2.5µm, can be built. The model combines data from many diﬀerent data sources, such as demographic data, pollutant data, time-use data, temperature data, and labour force data, to model the exposure as accurately as possible. First, individuals are sampled from a simulated population of Devon depending on user-deﬁned characteristics including age, sex, employment status, and whether the individual smokes. Then, activity sequences that indicate which microenvironments an individual visits during a day are sampled from time-use data for each individual. The model diﬀerentiates between three cases: (i) an individual works during the day, (ii) and individual works during the night, and (iii) an individual does not work on the given day. Depending on which scenario applies for the day and the daily average time spent in each microenvironment, unique activity sequences are sampled for each day and individual. Next, the concentration of PM2.5 in these microenvironments, dependent on the geographical location and the ambient concentrations, is estimated. In this process, individual microenvironments are modelled in diﬀerent ways: While most microenvironments are modelled as a linear transformation of the ambient concentrations, the microenvironment ‘home’ is modelled by a mass balance equation that also includes non-ambient PM2.5 generating sources, such as cooking and smoking. Lastly, the personal exposure can be extracted for each individual in hourly time steps, enabling further analysis.

By using papers by Zidek et al. (2003, 2005, 2007) as the basis for the model, a probabilistic structure that allows for statistical inference is introduced. This enables the estimation of days that exceed a speciﬁc threshold along with a measure of uncertainty. The measure of

uncertainty combines standard errors and a measure of the correlation between the days that exceed the threshold. In this dissertation, thresholds from the Daily Air Quality Index (Defra, n.d.) are analysed and compared to the ambient concentration. This revealed that taking the ambient concentrations as a proxy for personal exposure, as discussed in chapter 2, would overestimate the days on which the exposure exceeds low limits by more than double for April (0.9 vs 2). However, the use case presented in this dissertation mainly serves illustrative purposes. Due to hardware limitations, it was only possible to compute a small sample size. To draw conclusions from the model, a simulation using a signiﬁcantly higher number of individuals is recommended.

Nevertheless, the results show that personal exposures vary greatly in space and time. The exposure for the microenvironment ‘home’ is lowest in both cases, followed closely by the exposure in ‘indoor not home’ and ‘outdoor’. The exposure in ‘transport’, however, is much higher than the rest. This shows that it is important to include spatial variability not only in the sense of geographical location but also in the sense of diﬀerent microenvironments that are characterised by properties unique to each environment. Additionally, exposures are generally low during the night and higher during the day when individuals leave their home. This underpins the importance of temporal variability that is included in the model through the hourly ambient concentrations that introduce temporal variability external to the individuals and the activity sequence that introduces temporal variability that is internal to the individuals. The consideration of variability in space and time is one of the main advantages of agent-based models. This is especially important if heterogeneous pollutants, such as PM2.5, are analysed as they are characterised by a higher variability than homogeneous pollutants. Many of the other methods described in chapter 2 are unable to accurately model heterogeneous pollutants as they do not account for either the spatial variability or the temporal variability required. This highlights the beneﬁts that agent-based modelling has over the other methods.

The model presented here makes it very easy to analyse the eﬀect of diﬀerent public health policies on personal exposure to PM2.5 of the population. For example, in this dissertation, the use case of a decrease in ambient pollution by 20% is implemented. The results show that there is no real diﬀerence noticeable between the two scenarios (i) full ambient pollution

and (ii) 80% ambient pollution. One reason for this could be that individuals spend most of their time indoors and outdoor concentration is only one factor in calculating indoor concentrations. Another reason might be that the ambient concentrations in Devon are very low in general so that a decrease of 20% does not aﬀect the personal exposure noticeably. Lastly, the simulated population might just be too small to eﬀectively measure the eﬀect of the policy change. Nevertheless, the use case demonstrates how such an analysis can be implemented with ease.

However, there is still more research to be done and more accurate data needed to maximise the full potential of the model. Some parameters used in the calculation of the microenvironments are older and from North America. Having more recent parameters, such as emissions from cooking and smoking or the air exchange rate, that are estimated from data sourced in the UK, would yield more accurate estimates speciﬁc to the UK. Also, data about residential housing volume would be beneﬁcial and would replace the need for calculating the volume from sampled square footage and ceiling heights. Additionally, longer and more speciﬁc time-use data would enable more realistic modelling of the activity sequences. If the activity diaries also included the geographical location of the activity, such as the location of the workplace and shops, the model could be made much more accurate. The same is true for ﬁner ambient concentration ﬁelds. Coarse ﬁelds cannot give a good estimate of the ambient concentration, especially for heterogeneous pollutants. While agent-based models can add some spatial variability through the diﬀerent calculations of the concentration in individual microenvironments, they still are limited by relying on accurate ambient concentrations. Finally, having validation data would be immensely beneﬁcial to ﬁne-tune the model and ensure the validity of the outputs, however, this would require real individuals to carry mobile modules that record the geographical location and exposure, which comes with its diﬃculties, as discussed in chapter 2.6. Overall, the agent-based modelling approach applied in this dissertation is able to introduce individual human behaviour in the estimation of personal exposure to PM2.5, yielding results that more accurately reﬂect the natural variability of heterogeneous pollutants in space and time.

# Bibliography

Baxter, L.K., Bure, J., Lunden, M., Turpin, B.J., Rich, D.Q., Thevenet-Morrison, K., Hodas,

N. and Özkaynak, H. (2013) Inﬂuence of human activity patterns, particle composition, and residential air exchange rates on modeled distributions of PM2.5 exposure compared with central-site monitoring data. *Journal of Exposure Science & Environmental Epidemiology*, 23:241-247. Available at: https://doi.org/10.1038/jes.2012.118 [Accessed 19 August 2020]

Beelen, R., Hoek, G., Viennau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M., Künzli, N., Schikowski, T., Marcon, A., Eriksen, K.T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., von Klot, S., Nádor, G., Varró, M.J., Dedele, A., Grazuleviciene, R., Mölter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoﬀmann, B., Nonnemacher, M., Krämer, U., Kuhlbusch, T., Cirach, M., de Nazelle, A., Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Strömgren, M., Dons, E., Jerett, M., Fischer, P., Wang, M., Brunekreef, B., and de Hoogh, K. (2013) Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmospheric Environment*, 72:10-23. Available at: https://doi.org/10.1016/j.atmosenv.2013.

02.037 [Accessed 19 August 2020]

Ben, Y., Ma, F., Wang, H., Hassan, M.A., Yevheniia, R., Fan, W., Li, Y. and Dong, Z. (2019) A spatio-temporally weighted hybrid model to improve estimates of personal PM2.5 exposure: Incorporating big data from multiple data sources. *Environmental Pollution*, 253:403-411. Available at: https://doi.org/10.1016/j.envpol.2019.07.034 [Accessed 19 August 2020]

Brauer, M. (2010) How much, how long, what, and where: air pollution exposure assessment for epidemiologic studies of respiratory disease. *Proceedings of the American Thoracic Society*, 7(2):111-5. Available at: https://doi.org/10.1513/pats.200908-093RM [Accessed 06 April 2020].

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R.V., Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J.S, Balakrishnan, K., Barregard, L., Broday, D., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett,

R.T. and Cohen, A. (2016) Ambient air pollution exposure estimation for the global burden of disease 2013. *Environmental Science & Technology*, 50(1):79-88. Available at: https:

//doi.org/10.1021/acs.est.5b03709 [Accessed 06 April 2020].

Breen, M.S., Long, T.C, Schultz, B.D, Crooks, J., Breen, M., Langstaﬀ, J.E., Isaacs, K.K, Tan, Y., Williams, R.W., Cao, Y., Geller, A.M., Devlin, R.B., Batterman, S.A. and Buckley, T.J. (2015a) 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 & Environmental Epidemiology*, 24:412–420. Available at: https://doi.org/10.1038/jes.2014.13 [Accessed 19 August 2020].

Breen, M.S., Long, T.C, Schultz, B.D., Williams, R.W., Richmond-Bryant, J., Breen, M., Langstaﬀ, J.E., Devlin, R.B., Schneider, A., Burke, J.M, Batterman, S.A., Meng, Q.Y. (2015b) Air pollution exposure model for individuals (EMI) in health studies: evaluation for ambient

PM2.5 in Central North Carolina. *Environmental Science & Technology*, 49(24):14184-94. Available at: https://doi.org/10.1021/acs.est.5b02765 [Accessed 06 April 2020].

Burke, J.M., Zufall, M.J., and Özkaynak, H. (2001) A population exposure model for particulate matter: case study results for PM2.5 in Philadelphia, PA. *Journal of Exposure Science & Environmental Epidemiology*, 11:470–489. Available at: https://doi.org/10.1038/sj. jea.7500188 [Accessed 18 August 2020]

Chang, H.H, Fuentes, M. and Frey, H.C. (2012) Time series analysis of personal exposure to ambient air pollution and mortality using an exposure simulator. *Journal of Exposure Science & Environmental Epidemiology*, 22:483–488. Available at: https://doi.org/10.1038/jes.2012.53 [Accessed 06 April 2020].

Copernicus (2020) *CAMS global reanalysis (EAC4)* [online] (6 February 2020). Available at: https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-global-reanalysis-eac4? tab=overview [Accessed 19 August 2020]

Defra (n.d.) *What is the Daily Air Quality Index?* [online]. Available at: https://uk-air.defra. gov.uk/air-pollution/daqi?view=more-info&pollutant=pm25#pollutant [Accessed 19 August 2020]

Delgado-Saborit, J.M., Aquilina, N.J., Meddings, C., Baker, S. and Harrison, R.M. (2011) Relationship of personal exposure to volatile organic compounds to home, work and ﬁxed site outdoor concentrations. *Science of the Total Environment*, 409(3):478-488. Available at: https://doi.org/10.1016/j.scitotenv.2010.10.014 [Accessed 19 August 2020]

Dias, D. and Tchepel, O. (2014) Modelling of human exposure to air pollution in the urban environment: a GPS-based approach *Environmental Science and Pollution Research*, 21:3558–3571. Available at: https://doi.org/10.1007/s11356-013-2277-6 [Accessed 06 April 2020].

Dionisio, K.L., Isakov, V., Baxter, L.K., Sarnat, J.A., Sarnat, S.E., Burke, J., Rosenbaum, A., Graham, S.E., Cook, R., Mulholland, J. and Özkaynak, H. (2013) Development and evaluation of alternative approaches for exposure assessment of multiple air pollutants in Atlanta, Georgia. *Journal of Exposure Science & Environmental Epidemiology*, 23:581–592. Available at: https://doi.org/10.1038/jes.2013.59 [Accessed 06 April 2020].

European Comission (2003) *Indoor air pollution: new EU research reveals higher risks than previously thought* [online] (22 September 2003). Available at: https://ec.europa.eu/ commission/presscorner/detail/en/IP\_03\_1278 [Accessed 19 August 2020]

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliber, J., Fischer, P. and Briggs,

D. (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561-7578. Available at: https://doi.org/10. 1016/j.atmosenv.2008.05.057 [Accessed 19 August 2020]

ICF International (2020) *An introduction to APEX* [online] Research Triangle Park, U.S. Environmental Protection Agency. Available at: https://[www.epa.gov/sites/production/ﬁles/](http://www.epa.gov/sites/production/ﬁles/) 2019-11/documents/apex5\_introduction\_document.pdf [Accessed 06 April 2020].

Jerret, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J. and Giovis, C. (2004) A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science & Environmental Epidemiology volume*, 15:185-204. Available at : https://doi.org/10.1038/sj.jea.7500388 [Accessed 19 August 2020]

Johannesson, S., Rappaport, S.M. and Sallsten, G. (2011) Variability of environmental exposure to ﬁne particles, black smoke, and trace elements among a Swedish population. *Journal of Exposure Science & Environmental Epidemiology*, 21:506-514. Available at: https:

//doi.org/10.1038/jes.2011.13 [Accessed 19 August 2020]

Kim, K., Kabir, E., and Kabir, S. (2015) A review on the human health impact of airborne particulate matter. *Environmental International*, 74:136-143. Available at: https://doi.org/ 10.1016/j.envint.2014.10.005 [Accessed 18 August 2020]

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P.,Behar, J.V., Hern, S.C. and Engelmann, W.H. (2001) The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemology*, 11:231-252. Availablet at: https://doi.org/10.1038/sj. jea.7500165 [Accessed 19 August 2020]

Koehler, K.A. and Peters, T.M. (2015) New methods for personal exposure monitoring for airborne particles. *Current Environmental Health Reports*, 2:399-411. Available at: https://doi.org/10.1007/s40572-015-0070-z [Accessed 19 August 2020]

Liang, L., Gong, P., Cong, N., Li, Z., Zhao, Y. and Chen, Y. (2019) Assessment of per- sonal exposure to particulate air pollution: the ﬁrst result of City Health Outlook (CHO) project. *BMC Public Health*, 19:711. Available at: https://doi.org/10.1186/s12889-019-7022-8 [Accessed 06 April 2020].

