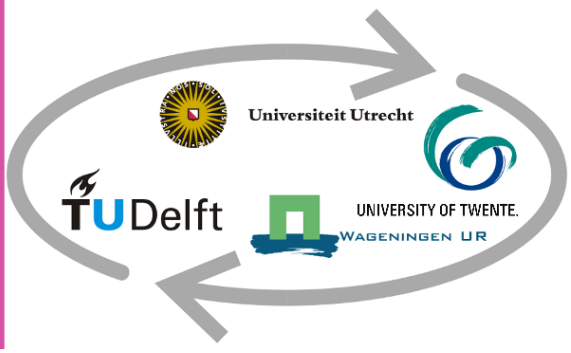
GIMA
Geographical Information Management and Applications

Modelling air pollution and personal exposure in Bangkok and Mexico City using a land use regression model

GIMA Master's thesis
February 2016

Author: Patrick van den Ende
Supervisors: Dr. D. Karssenberg & Dr. O. Schmitz



Title:	Modelling air pollution and personal exposure in Bangkok and Mexico City using a land use regression model
Author:	Patrick van den Ende
Student number:	S6020593 (University of Twente), 3859126 (Utrecht University)
Email:	patrickvandenende@gmail.com
Supervisors:	Dr. D. Karssenberg & Dr. O. Schmitz
Responsible professor:	Prof. Dr. S. de Jong
Universities:	TU Delft, University of Twente, Utrecht University & Wageningen University
Course:	GIMA Module 7 MSc thesis research
Date:	February 2016

Preface

This thesis is the result of 5 months of research in the context of the MSc GIMA Master's program. It was an interesting period which brought me a lot of insights in air pollution modelling, land use regression models and academic skills.

When I started this thesis the subject of air pollution modelling got my attention because of the relevance of the subject and the practical GIS activities that would be possible with this subject. Looking back, I can say that I am happy that I have chosen this subject. Despite the (software) issues that crossed my path during the thesis period, I was satisfied with the results at the end. It was an interesting journey to go from regression equations and small amounts of data to results, like output maps, which could subsequently be interpreted and discussed.

I want to thank my supervisors for their help and useful feedback on my thesis. The comments of Derek Karssenberg on my thesis helped me to keep the right focus. Also the face to face meetings were interesting, encouraged me to find things out and gave me inspiration to carry on. Oliver Schmitz was always prepared to give me the technical support I needed and gave useful feedback on how to properly use the software and calculate the land use regression equations.

Summary

The objective of this research is to examine the possibilities of modelling air pollution concentrations in cities outside Europe at a high spatial resolution (five to ten meters). This is ideally done with datasets that are available globally, which would ultimately allow comparison of different cities. This is done by using a $PM_{2.5}$ land use regression equation which was developed in the ESCAPE (European Study of Cohorts for Air Pollution Effects) project for the city of London. Also personal exposure to air pollution is examined in this study. The main question of this research is: 'To what extent is it possible to model air pollution concentrations in cities outside Europe using the London ESCAPE LUR model, how valid is the model and to what extent can the personal exposure of the population to air pollution be modelled?' The main question is divided in three sub-questions which deal with the input data of the models, the validation and sensitivity analysis of the models and the personal exposure of the population.

This paragraph deals with the methodology of this research. Land use regression models were used to model the relationship between the response variable air pollution and two or more explanatory variables. These explanatory variables in the ESCAPE project were for instance land use, traffic density and topography. The regression equation used in this study consists of two variables: '*INTMAJORINVDIST*' (*'i')* and *'ROADLENGTH_500' ('l')*. The former is the product of the number of cars on the nearest major road and the inverse of distance to this nearest major road. The latter is the total amount of roads within a buffer of 500 meters. The road network data used for this regression equation is OpenStreetMap (OSM, 2015). The number of cars on the roads are estimated by assuming that all registered cars within a city are driving while they are counted. This assumption was needed because traffic intensity data was not available. Two parts of the cities Bangkok and Mexico City were modelled in this research project. Validation and sensitivity analysis were carried out to examine model errors. To examine the exposure of the population, the population numbers of the neighbourhoods were distributed across the modelled areas with a global population density layer grid used as a weighting layer. Then then the personal exposure could be measured.

This paragraph discusses the results of the first question, which is formulated as follows: 'which input data can be used to compare different cities outside Europe using the model?' The results have shown that it is possible to model $PM_{2.5}$ values with the London regression equation. However, a traffic intensity dataset was not available. This is why assumptions needed to be made about the number of cars on a road. The $PM_{2.5}$ values of the Bangkok area varied from 7.2 to 37066 microgram per cubic meter. For Benito Juárez (Mexico City) this ranged from 13.8 to 183727. The high values for both cities were all located on and near the major roads. This is why another output was generated without these major road locations and a buffer of 10 meter around these major roads. With this output the values range from 7.2 to 38.7 for Bangkok and from 13.8 to 46.3 for Benito Juárez.

The results of the second sub-question are discussed in this paragraph. The question was formulated as follows: 'to what extent do model errors occur when the model is applied in a study area outside Europe and what causes these errors?' The input data validation was done by comparing two model outputs with different input data. One model output modelled the air pollution with data used in the ESCAPE project and one model output modelled air pollution with data used in this research project. A direct comparison between the input data was not possible, because the data used in the ESCAPE project was not available as open data. By comparing the two model outputs the effects of using different input data can be analyzed. The input data validation showed a weak positive correlation between the model outputs for Rotterdam of this study and the model output of the ESCAPE project. There was an underestimation of the $PM_{2.5}$ values, but the model errors were not large. This was concluded based on the small difference between the Root Mean Squared Error and the Mean Absolute Error. The validation of the models for Bangkok and Mexico City showed that the average

model outputs were higher than the remote sensing data. This validation also showed that the London regression equation predicted the average $PM_{2.5}$ values better for Bangkok than for Mexico City. The differences in model output between Bangkok and Mexico City can probably be explained by the distribution of the major roads. The major roads of Benito Juárez are more evenly distributed than the roads in Bangkok. The consequence is that for each grid cell in Benito Juárez a major road is, on average, closer than for each grid cell in Bangkok. This could lead to higher $PM_{2.5}$ values predicted by the model. The sensitivity analysis showed that the '*l*' variable had more influence on the model output than the '*i*' variable.

This paragraph discusses the results of the third sub-question, which was formulated as follows: 'to what extent is it possible to model the personal exposure of the population to air pollution?' The exposure results showed that the population in Mexico City was exposed to higher $PM_{2.5}$ values than the population of Bangkok. The distribution of the population raised the average personal $PM_{2.5}$ for both populations, because the population was concentrated at locations with higher $PM_{2.5}$ values, like roads. A comparison of multiple locations at 50 meters from a major road and locations at 200 meters from a major road showed that the differences between these locations vary from 1.6 to 4.4 microgram per cubic metre.

This paragraph deals with the conclusions of this research. With the available global datasets it is possible to model air pollution concentrations to a certain extent. The road network dataset can be used without much data pre-processing. The traffic intensity dataset is the most problematic one and should be obtained locally or be created using assumptions. The model errors are partly explained by the input data. The validation of the input data showed an underestimation of the $PM_{2.5}$ values in Rotterdam. The absolute errors showed that there were not many large errors in the predictions. This implies that there is a constant underestimation of $PM_{2.5}$ values in Rotterdam. The model errors are also caused by the model itself. The differences between Bangkok and Mexico City showed that local calibration would be a suitable solution to take city-specific characteristics into account. This will lead to smaller prediction errors. The sensitivity analysis showed that especially the *'l'* variable has a substantial influence on the model output, which means that it is important to use an accurate and unambiguously mapped road network dataset. The results of the third sub-question have shown that the personal exposure of the population in Bangkok and Mexico City can be estimated with the regression equation. For any location in the research area the $PM_{2.5}$ can be estimated. However, the air pollution map and the distribution of the population caused some uncertainties.

Contents

List with figures and tables

1. Introduction

Air pollution has been a problem for a long time and several news sources reported on this issue lately, especially on the air pollution in large cities throughout the world. Examples of cities that were mentioned are Beijing and New Delhi, but also European cities, like Madrid and Milan. This shows that air pollution is a worldwide problem in large cities. However, it is not easy to make estimations of air pollution concentrations within cities in a quick way. This is one of the reasons that this research project was carried out. This section provides information on the context of this research project and subsequently more details are given about this study.

This study was done in the context of the ESCAPE project. The European Study of Cohorts for Air Pollution Effects (ESCAPE) was carried out to better understand the health effects of long-term air pollution exposure. The study in the European context was needed because the impact of air pollution on health in Europe was mainly based on research carried out in North America (Beelen et al., 2013). It is important to carry out research like this, because air pollution related to traffic can have a negative influence on health of citizens. These health effects are caused by long-term exposure to certain air pollution concentrations (Beelen et al., 2013). Several epidemiological studies have shown that it is important to account for spatial and temporal variation in air pollution concentrations within cities (Brauer et al., 2003; Jerrett et al., 2005). Land use regression models are able to take this within-city variability of air pollution concentrations into account (Marshall, Nethery, & Brauer, 2008). Land use regression models use multiple linear regression to model the relationship between the response variable air pollution (dependent variable) and two or more explanatory variables (independent variables). These explanatory variables are for instance land use, traffic density and topography (Beelen et al., 2013). By using these predictor variables, air pollution can be estimated for locations without air pollution measurement instruments (Johnson, Isakov, Touma, Mukerjee, & Özkaynak, 2010). The next step is to calculate the personal exposure of persons to these air pollution concentrations, for instance at their home location (Beelen et al., 2013; Ryan & LeMasters, 2007). LUR models are specifically suitable for calculating personal exposure because of the detailed spatial resolution they provide, which is needed to take the variability of air pollution concentrations within cities into account (Marshall, Nethery, & Brauer, 2008). Other strong points of land use regression models are their empirical basis, which means they can be adapted to local areas without adding extra monitoring data and the relatively low costs (Jerrett et al., 2005). Most of the LUR models were applied in Europe and North America (Hoek et al., 2008), including those developed in the ESCAPE project. This is also one of the reasons that the applicability of LUR models will be tested outside Europe and North America.

As stated, the models of the ESCAPE project were only developed and used in Europe. It would however also be interesting to apply the models to cities outside Europe because the adverse health effects of exposure to air pollution do of course not only apply to cities within Europe. Besides, it is also valuable to have a land use regression model which can be applied to several cities by using the same data source which is available for all the cities. This makes it easier to compare cities regarding air pollution and personal exposure with each other and gives an overall idea of the air pollution concentrations in cities throughout the world in a quick way. Because land use regression models were mainly applied within European and North American cities, it is the objective of this research to examine the possibilities and to evaluate what would currently be possible in modelling air pollution concentrations in cities outside Europe. This was done by using one of the LUR models of the ESCAPE project. Also the exposure of the population to air pollution was examined in this study. All these issues described in this introduction led to the following main objective:

‘Gain insight in the possibilities to model air pollution concentrations in cities outside Europe using the London ESCAPE LUR model, by applying ubiquitous datasets, to validate the model and to calculate and compare the personal exposure of the population in these cities.’

This main objective is divided in the following sub-objectives:

1. Identify and evaluate relevant and ubiquitous datasets which can be used as input for the model**.**
2. Identify the deviations that occur when the model is applied to different cities.
3. Calculate and compare the personal exposure of the population.

Related to these objectives the research questions were formulated. The main research question is:

‘To what extent is it possible to model air pollution concentrations in cities outside Europe using the London ESCAPE LUR model, how valid is the model and to what extent can the personal exposure of the population to air pollution be modelled?’

This main question is divided in the following sub-questions:

1. Which input data can be used to compare different cities outside Europe using the model?
2. To what extent do model errors occur when the model is applied in a study area outside Europe and what causes these errors?
3. To what extent is it possible to model the personal exposure of the population to air pollution?

The second chapter describes the theoretical framework of this study. This chapter is the basis for the following chapter which elaborates on the methodology to research the objectives and to answer the research questions. The fourth chapter presents the results of this study. Chapter five discusses the results and the final chapter deals with the conclusions based on this research.

2. Theoretical framework

This chapter deals with the theoretical underpinning of this research. The chapter first provides background information on the ESCAPE project. The next section elaborates on different methods to model air pollution to highlight the specific characteristics of land use regression models in comparison with others. The subsequent sections discuss LUR models. Issues that are covered are: the advantages and limitations of LUR models and transferability of LUR models. Section 2.6 discusses the validation of land use regression models. The subsequent section deals with the exposure of the population to air pollution. The chapter results in a synthesis with the expectations for this research based on the literature.

2.1 Background information on the ESCAPE project

The European Study of Cohorts for Air Pollution Effects (ESCAPE) project was carried out to describe health effects of long-term air pollution exposure. The study in the European context was needed because studies on the health effects of air pollution exposure were mainly carried out in North America (Beelen et al., 2013). Several air pollution concentrations were studied in the ESCAPE project. Eeftens et al. (2012) describe the research on the effects of particulate matter (PM) on health. The particulates that were studied are: $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{COARSE}. Besides particulate matter, also the effect of nitrogen oxides on health were studied. The nitrogen oxides that were taken into account in the project are NO_X and NO_2 (Beelen et al., 2013). The health data of individuals involved in the ESCAPE project was derived from existing cohort studies. The exposure of individuals to air pollution concentrations was measured at their home addresses. The home address of an individual in a cohort explains a lot of the differences in exposure between individuals (Cyrys et al., 2012).

In 36 areas in Europe the air pollution concentrations were measured, although not all the types of air pollution were measured in all areas. In 20 study areas particulate matter and nitrogen oxides (NO_2 and NO_X) were measured and in 16 areas only NO_X was measured (Beelen et al., 2013). The areas consisted most of the time of a large city and its surroundings. However, also larges areas were taken into account, for instance in the Netherlands and Belgium where the entire country was modelled (Beelen et al., 2013; Cyrys et al., 2012). There was a large variation between the different study areas, because the population of the study areas varied between 100000 inhabitants to millions of inhabitants for large cities like London or Paris (Cyrys et al., 2012). The outputs of the model which are in microgram per cubic meter ($\mu g/m^3$) represent the annual mean of air pollution concentrations at a certain location (Eeftens et al., 2012).

The data that is used in the ESCAPE project is described by Beelen et al. (2013). The data is divided in two parts: central GIS data and local GIS data. The central GIS data consists of 4 datasets: a 1:10000 digital road network from Eurostreets, land use data from the CORINE land cover dataset, population density data at a 100m grid and height data from SRTM 90m. The local datasets consist of a local digital road network in combination with data about traffic intensity, local land use data with more specific local land use types, population density data (which is not modelled, in contrast to the central GIS dataset on population density), altitude data (which was only used when local data was better than the central dataset) and local data which is specific for a certain study area. Examples of these specific data are: *'information about wood smoke, distance to sea/lake and distance to major air pollution sources'* (Beelen et al., 2013, p. 13). The focus in the ESCAPE project was specifically on air pollution related to traffic (Eeftens et al., 2012). The association between air pollution caused by traffic and health issues is dealt with in the next section.