Michanowicz, D.R., Shmool, J.L.C., Trunno, B.J., Tripathy, S., Gillooly, S., Kinnee, E. and Clougherty, J.E. (2016) A hybrid land use regression/AERMOD model for predicting intra-urban variation in PM2.5 *Atmospheric Environment*, 131:307-315. Available at: https:

//doi.org/10.1016/j.atmosenv.2016.01.045 [Accessed 19 August 2020]

Morello-Frosch, R., Jesdale, B.M., Sadd, J.L. and Pastor, M. (2010) Ambient air pollution exposure and full-term birth weight in California. *Environmental Health*, 9:44. Available at: https://doi.org/10.1186/1476-069X-9-44 [Accessed 06 April 2020].

Murray, D.M., and Burmaster, D.E. (1995) Residential air exchange rates in the United States: empirical and estimated parametric distributions by season and climatic region. *Risk Analysis*, 15(4):459-465. Available at: https://doi.org/10.1111/j.1539-6924.1995.tb00338.x [Accessed 18 August 2020]

OnAverage (n.d.) *Average Height Ceiling* [online]. Available at: https://[www.onaverage.co.](http://www.onaverage.co/) uk/other-averages/average-height-ceiling [Accessed 24 August 2020]

ONS (2017) *Leisure time in the UK: 2015* [online] (24 October 2017). Available at: https://[www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/articles/](http://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/articles/) leisuretimeintheuk/2015 [Accessed 19 August 2020]

ONS (2020) *Dataset: Adult smoking habits in Great Britain* [online] (07 July 2020). Availalbe at: https://[www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/](http://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/) drugusealcoholandsmoking/datasets/adultsmokinghabitsingreatbritain [Accessed 24 August 2020]

Özkaynak, H., Xue, J., Spengler, J.D., Wallace, L.A., Pellizzari, E.D., and Jenkins, P. (1996) Personal exposure to airborne particles and metals: results from the Particle TEAM Study in Riverside, California. *Journal of Exposure Analysis and Environmental Epidemiology*, 6(1):57-78. Available at: https://[www.researchgate.net/proﬁle/Lance\_Wallace/publication/](http://www.researchgate.net/proﬁle/Lance_Wallace/publication/) 14426195\_Personal\_Exposure\_to\_Airborne\_Particles\_and\_Metals\_Results\_from\_ the\_Particle\_TEAM\_Study\_in\_Riverside\_California/links/5c192bc5299bf12be388cdfe/

Personal-Exposure-to-Airborne-Particles-and-Metals-Results-from-the-Particle-TEAM-Study-in-Riverside- pdf [Accessed 18 August 2020]

Özkaynak, H., Baxter, L.K., Dionisio, K. and Burke, J. (2013) Air pollution exposure prediction approaches used in air pollution epidemiology studies. *Journal of Exposure Science & Environmental Epidemiology*, 23:566–572. Available at: https://doi.org/10.1038/jes.2013.15 [Accessed 06 April 2020].

Rice, M.B., Rifas-Shiman, S.L., Litonjua, A.A., Oken, E., Gillman, M.W., Kloog, I., Luttmann- Gibson, H., Zanobetti, A., Coull, B.A., Schwartz, J., Koutrakis, P., Mittleman, M.A. and Gold, D.R. (2015) Lifetime exposure to ambient pollution and lung function in children. *American Journal of Respiratory an Critical Care Medicine*, 193(8):881-888. Available at: https://doi.org/10.1164/rccm.201506-1058OC [Accessed 19 August 2020]

Sloan, C.D., Philipp, T.J., Bradshaw, R.K., Chronister, S., Barber, W.B. and Johnston, J.D. (2016) Applications of GPS-tracked personal and ﬁxedlocation PM2.5 continuous exposure monitoring. *Journal of the Air & Waste Management Association*, 66(1):53-65. Available at: https://doi.org/10.1080/10962247.2015.1108942 [Accessed 19 August 2020]

Smith, J.D., Mitsakou, C., Kitwiroon, N., Barratt, B.M., Walton, H.A., Tylor, J.G., Anderson, H.R., Kelly, F.J and Beevers, S.D. (2016) London hybrid exposure model: improving human exposure estimates to NO2 and PM2.5 in an urban setting. *Environmental Science & Techonology*, 50(21):11760-11768. Available at: https://doi.org/10.1021/acs.est.6b01817 [Accessed 19 August 2020]

Steinle, S., Reis, S. and Sabel, C.E. (2013) Quantifying human exposure to air pollution — Moving from static monitoring to spatio-temporally resolved personal exposure assessment. *Sciene of Total Environment*, 443:184-193. Available at: https://doi.org/10.1016/j.scitotenv. 2012.10.098 [Accessed 19 August 2020]

Steinle, S., Reis, S., Sabel, C.E., Semple, S., Twigg, M.M., Braban, C.F., Leeson, S.R., Heal, M.R., Harrison, D., Lin, C. and Wu, H. (2015) Personal exposure monitoring of PM2.5 in indoor and outdoor microenvironments. *Science of The Total Envrionment*, 508:383-394. Availalbe at: https://doi.org/10.1016/j.scitotenv.2014.12.003 [Accessed 19 August 2020]

Su, J.G., Jerret, M., Meng, Y., Pickett, M. and Ritz, B. (2015) Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure

assessment. *Science of The Total Environment*, 506-507:518-526. Available at: https:

//doi.org/10.1016/j.scitotenv.2014.11.022 [Accessed 19 August 2020]

Sun, Q., Zhuang, J., Du, Y., Xu, D. and Li, T. (2018) Design and application of a web-based real-time personal PM2.5 exposure monitoring system. *Science of The Total Environment*, 637:852-859. Available at: https://doi.org/10.1016/j.scitotenv.2018.01.299 [Accessed 19 August 2020]

The Alan Turing Institute (n.d.) *Synthetic population estimation and scenario pro- jection* [online]. Available at: https://[www.turing.ac.uk/research/research-projects/](http://www.turing.ac.uk/research/research-projects/) synthetic-population-estimation-and-scenario-projection [Accessed 19 August 2020]

U.S. Environmental Protection Agency (2019a) *Air pollutants exposure model documenta- tion (APEX, Version 5.2) Volume I: user’s guide* [online] Research Triangle Park, North Carolina, U.S. Environmental Protection Agency. Available at: https://[www.epa.gov/sites/](http://www.epa.gov/sites/) production/files/2019-12/documents/apex52\_usersguide\_vol1\_october2019.pdf [Accessed 06 April 2020].

U.S. Environmental Protection Agency (2019b) *Air pollutants exposure model documentation (APEX, Version 5.2) Volume II: technical support document* [online] Research Triangle Park, North Carolina, U.S. Environmental Protection Agency. Available at: https://[www.epa.gov/](http://www.epa.gov/) sites/production/files/2019-12/documents/apex52\_usersguide\_vol2\_october2019.pdf [Ac- cessed 06 April 2020].

Wallace, L.A., Özkaynak, H., Spengler, J.D., Pellizzari, E.D., Jenkins, P. (1993) *Indoor, outdoor, and personal air exposures to particles, elements, and nicotine for 178 residents of Riverside, California.* Washington, D.C., U.S. Environmental Protection Agency. Available at: https://cfpub.epa.gov/si/si\_public\_record\_report.cfm?dirEntryId=47120&Lab=ORD& keyword=jenkins&actType=&TIMSType=+&TIMSSubTypeID=&DEID=&epaNumber= &ntisID=&archiveStatus=Both&ombCat=Any&dateBeginCreated=&dateEndCreated= &dateBeginPublishedPresented=&dateEndPublishedPresented=&dateBeginUpdated= &dateEndUpdated=&dateBeginCompleted=&dateEndCompleted=&personID=&role= Any&journalID=&publisherID=&sortBy=revisionDate&count=50 [Accessed 29 June 2020]

WHO (2018) *Ambient (outdoor) air pollution* [online] (2 May 2018). Available at: https:

//[www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health](http://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health) [Ac- cessed 19 August 2020]

Zanobetti, A. and Schwartz, J. (2009) The effect of fine and coarse particulate air pollution on mortality: a national analysis. *Environmental Health Perspectives*, 117(6):898-903. Available at: https://doi.org/10.1289/ehp.0800108 [Accessed 19 August 2020]

Zidek, J.V., Meloche, J., Shaddick, G., Chatfield, C. and White, R. (2003) A computational model for estimating personal exposure to air pollutants with application to London’s PM10 in 1997. *Technical Report of the Statistical and Applied Mathematical Sciences Institute*, 3:1-36. Available at: https://[www.researchgate.net/publication/255668378\_A\_](http://www.researchgate.net/publication/255668378_A_) Computational\_Model\_for\_Estimating\_Personal\_Exposure\_to\_Air\_Pollutants\_with\_ Application\_to\_London’s\_PM10\_in\_1997 [Accessed 06 April 2020].

Zidek, J.V., Shaddick, G., White, R., Meloche, J., and Chatfield, C. (2005) Using a prob- abilistic model (pCNEM) to estimate personal exposure to air pollution. *Environmetrics*, 16(5):481–493. Available at: https://doi.org/10.1002/env.716 [Accessed 06 April 2020].

Zidek, J.V., Shaddick, G., Meloche, J., Chatfield, C. and White, R. (2007) A framework for predicting personal exposures to environmental hazards. *Environmental and Ecological Statistics*, 14:411–431. Available at: https://doi.org/10.1007/s10651-007-0028-x [Accessed 06 April 2020].

Zou, B., Wilson, J.G., Zhan, F.B. and Zeng, Y. (2009) Air pollution exposure assessment methods utilized in epidemiological studies. *Journal of Environmental Monitoring*, 11(3):475-

490. Available at: https://doi.org/10.1039/b813889c [Accessed 19 August 2020]

# Appendix: Code

## Packages

*# import libraries* **library**(tidyverse) **library**(raster) **library**(sf) **library**(maptools)

**library**(lubridate) *# to work with dates*

**library**(EnvStats) *# for triangular distribution*

**library**(truncnorm) *# to truncate Normal distributions from Burke et al. (2001)*

**library**(lhs) *# for Latin Hypercube Sampling*

## Population data

*##################################################*

*############ import and wrangle data ############# ##################################################*

*# read population data*

devon\_data <- **read.csv**("data/final/Devon\_simulated\_TU\_keyworker\_health.csv")

*# add id column to Devon dataset and order columns*

devon\_data**$**id <- 1**:length**(devon\_data**$**area)

devon\_data <- dplyr**::select**(devon\_data, id, **everything**())

*# subset with relevant columns*

population\_data <- dplyr**::select**(devon\_data, id, area, hid, pid, Sex, Age1, pwkstat,

smoke, punknown**:**ptransport, pid\_tus)

*# add rounded hours spent in each ME*

population\_data <- population\_data **%>%**

**mutate**(rhome = **ceiling**(phome **\*** 24 **+** pworkhome **\*** 24),

rindoor = **ceiling**(pwork **\*** 24 **+** pschool **\*** 24 **+** pshop **\*** 24 **+** pservices **\*** 24 **+**

pescort **\*** 24), rleisure = **ceiling**(pleisure **\*** 24), rtransport = **ceiling**(ptransport **\*** 24))

*# add total of rounded hours*

population\_data <- population\_data **%>% mutate**(rtotal = **rowSums**(.[20**:**23]))

*# get all time use patterns*

TUS <- population\_data **%>%**

**group\_by**(pid\_tus, Sex, Age1, pwkstat, punknown, phome, pworkhome, pwork, pschool, pshop, pservices, pleisure, pescort,

ptransport, rhome, rindoor, rtransport, rleisure, rtotal) **%>%**

**count**()

*# subset of al complete time use patterns*

TUS\_complete <- **subset**(TUS, rtotal **>=** 24 **&** rhome **>=** 8)