2.1.1 Relevance for epidemiologic studies

One of the main reasons the ESCAPE project was carried out are the concerns regarding the health effects of air pollution. The air pollution that is meant here deals mainly with air pollution caused by motorized road traffic (Eeftens et al., 2012). These health effects are shortly described in this section to get an idea of the possible effects of long-term exposure to certain air pollution concentrations.

In a review on articles that deal with the effects of particulate matter (PM) on health it is stated that there are associations between PM and deaths related to heart and lungs (Pope & Dockery, 2006). Especially the long-term exposure to PM seems to have a major impact on mortality. The long-term exposure is linked with issues such as cardiovascular illness and arteriosclerosis. Short-term exposure to air pollution is linked with issues like hospitalization for heart problems, pneumonia and strokes (Pope & Dockery, 2006). It is estimated that variations in PM_{10} concentrations cause at least 2100 deaths per year in the Netherlands. This is 1.5% of the total number of deaths per year in the Netherlands and is almost twice the number of deaths due to traffic incidents (Brunekreef & Holgate, 2002). In the same article it is estimated that 40000 deaths are caused by air pollution in Switzerland, Austria and France. About half of these deaths are caused by air pollution from road traffic. Another illness which is associated with traffic related air pollution (TRAP) is asthma (Brauer et al., 2003). TRAP is an important factor for health effects, therefore it is studied in the ESCAPE project. This is evident from the fact that several model indicators are (in)directly related to traffic, like traffic intensity data and distance to nearest road (Beelen et al., 2013).

The World Health Organization (WHO) has set up guidelines about air quality in 2006 (World Health Organization, 2006). The goal of these guidelines is to reduce the impact of air pollution on health. Policy-makers can use these guidelines as targets in order to manage the air quality. Besides the guidelines the report also presents interim targets. These interim targets are especially meant for areas with high air pollution concentrations. By using these interim targets a shift can be made from high concentrations with severe health impacts, to lower air pollution concentrations. However, the report states that the guidelines should always be the main objective. The guidelines for $PM_{2.5}$ are 10 µg/m^3 (microgram per cubic meter) per annual mean, which is the long-term guideline and 25 µg/m^3 per 24-hour mean, which is the short-term guideline. The guidelines for PM_{10} are 20 µg/m^3 per annual mean and 50 µg/m^3 per 24-hour mean. The annual mean of 10 microgram per cubic meter was chosen because health effects are likely to occur when the annual $PM_{2.5}$ concentrations are between 11 and 15 µg/m^3. Table 1 shows the interim targets and guidelines for $PM_{2.5}$ and PM_{10}.

WHOair quality guidelines and interim targets for particulate matter: annual mean concentrations[a]

	PM_{10} ($\mu g/m^3$)	$PM_{2.5}$ ($\mu g/m^3$)	Basis for the selected level
Iinterim target-1 (IT-1)	70	35	These levels are associated with about a 15% higher long-term mortality risk relative to the AQG level.
Interim target-2 (IT-2)	50	25	In addition to other health benefits, these levels lower the risk of premature mortality by approximately 6% [2–11%] relative to theIT-1 level.
Interim target-3 (IT-3)	30	15	In addition to other health benefits, these levels reduce the mortality risk by approximately 6% [2-11%] relative to the -IT-2 level.
Air quality guideline (AQG)	**20**	**10**	These are the lowest levels at which total, cardiopul-monary and lung cancer mortality have been shown to increase with more than 95% confidence in response to long-term exposure to $PM_{2.5}$

Table 1. WHO air quality guidelines. Especially the third and fourth column are relevant, because $PM_{2.5}$ values are modelled in this research. These columns indicate respectively the $PM_{2.5}$ threshold values and the risks of exposure to these levels of air pollution.
Source: WHO (2006)

2.2 General air pollution concentration measurement methods

Before going into detail on land use regression methods, this section deals with several other methods which can be used to model air pollution concentrations. Hystad et al. (2011) mention several methods which can be used to model air pollution in larger areas, like '*interpolation of fixed-site government monitoring data, dispersion modelling, satellite remote sensing, land use regression (LUR), and proximity and deterministic methods*' (Hystad et al., 2011, p. 1123). The most important methods, in the context of this study, are discussed below to determine their characteristics and their strong and weak points. Each sub section ends with a short argument why the specific approach is not suitable for this study.

2.2.1 Proximity models

According to Jerrett et al. (2005) the most basic approach to model spatial and temporal variation in air pollution within a city are proximity models. These models are based on the assumption that a person's health is influenced by the proximity to an emission source, like roadways and industrial areas (Hystad et al., 2011; Jerrett et al., 2005). Proximity models are for instance used to assess the effect of traffic-related air pollution on the respiratory tract of children (Jerrett et al., 2005).

Two advantages of proximity metrics are '*their clear policy relevance*' and pollution measurements are not necessarily required (Allen, Amram, Wheeler, & Brauer, 2011, p. 369). One of the major drawbacks of proximity models are the simple assumptions that are used. One example is given by Jerrett et al. (2005) which is about a respiratory health survey which is based on the proximity to major roadways. Respondents which live within a certain buffer around the roadway were assigned a '1' and people living outside this buffer were assigned a '0'. This is however a very simplistic method because it assumes an isotropic dispersion (Ryan & LeMasters, 2007) (i.e. people within the buffer all have the same quantity of exposure) and it assumes that people outside this buffer are not exposed to certain values of air pollution concentrations (Jerrett et al., 2005). Especially seen in the context of the ESCAPE project this is not a suitable approach to measure the personal exposure of the population, because the air pollution needs to be known for any address.

2.2.2 Geostatistical interpolation methods

Another method to model air pollution concentrations is geostatistical interpolation (Beelen et al., 2013). Interpolation techniques to estimate air pollution concentrations are often used in combination with monitoring data (Marshall et al., 2008). Interpolation in this context means that the monitoring data is used to predict the values at unknown points (i.e. points where monitoring data is not available for). The values of these unknown points are based on the surrounding monitoring data points.

A problematic characteristic of interpolation methods is that a '*smoothly varying concentration field*' is created. This means that hot spots of air pollution, such as roadways, are badly covered with these models. This is especially problematic in urban areas, where spatial variability is often caused by these kind of hot spots (Hoek et al., 2008). This is also linked to what is mentioned by Brauer et al. (2003). They state that interpolation is a suitable method to model regional pollution patterns, but it fails to capture the variations of air pollution concentrations on a smaller scale. This is often caused by the density of a monitoring network and how traffic sources are spatially distributed. The problem described in this paragraph makes it very hard to accurately calculate the personal exposure which is of importance in the ESCAPE project. Because this method badly covers the hot spots, which especially occur in urban areas, it is hardly possible to make an accurate estimation of the personal exposure of the population. This certainly applies to people living near roads.

2.2.3 Dispersion models

Dispersion models use different kinds of data to model air pollution concentrations, to be specific: topography, emission and meteorological data. The spatial distribution is modelled by using assumptions about deterministic processes which deal with the data mentioned before (Jerrett et al., 2005). A few advantages of this approach are mentioned by Jerrett et al. (2005). Dispersion models can capture spatial and temporal differences in air pollution concentrations, without having a dense monitoring network. The models can also be used at different geographical scales and they can easily be adjusted to be used in different study areas.

Even though a dense monitoring network is not needed, dispersion models are very data intensive and using these models requires months of training (Ross et al., 2006). Jerret et al. (2005) mention a number of other disadvantages of dispersion models: they require expensive input data, the assumptions about dispersion patterns are not realistic, data of different time periods can cause estimate errors and extensive cross validation is needed. In contrast to the models mentioned above, this model could be suitable to calculate personal exposure of a population. However, the reason that expensive input data is needed and that the models require months of training, makes this approach less suitable for this research.

2.2.4 Hybrid models

Hybrid models are a combination of two or more air pollution models. Jerrett et al. (2005) give the example of personal monitoring combined with regional monitoring. Personal monitoring is done by people who wear measuring equipment on their clothes. This kind of measurements can then be compared or combined with measurements from outdoor stations at fixed locations. Besides combining personal and regional monitoring, hybrid methods also consist of combining two or more air pollution concentration models (Beckerman et al., 2013), for instance two models that are mentioned in the paragraphs above.

In the same review article Jerrett et al. (2005) state that personal monitoring is a more accurate method to measure exposure of individuals to air pollution concentrations. The reason is that people spend most of their time indoors, while fixed monitoring stations are often located outside. The combination of regional monitoring and personal monitoring would be a suitable approach within the ESCAPE project, because this methodology also takes indoor air pollution into account. However, for this research project one of the objectives is to model air pollution in a quick way, preferably with datasets with a global coverage. Seen in the context of this objective it makes this methodology less suitable for this study.

2.2.5 Remote sensing

Remote sensing is a relatively new method to estimate air pollution concentrations. Satellites can be used to predict air pollution concentrations over large areas. However, they are not suitable for applications on a smaller scale, like cities, because of spatial resolution limits (Hystad et al., 2011). Remote sensing data can also be used as an input for land use regression models. In a research on estimating the variability of $PM_{2.5}$ in the United States remote sensing data on $PM_{2.5}$ concentrations was used as input for a land use regression model. By comparing two LUR models, one with and one without this remote sensing data, it seemed that remote sensing estimates are strong predictors of $PM_{2.5}$ variance (Beckerman et al., 2013; Jerrett et al., 2007). Similar findings are presented by Liu, Paciorek and Koutrakis (2009). They found that the regression model with remote sensing data predicted the PM2.5 values better than the model without this data. They concluded this by comparing the models with $PM_{2.5}$ data from ground monitoring sites.

One of the advantages of remote sensing is the possibility to measure air pollution concentrations in areas where ground monitoring stations are not available. This can for example be done in

developing countries where no money is available for ground monitoring of air pollution. Especially in developing countries where it is most needed, for instance in countries with large populations and high air pollution levels remote sensing can be a suitable method to measure air pollution concentrations (Donkelaar et al., 2010). Another advantage of using this technique in assessing air pollution concentrations is the availability of these data (Jerrett et al., 2005). Some methods, like ground monitoring sites, have issues with obtaining enough values which are representative for a larger area. Remote sensing methods are not affected by this issue, because these methods can capture air pollution over large areas (Donkelaar et al., 2010).

According to Jerrett et al. (2005) there are two drawbacks of estimating air pollution concentrations with remote sensing. The first disadvantage they mention is that it is hard to classify the type of air pollution that is derived from remote sensing. The other shortcoming is that an accepted means to assess estimate errors does not exist (Jerrett et al., 2005). In particular the spatial resolution limitations of the remote sensing methodology makes it less suitable for this study. The spatial resolution of remote sensing data would be too low to estimate the personal exposure of the population in an accurate way. However, remote sensing data is used in this study for validation of the model, because data from ground monitoring stations was not available.

2.3 Land use regression models

Land use regression models use multiple linear regression to model the relationship between the response variable air pollution (dependent variable) and one or more explanatory variables (independent variables). These explanatory variables are for instance land use, traffic density and topography (Beelen et al., 2013). The predictor values are used to explain air pollution concentrations at locations without air pollution samplers (Johnson, Isakov, Touma, Mukerjee, & Özkaynak, 2010). The air pollution levels can then be predicted for locations like homes (Ryan & LeMasters, 2007), which can be used for epidemiological studies which use home addresses of participants in birth cohorts to estimate air pollution exposure (Beelen et al., 2013). Often the predictor variables, which are mentioned above, explain the spatial variation in pollution concentration fairly well (Beckerman et al., 2013).

Not only the land use types at the monitoring sites are taken into account, but also land use types that are within a buffer around these sites (Jerrett et al., 2005). The buffers are included in every LUR model, but the size of the buffers vary per study and per variable (Ryan & LeMasters, 2007). The response variable, air pollution, is often measured at multiple monitoring locations (Beelen et al., 2013) to identify the regression. Marshall et al. (2008) provide a clear stepwise description of how LUR models are used. The description is as follows: the first step is to measure air pollution concentrations at many locations in the study area. Then land use measures within a buffer of each monitoring site are needed (e.g. traffic intensity on nearest road within a 1-km buffer). Subsequently a regression equation on air pollution concentrations needs to be developed, based on the land use in the proximity of monitoring sites. The last step is to apply the regression equation to a raster grid in the study area to estimate air pollution for any location in the area.

The capability of a LUR model to predict air pollution can be improved when more sampling sites are added (step 1 as described by Marshall et al. (2008)) (Wang et al., 2014). However, by other authors this argument is partly rejected. They state that the variability of sampling locations is more important than the number of sampling locations. This means that a land use regression model can be more suitable when a larger variety of land use characteristics are captured by the sampling network (Ryan & LeMasters, 2007).

Land use regression models have often been used for epidemiological studies (Beelen et al., 2013; Cyrys et al., 2012; Eeftens et al., 2012; Gilbert, Goldberg, Beckerman, Brook, & Jerrett, 2005; Slama et

al., 2007). Land use regression models offer the possibility to retrieve air pollution values from any point in the study area, based on several samples. This can provide valuable information on the effects of traffic of current and future roads on health of citizens (Gilbert et al., 2005). One of the main advantages of LUR models for epidemiologic studies is the opportunity to take within-city variability of air pollution concentrations into account which can be especially advantageous for traffic-related air pollution (TRAP) (Poplawski et al., 2009).

2.3.1 Advantages of LUR models

In comparison with dispersion models (see section 2.2.3 for a description) land use regression models are more favourable because they are less data intensive. Besides, LUR models seem to predict well (Ross et al., 2006). Another advantage of LUR models that is mentioned by Ross et al. (2006) is that they integrate more factors than proximity methods. This argument is further explained by Ryan and LeMasters (2007) who state that within a buffer the exposure can be differentiated by using extra land-use variables. This cannot be done with a proximity method, because then only one value for a whole buffer is accounted for, which is also mentioned in section 2.2.1.

In their evaluation of a land use regression model to predict concentrations of NO_2 Ross et al. (2006, p. 113) found that it is a robust method and '*was relatively simple in terms of data inputs required and analysis'.* Transferability of land use regression models is also an advantage, because different studies can be compared and it will save money that would otherwise be spend on monitoring in multiple areas instead of only one area (Hoek et al., 2008). This is further explained in section 2.4. In terms of cost-effectiveness LUR models are also performing well, especially seen in the light of budgets for large epidemiological studies (Hoek et al., 2008). Some advantages in comparison to other air pollution modelling methods are that LUR models require less data than dispersion models, but on the other hand the models can be more accurate than proximity methods, because more factors are included that affect exposure (Ross et al., 2006).

2.3.2 Limitations of LUR models

Although transferability was mentioned as an advantage in the previous section, it can only be done to a certain extent. According to Hoek et al. (2008) the transferability of LUR models depends on the similarity between two areas in terms of land use. In another article it is even stated that these models are not even transferable most of the time (Johnson et al., 2010). This will be further discussed in section 2.4.