*# subset of all broken time use patterns*

TUS\_broken <- **subset**(TUS, **!**(rtotal **>=** 24 **&** rhome **>=** 8))

*##### find out if all broken time use patterns can be resampled from same strata # this is possible if there are more timeuse patterns in strata than are broken # for the strata*

*# find number of all unique patterns for each strata*

overall\_timeuse <- population\_data **%>% group\_by**(Sex, Age1, pwkstat) **%>% mutate**(unique\_types = **n\_distinct**(pid\_tus))

overall\_timeuse <- overall\_timeuse **%>% group\_by**(Sex, Age1, pwkstat) **%>% summarise**(total\_unique = **mean**(unique\_types))

*# find number of broken unique patterns for each strata*

broken\_timeuse <- **subset**(population\_data, **!**(rtotal **>=** 24 **&** rhome **>=** 8)) **%>% group\_by**(Sex, Age1, pwkstat) **%>%**

**mutate**(unique\_types = **n\_distinct**(pid\_tus))

broken\_timeuse <- broken\_timeuse **%>% group\_by**(Sex, Age1, pwkstat) **%>% summarise**(broken\_unique = **mean**(unique\_types))

*# compare number of all patterns to number of broken and see which strata cannot be # resampled*

compare\_timeuse <- **left\_join**(overall\_timeuse, broken\_timeuse) **%>% mutate**(remaining\_unique = total\_unique **-** broken\_unique)

*# there are five individuals for which the time use pattern is broken and cannot be # resampled*

*# save ids of those individuals*

badid <- **c**(291063, 186008, 444492, 660272, 184296)

*# exlude bad ids from dataset*

pop\_data <- **subset**(population\_data, **!**(id **%in%** badid))

*# split in two datasets: 1: subset with broken patterns, 2: subset with complete # patterns*

pop\_data1 <- **subset**(pop\_data, pid\_tus **%in%** TUS\_broken**$**pid\_tus) pop\_data2 <- **subset**(pop\_data, **!**(pid\_tus **%in%** TUS\_broken**$**pid\_tus))

*# sample new time use patterns for bad time use patterns*

**for** (i **in** 1**:nrow**(pop\_data1)){

*# subset of possible complete time use patterns of strata*

temp <- **as.data.frame**(**subset**(TUS\_complete, Sex **==** pop\_data1**$**Sex[i] **&**

Age1 **==** pop\_data1**$**Age1[i] **&**

pwkstat **==** pop\_data1**$**pwkstat[i]))

*# select random time use pattern*

pattern <- **sample\_n**(temp, 1)

*# fill in new pattern*

pop\_data1[i, 9**:**24] <- **c**(pattern**$**punknown, pattern**$**phome, pattern**$**pworkhome,

pattern**$**pwork, pattern**$**pschool, pattern**$**pshop, pattern**$**pservices, pattern**$**pleisure, pattern**$**pescort, pattern**$**ptransport, pattern**$**pid\_tus,

pattern**$**rhome, pattern**$**rindoor, pattern**$**rtransport, pattern**$**rleisure, pattern**$**rtotal)

}

*# merge the two sets and organise, and duplicate pwkstat column for next steps*

population\_data <- **bind\_rows**(pop\_data1, pop\_data2) **%>% mutate**(work = pwkstat) **%>%**

dplyr**::select**(id, area, hid, pid, Sex, Age1, pwkstat, work, **everything**()) **%>% arrange**(id)

*##### now a column specifying if an individual works is created # change levels of work column to 1 if works and 0 if not*

**levels**(population\_data**$**work) <- **c**(0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0)

*# change column class to factor and adjust values so that 1 if works and 0 if not*

population\_data**$**work <- **as.numeric**(population\_data**$**work)

population\_data <- **mutate**(population\_data, work = work-1)

*# check if dataset tidy*

**summary**(population\_data)

*# fill missing smoking status with 0*

population\_data[**is.na**(population\_data)] <- 0

*##################################################*

*############ add coordinates of area ############# ##################################################*

*# import shapefile, converte to longitude/latitude coordinates and organise data*

shape <- rgdal**::readOGR**(dsn = "Data/Middle\_Layer\_Super\_Output\_Areas December\_2011\_

\_Population\_Weighted\_Centroids-shp",

layer = "Middle\_Layer\_Super\_Output\_Areas December\_2011\_

\_Population\_Weighted\_Centroids")

**proj4string**(shape)

shapeLL = **spTransform**(shape, **CRS**("+init=epsg:4326")) converted\_coordinates <- **as.data.frame**(shapeLL) **colnames**(converted\_coordinates) <- **c**("id", "msoa", "name", "long", "lat")

*# get unique values for MSOAs from population dataset*

devon\_msoa <- **unique**(population\_data**$**area)

*# get corresponding coordinates in Devon*

devon\_msoa\_coordinates <- **subset**(converted\_coordinates, msoa **%in%** devon\_msoa)

*# create longitude and latitude column in population dataset and organise*

population\_data**$**Longitude <- NA population\_data**$**Latitude <- NA

population\_data <- dplyr**::select**(population\_data, id, area, Longitude, Latitude, hid,

pid, Sex, Age1, pwkstat, work, **everything**())

*# add coordinates to population\_data*

**for** (i **in** 1**:nrow**(population\_data)) {

population\_data**$**Longitude[i] <- **subset**(devon\_msoa\_coordinates, msoa **==**

**as.character**(population\_data**$**area[i]))**$**long population\_data**$**Latitude[i] <- **subset**(devon\_msoa\_coordinates, msoa **==**

**as.character**(population\_data**$**area[i]))**$**lat

}

*# save complete dataset*

**save**(population\_data, file = "population\_data\_location.RData")

## Pollutant data

*##################################################*

*############ import and wrangle data ############# ##################################################*

*# read in data*

pm\_full\_brick <- **brick**("Data/final/pm2.5\_UK\_2019.grib")

*# get coords for region just bigger than Devon and crop data*

newextent <- **c**(**-**7, -1, 48, 53)

pm\_brick <- **crop**(pm\_full\_brick, newextent) pm\_data <- **as.data.frame**(pm\_brick, xy = TRUE)

*### create tidy df*

*# collapse and name columns*

pm\_tidy <- **pivot\_longer**(pm\_data, **starts\_with**("pm"), names\_to = "tmp", values\_to = "PM2.5")

**colnames**(pm\_tidy) <- **c**("Longitude", "Latitude", "Tmp", "PM2.5")

*# add dates*

Dates <- **seq**(**as.Date**("2018/12/25"), **as.Date**("2019/12/31"), by = "day") pm\_tidy**$**Date <- **rep**(Dates, each = 8)

*# add time column*

TimeSteps <- **c**(0, 3, 6, 9, 12, 15, 18, 21)

pm\_tidy**$**Time <- TimeSteps

*# transform unit*

pm\_tidy <- **mutate**(pm\_tidy, PM2.5 = PM2.5 **\*** 1000000000)

*# select and order relevant columns*

pm\_tidy <- dplyr**::select**(pm\_tidy, Longitude, Latitude, Date, Time, PM2.5)

*##################################################*

*############### space interpolation ############## ##################################################*

*# define function to interpolate space*

*# long\_int and lat\_int specify how fine interpolation*

interpolate\_space <- **function**(PM\_data, long\_int, lat\_int) {

*# create vectors of unique date and time values to loop through*

dates <- **unique**(PM\_data**$**Date) times <- **unique**(PM\_data**$**Time)

*# initilaise final dataframe*

df <- **data.frame**()

*# looping through days*

**for** (i **in** 1**:length**(dates)) {

*# select rows that correspond to date*

df\_by\_date <- **filter**(PM\_data, Date **==** dates[i])

*# looping through times*

**for** (j **in** 1**:length**(times)) {

*# select rows that correspond to time*

df\_by\_date\_and\_time <- **filter**(df\_by\_date, Time **==** times[j])

*# interpolate*

int <- akima**::interp**(x = df\_by\_date\_and\_time**$**Longitude,

y = df\_by\_date\_and\_time**$**Latitude, z = df\_by\_date\_and\_time**$**PM2.5,

nx = long\_int, ny = lat\_int)

*### create dataframe*

*# extract longitude and latitude from interpolated list object*

lon <- int[[1]]

lat <- int[[2]]

*# start dataframe with interpolated PM values*

PM <- **as.data.frame**(int[[3]])

*# add longitude to dataframe*

PM**$**Longitude <- lon

*# transform columns into rows and add latitude*

PM <- **pivot\_longer**(PM, **starts\_with**("V"), names\_to = "Latitude", values\_to = "PM2.5")

PM**$**Latitude <- lat

*# add date and time*

PM**$**Date <- dates[i]

PM**$**Time <- times[j]

*# sort columns*

PM <- dplyr**::select**(PM, Longitude, Latitude, Date, Time, PM2.5)

*# merge to final dataframe*

df <- **bind\_rows**(df, PM)

}

}

*# return final dataframe*

**return**(df)

}

*##################################################*

*############### time interpolation ############### ##################################################*

*# define function to interpolate between 3 hours*

interpolate\_time <- **function**(PM\_data) {

*### create dataframe with missing hours # define vectors*

times\_missing <- **c**(1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23)

dates <- **unique**(PM\_data**$**Date) lons <- **unique**(PM\_data**$**Longitude) lats <- **unique**(PM\_data**$**Latitude)

*# create rows from all combinations of the above and rename columns* missing\_df <- **expand.grid**(lons, lats, dates, times\_missing) **colnames**(missing\_df) <- **c**("Longitude", "Latitude", "Date", "Time")

*# merge dataframe with PM measurements and dataframe with missing values # and arrange*

total\_df <- **bind\_rows**(PM\_data, missing\_df)

total\_df <- **arrange**(total\_df, Longitude, Latitude, Date, Time)

*# add id column to interpolate missing values*

total\_df**$**id <- 1**:length**(total\_df**$**Longitude)

*# transform into zoo object and interpolate*

total\_zoo <- zoo**::zoo**(total\_df**$**PM2.5, total\_df**$**id)

total\_zoo <- imputeTS**::na\_interpolation**(total\_zoo, option = "linear")

*# create final dataframe*

tmp <- zoo**::fortify.zoo**(total\_zoo) total\_df**$**PM2.5 <- tmp**$**total\_zoo total\_df[, 1**:**5]

}

*##################################################*

*################ interpolate data ################ ##################################################*

space\_int\_df <- **interpolate\_space**(pm\_tidy, 52, 37) space\_time\_int\_df <- **interpolate\_time**(space\_int\_df)

*# round values and select only for Devon*

PM\_data <- space\_time\_int\_df **%>%**

**mutate**(Longitude = **round**(Longitude, 5), Latitude = **round**(Latitude, 5)) **%>% filter**(Longitude **<** -2 **&** Longitude **>** -6 **&** Latitude **<** 51.5 **&** Latitude **>** 49.5)

*# save data*

**save**(PM\_data, file = "PM\_2019\_fine.RData")

*################################################## ################### create plot ################## ##################################################*

*# This is a data set from the maptools package*

**data**(wrld\_simpl)

*# Create a data.frame object for ggplot. ggplot requires a data frame.*

mymap <- **fortify**(wrld\_simpl)

**ggplot**() **+**

**geom\_tile**(data = PM\_data[ **which**( PM\_data**$**Date **== as.Date**("2019/05/01") **&**

PM\_data**$**Time **==** 12) , ],

**aes**(x = Longitude, y = Latitude, fill = PM2.5)) **+ geom\_point**(data = PM\_data[ **which**( PM\_data**$**Date **== as.Date**("2019/05/01") **&**

PM\_data**$**Time **==** 12) , ],

**aes**(x = Longitude, y = Latitude), pch = 20) **+**

**geom\_map**(data = mymap, map = mymap, **aes**(x = long, y = lat, map\_id = id), alpha = 0, fill = "white", color = **grey**(0)) **+**

**scale\_fill\_gradientn**(colours = **c**("#9CFF9C", "#9CFF9C", "#31FF00", "#31CF00",

"#FFFF00", "#FFCF00", "#FF9A00", "#FF6464",

"#FF0000", "#990000"),

values = **c**(0, 11**/**70, 23**/**70, 35**/**70, 41**/**70, 47**/**70, 53**/**70, 58**/**70, 64**/**70, 1),

breaks = **c**(0, 11, 23, 35, 41, 47, 53, 58, 64),

limits = **c**(0.0, 70), oob = scales**::**squish, name = "µg/m3") **+**

**scale\_x\_continuous**(limits = **c**(**-**6, -1.95), expand = **c**(0, 0)) **+ scale\_y\_continuous**(limits = **c**( 49.515, 51.5265), expand = **c**(0, 0)) **+ coord\_equal**() **+**

**guides**(fill = **guide\_colorbar**(barheight = 9)) **+**

**labs**(title = "Ambient pollution field of PM2.5", subtitle = "01/05/2019, 12:00", x = "Longitude", y = "Latitude") **+**

**theme**(panel.grid = **element\_blank**())

*##################################################*

*############## create rollback data ############## ##################################################*

*# decrease pollution by 20%*

PM\_rollback\_data <- PM\_data **%>% mutate**(PM2.5 = PM2.5 **\*** 0.8)

*# save data*

**save**(PM\_rollback\_data, file = "PM\_rollback\_2019\_fine.RDATA")

## Temperature data

*##################################################*

*############ import and wrangle data ############# ##################################################*

*# read in data*

temp\_full\_brick <- **brick**("Data/final/temperature\_UK\_2019.grib")

*# get coords for region just bigger than Devon and crop data*

newextent <- **c**(**-**7, -1, 48, 53)

temp\_brick <- **crop**(temp\_full\_brick, newextent) temp\_data <- **as.data.frame**(temp\_brick, xy = TRUE)

*### create tidy df*

*# collapse and name columns*

temp\_tidy <- **pivot\_longer**(temp\_data, **starts\_with**("temp"), names\_to = "Tmp",

values\_to = "Temp\_Kelvin")

**colnames**(temp\_tidy) <- **c**("Longitude", "Latitude", "Tmp", "Temp\_Kelvin")

*# add dates*

Dates <- **seq**(**as.Date**("2018/12/25"), **as.Date**("2019/12/31"), by = "day") temp\_tidy**$**Date <- **rep**(Dates, each = 8)

*# add time column*

TimeSteps <- **c**(0, 3, 6, 9, 12, 15, 18, 21)

temp\_tidy**$**Time <- TimeSteps

*# transform unit*

temp\_tidy <- **mutate**(temp\_tidy, Temp\_Celsius = Temp\_Kelvin **-** 273.15)

*# select and order relevant columns*

temp\_tidy <- dplyr**::select**(temp\_tidy, Longitude, Latitude, Date, Time, Temp\_Celsius)

*##################################################*

*############### space interpolation ############## ##################################################*

*# define function to interpolate space*

*# long\_int and lat\_int specify how fine interpolation*

interpolate\_space <- **function**(temp\_data, long\_int, lat\_int) {

*# create vectors of unique date and time values to loop through*

dates <- **unique**(temp\_data**$**Date) times <- **unique**(temp\_data**$**Time)

*# initilaise final dataframe*

df <- **data.frame**()

*# looping through days*

**for** (i **in** 1**:length**(dates)) {

*# select rows that correspond to date*

df\_by\_date <- **filter**(temp\_data, Date **==** dates[i])

*# looping through times*

**for** (j **in** 1**:length**(times)) {

*# select rows that correspond to time*

df\_by\_date\_and\_time <- **filter**(df\_by\_date, Time **==** times[j])

*# interpolate*

int <- akima**::interp**(x = df\_by\_date\_and\_time**$**Longitude,

y = df\_by\_date\_and\_time**$**Latitude,

z = df\_by\_date\_and\_time**$**Temp\_Celsius, nx = long\_int, ny = lat\_int)

*### create dataframe*

*# extract longitude and latitude from interpolated list object*

lon <- int[[1]]

lat <- int[[2]]

*# start dataframe with interpolated temp values*

temp <- **as.data.frame**(int[[3]])

*# add longitude to dataframe*

temp**$**Longitude <- lon

*# transform columns into rows*

temp <- **pivot\_longer**(temp, **starts\_with**("V"), names\_to = "Latitude", values\_to = "Temp\_Celsius")

temp**$**Latitude <- lat

*# add date and time* temp**$**Date <- dates[i] temp**$**Time <- times[j]

*# sort columns*

temp <- dplyr**::select**(temp, Longitude, Latitude, Date, Time, Temp\_Celsius)

*# merge to final dataframe*

df <- **bind\_rows**(df, temp)

}

}

*# return final dataframe*

**return**(df)

}

*##################################################*

*############### time interpolation ############### ##################################################*

*# define function to interpolate between 3 hours*

interpolate\_time <- **function**(temp\_data) {

*### create dataframe with missing hours # define vectors*

times\_missing <- **c**(1, 2, 4, 5, 7, 8, 10, 11, 13, 14, 16, 17, 19, 20, 22, 23)

dates <- **unique**(temp\_data**$**Date) lons <- **unique**(temp\_data**$**Longitude) lats <- **unique**(temp\_data**$**Latitude)

*# create rows from all combinations of the above and rename columns* missing\_df <- **expand.grid**(lons, lats, dates, times\_missing) **colnames**(missing\_df) <- **c**("Longitude", "Latitude", "Date", "Time")

*# merge dataframe with temp measurements and dataframe with missing values # and arrange*

total\_df <- **bind\_rows**(temp\_data, missing\_df)

total\_df <- **arrange**(total\_df, Longitude, Latitude, Date, Time)

*# add id column to interpolate missing values*

total\_df**$**id <- 1**:length**(total\_df**$**Longitude)

*# transform into zoo object and interpolate*

total\_zoo <- zoo**::zoo**(total\_df**$**Temp\_Celsius, total\_df**$**id)

total\_zoo <- imputeTS**::na\_interpolation**(total\_zoo, option = "linear")

*# create final dataframe*

tmp <- zoo**::fortify.zoo**(total\_zoo) total\_df**$**Temp\_Celsius <- tmp**$**total\_zoo total\_df[, 1**:**5]

}

*##################################################*

*################ interpolate data ################ ##################################################*

space\_int\_df <- **interpolate\_space**(temp\_tidy, 52, 37) space\_time\_int\_df <- **interpolate\_time**(space\_int\_df)

*# round values and select only for Devon*

temp\_data <- space\_time\_int\_df **%>%**

**mutate**(Longitude = **round**(Longitude, 5), Latitude = **round**(Latitude, 5)) **%>%**

**filter**(Longitude **<** -2 **&** Longitude **>** -6 **&** Latitude **<** 51.5 **&** Latitude **>** 49.5)

*# find daily averages*

temp\_avg\_Devon\_data <- temp\_data **%>% group\_by**(Longitude, Latitude, Date) **%>% summarise**(Average = **mean**(Temp\_Celsius))

*# save to data file*

**save**(temp\_avg\_Devon\_data, file = "temp\_avg\_2019\_Devon\_fine.RData")

*################################################## ################## create plot ################### ##################################################*

*# This is a data set from the maptools package*

**data**(wrld\_simpl)

*# Create a data.frame object for ggplot. ggplot requires a data frame.*

mymap <- **fortify**(wrld\_simpl)

*# plot*

**ggplot**() **+**

**geom\_tile**(data = temp\_avg\_Devon\_data[ **which**( temp\_avg\_Devon\_data**$**Date **==**

**as.Date**("2019/02/03")) , ],

**aes**(x = Longitude, y = Latitude, fill = Average)) **+ geom\_point**(data = PM\_data[ **which**( PM\_data**$**Date **== as.Date**("2019/02/03") **&**

PM\_data**$**Time **==** 12) , ],

**aes**(x = Longitude, y = Latitude), pch = 20) **+**

**geom\_map**(data = mymap, map = mymap, **aes**(x = long, y = lat, map\_id = id), alpha = 0, fill = "white", color = **grey**(0)) **+**

**scale\_fill\_gradientn**(colours = **c**("blue", "green", "yellow", "orange", "red"), breaks = **c**(0, 5, 10, 15, 20), limits = **c**(0.0, 25),

oob = scales**::**squish, name = "°C") **+ scale\_x\_continuous**(limits = **c**(**-**6, -1.95), expand = **c**(0, 0)) **+ scale\_y\_continuous**(limits = **c**( 49.5, 51.53), expand = **c**(0, 0)) **+ coord\_equal**() **+**

**guides**(fill = **guide\_colorbar**(barheight = 9)) **+**

**labs**(title = "Spatial field of average daily temperature", subtitle = "03/02/2019",

x = "Longitude", y = "Latitude") **+ theme**(panel.grid = **element\_blank**())

## Model building

*################################################## ################## import data ################### ##################################################*

*# read population data*

**load**("Data/final/population\_data\_location.RDATA")

*# load particulate matter data* **load**("Data/final/PM\_2019\_fine.RDATA") **load**("Data/final/PM\_rollback\_2019\_fine.RDATA")

*# load temperature data*

**load**("Data/final/temp\_avg\_2019\_Devon\_fine.RDATA")

*##################################################*

*############### match coordinates ################ ##################################################*

*##### match coordinates to of population data to PM data # get coordinates of PM data*

long\_pm <- **unique**(PM\_data**$**Longitude)

lat\_pm <- **unique**(PM\_data**$**Latitude)

*# get coordinates of population dataa* long\_pop <- **unique**(population\_data**$**Longitude) lat\_pop <- **unique**(population\_data**$**Latitude)

*# create dataframe to match coords* match\_long <- **data.frame**(long\_pop) match\_lat <- **data.frame**(lat\_pop)

*# match longitude (closest distance) and fill in dataframe*

**for** (i **in** 1**:nrow**(match\_long)) {

dist <- **sqrt**((match\_long[i, 1] **-** long\_pm)**^**2)

match\_long[i, 2] <- long\_pm[**match**(**min**(dist), dist)]

}

*# match latitude (closest distance) and fill in dataframe*

**for** (i **in** 1**:nrow**(match\_lat)) {

dist <- **sqrt**((match\_lat[i, 1] **-** lat\_pm)**^**2)

match\_lat[i, 2] <- lat\_pm[**match**(**min**(dist), dist)]

}

*# change column names* **colnames**(match\_long)[2] <- "long\_match" **colnames**(match\_lat)[2] <- "lat\_match"

*# add column in population data for coords match* population\_data**$**Longitude\_match <- NA population\_data**$**Latitude\_match <- NA

*# match coordinates of PM data to population\_data*

**for** (i **in** 1**:nrow**(population\_data)) { population\_data**$**Longitude\_match[i] <-

**subset**(match\_long, long\_pop **==** population\_data**$**Longitude[i])**$**long\_match population\_data**$**Latitude\_match[i] <-

**subset**(match\_lat, lat\_pop **==** population\_data**$**Latitude[i])**$**lat\_match

}

*# arrange columns and save data*

population\_data <- dplyr**::select**(population\_data, id, area, Longitude, Latitude,

Longitude\_match, Latitude\_match, hid, pid, Sex, Age1, pwkstat, work, **everything**())

**save**(population\_data, file = "population\_data\_location\_matched.RData")

*##################################################*

*############# open microenvironments ############# ##################################################*

*# define function to calculate local concentration of microenvironment outdoor*

calculate\_outdoor <- **function**(lon, lat, date, time) {

*# parameters (Zidek et al. (2003))*

a <- 0

b <- 1

*# extract ambient concentration*

tmp <- **subset**(PM\_data, Longitude **==** lon **&** Latitude **==** lat **&**

Date **== as.Date**(date) **&** Time **==** time) AmbCon <- tmp**$**PM2.5

*# calculate local concentration*

a **+** b **\*** AmbCon

}

*##################################################*

*# define function to calculate local concentration of microenvironment indoor*

calculate\_indoor <- **function**(lon, lat, date, time) {

*# parameters (Normal from Burke et al. (2001) truncated)*

a <- **rtruncnorm**(n = 1, a = 0, mean = 6.467, sd = 2.1) b <- **rtruncnorm**(n = 1, a = 0, mean = 0.507, sd = 0.11)

*# extract ambient concentration*

tmp <- **subset**(PM\_data, Longitude **==** lon **&** Latitude **==** lat **&**

Date **== as.Date**(date) **&** Time **==** time) AmbCon <- tmp**$**PM2.5

*# calculate local concentration (Normal from Burke et al. (2001) truncated)*

**rtruncnorm**(n = 1, a = 0, mean = a **+** b **\*** AmbCon, sd = 3.467)

}

*##################################################*

*# define function to calculate local concentration of microenvironment car*

calculate\_transport <- **function**(lon, lat, date, time) {

*# parameters (Normal from Burke et al. (2001) truncated)*

a <- **rtruncnorm**(n = 1, a = 0, mean = 33, sd = 7.2)

b <- **rtruncnorm**(n = 1, a = 0, mean = 0.26, sd = 0.14)