Another drawback of these models is that it is hard to make a distinction between the influence of the different air pollutants. The reason for this is that some important air pollutants, like NO_2 and $PM_{2.5}$, are often highly correlated. The inability of LUR models to represent large variations in air pollution concentrations over a short distance near, for instance, major roads is the second disadvantage discussed by Hoek et al. (2008). Another limitation that has to be taken into account, especially for epidemiological studies, is the amount of outdoor air that infiltrates the homes of people. This is an important issue, because people spend a lot of their time at home, while the models are based on the air pollution outside. This means that other factors, like daily patterns of people, are also important to determine their exposure to air pollution concentrations. The last limitation of land use regression models discussed by Hoek et al. (2008) is the problem of confounding which can appear when the models are used in epidemiological studies. Hoek et al. (2008) explain this by an example where population density is included in a LUR model. This can be problematic because population density can also be correlated with low socio-economic status, which can also have an influence on the disease that is studied. However, the problem of confounding could also occur in other methods.

Another limitation of land use regression models is that the output results are influenced by the quality of the input data. For example traffic density data is often collected before the sampling period. It would however be more ideal when this data is collected during the sampling period. Besides, the desired data is not always available (Ryan & LeMasters, 2007).

With land use regression models it is hardly possible to say which emission source is responsible for certain air pollution concentrations. This makes it hard to define the right policy to prevent air pollution (Johnson et al., 2010). The authors of this article also state that most of the time LUR models only address one pollutant at a time. Besides, large numbers of monitoring sites are needed as well as accurate data.

2.4 LUR model transferability

According to Jerrett et al. (2005) it is to a certain extent possible to use land use regression models in different cities. However, this can become problematic when moving to cities or study areas with deviating topography and land use. In their article it is shown that in a certain study a model was used in an area with a very different landscape structure. This resulted in a model which showed very few spatial variation and also lacked correlation with measured pollution data. This means that transferring LUR models is possible, however, only when there is a similar geographic structure with comparable land use and transportation characteristics. Besides, often it requires a lot of samples to measure air pollution concentrations (Jerrett et al., 2005). Additionally another article states that for each pollutant and urban area the right variables have to be chosen. This is explained by an example which uses two cities with different topography. The result is that elevation is included for one city, but is left out for the other city. Also distance to the ocean was included for one city, but was of course not relevant for the non-coastal city (Ryan & LeMasters, 2007).

In a study which focused specifically on LUR model transferability it was concluded that locally calibrated models performed better than transferred models (Allen et al., 2011). This is mainly due to the empirical basis of land use regression models, which means that it is tailor-made for a specific area (Jerrett et al., 2005). Thus, in epidemiological studies it is better to develop models for a specific area of interest, because this will lead to more accurate results. However, it was also found that transferred LUR models performed better than proximity metrics (which are discussed before, in section 2.2.1). This means that applying a LUR model in another city can be a good compromise between price and quality (Allen et al., 2011).

Transferring land use regression models between different countries can be problematic due to several causes. Air pollution concentrations could be worse modelled with transferred models because of *'hidden inconsistencies in the data'* or because of fundamental differences in the determining factors of air pollution concentrations (Vienneau et al., 2010, p. 9). Therefore, these authors state that transferring LUR models to other countries or areas needs to be done with care. In fact Vienneau et al. (2010) mention two causes for errors which can occur when land use regression models are transferred: the data and the model. On the one hand a LUR model can fail in predicting air pollution concentrations because the input data contains errors or the quality is not high enough. On the other hand errors can occur because the model is not suitable for the area, i.e. the model does not have the right independent variables to predict air pollution concentrations.

Transferring a LUR model between different areas can be successful, although input data needs to be from the same source. Another condition that can help in achieving successful model transfers is by modelling two areas with similar spatial structure (Poplawski et al., 2009). However, it is also possible to achieve a successful transfer when the input data is not from the same source, as shown by Poplawski et al. (2009). It should nevertheless be noted that calibration was used when these models were transferred. This means that data needs to be available from field monitoring. According to

Briggs et al. (2000) it is not needed to have a lot of field monitoring sites for calibration. It is more important that the samplers are located at the right positions, i.e. that they *'reflect the range of actual values in the study area'* (Briggs et al., 2000, p. 161). Achieving a successful model transfer can be improved by focusing less on achieving the highest R^2. This means that the predictability of the model will be worse, but the model has more transfer potential. In addition to this transferability of a model can also be improved by using centralized and uniform data (Hoek et al., 2008)

One citation which summarizes the section above very well comes from Jerrett et al. (2007, p. 209), they state that: '*the more the model is refined to specific conditions in one locale, the less transferable and operational it becomes'*. It can be concluded that transferring land use regression models is possible, however only to a certain extent and it should be done with care. Local characteristics of specific cities have to be taken into account and better results can be achieved when cities are chosen with similar characteristics. Also the input data has a large influence on the success of a transfer: input data from one, centralized source will probably lead to more success than using different local data sources.

2.5 Data availability

This section deals briefly with the availability of data as input for the land use regression models of the ESCAPE project. The focus is mainly on the central datasets as described in section 2.1: a digital road network, land use data, population density and height data.

2.5.1 Height data

The height data that is used in the ESCAPE project is the SRTM90 dataset, which stands for Shuttle Radar Topography Mission (Wang et al., 2014) with a 90 meter resolution at the equator. The data that is collected by the SRTM is available globally (Eeftens et al., 2012). Since September 2014 for some areas the SRTM 1 Arc-Second (30 meter resolution) data have been released (USGS, 2015a). Depending on the availability of the data for the area of interest it can be used in this research to have more detailed height data.

2.5.2 Land cover data

For the ESCAPE project CORINE (COoRdination of INformation on the Environment) (Eeftens et al., 2012) land cover data is used. As CORINE only contains European land cover (Wang et al., 2014) this dataset is not suitable for this research. Below a number of alternative datasets are discussed with a global land cover.

- GlobCover is an initiative of the European Space Agency (ESA) to obtain a global land cover map. The spatial resolution of GlobCover is 300 meters and is collected by using the MERIS (MEdium Resolution Imaging Spectrometer). The classification is done according to the UN Land Cover Classification (Arino et al., 2007).
- The Chinese alternative for a global land cover map is Globeland30. Globeland30 has a 30 meter resolution and has 10 land cover classes: cultivated land, forest, grassland, shrub-land, wetland, water bodies, tundra, artificial surfaces, bare land, permanent snow and ice (Ran & Li, 2015, p. 1678).

2.5.3 Road network

Search engines with scientific articles did not show much results about a global road dataset. It is however also possible to extract road data from OpenStreetMap. OpenStreetMap makes it possible for everyone to edit and add geographic information (C. Liu, Xiong, Hu, & Shan, 2015). Other research has shown that OpenStreetMap data is pretty accurate: already in 2008 the OpenStreetMap had circa 80% overlap with the Ordnance Survey dataset of England (Haklay, 2010).

2.5.4 Population density

One of the global population density maps is the Gridded Population of the World (GPW). This map is based on administrative units and distributes the population uniformly over administrative areas (Pozzi, Small, & Yetman, 2002). This uniform distribution is a disadvantage because it does not take into account that some parts of an administrative area will be more densely populated than other parts. This can be especially problematic when the data is needed for land use regression models because in some cases these models require population density data on a city scale.

Another global population density map is the Global Rural-Urban Mapping Project (GRUMP). This map has a 927m resolution and is created by the Center for International Earth Science Information Network (CIESIN), just like the GPW which was mentioned above (Schneider, Friedl, & Potere, 2010). This dataset has the same basis as the GPW, although it distinguishes between urban and rural areas. The urban areas are identified by capturing night-time light of cities with satellites (SEDAC, 2000).

An alternative global dataset with a higher spatial resolution than the two datasets mentioned above is the World Population Estimate (ESRI, 2015b). This dataset has a spatial resolution of approximately 250 meters. The map is created by using satellite imagery to detect places where people do not live, like places with water and permanent snow. A lot of texture often indicates places where a lot of buildings and road are located, which is an indication that people live there. This is combined with additional information, like road intersections and areas with extreme climates. Places with a lot of road intersections indicate that people are likely to live there. It is less likely that a lot of people live in areas with extreme climates. Combined with census data each grid cell is populated with a certain number of people, based on the data mentioned in this paragraph (ESRI, 2015c).

2.6 Validation

To decide whether a LUR model performs well, validation techniques need to be carried out to prove this. A broad range of articles mentioned in this chapter describe the validation technique that was used in the research, including the articles about the LUR models of the ESCAPE project (Beelen et al., 2013; Cyrys et al., 2012; Eeftens et al., 2012; Wang et al., 2014). The validation method used by Beelen et al. (2013) is the leave-one-out-cross-validation, which means that a model is run with one monitoring site left out of the model run. The predicted values for this monitoring site are then compared with the measured concentration at this monitoring site. This is repeated for all monitoring sites which results in a measure which indicates the performance of the model (Hoek et al., 2008). An example of a measure of model performance is the Root Mean Squared Error (RMSE). The RMSE calculates the squared difference between the observed and the predicted value. All these values are then summed up and divided by the total number of observations. The last step is to calculate the square root of the last calculated value (Brauer et al., 2003). Another type of validation which uses one dataset is to split this dataset in one part to develop the model and one part to validate the model (Briggs et al., 1997, as cited in Hoek et al., 2008).

A different validation method is mentioned by Henderson, Beckerman, Jerrett, & Brauer (2007). In their article about applying a land use regression model in Vancouver they validate their model by using pollution data from other monitoring stations which were not used to develop the model. With leave-one-out-cross-validation the model error was estimated. Hoek et al. (2008, p. 7572) call this approach a '*comparison with databases that have not been used in model development*'. Besides other monitoring stations, which were not used to develop the model, also measurements from personal monitoring can be used to validate a model. Personal monitoring means that people wear air samplers on their clothes which measure the air pollution concentrations at locations where these people are (Jerrett et al., 2005). Also remote sensing data (which is dealt with in section 2.2.5) could be used to validate land use regression models because it can predict air pollution concentrations over large areas, as stated in that same section. However, on smaller scales, like

cities, it is not possible to make detailed predictions because the spatial resolution is limited (Hystad et al., 2011). This means that validation can only be roughly done with remote sensing data.

2.7 Exposure

As stated in the section about the ESCAPE project the personal exposure was measured by using home addresses of people in a cohort. These cohorts with health data were already available from previous cohort studies (Cyrys et al., 2012; Eeftens et al., 2012). However, these cohorts are often not publicly available. Therefore this section shortly describes some datasets which can be used to estimate the personal exposure of the population.

Population density maps, like GPW and GRUMP (Pozzi et al., 2002; SEDAC, 2000) mentioned in section 2.5.4, could be used to estimate the exposure of the population. These datasets contain a global grid with the estimated population per grid. With a random distribution of the population within these grids the population can be spread over the surface and be combined with air pollution data. However, this gives just an overall and rough picture of the population distribution in a certain area.

Another methodology to get a more accurate distribution of population is to combine census data with satellite pictures of settlements (Linard, Gilbert, Snow, Noor, & Tatem, 2012). But then there is still the problem that the population needs to be distributed within these settlements. Population census data can however more accurately be distributed by combing these data with building footprints, building volume and the number of building floors. Lwin and Murayama (2009) wrote in their article that this methodology provides good results which can be used in micro-spatial analysis. Examples of micro-spatial analysis they give are disaster management, consumer market analysis and public health programs. Based on this article this seems to be a suitable approach to estimate home locations of the population which can subsequently be used to calculate personal exposure. Besides the suitability of this methodology for an accurate population distribution, Lwin and Murayama (2009) write that some problems still need to be solved. Examples of problems are the estimation of how much percent of a building has a residential function and the building status, because buildings can be under construction or abandoned. Another likely problem, which is not mentioned in their article, is the availability of the data. It is conceivable that a lot of countries lack the data about the locations of all buildings with additional attributes, like the number of floors and the types of those buildings (e.g. residential or commercial).

2.8 Synthesis

This synthesis deals with several expectations for this research based on the literature study above. With the research objectives and questions in mind, the literature above is used to formulate some expected outcomes for this study.

Transferability of a model can be improved by using centralized and uniform data from one source (Hoek et al., 2008; Poplawski et al., 2009). Because in this study cities will be compared which are located on different continents, transferring of the LUR model will be harder. However, the prediction capacity of the models can be improved by calibration (Poplawski et al., 2009). This is however dependent on the availability of existing air pollution data. Another issue when using data is that it is preferably collected in the same year (Ryan & LeMasters, 2007). It is expected that in this study data will be from different years. Although this will worsen the model results, using this data can provide insight in the transferability of the ESCAPE LUR model, which has more priority in this research than achieving the highest R^2.

Besides the data, it is also expected that the difference in land use and topography of the concerned cities will play a role in the transferability of the LUR model in this study (Jerrett et al., 2005). The

differences which can occur due to land use and topography between the London model and the two cities that are using this study can lead to deviations which will be researched in this study. This underpinning from literature gives useful insights in what can be taken into account beforehand. What is mentioned above comes down to the two error sources mentioned by Vienneau et al. (2010). These two error sources are a model which is not suitable for a specific area and the input data contains errors or the quality of this data is not high enough.

The last two objectives of this research about population exposure and model validation rely heavily on the availability of the data. When (validation) data is not available at all it will be hard, if not impossible to examine these questions. However, when data is available the quality of this data will play an important role in achieving accurate results (Ryan & LeMasters, 2007). It is for instance possible to make accurate estimations of exposure to air pollution when home addresses of inhabitants are available. However, the coarser the data is, the less accurate exposure results can be achieved.

3. Methodology

This section is about the methodology that is used in this study. The research has mainly a quantitative approach because land use regression models are quantitative ways to model air pollution concentrations (Hoek et al., 2008). The first section deals with the input data and the regression equations of the land use regression models. The next section describes which steps need to be taken to achieve the objectives. The third section deals with the software used in this study. The last part of this methodology chapter deals with the selection of the cities which are modelled during this research project.

3.1 Data and the regression equations

In the ESCAPE model several datasets are included to model air pollution concentrations. Beelen et al. (2013) mention the required datasets. The authors make a distinction between datasets that were available for all the study areas and datasets that were only used when they were available for the specific area. The four central GIS datasets, which were available for all areas are:

- digital road network data
- land use data
- population density data
- altitude data

The local GIS data consists of the following five datasets:

- local road network in combination with data about traffic intensity
- local land use data with more specific local land use types
- population density data (which is not modelled, in contrast to the central GIS dataset on population density)
- altitude data (only when local data was better than the central dataset)
- local data which is specific for a certain study area. Examples of these specific data are: '*information about wood smoke, distance to sea/lake and distance to major air pollution sources'* (Beelen et al., 2013, p. 13).