*# extract ambient concentration*

tmp <- **subset**(PM\_data, Longitude **==** lon **&** Latitude **==** lat **&**

Date **== as.Date**(date) **&** Time **==** time) AmbCon <- tmp**$**PM2.5

*# calculate local concentration (Normal from Burke et al. (2001) truncated)*

**rtruncnorm**(n = 1, a = 0, mean = a **+** b **\*** AmbCon, sd = 12)

}

*##################################################*

*############ closed microenvironments ############ ##################################################*

*##### find region for air exchange rate (Murray and Burmaster (1995)) # calculate annual heating degree days for Devon*

temp <- temp\_avg\_Devon\_data **%>%**

**mutate**(Average = (Average**\***9**/**5)**+**32) **%>% group\_by**(Date) **%>%**

**summarise**(Average = **mean**(Average)) **%>% filter**(Date **>= as.Date**("2019-01-01")) **%>% filter**(Average **<** 65) **%>%**

**mutate**(Diff = 65 **-** Average)

**sum**(temp**$**Diff) *# 4170 -> use region 3*

*# calculate annual heating degree days for Exeter*

temp <- temp\_avg\_Devon\_data **%>% mutate**(Average = (Average**\***9**/**5)**+**32) **%>% group\_by**(Date) **%>%**

**filter**(Latitude **==** 50.72917 **&** Longitude **==** -3.55882) **%>% filter**(Date **>= as.Date**("2019-01-01")) **%>%**

**filter**(Average **<** 65) **%>% mutate**(Diff = 65 **-** Average)

**sum**(temp**$**Diff) *# 4618 -> use region 3*

*##### calculate average consumption of cigarettes per day # (ONS, Adult Smoking Habits in GB (2019))*

*### MALE*

male\_age\_1 <- 4.0 *# age 16-24*

male\_age\_2 <- 5.1 *# age 25-34*

male\_age\_3 <- 10.9 *# age 35-49*

male\_age\_4 <- 12.4 *# age 50-59*

male\_age\_5 <- 12.9 *# age 60+ # adjust to model age groups*

*# under 19*

male\_age\_1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *# 19*  (5 **\*** | *- 29*  male\_age\_1 | | **+** | 5 **\*** male\_age\_2)**/**10 | | | |
| *# 30*  (5 **\*** | *- 44*  male\_age\_2 | | **+** | 10 | **\*** male\_age\_3)**/**15 | | |
| *# 45*  (5 **\*** | *- 59*  male\_age\_3 | | **+** | 10 | **\*** male\_age\_4)**/**15 | | |
| *# 60* | *- 74* | |  |  |  | | |
| male\_age\_5 | |  |  | |  |  |  |
| *# over 74*  male\_age\_5 | |  |  | |  |  |  |
| *### FEMALE*  female\_age\_1 | | <- | 5 | | *#* | *age* | *16-24* |
| female\_age\_2 | | <- | 8.8 | | *#* | *age* | *25-34* |
| female\_age\_3 | | <- | 8.1 | | *#* | *age* | *35-49* |
| female\_age\_4 | | <- | 11.7 | | *#* | *age* | *50-59* |
| female\_age\_5 | | <- | 10.8 | | *#* | *age* | *60+* |

*# adjust to model age groups*

*# under 19*

female\_age\_1

*# 19 - 29*

(5 **\*** female\_age\_1 **+** 5 **\*** female\_age\_2)**/**10

*# 30 - 44*

(5 **\*** female\_age\_2 **+** 10 **\*** female\_age\_3)**/**15

*# 45 - 59*

(5 **\*** female\_age\_3 **+** 10 **\*** female\_age\_4)**/**15

*# 60 - 74*

female\_age\_5

*# over 74*

female\_age\_5

*##### define function to calculate local concentration of microenvironment home*

calculate\_home <- **function**(lon, lat, date, time, housetype, smoker, sex,

age, Cbefore) {

*# penetration factor (Özkaynak et al. (1996))*

Fp <- **rnorm**(n = 1, mean = 1, sd = 0.055)

*# deposition rate (Özkaynak et al. (1996))*

Fd <- **rnorm**(n = 1, mean = 0.39, sd = 0.0825)

*# emission generating source (Özkaynak et al. (1996))* SCigarette <- **rnorm**(n = 1, mean = 13.8, sd = 1.775) SCooking <- **rnorm**(n = 1, mean = 1.7, sd = 0.325) SOther <- **rnorm**(n = 1, mean = 1.1, sd = 0.525)

*# parameter for cooking AND emitting PM2.5*

Cook\_and\_pm <- 0.1

*# define variable if individual is cooking AND cooking emits PM2.5*

cooking <- **sample**(0**:**1, 1, prob = **c**(1 **-** Cook\_and\_pm, Cook\_and\_pm))

*# calculate total emissions by source (ONS, Adult Smoking Habits in GB (2019)) # assumption: no second hand smoke*

*# for smokers*

S <- **if** (smoker **==** 1) {

*# for males*

**if** (sex **==** 1) {

*# aged 16 to 24*

**if** (age **==** 1) {

SSmoking <- **rpois**(1, 4**/**15) **\*** SCigarette

S <- SSmoking **+** cooking **\*** SCooking **+** SOther

*# aged 25 to 34*

} **else if** (age **==** 2) {

SSmoking <- **rpois**(1, 4.55**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther *# aged 35 to 49*

} **else if** (age **==** 3) {

SSmoking <- **rpois**(1, 8.97**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther *# aged 50 to 59*

} **else if** (age **==** 4) {

SSmoking <- **rpois**(1, 11.9**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther *# aged 60 or older*

} **else if** (age **==** 5 **|** age **==** 6) {

SSmoking <- **rpois**(1, 12.9**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther

}

*# for females*

} **else if** (sex **==** 0) {

*# aged 16 to 24*

**if** (age **==** 1) {

SSmoking <- **rpois**(1, 5**/**15) **\*** SCigarette

S <- SSmoking **+** cooking **\*** SCooking **+** SOther

*# aged 25 to 34*

} **else if** (age **==** 2) {

SSmoking <- **rpois**(1, 6.9**/**15) **\*** SCigarette

S <- SSmoking **+** cooking **\*** SCooking **+** SOther

*# aged 35 to 49*

} **else if** (age **==** 3) {

SSmoking <- **rpois**(1, 8.33**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther *# aged 50 to 59*

} **else if** (age **==** 4) {

SSmoking <- **rpois**(1, 10.5**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther *# aged 60 or older*

} **else if** (age **==** 5 **|** age **==** 6) {

SSmoking <- **rpois**(1, 10.8**/**15) **\*** SCigarette S <- SSmoking **+** cooking **\*** SCooking **+** SOther

}

}

*# for non-smokers*

} **else if** (smoker **==** 0) {

S <- cooking **\*** SCooking **+** SOther

}

*# seasons*

Winter <- **c**(12, 1, 2)

Spring <- **c**(3, 4, 5)

Summer <- **c**(6, 7, 8)

Autumn <- **c**(9, 10, 11)

*# time*

Night <- **c**(22**:**23, 0**:**5) Day <- **c**(6**:**21)

*# calculate air exchange rate (Murray and Burmaster (1995) Region 3)*

v <- **if** (**month**(**as.Date**(date)) **%in%** Winter) {

**rlnorm**(n = 1, meanlog = -0.958, sdlog = 0.589)

} **else if** (**month**(**as.Date**(date)) **%in%** Spring) {

**rlnorm**(n = 1, meanlog = -0.802, sdlog = 0.782)

} **else if** (**month**(**as.Date**(date)) **%in%** Summer) {

**rlnorm**(n = 1, meanlog = -0.588, sdlog = 0.612)

} **else if** (**month**(**as.Date**(date)) **%in%** Autumn) {

**rlnorm**(n = 1, meanlog = -0.787, sdlog = 0.453)

}

*# calculate volume of home (zoopla + onaverage.co.uk)*

V <- **if** (housetype **==** "detached") {

**rtri**(1, min = 81, max = 214, mode = 159) **\* runif**(1, min = 2.1, max = 2.6)

} **else if** (housetype **==** "semi-detached") {

**rtri**(1, min = 56, max = 204, mode = 84) **\* runif**(1, min = 2.1, max = 2.6)

} **else if** (housetype **==** "terrace") {

**rtri**(1, min = 33, max = 155, mode = 59) **\* runif**(1, min = 2.1, max = 2.6)

} **else if** (housetype **==** "flat") {

**rtri**(1, min = 34, max = 106, mode = 41) **\* runif**(1, min = 2.1, max = 2.6)

}

*# extract ambient concentration*

tmp <- **subset**(PM\_data, Longitude **==** lon **&** Latitude **==** lat **&**

Date **== as.Date**(date) **&** Time **==** time)

Cout <- tmp**$**PM2.5

*# concentration added*

Cadd <- S **/** V **+** v **\*** Fp **\*** Cout

*# calculate indoor concentration (Zidek et al. (2007)) # Cbefore \* (1 - v - Fd) + Cadd*

(Cadd **/** (v **+** Fd)) **+** (Cbefore **-** (Cadd **/** (v **+** Fd))) **\* exp**(**-**(v **+** Fd) **\*** 1)

}

*##################################################*

*############### calibrate outdoors ############### ##################################################*

*# on average individuals spend 85-90% of their time indoors # (European Comission, 2003)*

*# define threshold for cold and warm days*

**median**(temp\_avg\_Devon\_data**$**Average)

**hist**(temp\_avg\_Devon\_data**$**Average, main = "Median average temperature in Devon", xlab = "Temperature (in °C)")

**abline**(v = 14, col = "red", lwd = 3)

*# find quantile of 14°C*

**quantile**(temp\_avg\_Devon\_data**$**Average, probs = 0.7)

*# balance time of leisure to MEs indoor not home and outdoors*

total\_hours <- **sum**(population\_data**$**rtotal)

indoor\_hours <- **sum**(population\_data**$**rhome, population\_data**$**rindoor) transport\_hours <- **sum**(population\_data**$**rtransport)

leisure\_hours <- **sum**(population\_data**$**rleisure)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| indoor\_hours**/**total\_hours | | *#* | *85.4%* | |
| transport\_hours**/**total\_hours | | *#* | *9.0%* | |
| leisure\_hours**/**total\_hours | | *#* | *5.6%* | |
| *#*  a | *proportion of leisure time spent indoors on cold*  <- 0.7 | | | *day* |

*# proportion of leisure time spent indoors on warm day*

b <- 0.2

85.4 **+** 5.6 **\*** a **\*** 0.7 **+** 5.6 **\*** b **\*** (1-0.7) *# 88.5%*

*##################################################*

*############ activity pattern sampler ############ ##################################################*

*# define function to sample activity pattern*

sample\_activity <- **function**(df, i, date) {

*### divide leisure into outdoor and indoor depending on temperature # get daily average temperature*

temp <- **subset**(temp\_avg\_Devon\_data, Longitude **==** df**$**Longitude\_match[i] **&**

Latitude **==** df**$**Latitude\_match[i] **&** Date **== as.Date**(date))**$**Average

*# define if warm or cold day*

**if** (temp **>** 14) { warm\_day <- 1

} **else** { warm\_day <- 0

}

*# split leisure into indoor and outdoor (assumptions that overall*

*# time spent outdoors of popultion 85-90% (European Comission (2003)))*

*# parameter: proportion of leisure time spent indoors on warm day*

prop\_leis\_warm <- 0.2

*# parameter: proportion of leisure time spent indoors on cold day*

prop\_leis\_cold <- 0.7

*# CASE 1: it is a warm day*

**if** (warm\_day **==** 1) {

rleisure\_outdoor <- df**$**rleisure[i] **\*** (1 **-** prop\_leis\_warm) rleisure\_indoor <- df**$**rleisure[i] **\*** prop\_leis\_warm

*# CASE 2: it is a cold day*

} **else** {

rleisure\_outdoor <- df**$**rleisure[i] **\*** (1 **-** prop\_leis\_cold) rleisure\_indoor <- df**$**rleisure[i] **\*** prop\_leis\_cold

}

*# set up vector with MEs to sample from*

MEs <- **c**(**rep**("home", **ceiling**(df**$**rhome[i] **-** 8)), *# subtract 8 as we allocate 8*

*# hours to home independent from # sampling*

**rep**("indoor", **ceiling**(df**$**rindoor[i] **+** rleisure\_indoor)), **rep**("outdoor", **ceiling**(rleisure\_outdoor)), **rep**("transport", **ceiling**(df**$**rtransport[i])))

*# get hours spent at work and total hours without work*

work <- (df**$**pwork[i] **+** df**$**pworkhome[i]) **\*** 24

total <- df**$**rhome[i] **+** df**$**rindoor[i] **-** work **+** df**$**rleisure[i] **+** df**$**rtransport[i]

*# set up vector with MEs to sample from for weekend # (hours spent at work are distributed to other MEs)*

MEs\_weekend <- **c**(**rep**("home", **ceiling**(df**$**rhome[i] **-** 8 **+** df**$**rhome[i]**/**total **\***

work)),

*# subtract 8 as we allocate 8 hours to home independent from # sampling*

**rep**("indoor", **ceiling**(df**$**rindoor[i] **+** rleisure\_indoor **-** work **+**

(df**$**rindoor[i] **+** rleisure\_indoor)**/** total **\*** work)),

**rep**("outdoor", **ceiling**(rleisure\_outdoor **+** rleisure\_outdoor**/**

total **\*** work)),

**rep**("transport", **ceiling**(df**$**rtransport[i] **+**

df**$**rtransport[i]**/**total **\*** work)))

*# indicator variable to decide if work on specific day # (Labour Force Survey (2020))*

*# don’t work == 0, work == 1*

work\_day <-

*# CASE 1: individual employed*

**if** (df**$**work[i] **==** 1) {

**if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Mon")) {

**sample**(0**:**1, 1, prob = **c**(0.131, 0.869))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Tue")) {

**sample**(0**:**1, 1, prob = **c**(0.106, 0.894))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Wed")) {

**sample**(0**:**1, 1, prob = **c**(0.109, 0.891))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Thu")) {

**sample**(0**:**1, 1, prob = **c**(0.116, 0.884))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Fri")) {

**sample**(0**:**1, 1, prob = **c**(0.157, 0.843))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Sat")) {

**sample**(0**:**1, 1, prob = **c**(0.758, 0.242))

} **else if** ((**weekdays**(**as.Date**(date), TRUE) **==** "Sun")) {

**sample**(0**:**1, 1, prob = **c**(0.823, 0.177))

}

*# CASE 2: individual unemployed*

} **else** { 0

}

*# indicator variable to decide if work at day (1) or night (0) # (Labour Force Survey (2020))*

nightshift <- **sample**(0**:**1, 1, prob = **c**(0.895, 0.105))

*# CASE 1: individual works during the day*

**if** (work\_day **==** 1 **&&** nightshift **==** 0) { pattern <- **rep**("home", 6)

pattern <- **append**(pattern, **sample**(MEs, size = 16), 6) pattern <- **append**(pattern, **rep**("home", 2), 22) pattern

*# CASE 2: individual works during the night*

} **else if** (work\_day **==** 1 **&&** nightshift **==** 1) { pattern <- **sample**(MEs, size = 6)

pattern <- **append**(pattern, **rep**("home", 8), 6)

pattern <- **append**(pattern, **sample**(MEs, size = 10), 14) pattern

*# CASE 3: individual does not work*

} **else if** (work\_day **==** 0) { pattern <- **rep**("home", 6)

pattern <- **append**(pattern, **sample**(MEs\_weekend, size = 16), 6) pattern <- **append**(pattern, **rep**("home", 2), 22)

pattern

}

}

*################################################## ################# stratification ################# ##################################################*

*# define function for stratification*

stratify <- **function**(Data, age, sex, Work, Smoke) {

**subset**(Data, Age1 **==** age **&** Sex **==** sex **&** work **==** Work **&** smoke **==** Smoke)

}

*##################################################*

*############### assign housingtype ############### ##################################################*

assign\_housingtype <- **function**(data){

*# combine area and house id to create unique id*

*# (o/w same household could get assigned different housing types)*

tmp\_data <- data **%>%**

**mutate**(areahid = **paste**(area, hid, sep = ""))

*# create housetype column*

housetype\_df <- **data.frame**(areahid = **unique**(tmp\_data**$**areahid),

housetype = NA)

*# initialise vector to fill with loop*

housetype <- NULL

*# (Office of National Statistics, Census data (2011))*

**for** (i **in** 1**:length**(housetype\_df**$**areahid)) {

housetype[i] <- **sample**(**c**("detached", "semi-detached", "terrace", "flat"),

prob = **c**(0.3000, 0.2746, 0.2340, 0.1914), size = 1)

}

housetype\_df**$**housetype <- housetype

*# add housetype to tmp\_data*

tmp\_data <- **left\_join**(tmp\_data, housetype\_df, by = "areahid")

*# order columns*

tmp\_data <- dplyr**::select**(tmp\_data, id, area, Longitude, Latitude,

Longitude\_match, Latitude\_match, hid, pid, Sex, Age1, pwkstat, work, housetype, **everything**())

tmp\_data <- dplyr**::select**(tmp\_data, **-**areahid)

}

*##################################################*

*############### calculate exposure ############### ##################################################*

calculate\_exposure <- **function**(Pop\_data, age, sex, Work, Smoke, Pol\_data, rep,

Start, End) {

*##### create dataframe with individuals*

strata <- **stratify**(Pop\_data, age, sex, Work, Smoke) individuals <- **sample\_n**(strata, rep)

*##### create dataframe with exposure*

days <- **seq**(**as.Date**(Start), **as.Date**(End), by = "days")

exposure <- **expand.grid**(Date = days, Time = 0**:**23) **%>% arrange**(Date, Time)

*##### create data frame for home concentrations*

*# begins 3 days earlier than timeframe to let concentration indoor get settled*

home\_concentration <- **expand.grid**(

Date = **unique**(**subset**(Pol\_data, Date **<= as.Date**(End) **&**

Date **>** (**as.Date**(Start) **-** 4))**$**Date), Time = **unique**(Pol\_data**$**Time)

) **%>%**

**arrange**(Date, Time)

*##### fill in exposure for individuals # loop over individuals*

**for** (k **in** 1**:**rep) {

**print**(**paste**("Calculating individual", k, sep = " "))

*# initialise / empty vector*

test\_activitypattern <- NULL

*# simulate activity pattern*

**for** (h **in** 1**:length**(days)) {

test\_activitypattern <- **append**(test\_activitypattern,

**sample\_activity**(individuals, k,

**as.Date**(days[h])))

}

*# number of columns for indices*

j <- **ncol**(exposure)

l <- **ncol**(home\_concentration)

**print**("Activity done")

*# Calculate indoor concentrations for the entire time frame of the pollution*

*# dataset. Has to be done in one piece as it is a mass balance equation # concentrations from the hour before are needed to calculate current*

*# concentrations. Startpoint for the the first hour is median indoor # concentration from Wallace et al. (1993).*

**for** (i **in** 1**:length**(home\_concentration**$**Date)) {

*# find concentration of hour before*

Cbefore <- **if** (i **==** 1) {

*# concentration starting point (Wallace et al. (1993))*

**rnorm**(n = 1, mean = 26, sd = 2)

} **else** {

home\_concentration[i-1, l**+**1]

*and*

}

*# calculate concentration of current hour*

home\_concentration[i, l**+**1] <- **calculate\_home**(individuals**$**Longitude\_match[k],

individuals**$**Latitude\_match[k], home\_concentration**$**Date[i], home\_concentration**$**Time[i], individuals**$**housetype[k],

individuals**$**smoke[k], individuals**$**Sex[k], individuals**$**Age1[k], Cbefore)

}

*# rename added column*

**colnames**(home\_concentration)[l**+**1] <- **paste**("Home", k, sep = "")

**print**("Home concentration done")

*# loop over hours*

*# calculate exposure for MEs transport, indoor, and outdoor and take #concentration for home from home\_concentration dataframe*

**for** (i **in** 1**:length**(test\_activitypattern)) {

**if** (test\_activitypattern[i] **==** "transport") { exposure[i, j**+**1] <- test\_activitypattern[i]

exposure[i, j**+**2] <- **calculate\_transport**(individuals**$**Longitude\_match[k],

individuals**$**Latitude\_match[k], exposure**$**Date[i], exposure**$**Time[i])

} **else if** (test\_activitypattern[i] **==** "home") { exposure[i, j**+**1] <- test\_activitypattern[i]

exposure[i, j**+**2] <- **subset**(home\_concentration, Date **==** exposure**$**Date[i] **&**

Time **==** exposure**$**Time[i])[, l**+**1]

} **else if** (test\_activitypattern[i] **==** "indoor") { exposure[i, j**+**1] <- test\_activitypattern[i]

exposure[i, j**+**2] <- **calculate\_indoor**(individuals**$**Longitude\_match[k],

individuals**$**Latitude\_match[k], exposure**$**Date[i], exposure**$**Time[i])

} **else if** (test\_activitypattern[i] **==** "outdoor") { exposure[i, j**+**1] <- test\_activitypattern[i]

exposure[i, j**+**2] <- **calculate\_outdoor**(individuals**$**Longitude\_match[k],

individuals**$**Latitude\_match[k], exposure**$**Date[i], exposure**$**Time[i])

}

}

*# rename added column*

**colnames**(exposure)[j**+**1] <- **paste**("Activity", k, sep = "")

**colnames**(exposure)[j**+**2] <- **paste**("Exposure", k, sep = "")

**print**("Exposure done")

}

*# return dataframe*

returnlist <- **list**(exposure, individuals, home\_concentration)

**return**(returnlist)

}

## Sensitivity analysis

*##################################################*

*################ create hypercube ################ ##################################################*

*########## space filling design: maximin latin hypercube # create maximin latin hypercube*

HypCube <- **maximinLHS**(5, 3)

*# plot hypercube*

**plot**(HypCube[, 1**:**2], pch = 21, xlim = **c**(0,1), ylim = **c**(0,1), col = "#3B3B3BFF", bg = "#CD534CFF", main = "Latin hypercube sampling", xlab = "Variable 1", ylab = "Variable 2")

**abline**(v = **seq**(0.2, 0.8, by = 0.2), col = "gray50", lty = 3)

**abline**(h = **seq**(0.2, 0.8, by = 0.2), col = "gray50", lty = 3)

*################################################## ################### run model #################### ##################################################*

*# assign housting type*

Model\_data <- **assign\_housingtype**(population\_data)

*# pick individual for sensitivity analysis*

individual <- **subset**(Model\_data, id **==** 16602)

*# run model and save with all five sets of parameters # change parameters of function in between model runs*

exposure1\_df <- **calculate\_exposure**(individual, 3, 0, 0, 0, PM\_data, 1,

"2019-01-01", "2019-12-31")

exposure2\_df <- **calculate\_exposure**(individual, 3, 0, 0, 0, PM\_data, 1,

"2019-01-01", "2019-12-31")

exposure3\_df <- **calculate\_exposure**(individual, 3, 0, 0, 0, PM\_data, 1,

"2019-01-01", "2019-12-31")

exposure4\_df <- **calculate\_exposure**(individual, 3, 0, 0, 0, PM\_data, 1,

"2019-01-01", "2019-12-31")

exposure5\_df <- **calculate\_exposure**(individual, 3, 0, 0, 0, PM\_data, 1,

"2019-01-01", "2019-12-31")

exposure\_list <- **list**(exposure1\_df, exposure2\_df, exposure3\_df, exposure4\_df, exposure5\_df, HypCube)

*# extract exposure from each run*

exp1 <- exposure1\_df[[1]] exp2 <- exposure2\_df[[1]] exp3 <- exposure3\_df[[1]] exp4 <- exposure4\_df[[1]] exp5 <- exposure5\_df[[1]]

*##################################################*

*############# format for regression ############## ##################################################*

*# run 1*

exp1a <- dplyr**::select**(exp1, Date, Time, **contains**("Act"))

exp1a <- **pivot\_longer**(exp1a, cols = **starts\_with**("Act"), names\_to = "Individual", values\_to = "Exposure")

exp1b <- dplyr**::select**(exp1, Date, Time, **contains**("Exp"))

exp1b <- **pivot\_longer**(exp1b, cols = **starts\_with**("Exp"), names\_to = "Individual", values\_to = "Exposure")

exp1full <- **data.frame**(exp1a**$**Date, exp1a**$**Time, exp1a**$**Individual, exp1a**$**Exposure, exp1b**$**Exposure)