These local datasets were used in the ESCAPE project if European data was not available for a specific region, the local data was more up-to-date or more precise than the central datasets (Eeftens et al., 2012). In this study the road network dataset can be regarded as a central dataset, because OpenStreetMap is used, which is discussed in section 4.1.3. The traffic intensity data is regarded as a local dataset, which is discussed in section 4.1.5. It depends on the regression equation which dataset is required. For this study there are two important data inputs for the models: the digital road network and the traffic intensity data. The traffic intensity data consists of all the vehicles which pass by per 24 hours on a road (Eeftens et al., 2012). Because traffic intensity datasets were not available for the cities modelled in this study, an assumption was made to estimate the traffic intensity. This assumption can be found in section 4.1.5. Appendix 2 (section 8.2) provides more information on how the traffic intensity data is exactly calculated.

The London model which is applied to Bangkok and Mexico City requires the digital road network dataset and the traffic intensity data on those roads. The equation (1) to estimate $PM_{2.5}$ as it was developed for the London area is (Eeftens et al., 2012):

$PM_{2.5} = a + bi + cl$ (1)

The values of the parameters can be found in table 2. The *'i*' in the first equation stands for '*INTMAJORINVDIST*' and the '*l'* stands for '*ROADLENGTH_500'*.

Parameter	Value
a	7.19
b	$1.38*10^{-3}$
c	$2.65*10^{-4}$

Table 2. The parameters of the London $PM_{2.5}$ equation with the corresponding values.

The '*i*' variable is the product of the traffic intensity on the nearest major road and the inverse of distance to the nearest major road. The '*l*' is the road length of all roads in a buffer of 500 meters.

The Dutch model, which was used to validate the input datasets also required the road network dataset and the traffic intensity data, in addition to the regional estimate variable. The Dutch regression equation (2) to estimate $PM_{2.5}$ is (Eeftens et al., 2012):

$PM_{2.5} = a + br + cm + dt$ (2)

The values of the parameters can be found in table 3. The '*r*' in the second equation stands for '*REGIONALESTIMATE*', the *'m'* stands for '*MAJORROADLENGTH_50*' and the *'t'* stands for *'TRAFMAJORLOAD_1000'*.

Parameter	Value
a	9.46
b	0.42
c	0.01
d	$2.28 * 10^{-9}$

Table 3. The parameters of the Dutch $PM_{2.5}$ equation with the corresponding values.

The '*r*' variable is a background variable which measures the $PM_{2.5}$ values at locations which are not in the proximity of emission sources, like roads and harbours. This variable was used to explain the variation in air pollution which could not be explained by the other variables because of the limited buffer size of 5000 meters (Eeftens et al., 2012). The data for this background variable was collected at luchtmeetnet.nl which contains a collection of different monitoring sites, for instance from the Ministry of Infrastructure and the Environment and the National Institute for Public Health and the Environment (Luchtmeetnet, 2016).

The *'m'* variable consists of the total length of all major roads within a 50 meter buffer. The *'t'* variable is the sum of the traffic load (traffic intensity on a major road * the length of the major road) of major roads in a buffer of 1000 meters. Appendix 2 (section 8.2) provides more information on how the variables of the London and Dutch equation were calculated.

The road network dataset, which is discussed in section 4.1.3 was downloaded from Geofabrik.de (Geofabrik, 2015). With this website it is possible to download the required road data in a shapefile format. The fourth chapter also discusses the traffic intensity data which was used in this research project.

3.2 Methodology per objective

Below a stepwise explanation per sub-objective is given of the steps that are required to achieve the main objective.

3.2.1 Input data for the regression equations

The main objectives is divided in three sub-objectives. The first sub-objective is: 'Identify and evaluate relevant and ubiquitous datasets which can be used as input for the model.'

Figure 1 shows the steps that were followed to examine this first sub-objective. The first step is doing a literature study to identify relevant input data for the land use regression model. After the data identification the data had to be searched for. When suitable data was found it was discussed and described to know what input data should be used in the model. This is important because the input data influences the model results. Also data pre-processing (when necessary) was done for the first sub-objective.

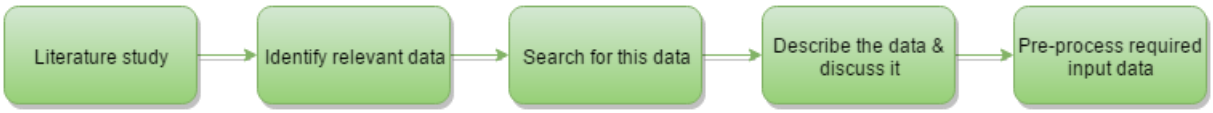


Figure 1. Stepwise description of sub-objective 1.

3.2.2 Validation & sensitivity analysis

The second sub-objective of this research is: 'Identify the deviations that occur when the model is applied to different cities.' After the model was run the deviations of the model needed to be identified. In the theoretical framework chapter it is already mentioned that two errors can occur when transferring LUR models: the model is not correct or unsuitable data was used. The model validation is done by comparing the model outcomes with data from remote sensing, which is also described in the theoretical framework (section 2.6). This section in the theoretical framework also describes the possibility to validate a model with data from monitoring sites which are not used for the development of the model. It was however not possible to find any $PM_{2.5}$ monitoring site in the areas which were modelled. The remote sensing data used in this research have a spatial resolution of approximately 10 kilometres at the equator and represent the average $PM_{2.5}$ values in the period between 2001 and 2010 (SEDAC, 2012). For the comparison with the remote sensing data two model outputs are used: one 'standard' (i.e. unchanged) output and one output which does not include major road locations and buffers of 10 meter around major roads. The reason is that these locations contain very high $PM_{2.5}$ values (up to 37000 µg/m^3 for Bangkok and 184000 µg/m^3 in Mexico City). These high values would have a lot of influence on the average $PM_{2.5}$ values of the models. By comparing both outputs of the models the influence of the major roads and their surroundings can be assessed. However, for the sensitivity analysis and the exposure modelling the model output is used without the major roads and the 10 meter buffers around these major roads. The reasons is that these model output values are more similar to the remote sensing data values than the 'standard' output values.

The input data validation was done by using data from the same source as the datasets used in this research and apply this data to locations in the Netherlands. This was done for locations which were already modelled in the ESCAPE project in order to compare those two outcomes. The reason to compare those two outcomes is that in the ESCAPE project European and local datasets were used. In this study global data is used and the traffic intensity data is based on assumptions (which is described in section 4.1.5). This means that less specific and worse data is used in this study. To identify the differences between input datasets, there are two options: to compare the input datasets and to compare the outputs of the models with the different input data. In this study only the outputs of the models with different input data are compared. The reason is that the road network data used in the ESCAPE project is not available as open data.

A part of Rotterdam was used to validate the models with the different input data. The source of the datasets that was used in Bangkok and Mexico City was also used in combination with the ESCAPE model for the Netherlands. For example, the road network data for Rotterdam is derived from OpenStreetMap, just like the road network data for Bangkok and Mexico City. In the ESCAPE project the Eurostreets dataset was used for the road network data (Eeftens et al., 2012). By comparing these model outputs the effect of the different input datasets can be measured.

The second sub-objective also contains a section with a sensitivity analysis. The section describes which variable has most influence on the model output. This can help in determining the importance of the quality of the input data. When a variable has a lot of influence on the model output, it is important to have high quality input data. This is less important when the variable has little influence on the model output. It means that the sensitivity analysis is not meant to identify deviations, but to identify which variables have the most influence to cause the deviations. Figure 2 shows the steps which need to be taken to achieve the second sub-objective.

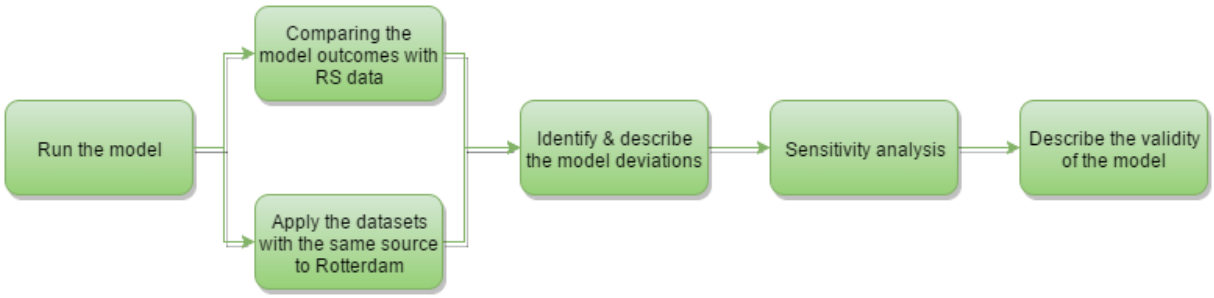


Figure 2. Stepwise description of sub-objective 2.

For the Dutch regression equation one of the inputs consists of a background variable. These background variables take the pollution into account that is caused by sources which are not within a distance of the maximum buffer distance used in the regression equations (5000 meters). In the Dutch model of the ESCAPE project this was modelled by interpolating the data of 10 regional background sites (Eeftens et al., 2012, supplementary information).

This study also used background variables which were downloaded from the website of luchtmeet.nl (Luchtmeetnet, 2016). In comparison to the background data used in the ESCAPE project this research project also used city background variables to interpolate the regional background variable. The reason is that it was not possible to cover the research area of Rotterdam with just the background variables in the Netherlands. The model was run with two background variables, one with all the (city) background measuring sites included and one with two of these measuring sites left out. The reason for this choice is that these two background measuring sites are located in and very near the research area which could cause too much influence on the model results. For better readability the model with all (city) background measuring sites is called 'model A1' and the model with two measuring sites left out is called 'model A2'. Additional information on the regional background variable can be found in section 3.1.

The input data validation was done by comparing the $PM_{2.5}$ values at 1226 points. These 1226 points were the result of the official ESCAPE model and represent address points. The model output data was provided by Utrecht University. To research the correlation between the official ESCAPE model and the models created in this research Pearson's was used. Also the absolute model errors were analyzed by applying the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). These statistics were normalized in order to compare the RMSE and the MAE of the two model outputs.

3.2.3 Exposure modelling

The last sub-objective of this research is: 'Calculate and compare the personal exposure of the population.' Personal in this research means the exposure at the home address of a person. The first step of this third objective was to do a literature study to figure out which data is needed to model personal exposure and how such an assessment needs to be done. The next step was to find the required population data and pre-process it when needed. The last step was to combine the population data with the results of the land-use regression model to calculate the exposure of the

population to air pollution. The steps are also shown in figure 3. As expected the health data and cohorts were not publicly available, like in the ESCAPE project. This is the reason that a less accurate exposure assessment was done. To get an overall idea of the exposure of the population the amount of people per neighbourhood were randomly distributed over the neighbourhood. A global world population layer with a spatial resolution of approximately 250 meters, the World Population Estimate (WPE) (ESRI, 2015b), was used as a weight layer to get a more accurate distribution. This means that for each neighbourhood the amount of people were distributed based on the weighting of a WPE grid cell within that neighbourhood. For example, a neighbourhood with 1000 people overlaps with 4 WPE grid cells with a weighting of 0.3, 0.4, 0.1 and 0.2. The first grid cell within that neighbourhood gets assigned 300 people, the second grid cell 400 people, the third grid cell 100 people and the fourth grid cell 200 people. Subsequently the people were randomly assigned a location within each grid cell. The reason that not just the WPE is used to distribute the population is the large differences between the total amount of people in a neighbourhood according to the WPE and the census data. However, by using the WPE as a weight layer a more accurate distribution could be made. The 2010 census data of Mexico City can be found at the website of the National Institute of Statistics and Geography (INEGI, 2010). The total population of Mexico City (Federal District) in 2010 was 8851080. The Benito Juárez borough had 385439 inhabitants in 2010. The 2010 census data of Bangkok can be found at the website of Citypopulation.de (Citypopulation, 2010). This website derived the census data from the National Statistical Office of Thailand. The total population of Bangkok in 2010 was 8305218**.** The six districts of Bangkok used in this research had 489893 inhabitants in 2010.

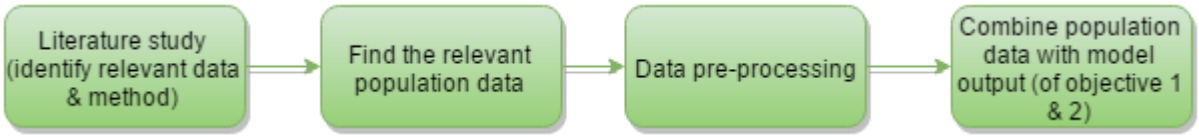


Figure 3. Stepwise description of sub-objective 3.

For each randomly distributed person in the modelled areas his or her $PM_{2.5}$ exposure value was derived by taking the $PM_{2.5}$ model output at the location of the person. The model outputs used for the exposure modelling are the outputs with the NoData values for the major roads and the buffers of 10 meters around these roads, which is also explained in section 3.2.2. People that were randomly distributed on these locations were appointed the same value as the nearest person which was outside the major road area. After the people were assigned a value, based on their location, the population was classified. The classification table of the WHO, presented in section 2.1.1 of the theoretical framework was used to determine in which class each person would fall. Because this classification table only provides targets and not ranges, the ranges of the classes are based on the report of the WHO (World Health Organization, 2006). The lowest class ranges from 1 to 10 µg/m^3, the second class ranges from 10.1 to 15, the third class ranges from 15.1 to 25 and the highest class ranges from 25.1 to 35. The first class (the Air Quality Guideline class) is provided as an example of how the WHO came up with this classification. Most of the health effects start to occur when the annual mean of µg/m^3 is between 11 and 15. That is why the WHO chose the guideline of 10 µg/m^3 (World Health Organization, 2006). For this research it means that people assigned to this class have little chance to get health effects due to air pollution based on their home location. An average exposure value for the whole population was calculated in order to compare this value with the average air pollution value of the models. This made it possible to analyse the effect of the population distribution on the personal exposure. Also the differences between someone located at 50 meters from a major road and someone at 200 meters from a major road were analyzed. This made it possible to estimate the influence of the population distribution on the personal exposure of the population.

3.3 Software

The data pre-processing is mainly done with ArcGIS 10.2 and PCRaster '*a collection of software targeted at the development and deployment of spatio-temporal environmental models'* (PCRaster, 2015). The PCRaster software is also used to calculate the final land use regression models. The python scripts used in the PCRaster software with additional explanation can be found in appendix 3 (section 8.3). Appendices 1 (section 8.1) and 2 (section 8.2) describe how the road network was reclassified and how the variables were pre-processed and calculated. PCRaster was also used to do the buffer operations for the regression equations. The 'windowtotal' operator uses a square window to sum all the values within the defined distance. In the case of the *'l'* variable this distance is 50 cells of 10 meters. Appendix 2.1 (section 8.2.1) provides the python script of the 'windowtotal' operation. SPSS was used to do the correlation calculations (IBM, 2016).

3.4 City selection

In London both particulate matter and nitrogen were modelled (Beelen et al., 2013; Eeftens et al., 2012). The $PM_{2.5}$ regression equation was used to model the air pollution concentrations in the two cities outside Europe. Below the selection of these cities is dealt with. The city of London was chosen as the land use regression model which is used in this study. One of the reasons is that London is one of the larger cities modelled in the ESCAPE project. By using this model it can be applied to other large cities throughout the world. The idea behind it is that data availability will probably be better in larger cities in the somewhat more developed countries. Besides, the $PM_{2.5}$ equation for London contains two variables which are directly related to roads and traffic. This means that it is probably a suitable model to estimate air pollution in cities with a lot of air pollution problems and a dense road network.

The cities that were used in this study are Mexico City (Mexico) and Bangkok (Thailand). These cities are selected by using the Wikipedia 'List of cities proper by population' (Wikipedia, 2015). Even though Wikipedia is not known as a very reliable source, it gave a good indication of the population densities in these cities. The population density of the city of London was used as a guidance to find cities with similar population densities. Although it was tried to have similar population densities for the cities in this study, there are some differences in population density. The main reason was that data availability seemed to be worse for cities with a more similar population density. Data availability was checked by doing a quick scan on the web. Bangkok and Mexico City seem to be both cities with a dense road network, a lot of motorized road traffic and serious air pollution problems. A small part of both cities, Mexico City and Bangkok, was analyzed with the model. For Mexico City the borough Benito Juárez was chosen which has a total surface of 28.5 km^2. In Bangkok an area was selected with a similar surface size. Six districts were selected with a total surface size of 31.4 km^2. The reason that not the entire cities were modelled is the amount of calculation time that would take. An overview of the cities and the modelled districts can be found in figure 4 and 5. In the rest of this thesis Benito Juárez and Mexico City are used interchangeably. The six districts of Bangkok are referred to as Bangkok or the Bangkok area.

As stated in section 3.2.2 the source of the road network dataset (OSM) used for Bangkok and Mexico City was also used for the Dutch model. Also the traffic intensity data was estimated in the same way for both models. This was done in order to validate the input data. By using cities which were already modelled in the ESCAPE project the outcomes of the models could be compared. One of the cities that was assessed in the ESCAPE project is Rotterdam. Thus, in this study the road network required to model air pollution in Rotterdam was also OpenStreetMap and an assumption was made about the traffic intensity data. The output of this model was compared to the original output of the ESCAPE project. The area of Rotterdam has a similar surface area (31.6 km^2) as the areas of Bangkok and Benito Juárez. The area consists of 4 neighbourhoods: Rotterdam-Centrum, Noord, Kralingen-Crooswijk and Feijenoord. Figure 6 shows an overview of the area.

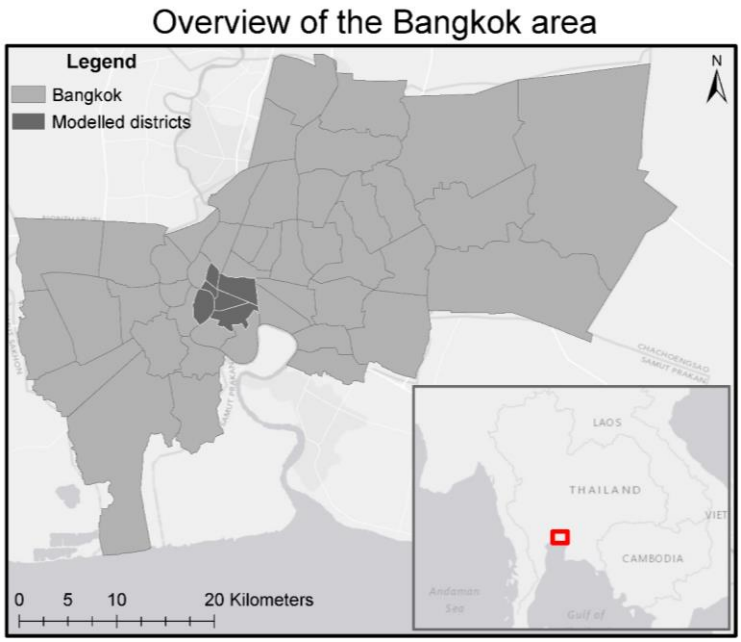


Figure 4 Overview of Bangkok and the location of the six districts. The inset map shows the location of Bangkok within the region. Sources: ESRI, HERE, Delorme, MapmyIndia, © OpenStreetMap contributors, GIS user community. These sources are derived from the ESRI service layer and are not included in the bibliography. Own source: ESRI (2015a).

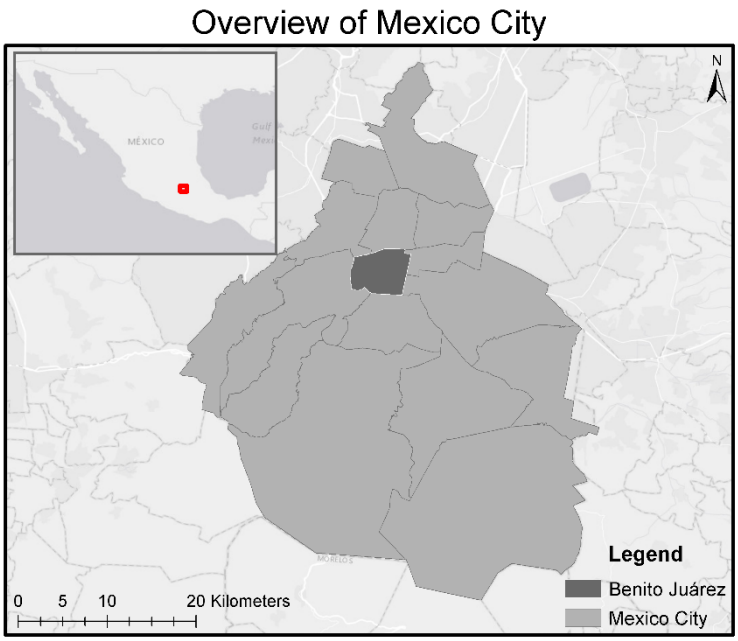


Figure 5 Overview of Mexico City and the Benito Juárez borough. The inset map shows the location of Mexico City within the country. Sources: ESRI, HERE, Delorme, MapmyIndia, © OpenStreetMap contributors, GIS user community. These sources are derived from the ESRI service layer and are not included in the bibliography. Own source: ESOC (2010).

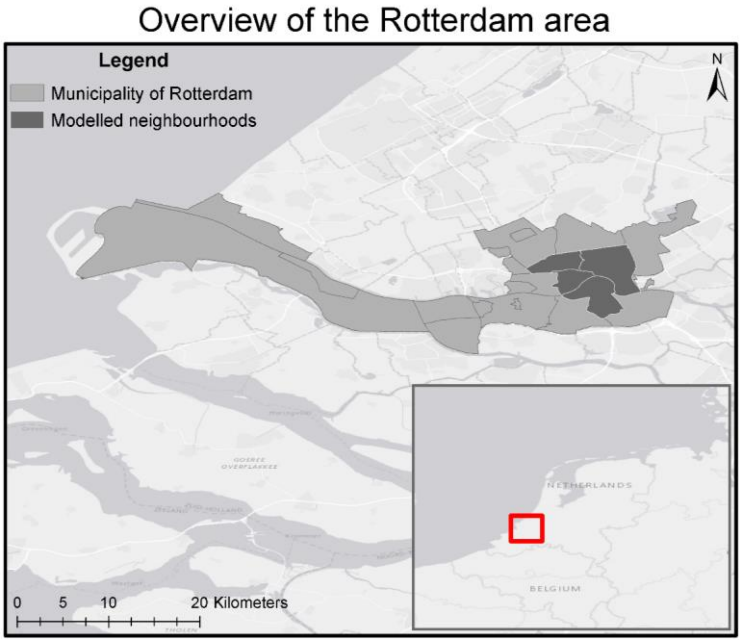


Figure 6 Overview of Rotterdam and the four modelled neighbourhoods. The inset map shows the location of Rotterdam within the Netherlands.

Sources: ESRI, HERE, Delorme, MapmyIndia, © OpenStreetMap contributors, GIS user community. These sources are derived from the ESRI service layer and are not included in the bibliography. Own source: CBS (2015).

4. Results

This chapter contains the results of this study. The chapter presents the results separately for each sub-objective.

4.1 Input data for the models

This section presents the results of the first sub-question and the related sub-objective. The first sub-question was formulated as follows: 'Which input data can be used to compare different cities outside Europe using the model?' The objective of this question is to identify and evaluate relevant data which can be used as input for the model. The reason that this question is studied is the availability and quality of global datasets, which are probably less organized, of a lower quality than European datasets, and have presumably a lower spatial resolution. The identification part of this sub-objective was done by searching the data on the internet and by searching in scientific literature. The evaluation part of this sub-objective was done by comparing the characteristics of the datasets with the datasets used in the ESCAPE project. Examples of characteristics are the classification of the data, the resolution of the dataset and the data collection method. Subsequently the potential consequences of using these datasets in the model is dealt with.

The datasets that are discussed in this section are not all needed as input data for the London ESCAPE model or the Dutch ESCAPE model. These models only require the road network dataset and traffic intensity data. However, these datasets are also discussed in this section to get a good overview of all the datasets with a global coverage. This gives an idea of the possibilities to model air pollution concentrations in cities throughout the world and shows the potential issues that could occur when the datasets are used. As stated in the theoretical framework the central GIS data used in the ESCAPE project consist of 4 datasets: a road network dataset, land use data from the CORINE land cover dataset, population density modelled at a 100m grid and height data from the Shuttle Radar Topography Mission at a 90m resolution (Beelen et al., 2013). Of the local GIS data only the traffic intensity data is discussed because this dataset is required in the equations of the London and Dutch LUR model. For each dataset an example is shown with the contours of the modelled area of Rotterdam. When possible, the input data of the ESCAPE project is also shown in order to compare the datasets.

4.1.1 Elevation dataset

The height dataset used in the ESCAPE project is SRTM3 (3 arc-seconds), which stands for Shuttle Radar Topography Mission with a resolution of 90 meters at the equator (Beelen et al., 2013). However, in 2014 the SRTM1 (1 arc-second) data was made publicly available (NASA, 2014). This dataset has a spatial resolution of about 30 meters at the equator. A comparison of the two versions is shown in figure 7 and 8.

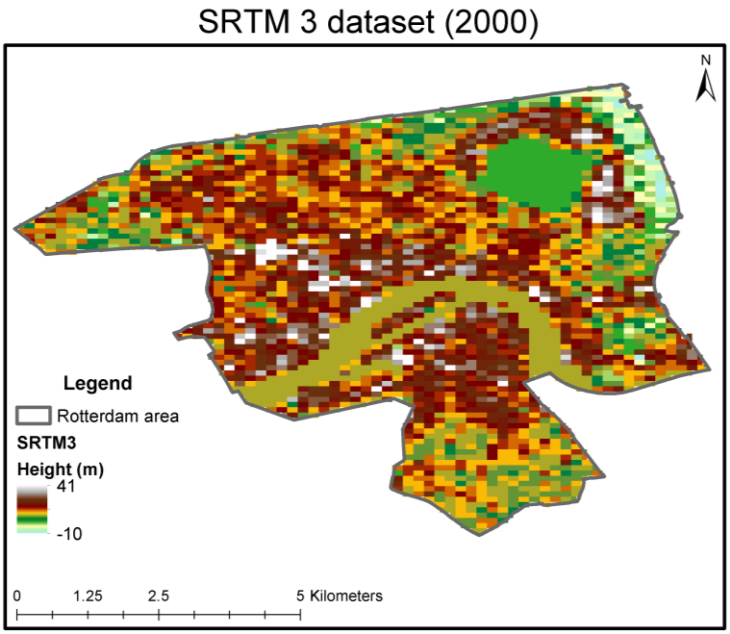


Figure 7. The SRTM 3 dataset with a spatial resolution of 90 meters. The extent of the area is the study area of Rotterdam used in this research project.
Source of the data: USGS (2015b).

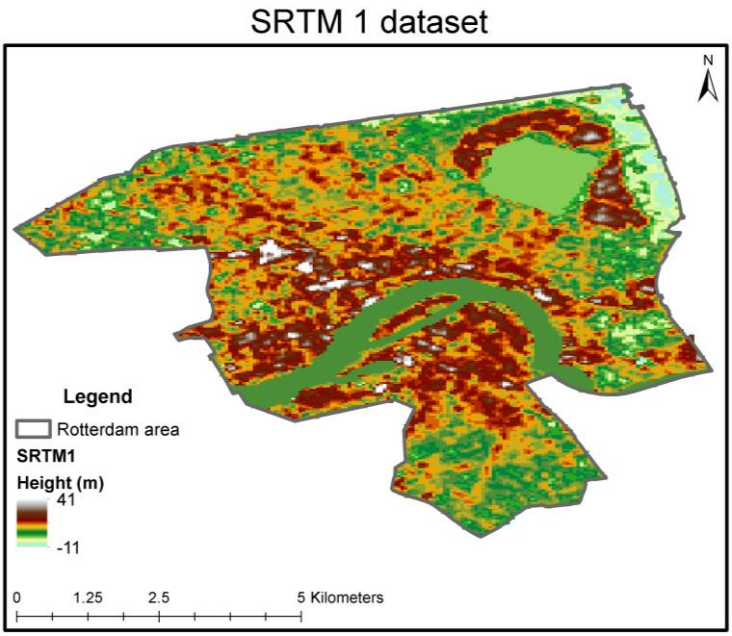


Figure 8. The SRTM 1 dataset with a spatial resolution of 30 meters. The extent of the area is the study area of Rotterdam used in this research project.
Source of the data: USGS (2015a).

When a specific LUR model requires land surface elevation data, it is possible to use the SRTM1 version of the elevation dataset. However, the regression equations of the ESCAPE project are based on the SRTM version with a spatial resolution of 90 meters. This could influence the output of the LUR models. Because the elevation data is not required for the Dutch and the London regression equation these data inputs are not validated. Before using the height data with a spatial resolution of 30 meters a validation should be done to discover the differences between model outputs using a 30 or 90 meter dataset. If the results do not show significant differences, the 30 meter version can be used to obtain a model output with a higher spatial resolution.

4.1.2 Land cover dataset

The land cover dataset used in the ESCAPE project is the CORINE land cover dataset, which is shown in figure 9. This is a dataset which provides the land cover of countries in Europe (Beelen et al., 2013). Because only the European countries are covered with this dataset another dataset had to be found which has a global coverage. One of the advantages of the CORINE land cover dataset in comparison with global land cover datasets is the broad class of artificial surfaces. This class contains for instance continuous urban fabric (at least 80% of the total surface is covered by artificial objects), discontinuous urban fabric (30 to 80% of the total surface is covered by artificial objects), port areas, airports and green urban areas (EIONET, 2012). Most global land cover datasets are much less comprehensive. Two examples are the GlobeLand30 and GlobCover V2 datasets. These datasets have both just one class representing urban areas (Ran & Li, 2015; Schneider et al., 2010): respectively ‘artificial surfaces’ and ‘artificial surfaces and associated areas’.