**colnames**(exp1full) <- **c**("Date", "Time", "Individual", "ME", "Exposure") exp1full**$**id <- 1**:**8760

exp1full**$**Cook <- HypCube[1, 1] exp1full**$**Warm <- HypCube[1, 2] exp1full**$**Cold <- HypCube[1, 3] exp1full**$**Run <- **as.character**(1)

*# run 2*

exp2a <- dplyr**::select**(exp2, Date, Time, **contains**("Act"))

exp2a <- **pivot\_longer**(exp2a, cols = **starts\_with**("Act"), names\_to = "Individual", values\_to = "Exposure")

exp2b <- dplyr**::select**(exp2, Date, Time, **contains**("Exp"))

exp2b <- **pivot\_longer**(exp2b, cols = **starts\_with**("Exp"), names\_to = "Individual", values\_to = "Exposure")

exp2full <- **data.frame**(exp2a**$**Date, exp2a**$**Time, exp2a**$**Individual, exp2a**$**Exposure, exp2b**$**Exposure)

**colnames**(exp2full) <- **c**("Date", "Time", "Individual", "ME", "Exposure") exp2full**$**id <- 1**:**8760

exp2full**$**Cook <- HypCube[2, 1] exp2full**$**Warm <- HypCube[2, 2] exp2full**$**Cold <- HypCube[2, 3] exp2full**$**Run <- **as.character**(2)

*# run 3*

exp3a <- dplyr**::select**(exp3, Date, Time, **contains**("Act"))

exp3a <- **pivot\_longer**(exp3a, cols = **starts\_with**("Act"), names\_to = "Individual", values\_to = "Exposure")

exp3b <- dplyr**::select**(exp3, Date, Time, **contains**("Exp"))

exp3b <- **pivot\_longer**(exp3b, cols = **starts\_with**("Exp"), names\_to = "Individual", values\_to = "Exposure")

exp3full <- **data.frame**(exp3a**$**Date, exp3a**$**Time, exp3a**$**Individual, exp3a**$**Exposure, exp3b**$**Exposure)

**colnames**(exp3full) <- **c**("Date", "Time", "Individual", "ME", "Exposure") exp3full**$**id <- 1**:**8760

exp3full**$**Cook <- HypCube[3, 1] exp3full**$**Warm <- HypCube[3, 2] exp3full**$**Cold <- HypCube[3, 3] exp3full**$**Run <- **as.character**(3)

*# run 4*

exp4a <- dplyr**::select**(exp4, Date, Time, **contains**("Act"))

exp4a <- **pivot\_longer**(exp4a, cols = **starts\_with**("Act"), names\_to = "Individual", values\_to = "Exposure")

exp4b <- dplyr**::select**(exp4, Date, Time, **contains**("Exp"))

exp4b <- **pivot\_longer**(exp4b, cols = **starts\_with**("Exp"), names\_to = "Individual", values\_to = "Exposure")

exp4full <- **data.frame**(exp4a**$**Date, exp4a**$**Time, exp4a**$**Individual, exp4a**$**Exposure, exp4b**$**Exposure)

**colnames**(exp4full) <- **c**("Date", "Time", "Individual", "ME", "Exposure") exp4full**$**id <- 1**:**8760

exp4full**$**Cook <- HypCube[4, 1] exp4full**$**Warm <- HypCube[4, 2] exp4full**$**Cold <- HypCube[4, 3] exp4full**$**Run <- **as.character**(4)

*# run 5*

exp5a <- dplyr**::select**(exp5, Date, Time, **contains**("Act"))

exp5a <- **pivot\_longer**(exp5a, cols = **starts\_with**("Act"), names\_to = "Individual", values\_to = "Exposure")

exp5b <- dplyr**::select**(exp5, Date, Time, **contains**("Exp"))

exp5b <- **pivot\_longer**(exp5b, cols = **starts\_with**("Exp"), names\_to = "Individual", values\_to = "Exposure")

exp5full <- **data.frame**(exp5a**$**Date, exp5a**$**Time, exp5a**$**Individual, exp5a**$**Exposure, exp5b**$**Exposure)

**colnames**(exp5full) <- **c**("Date", "Time", "Individual", "ME", "Exposure") exp5full**$**id <- 1**:**8760

exp5full**$**Cook <- HypCube[5, 1] exp5full**$**Warm <- HypCube[5, 2] exp5full**$**Cold <- HypCube[5, 3] exp5full**$**Run <- **as.character**(5)

*# combine all dataframes*

exp <- **bind\_rows**(exp1full, exp2full, exp3full, exp4full, exp5full)

*################################################## ################# model fitting ################## ##################################################*

*##### linear model*

*### linear model with all variables*

lm <- **lm**(Exposure **~** Date **+** Time **+** ME **+** Cook **+** Warm **+** Cold **+** Run, exp)

*# summary*

**summary**(lm)

*# residual plots* **par**(mfrow = **c**(2, 2)) **plot**(lm)

*### linear model with selected variables*

lm1 <- **lm**(Exposure **~** ME **+** Cook **+** Warm **+** Cold, exp)

*# summary*

**summary**(lm1)

*# residual plots* **par**(mfrow = **c**(2, 2)) **plot**(lm1)

*##### Gamma GLM*

*# add small noise value to 0s*

exp\_gamma <- exp

exp\_gamma[exp\_gamma **==** 0] <- 0.000001

*# fit model with all variables*

glm <- **glm**(Exposure **~** ME **+** Cook **+** Warm **+** Cold, family = **Gamma**(link = "log"), data = exp\_gamma)

*# summary*

**summary**(glm)

*# residual plots* **par**(mfrow = **c**(2, 2)) **plot**(glm)

*# repeatedly using drop1() yields the reduced model with only ME as a predictor # this is also true if quadratic and interaction terms are included in the model* **drop1**(glm, test = "LRT")

*# fit model with reduced parameters*

glm1 <- **glm**(Exposure **~** ME, family = **Gamma**(link = "log"), data = exp\_gamma)

*# summary*

**summary**(glm1)

*# residual plots* **par**(mfrow = **c**(2, 2)) **plot**(glm1)

*######### plots ##########*

*# calculate daily average exposure*

exp\_day <- exp **%>%**

**group\_by**(Date, Cook, Warm, Cold, Run) **%>% summarise**(Average = **mean**(Exposure))

*# plot daily average exposure for full year*

**ggplot**(exp\_day) **+**

**geom\_line**(**aes**(x = Date, y = Average, colour = Run)) **+**

**labs**(title = "Daily averages of exposure to PM2.5 for all runs", y = "Exposure (in µg/m³)") **+**

**theme\_bw**()

*# comparison plot of residuals*

**par**(mfrow=**c**(2, 2), mar = **c**(4,4,2,1) **+** 0.1)

**plot**(lm, which = 1, sub = "") **plot**(glm1, which = 1, sub = "") **plot**(lm, which = 2, sub = "") **plot**(glm1, which = 2, sub = "")

## Application

*# assign housting type*

Model\_data <- **assign\_housingtype**(population\_data)

*# normal run for april*

modelrun\_april <- **calculate\_exposure**(Model\_data, 2, 0, 1, 0, PM\_data, 30,

"2019-04-01", "2019-04-30")

**save**(modelrun\_april, file = "modelrun\_april.RDATA")

*# rollback run for april*

modelrun\_april\_rb <- **calculate\_exposure**(Model\_data, 2, 0, 1, 0, PM\_rollback\_data,

30, "2019-04-01", "2019-04-30")

**save**(modelrun\_april\_rb, file = "modelrun\_rb\_april.RDATA")

*# normal run for july*

modelrun\_july <- **calculate\_exposure**(Model\_data, 2, 0, 1, 0, PM\_data, 30,

"2019-07-01", "2019-07-31")

**save**(modelrun\_july, file = "modelrun\_july.RDATA")

*# rollback run for july*

modelrun\_july\_rb <- **calculate\_exposure**(Model\_data, 2, 0, 1, 0, PM\_rollback\_data, 30,

"2019-07-01", "2019-07-31")

**save**(modelrun\_july\_rb, file = "modelrun\_rb\_july.RDATA")

## Results

*###################################################################*

*################## calculating days exceeding x0 ################## ###################################################################*

*########## APRIL ##########*

*##### 1: personal exposure # combine runs*

modelrun <- modelrun\_april[[1]]

*# change column names*

**for** (i **in** 1**:**30) {

**colnames**(modelrun)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(modelrun)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*# predictive distribution over hours*

exposure <- dplyr**::select**(modelrun, Date, **contains**("Exp"))

exposure\_april <- **pivot\_longer**(exposure, cols = **starts\_with**("Ind"),

names\_to = "Individual", values\_to = "Exposure") **%>% group\_by**(Date, Individual) **%>%**

**summarise**(Exposure = **mean**(Exposure))

**nrow**(**subset**(exposure\_april, Exposure **>** 11))**/**30 *# 17.2 days* **nrow**(**subset**(exposure\_april, Exposure **>** 23))**/**30 *# 2.6 days* **nrow**(**subset**(exposure\_april, Exposure **>** 35))**/**30 *# 0.9 days* **nrow**(**subset**(exposure\_april, Exposure **>** 41))**/**30 *# 0.2 days* **nrow**(**subset**(exposure\_april, Exposure **>** 47))**/**30 *# 0.03 days* **nrow**(**subset**(exposure\_april, Exposure **>** 53))**/**30 *# 0.0 days* **nrow**(**subset**(exposure\_april, Exposure **>** 58))**/**30 *# 0.0 days* **nrow**(**subset**(exposure\_april, Exposure **>** 64))**/**30 *# 0.0 days* **nrow**(**subset**(exposure\_april, Exposure **>** 70))**/**30 *# 0.0 days*

*##### 2: ambient concentration*

individuals\_april <- modelrun\_april[[2]]

long <- **unique**(individuals\_april**$**Longitude\_match) lat <- **unique**(individuals\_april**$**Latitude\_match)

ambient <- **subset**(PM\_data, Date **>= as.Date**("2019-04-01") **&**

Date **<= as.Date**("2019-04-30"))

ambient <- **subset**(ambient, Longitude **%in%** long) ambient <- **subset**(ambient, Latitude **%in%** lat)

ambient\_Avg\_april <- ambient **%>% group\_by**(Date) **%>% summarise**(Average = **mean**(PM2.5))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 11)) | *#* | *21.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 23)) | *#* | *6.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 35)) | *#* | *2.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 41)) | *#* | *2.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 47)) | *#* | *1.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 53)) | *#* | *1.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 58)) | *#* | *1.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 64)) | *#* | *1.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_april, | Average | **>** | 70)) | *#* | *0.0* | *days* |
| *########## July ##########* |  |  |  |  |  |  |
| *##### 2: personal exposure* |  |  |  |  |  |  |
| *# combine runs*  modelrun <- modelrun\_july[[1]] |  |  |  |  |  |  |
| *# change column names*  **for** (i **in** 1**:**30) { |  |  |  |  |  |  |

**colnames**(modelrun)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(modelrun)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*# predictive distribution over hours*

exposure <- dplyr**::select**(modelrun, Date, **contains**("Exp"))

exposure\_july <- **pivot\_longer**(exposure, cols = **starts\_with**("Ind"),

names\_to = "Individual", values\_to = "Exposure") **%>% group\_by**(Date, Individual) **%>%**

**summarise**(Exposure = **mean**(Exposure))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 11))**/**30 | *#* | *6.9* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 23))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 35))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 41))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 47))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 53))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 58))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 64))**/**30 | *#* | *0.0* | *days* |
| **nrow**(**subset**(exposure\_july, | Exposure | **>** | 70))**/**30 | *#* | *0.0* | *days* |

*##### 2: ambient concentration*

individuals\_july <- modelrun\_july[[2]]

long <- **unique**(individuals\_july**$**Longitude\_match) lat <- **unique**(individuals\_july**$**Latitude\_match)

ambient <- **subset**(PM\_data, Date **>= as.Date**("2019-07-01") **&**

Date **<= as.Date**("2019-07-31"))

ambient <- **subset**(ambient, Longitude **%in%** long) ambient <- **subset**(ambient, Latitude **%in%** lat)

ambient\_Avg\_july <- ambient **%>% group\_by**(Date) **%>% summarise**(Average = **mean**(PM2.5))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 11)) | *#* | *8.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 23)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 35)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 41)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 47)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 53)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 58)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 64)) | *#* | *0.0* | *days* |
| **nrow**(**subset**(ambient\_Avg\_july, | Average | **>** | 70)) | *#* | *0.0* | *days* |