This is however a major shortcoming of these datasets in the context of the land use regression models. The reason is that the regression models sometimes need this distinction to predict air pollution concentrations. Some LUR models require for instance the amount of square meters of high density residential land within a buffer of 1000 meters (Eeftens et al., 2012). The fact that this distinction is missing, means that the regression equations which require this distinction cannot be used with global datasets. When a certain LUR model requires the variable artificial surface it is

preferred that the GlobeLand30 dataset is used because of the higher spatial resolution of 30 meters compared to the 300 meter spatial resolution of the GlobCover V2 dataset (Arino et al., 2007; Ran & Li, 2015). An example of the GlobeLand30 dataset is shown in figure 10.

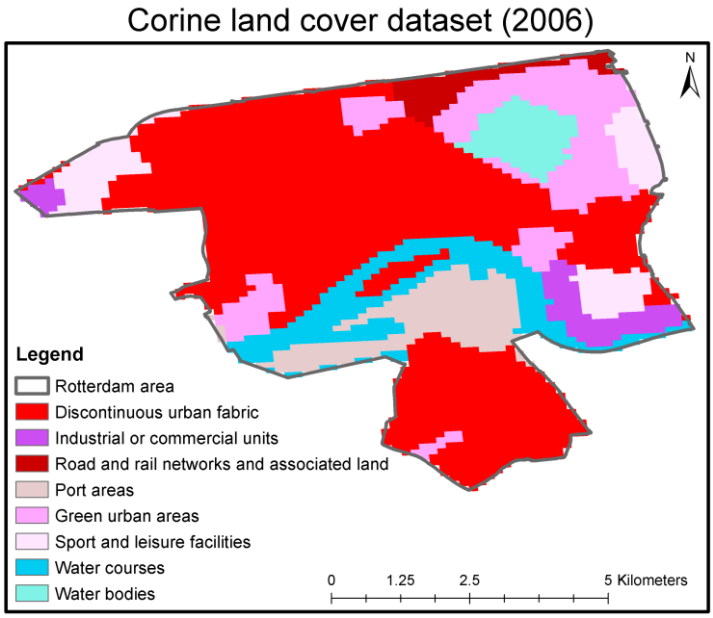


Figure 9. The Corine land cover dataset which was used in the ESCAPE project. The map has a spatial resolution of 100 meters. The extent of the area is the study area of Rotterdam used in this research project.
Source of the data: EEA (2010).

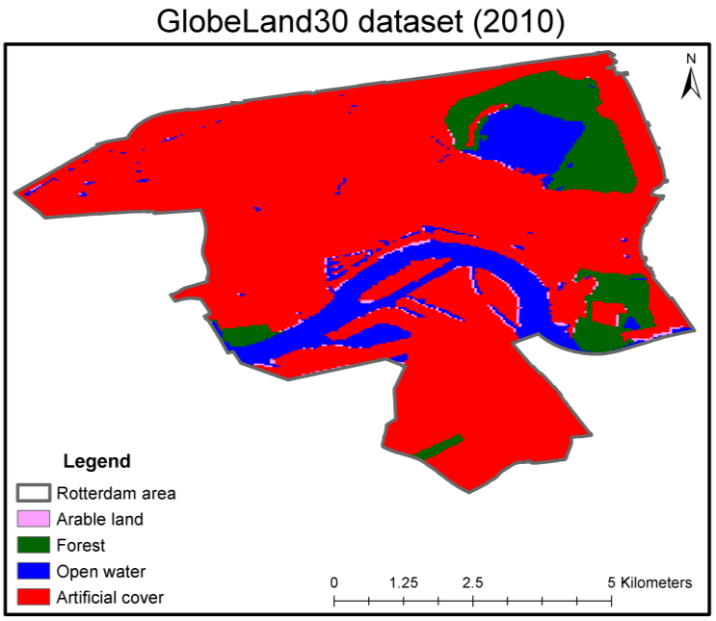


Figure 10. The Globeland30 dataset which can be used as a global alternative for the Corine land cover map. The map has a spatial resolution of 30 meters. The extent of the area is the study area of Rotterdam used in this research project.
Source of the data: GlobeLand30 (2010).

4.1.3 Road network dataset

For the road network dataset is chosen for OpenStreetMap (OSM). The main reason is that OSM has a global coverage. One alternative was found (gRoads), but the focus in this dataset was mainly on roads between settlements and not on streets (SEDAC, 2010). An example of the dataset can be found in figure 11. It was not possible to find an example of the dataset (Eurostreets) used in the ESCAPE project, because that dataset does not seem to be open data. The input road network datasets for Bangkok and Mexico City can be found in figure 12 and 13. The classification of these roads is based on the classification of Beelen et al. (2013. The reclassification scheme can be found in appendix 1 (section 8.1).

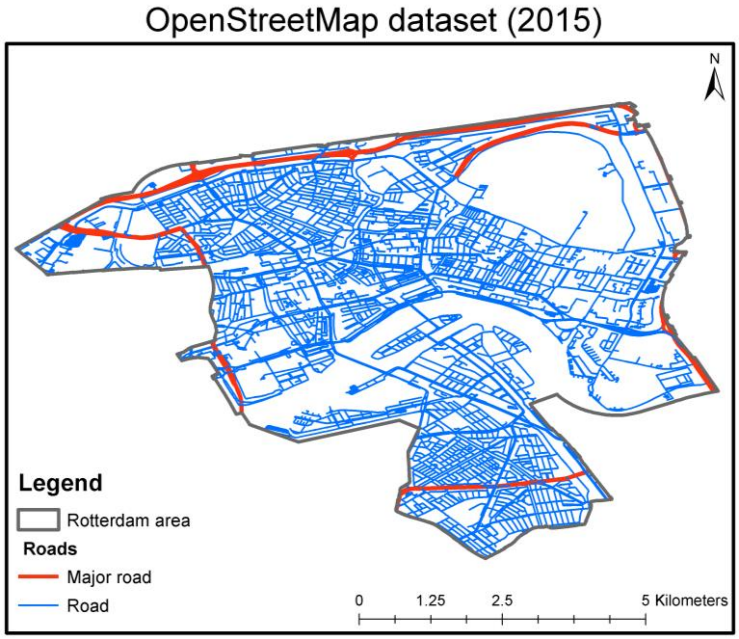


Figure 11. The OpenStreetMap dataset. The extent of the area is the study area of Rotterdam used in this research project.
Source of the data: Geofabrik (2015).

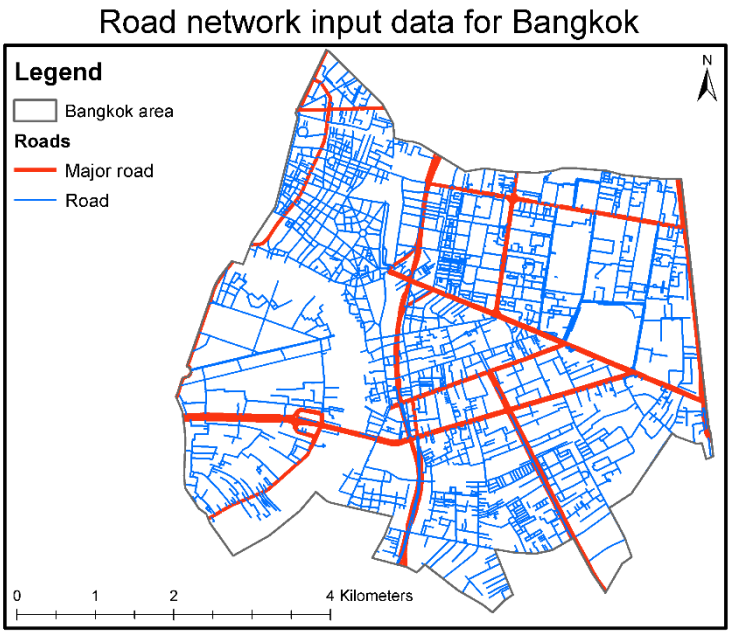


Figure 12. The OSM input data for the Bangkok area. The major roads are used for the 'i' variable and both road classes (major roads and roads) are used for the 'l' variable.

Source of the data: Geofabrik (2015) and OSM (2015).

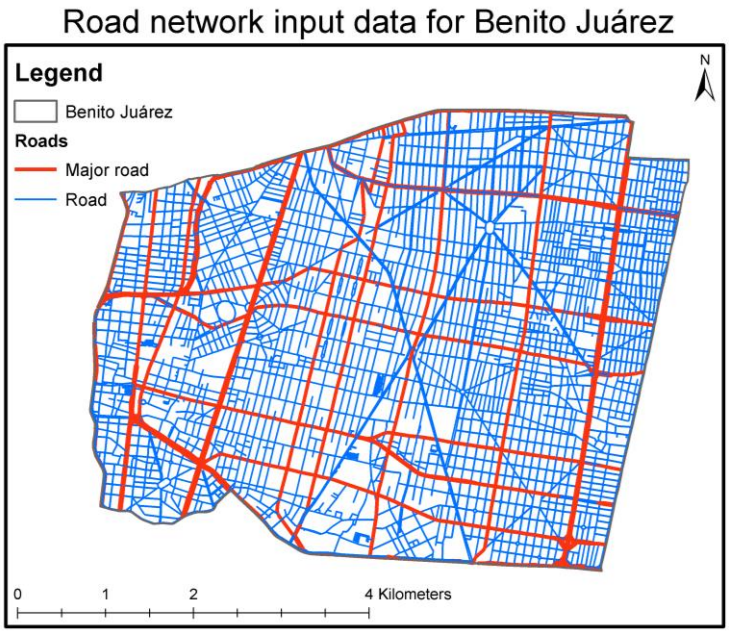


Figure 13. The OSM input data for the Benito Juárez. The major roads are used for the 'i' variable and both road classes (major roads and roads) are used for the 'l' variable.

Source of the data: Geofabrik (2015) and OSM (2015).

Although there can be some negative consequences of volunteered geographic information (VGI) this was the most preferable option, because of the global coverage of OSM. Some of the disadvantages of VGI can be quality inconsistency or the difference between places that are covered very well and places that are nearly not covered in OSM. One example given by Haklay (2010) is that the coverage is worse in deprived and rural areas. However, in England the OpenStreetMap dataset had already a 80% overlap with the dataset of the Ordnance Survey in 2008 (Haklay, 2010). Another article mentions that data provided by amateurs can also be credible (C. Liu et al., 2015).

Despite of the lower quality of OpenStreetMap it was used in this research. For a quick comparison of different cities this dataset can be convenient. When more accurate results should be achieved it is recommended to use a dataset which has a higher quality and a more consistent coverage. This can for instance be a dataset which is collected and maintained on a national level. Figure 14 shows one of the consequences using the OSM dataset. It can be clearly seen that the two diagonal roads in the centre of the background layer are not classified as a major road in the OSM dataset. This indicates that the roads are not uniformly classified.

PM2.5 values in Benito Juárez - comparison with underlying map

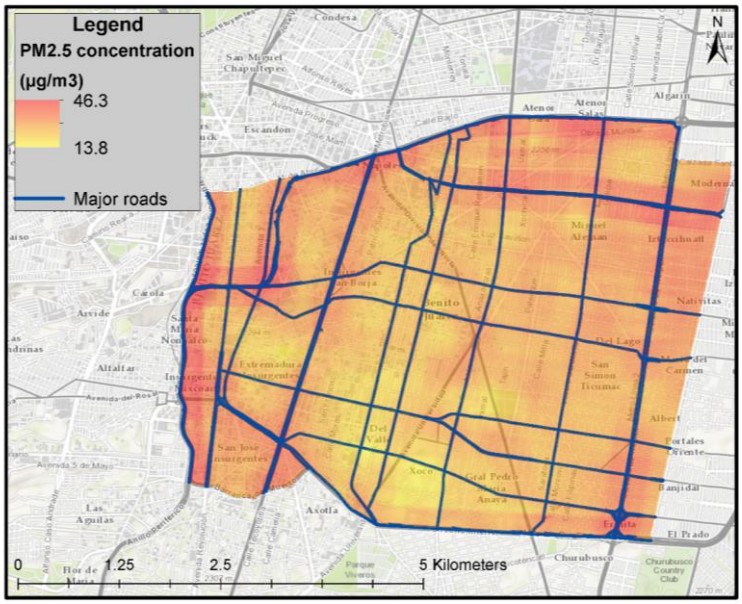


Figure 14. Comparison of the OSM road network dataset with the background layer, which is the topography service layer of ESRI.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography.

4.1.4 Population density dataset

Two datasets which are discussed in the theoretical framework are GRUMP and GPW, which stand respectively for Global Rural-Urban Mapping Project and Gridded Population of the World. The spatial resolution of the GPW is approximately 5 kilometres at the equator and the spatial resolution of the GRUMP is 927 meter (Schneider et al., 2010). This resolution is too coarse for applications, like LUR models, when calculating the personal exposure is kept in mind. Another global population dataset that is available is the World Population Estimate (WPE) (ESRI, 2015b). Figure 15 shows the dataset used in the ESCAPE project and figure 16 shows the WPE dataset for Benito Juárez. The WPE dataset is used as a weight layer to distribute the population in Bangkok and Mexico City, which was discussed in section 3.2.3.

Population density disaggregated with Corine land cover 2000

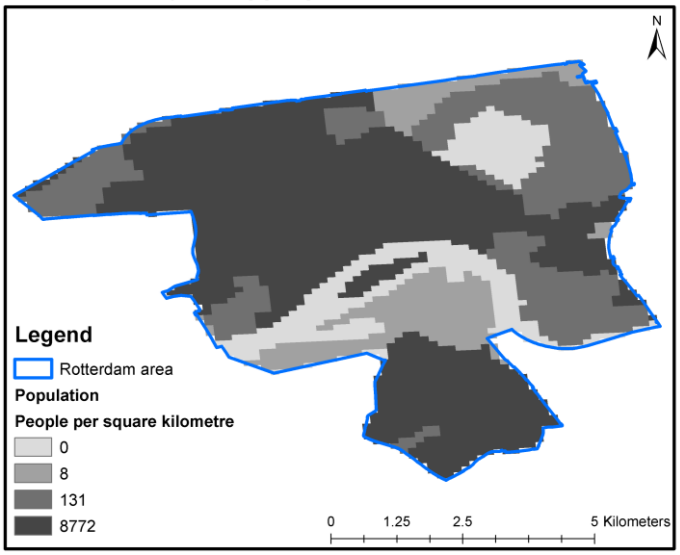


Figure 15. The population density layer used in the ESCAPE project. The spatial resolution is 100 meters. The data of this example consists of 4 classes.

Source: EEA (2009). Additional information on this dataset can be found in the article of Gallego (2010).

Population density dataset - Benito Juárez

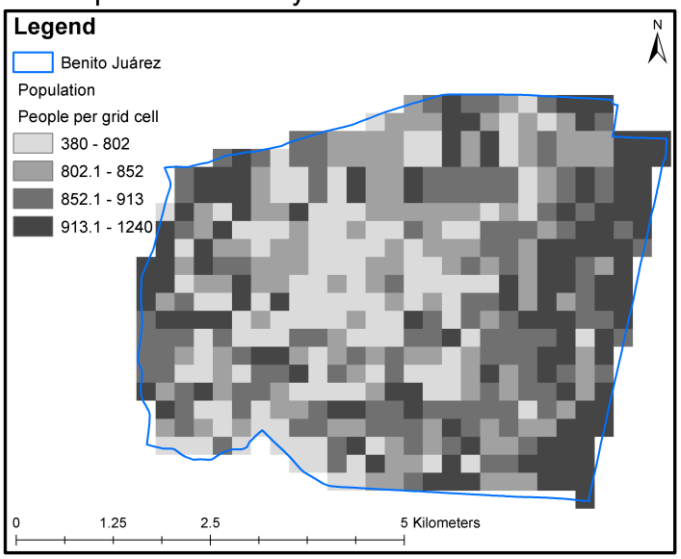


Figure 16. The World Population Estimate dataset. The spatial resolution is approximately 250 meters. The data is classified with a quantile classification, which means that each class contains an equal number of grid cells.

Source: ESRI (2015b).

The disadvantage of the WPE is that the resolution is still quite high (250 meters) this means that land use regression models which require population density as input produce an output with a spatial resolution of 250 meters. Another disadvantage is that there are quite some cells which contain the NoData value. This could however be solved by filling these cells with values based on the neighbouring cells with a neighbourhood operation. However, taken together, the WPE seems to be the best option to include in the land use regression models when population density data is required. The main reason for this is the spatial resolution which is much higher than the other datasets. Chapter five discusses how a higher spatial resolution can be achieved by combining population data with additional data, like building footprints, number of buildings floors and building volume.

4.1.5 Traffic intensity dataset

The only local dataset that is discussed here is the traffic intensity dataset. Local datasets were used in the ESCAPE project when the data was not available on European level, or were more suitable than the European datasets (Eeftens et al., 2012). In the case of this study a dataset with a global coverage of traffic intensity is not available. In fact these datasets were not even publicly available for Bangkok and Mexico City. The consequence is that assumptions needed to be made about the traffic intensity on the roads in cities which are modelled. Another consequence is that a uniform and quick comparison between cities is harder to make because a global and consistent dataset is not available.

In order to be able to compare Bangkok and Mexico City with the London LUR model a dummy dataset was created. This dummy dataset consists of all the vehicles that are registered in these cities and these vehicles are distributed over all the roads in Mexico City and Bangkok. Thus, the assumption is that all the vehicles are driving when the number of vehicles on the roads are 'counted'. This is a very basic and unrealistic assumption, but there was hardly any information available on traffic intensity for both cities. Besides, the vehicles which are registered outside the city are not taken into account. This can be compensated by assuming that all the registered vehicles within the city are driving during the 'measurement period'. Although these assumptions are very straightforward, they were needed in order to estimate the traffic intensity data. The next step was to assign a total number of cars per 24 hour period to each road segment. All major roads and roads consist of many road segments which together form the road network dataset. The road and major road classes are based on the classification of Beelen et al. (2013). The reclassification of the OpenStreetMap roads can be found in appendix 1 (section 8.1). All the major roads in a city got 75 percent of the total number of vehicles registered in the city and the roads got 25 percent of this number. Thus, in a city with 1000 registered vehicles, 750 vehicles are assigned to the major road class and 250 vehicles are assigned to the road class. Subsequently each road segment gets a proportion of the class (road or major road) it belongs to, depending on their length. For example, a major road segment has a length of 100 meters. The total length of all major road segments is 5000 meters. This means that the length of this road segment is two percent of the total length of all major road segments. The consequence is that this road segment gets two percent of the vehicles which are assigned to the major road class. In this case this is 15 vehicles (0.02 * 750). The data on the number of vehicles in Bangkok in 2010 was derived from UN-Habitat (2013). The data on the number of vehicles in Mexico City in 2001 was derived from FIMEVIC (2001). More current data was not available, or could not be found due to language issues. Appendix 2 (section 8.2) provides more information on how the traffic intensity data was used in the regression equations.

4.2 Outputs of the models

This section provides the outputs of the models. Figure 17 and 18 show the output results of the models for Bangkok and Mexico City at a 10 meter resolution. The classes are classified with a quantile classification, which means that all classes contain an equal number of grid cells.

PM2.5 values in the Bangkok area - including road concentrations

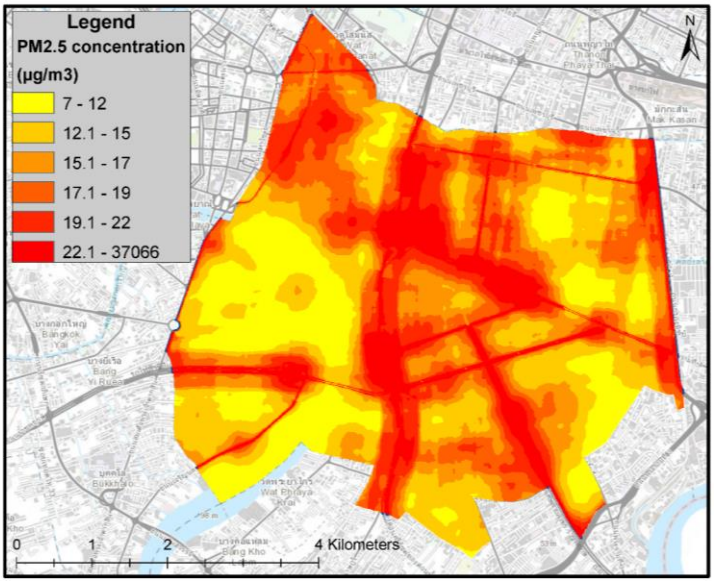


Figure 17. The $PM_{2.5}$ values in the Bangkok area, including the major road concentrations. Six classes are used which are defined by a quantile classification.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography.

PM2.5 values in Benito Juárez - including road concentrations

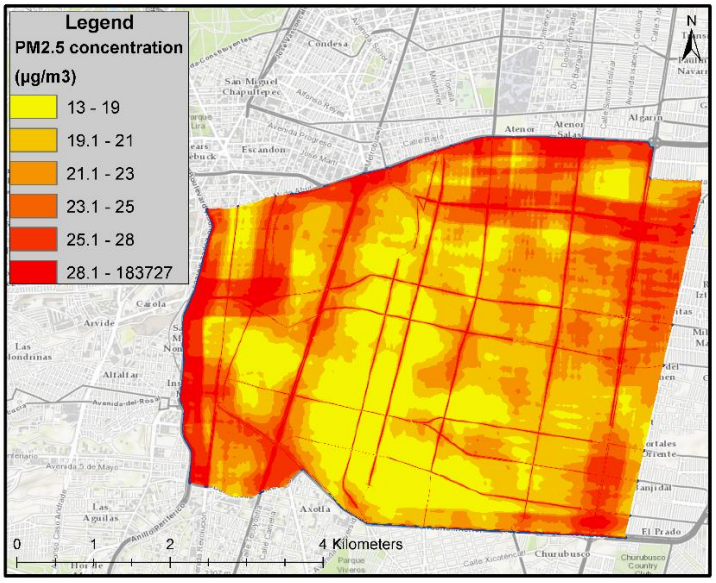


Figure 18. The $PM_{2.5}$ values in Benito Juárez, including the major road concentrations. Six classes are used which are defined by a quantile classification.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography.

Figures 19 and 20 show the $PM_{2.5}$ concentrations for the entire area without the values for the roads and a buffer of 10 meter around these roads (this is explained the methodology chapter in the first paragraph of section 3.2.2). It is interesting to see that some of the major roads seem to have higher air pollution values around it than other major roads. The cause is that some major roads are mapped with double lines and some of them are mapped with single lines. This has especially influence on the '*l*' variable which measures the total length of roads within a buffer of 500 meters.

PM2.5 values in the Bangkok area - without road concentrations

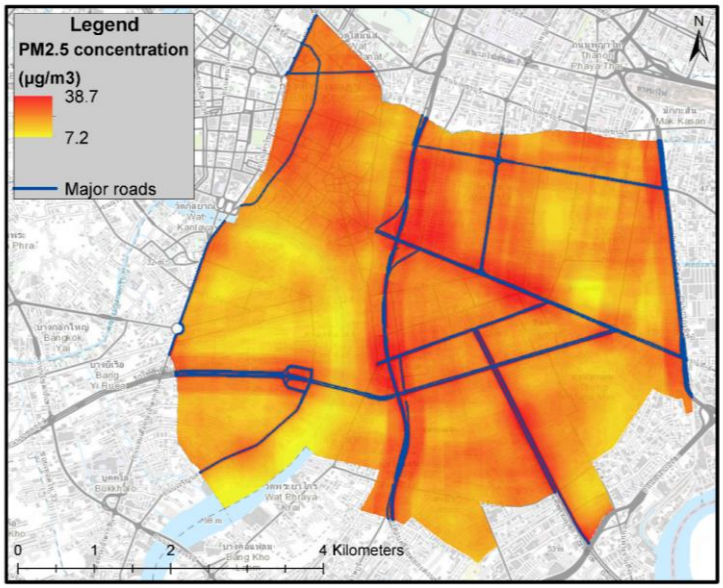


Figure 19. The $PM_{2.5}$ values in the Bangkok area. Yellow indicates the lowest values and red indicates the highest values. The major roads are drawn in blue.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography. Own sources: Geofabrik (2015), OSM (2015).

PM2.5 values in Benito Juárez - without road concentrations

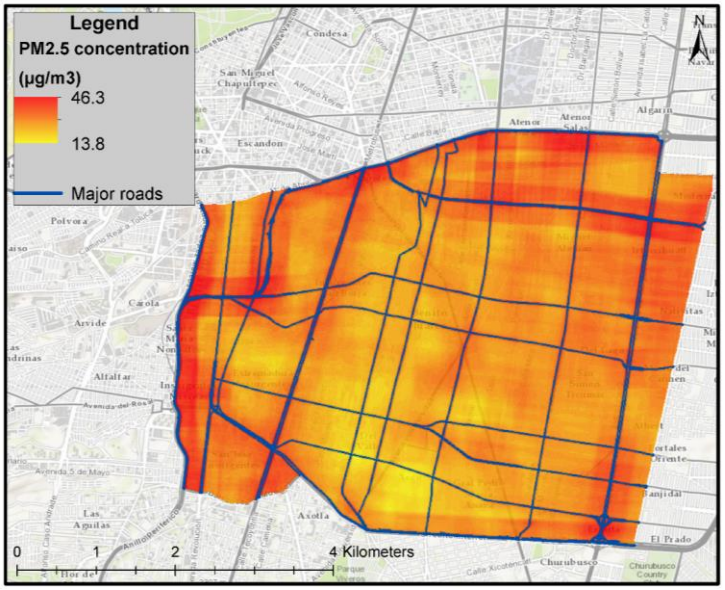


Figure 20. The $PM_{2.5}$ values in the Benito Juárez. Yellow indicates the lowest values and red indicates the highest values. The major roads are drawn in blue.

Sources: see figure 19.

Table 4 shows how much grid cells each air pollution class contains, based on the maps without the values for major roads and the buffer of 10 meters around these roads. The classification is based on the classification of the WHO, mentioned in the theoretical framework. Most grid cells of the Bangkok area have values between 10.1 and 25. Almost all the grid cells of Benito Juárez have a value above 15.1. The range of $PM_{2.5}$ values of the Bangkok area varies from 7.2 to 38.7. The $PM_{2.5}$ values of Benito Juárez range from 13.8 to 46.3. Further interpretation of the differences between the two cities can be found in section 4.3.2.

$PM_{2.5}$ ($\mu g/m^3$)	Grid cells - Bangkok area	%	Grid cells - Benito Juárez	%
7-10	21973	*8*	0	*0*
10.1-15	104441	*36*	3397	*1*
15.1-25	158200	*55*	221168	*84*
25.1-35	5184	*2*	38989	*15*
> 35	45	*0*	196	*0*
Total	**289843**	*100*	**263750**	*100*

Table 4. Number of grid cells per air pollution class. The first column shows the air pollution classes based on the classification of the WHO. The second and third column show the number and percentage of grid cells per air pollution class for Bangkok. The fourth and fifth column idem for Benito Juárez.

Figure 21 and 22 show the outputs of the Dutch $PM_{2.5}$ model applied to Rotterdam. Model A1 represents the model output with all the (city) background monitoring sites. Model A2 represents the model output with two (city) background monitoring sites left out of the analysis.

PM2.5 values in Rotterdam - model A1

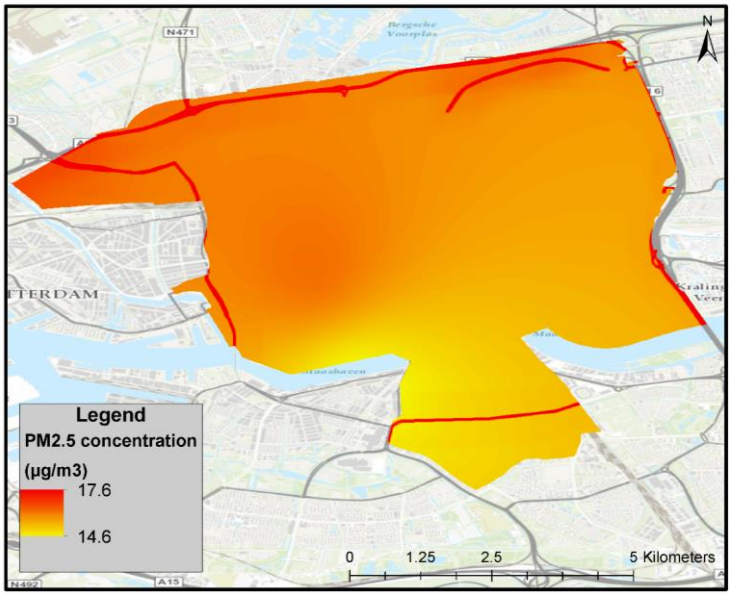


Figure 21. $PM_{2.5}$ values in Rotterdam of the A1 model. The A1 model is the model with all (city) background measuring sites.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography.

PM2.5 values in Rotterdam - model A2

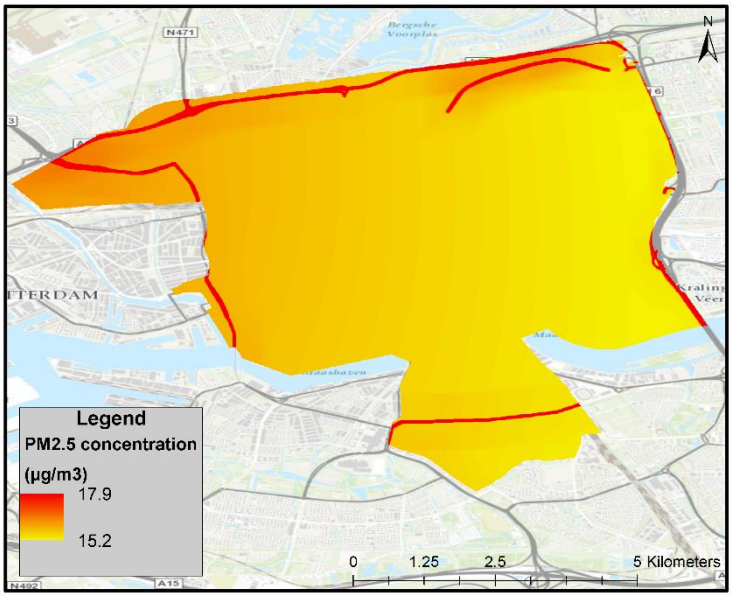


Figure 22. $PM_{2.5}$ values in Rotterdam of the A2 model. The A2 model is the model with a selection of (city) background measuring sites.