*##############################################################*

*############# calculating measure of uncertainty ############# ##############################################################*

*# define function for p\_i*

calculate\_pi <- **function** (N, exp, x0, day) {

1**/**N **\* sum**(**subset**(exp, Date **== as.Date**(day))**$**Exposure **>** x0)

}

*# definfe functon for p\_{j|i}*

calculate\_pji <- **function**(N, exp, x0, day\_j, day\_i) { pi <- **calculate\_pi**(N, exp, x0, day\_i)

pj <- **calculate\_pi**(N, exp, x0, day\_j)

pji <- 1**/**(N**^**2 **\*** pi) **\* sum**((**subset**(exp, Date **== as.Date**(day\_i))**$**Exposure **>** x0) **%o%**

(**subset**(exp, Date **== as.Date**(day\_j))**$**Exposure **>** x0))

*# check if pji is a number (if there are no exposures > x0 it returns NAN: not a num*

**if**(**is.nan**(pji)) { 0

} **else** { pji

}

}

*# set threshold*

x <- 11 *# 11, 23, 35, 41, 47, 53, 58, 64, 70*

*# adjust for month*

days <- **unique**(exposure\_april**$**Date) *# exposure\_july$Date or exposur\_april$Date*

exposure\_data <- exposure\_april *# exposure\_july or exposur\_april ###### calculate standard error term for full month*

*# initialise variable that store term total*

se <- 0

*# loop over all days*

**for** (i **in** 1**:**30) {

*#calculate pi for day*

pi <- **calculate\_pi**(30, exposure\_data, x, days[i])

*# and add to total*

se <- se **+** pi**\***(1 **-** pi)

}

*##### calculate correlation term for full month*

*# initialise variable to store term total*

dependance <- 0

*# loop over all days*

**for**(i **in** 1**:**30) {

*# find all days apart from current day for all possible combinations*

js <- 1**:**30

js <- js[**-**i]

*# for each day calculate all combinations with other days excluding where i = j*

**for**(j **in** js) {

pi <- **calculate\_pi**(30, exposure\_data, x, days[i]) pj <- **calculate\_pi**(30, exposure\_data, x, days[j])

pji <- **calculate\_pji**(30, exposure\_data, x, days[j], days[i])

*# add to total*

dependance <- dependance **+** pi**\***(pji **-** pj)

}

}

*# both terms included in measure of uncertainty*

se dependance

*# calculate full measure of uncertainty*

dependance **+** se

*##### APRIL*

*# for 11: 3.683333*

*# for 23: 0.2455556*

*# for 35: 0.2755556*

*# for 41: 0.1477778*

*# for 47: 0.03222222*

*# for 53: 0*

*##### JULY*

*# for 11: 3.146667*

*# for 23: 0*

*################################################## ################## result plots ################## ##################################################*

*########## JULY ##########*

*# extract results*

results\_july <- modelrun\_july[[1]]

*# change column names*

**for** (i **in** 1**:**30) {

**colnames**(results\_july)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(results\_july)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*##### PLOT 1: predictive distribution by hours*

*# prepare data*

hour\_data\_july <- dplyr**::select**(results\_july, Date, Time, **contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Time, Individual) **%>% summarise**(Exposure = **mean**(Exposure))

*# plot*

hour\_plot\_july <- **ggplot**(hour\_data\_july) **+ geom\_boxplot**(**aes**(x = **as.factor**(Time), y = Exposure)) **+ labs**(subtitle = "July",

x = "Hour", y = "Exposure (in µg/m³)") **+ scale\_x\_discrete**(breaks = **seq**(0, 23, by = 2)) **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 2: predictive distribution by MEs # prepare data*

exposure\_july <- dplyr**::select**(results\_july, Date, Time, **contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure")

activity\_july <- dplyr**::select**(results\_july, Date, Time, **contains**("act")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Activity")

ME\_data\_july <- **data.frame**(activity\_july, Exposure = exposure\_july**$**Exposure)

**group\_by**(Individual, Activity) **%>% summarise**(Exposure = **mean**(Exposure))

*# plot*

ME\_plot\_july <- **ggplot**(ME\_data\_july) **+ geom\_boxplot**(**aes**(x = Activity, y = Exposure)) **+**

**labs**(subtitle = "July", x = "Microenvironment", y = "Exposure (in µg/m³)")

**theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 3: exposure vs ambient by days*

**%>%**

### +

*# prepare data*

daily\_exposure\_july <- dplyr**::select**(results\_july, Date, Time, **contains**("Exp")) **%>%**

**pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual", values\_to = "Exposure") **%>%**

**group\_by**(Individual, Date) **%>% summarise**(Exposure = **mean**(Exposure))

*# calculate average ambient concentrations in areas of simulated individuals*

longs\_july <- **unique**(modelrun\_july[[2]]**$**Longitude\_match) lats\_july <- **unique**(modelrun\_july[[2]]**$**Latitude\_match)

ambient\_july <- **subset**(PM\_data, Date **>= as.Date**("2019-07-01") **&**

Date **<= as.Date**("2019-07-31")) **%>%**

**subset**(Longitude **%in%** longs\_july **&** Latitude **%in%** lats\_july) **%>% group\_by**(Date) **%>%**

**summarise**(Average = **mean**(PM2.5))

*# plot*

ambient\_plot\_july <- **ggplot**() **+**

**geom\_boxplot**(**aes**(x = Date, y = Exposure, group = Date), data = daily\_exposure\_july) **+**

**geom\_line**(**aes**(x = Date, y = Average), data = ambient\_july, colour = "red") **+ labs**(subtitle = "July", x = "Date", y = "Exposure (in µg/m³)") **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 4: rollback # extract results*

results\_july\_rb <- modelrun\_july\_rb[[1]]

*# change column names*

**for** (i **in** 1**:**30) {

**colnames**(results\_july\_rb)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(results\_july\_rb)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*# prepare data*

monthly\_exposure\_july <- dplyr**::select**(results\_july, Date, Time,

**contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Individual) **%>% summarise**(Exposure = **mean**(Exposure))

monthly\_exposure\_july\_rb <- dplyr**::select**(results\_july\_rb, Date, Time,

**contains**("Exp")) **%>%**

**pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual", values\_to = "Exposure") **%>%**

**group\_by**(Individual) **%>% summarise**(Exposure = **mean**(Exposure))

*# plot*

rollback\_plot\_july <- **ggplot**() **+**

**geom\_boxplot**(**aes**(x = "No rollback", y = Exposure), data = monthly\_exposure\_july) **+ geom\_boxplot**(**aes**(x = "Rollback", y = Exposure), data = monthly\_exposure\_july\_rb) **+ labs**(subtitle = "July", x = "", y = "Exposure (in µg/m³)") **+**

**theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*########## APRIL #########*

*# extract results*

results\_april <- modelrun\_april[[1]]

*# change column names*

**for** (i **in** 1**:**30) {

**colnames**(results\_april)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(results\_april)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*##### PLOT 1: predictive distribution by hours*

*# prepare data*

hour\_data\_april <- dplyr**::select**(results\_april, Date, Time, **contains**("Exp")) **pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Time, Individual) **%>% summarise**(Exposure = **mean**(Exposure))

*# plot*

hour\_plot\_april <- **ggplot**(hour\_data\_april) **+ geom\_boxplot**(**aes**(x = **as.factor**(Time), y = Exposure)) **+**

**labs**(subtitle = "April", x = "Hour", y = "Exposure (in µg/m³)") **+ scale\_x\_discrete**(breaks = **seq**(0, 23, by = 2)) **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 2: predictive distribution by MEs # prepare data*

**%>%**

exposure\_april <- dplyr**::select**(results\_april, Date, Time, **contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure")

activity\_april <- dplyr**::select**(results\_april, Date, Time, **contains**("act")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Activity")

ME\_data\_april <- **data.frame**(activity\_april, Exposure = exposure\_april**$**Exposure) **%>% group\_by**(Individual, Activity) **%>%**

**summarise**(Exposure = **mean**(Exposure))

*# plot*

ME\_plot\_april <- **ggplot**(ME\_data\_april) **+ geom\_boxplot**(**aes**(x = Activity, y = Exposure)) **+**

**labs**(subtitle = "April", x = "Microenvironment", y = "Exposure (in µg/m³)") **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 3: exposure vs ambient by days # prepare data*

daily\_exposure\_april <- dplyr**::select**(results\_april, Date, Time,

**contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Individual, Date) **%>% summarise**(Exposure = **mean**(Exposure))

*# calculate average ambient concentrations in areas of simulated individuals*

longs\_april <- **unique**(modelrun\_april[[2]]**$**Longitude\_match) lats\_april <- **unique**(modelrun\_april[[2]]**$**Latitude\_match)

ambient\_april <- **subset**(PM\_data, Date **>= as.Date**("2019-04-01") **&**

Date **<= as.Date**("2019-04-30")) **%>%**

**subset**(Longitude **%in%** longs\_april **&** Latitude **%in%** lats\_april) **%>% group\_by**(Date) **%>%**

**summarise**(Average = **mean**(PM2.5))

*# plot*

ambient\_plot\_april <- **ggplot**() **+**

**geom\_boxplot**(**aes**(x = Date, y = Exposure, group = Date), data = daily\_exposure\_april) **+**

**geom\_line**(**aes**(x = Date, y = Average), data = ambient\_april, colour = "red") **+**

**labs**(subtitle = "April", x = "Date", y = "Exposure (in µg/m³)") **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*##### PLOT 4: rollback # extract results*

results\_april\_rb <- modelrun\_april\_rb[[1]]

*# change column names*

**for** (i **in** 1**:**30) {

**colnames**(results\_april\_rb)[i**\***2**+**1] <- **paste**("Ind\_act\_", i, sep = "")

**colnames**(results\_april\_rb)[i**\***2**+**2] <- **paste**("Ind\_exp\_", i, sep = "")

}

*# prepare data*

monthly\_exposure\_april <- dplyr**::select**(results\_april, Date, Time,

**contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Individual) **%>% summarise**(Exposure = **mean**(Exposure))

monthly\_exposure\_april\_rb <- dplyr**::select**(results\_april\_rb, Date, Time,

**contains**("Exp")) **%>% pivot\_longer**(cols = **starts\_with**("Ind"), names\_to = "Individual",

values\_to = "Exposure") **%>% group\_by**(Individual) **%>% summarise**(Exposure = **mean**(Exposure))

*# plot*

rollback\_plot\_april <- **ggplot**() **+**

**geom\_boxplot**(**aes**(x = "No rollback", y = Exposure), data = monthly\_exposure\_april) **+**

**geom\_boxplot**(**aes**(x = "Rollback", y = Exposure), data = monthly\_exposure\_april\_rb) **+**

**labs**(subtitle = "April", x = "", y = "Exposure (in µg/m³)") **+ theme**(plot.subtitle = **element\_text**(hjust = 0.5))

*########## combine April and July in one plot ######### ##### PLOT 1: predictive distribution by hours*

hour\_plot <- ggpubr**::ggarrange**(hour\_plot\_april, hour\_plot\_july)

ggpubr**::annotate\_figure**(hour\_plot, top = ggpubr**::text\_grob**("Predictive distributions

of exposure by hour", size = 14))

*##### PLOT 2: predictive distribution by MEs*

ME\_plot <- ggpubr**::ggarrange**(ME\_plot\_april, ME\_plot\_july) ggpubr**::annotate\_figure**(ME\_plot, top = ggpubr**::text\_grob**("Predictive distributions

of exposure by microenvironment", size = 14))

*##### PLOT 3: exposure vs ambient by days*

ambient\_plot <- ggpubr**::ggarrange**(ambient\_plot\_april, ambient\_plot\_july) ggpubr**::annotate\_figure**(ambient\_plot, top = ggpubr**::text\_grob**("Exposure vs ambient

concentrations", size = 14))

*##### PLOT 4: rollback*

rollback\_plot <- ggpubr**::ggarrange**(rollback\_plot\_april, rollback\_plot\_july) ggpubr**::annotate\_figure**(rollback\_plot, top = ggpubr**::text\_grob**("Effects of a policy

change", size = 14))