Sources: ESRI, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN. These sources are derived from the ESRI service layer and are not included in the bibliography.

4.3 Validation and sensitivity analysis of the models

This sub-chapter deals with the following sub-question: 'To what extent do model errors occur when the model is applied in a study area outside Europe and what causes these errors?' The related sub-objective consists of three parts: validation of the input data, validation of the model output and a sensitivity analysis. The first part validates the input data. The ESCAPE $PM_{2.5}$ model output was used to do this part of the validation in order to compare the output with the model outputs of this research project. The second validation part compares the model results with remote sensing data. The sensitivity analysis was done by raising the variables with 20% to measure the influence on the average model output.

4.3.1 Validation of the input data

This section describes the validation of the input data. The validation was done by using the ESCAPE $PM_{2.5}$ regression equation for Rotterdam. OpenStreetMap, the source of the road network data which was used for Bangkok and Mexico City, was also used as the source for the road network data for Rotterdam. Also the traffic intensity data for Rotterdam was created in the same way as for Bangkok and Mexico City. The source of the data that was used to model the $PM_{2.5}$ values in Bangkok and Mexico City was also used to model the $PM_{2.5}$ values in Rotterdam with the Dutch ESCAPE model. Subsequently these model outputs were compared with the original model output which was created in the ESCAPE project. The road network input data for Rotterdam can be found in section 4.1.3 (figure 11). As mentioned in the methodology chapter it was not possible to compare the input datasets, because the dataset used in the ESCAPE project was not available as open data. The consequence is that this section does only discuss the outputs of the models. As was described in the methodology chapter two models were used to compare with the official ESCAPE output for Rotterdam: one model which uses all (urban) background monitoring sites and one model which uses a selection of (urban) background monitoring sites. To improve readability the former model is called 'model A1' and the latter is called 'model A2'.

The model outputs were compared at 1226 (spatial) points. The comparison of these points was done with a correlation measure (Pearson's r). The correlation measure of model A1 is 0.3. The correlation measure model A2 is 0.4. These numbers indicate that there is a weak positive correlation between the models with 'unofficial data' and the model which uses the 'original' ESCAPE datasets. The scatter plots below (figures 23 and 24) show that there is an underestimation of the $PM_{2.5}$ values in Rotterdam in comparison with the official ESCAPE output.

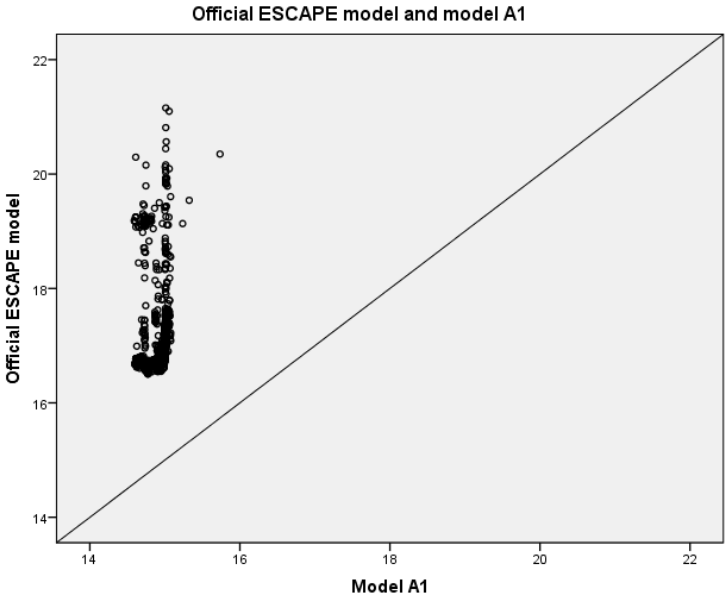


Figure 23. Scatter plot of the outputs of model A1 compared to the ESCAPE model output. The values indicate the micrograms per cubic meter. If all values would be on the line, the models would match exactly.

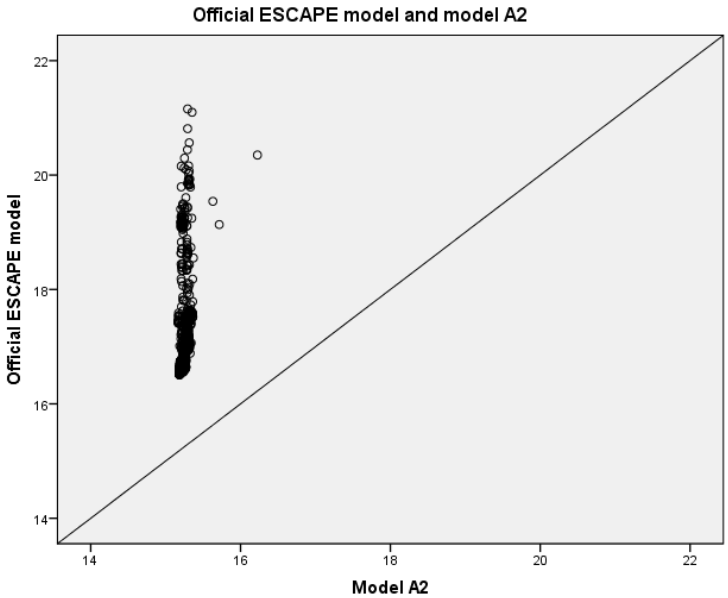


Figure 24. Scatter plot of the outputs of model A2 compared to the ESCAPE model output. The values indicate the micrograms per cubic meter. If all values would be on the line, the models would match exactly.

The absolute errors of the models can be measured with the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). These statistics were normalized in order to compare them. Table 5 shows that model A2 has slightly lower values for the RMSE, the MAE and the normalized statistics. This means that model A2 is slightly more accurate than the other model, i.e. there is less difference between the predicted and the observed values.

	Model A1	**Model A2**
Pearson's r	0.3	0.4
RMSE	2.4	2.1
NRMSE	16.4	13.7
MAE	2.3	1.9
NMAE	15.5	12.6

Table 5. Statistical measures of the input data validation.

Because the errors were squared, a higher weight is given to large errors in the RMSE. The result is that in any case the RMSE is equal to or larger than the MAE. However, when great differences occur, the individual errors have a greater variance. Because the RMSE and MAE are nearly equal for both models, the errors of the models are not large. The statistical measures mentioned above indicate that model A2 is more suitable to predict $PM_{2.5}$ values than the other model, because there is a stronger positive relationship between the two variables and there is a better model fit with more accurate predictions.

4.3.2 Validation of the model

Besides the validation of the input data, the model outputs of Bangkok and Mexico City were compared with 'external' data in order to validate the model. This means that this data was not used to develop the model. The external data for this validation consisted of remote sensing data. Table 6 shows three values for each city: the average $PM_{2.5}$ value of the model output, the average $PM_{2.5}$ value of the model output without the roads, including a buffer of 10 meters, and the $PM_{2.5}$ values of the remote sensing data. The values in the last column represent the values of the remote sensing data grid cell with the most overlap with the model output. The reason that this table also shows the values of the model without the major roads, including a 10 meter buffer, is that the values on and near roads were rather high. Besides, the model outputs are intended to use for the exposure modelling. This means that it is less relevant to know what the values are on the major roads, because people do not live there.

Table 6 shows that the six districts in Bangkok have an average $PM_{2.5}$ value of 19.48 when the roads are included in the analysis. When the roads with the 10 meter buffer are excluded, the average model output is 16.30. This is pretty much the same as the amount of microgram per cubic meter from the remote sensing imagery, which is 15.90. One very small part of the neighbourhoods of Bangkok overlaps with a remote sensing grid cell with a value of 13.70, but this overlap is negligible. The Benito Juárez borough in Mexico City has an average $PM_{2.5}$ value of 25.36, which is 8.56 microgram per cubic meter more than the remote sensing data. When the roads with the 10 meter buffer are excluded from the analysis this difference is a little bit smaller (5.16 $\mu g/m^3$). A small part of the borough overlaps with a remote sensing grid cell with a value of 16.20 which makes the difference between the model outputs and the remote sensing data a little bit larger.

	Average of model output (µg/m^3) – annual average	Average of model output, without roads (µg/m^3) – annual average	Remote sensing (µg/m^3)* – average of 2001-2010
Bangkok	19.48	16.30	15.90
Mexico City	25.36	21.96	16.80
*** This is the grid cell value with the most overlap with the model output. Source of the data: SEDAC (2012)**			

Table 6. Comparison of the model output with remote sensing data. The second column shows the average model output of the 'standard' model. The third column shows the model output without the major road $PM_{2.5}$ values and the 10 meter buffer values. The last column shows the $PM_{2.5}$ values of the remote sensing data.

Compared to the remote sensing data the London model seems to be able to predict $PM_{2.5}$ concentrations fairly well for Bangkok, but larger differences occur when the model is applied to Benito Juárez (Mexico City). The main reason for the differences between Bangkok and Mexico City is probably related to the major roads. Within the six districts of Bangkok there is a total of 93 kilometres of major roads. In Mexico City there is a total amount of 124 kilometres of major roads. Besides, the major roads in Mexico City are more evenly distributed across the surface than the major roads in Bangkok. Because one of the variables of the London model is based on the distance to the major road, this has an influence on the output map. When the roads are more evenly distributed across the surface, the distance to the nearest major road is smaller for every grid cell in the research area. This leads to higher output values, because the farther away a major road is, the lower the $PM_{2.5}$ value for a grid cell. Thus, the differences in the average model outputs between Bangkok and Mexico City seem mainly to be caused by the spatial distribution of the major roads. Besides, the equation is not developed for these specific cities and is most likely based on another distribution of major roads.

Another cause is that Benito Juárez contains a lot more kilometres of roads in comparison to the six districts of Bangkok. In Benito Juárez there is a total amount of 640 kilometres of roads and in the six districts of Bangkok there is a total amount of 428 kilometres of roads. This also has influence on the output of the models because the variable *'l'* calculates the amount of kilometres of roads within a buffer of 500 metres for every grid cell. In addition, the district of Mexico City is somewhat smaller than Bangkok, which means that Benito Juárez contains even more kilometres of roads per square kilometre. Another reason that Bangkok has lower $PM_{2.5}$ values is the river that flows through the modelled area. This means that there are less roads and less air pollution.

4.3.3 Sensitivity analysis

This section describes the sensitivity analysis. This section is meant to identify which of the variables of the London equation has most influence on the output of the model. The two variables of this model (*'i' and 'l'*) were raised with 20% to examine how much percent the output of the model would change. Table 7 provides the results of the sensitivity analysis. The second column shows the average model output of the London model. These are the results without the major roads and the 10 meter buffers around these roads. The third column shows the average model output with the 'i' variable raised with 20%. For both areas there was just a minor difference of 0.6 and 0.5 percent which is shown in the fourth column. The fifth column shows the output of the model with the 'l' variable raised with 20 percent. Changing this variable with 20 percent has more influences than changing the 'i', because the sixth column shows an increase of 11.2 and 13.5 percent in the average model output. The consequences of the differences between the two variables are discussed in chapter 5.

City	Average model output	Average model output with variable 'i' + 20%	*Difference*	*Difference (%)*	Average model output with variable 'l' +20 %	*Difference*	*Difference (%)*
Bangkok areas	16.3	16.4	*0.1*	*0.6*	18.1	*1.8*	*11.2*
Benito Juárez	22.0	22.1	*0.11*	*0.5*	24.9	*2.9*	*13.5*

Table 7. Sensitivity analysis of the variables. The second column shows the average model output of the model without the major road values and the 10 meter buffer values. The third and fourth column show the difference between the model outputs with variable 'i' raised with 20% and the model outputs of the second column. The fifth and sixth column show the difference between the model outputs with variable 'l' raised with 20% and the model outputs of the second column.

Also a sensitivity analysis was done for multiple locations in both cities. This is done in addition to the sensitivity analysis mentioned above, because that analysis only took average model outputs into account. Table 8 shows the results of this sensitivity analysis. This table shows that eight locations were analyzed. For each city two locations were analyzed at 50 meters of a major road and two locations at 200 meters of a major road. These locations were randomly picked and are perpendicular to the major roads. The 200 meter locations lie in one line with the 50 meter locations. The numbers in the first column show which pairs belong together. The differences in $PM_{2.5}$ between the model output without the major roads and the ten meter buffer (third column) and the model output with variable 'i' raised with 20% can be found in the fifth column. These differences are very small, as the largest difference is just 1%. The differences between the model output and the model output with variable 'l' raised with 20% are larger. These differences range from 4.7 to 11.5%. The differences between the influence of the two variables correspond with the differences of the variables presented in table 7.

City	Location	Model output	Variable '*i*' raised with 20%	*Difference (%)*	Variable '*l*' raised with 20%	*Difference (%)*
Bangkok 50m (1)	100°32'28.864"E 13°43'41.352"N	13.8	13.8	*0.0*	15.0	*8.7*
Bangkok 200m (1)	100°32'30.982"E 13°43'45.716"N	9.4	9.4	*0.0*	9.84	*4.7*
Bangkok 50m (2)	100°30'48.318"E 13°43'6.319"N	13.9	13.9	*0.0*	15.1	*8.6*
Bangkok 200m (2)	100°30'46.676"E 13°43'1.733"N	12.3	12.3	*0.0*	13.3	*8.1*
Benito Juárez 50m (3)	99°8'48.643"W 19°23'33.206"N	19.1	19.3	*1.0*	21.3	*11.5*
Benito Juárez 200 m (3)	99°8'53.486"W 19°23'33.549"N	17.4	17.4	*0.0*	19.4	*11.5*
Benito Juárez 50m (4)	99°8'42.602"W 19°23'54.976"N	22.5	22.7	*0.9*	25.3	*12.4*
Benito Juárez 200m (4)	99°8'37.802"W 19°23'54.617"N	20.1	20.1	*0.0*	22.6	*12.4*

Table 8. Sensitivity analysis of the variables at eight locations. The first two columns show the locations. The third column shows the model output on this location. The fourth and the fifth column show the difference at the locations between the normal output and variable 'i' raised with 20%. Column six and seven show the difference at the randomly picked locations between the normal output and variable 'l' raised with 20%.

4.4 Exposure modelling

This section presents the results of the third sub-objective of this research which deals with the exposure of the population. The sub-question related to this sub-objective is: 'To what extent is it possible to model the personal exposure of the population to air pollution?' Figures 25 and 26 show the maps with the personal exposure of the population. The classes are classified with a quantile classification, which means that each class contains an equal number of people. For a better visibility just one percent of the population is shown in this map.

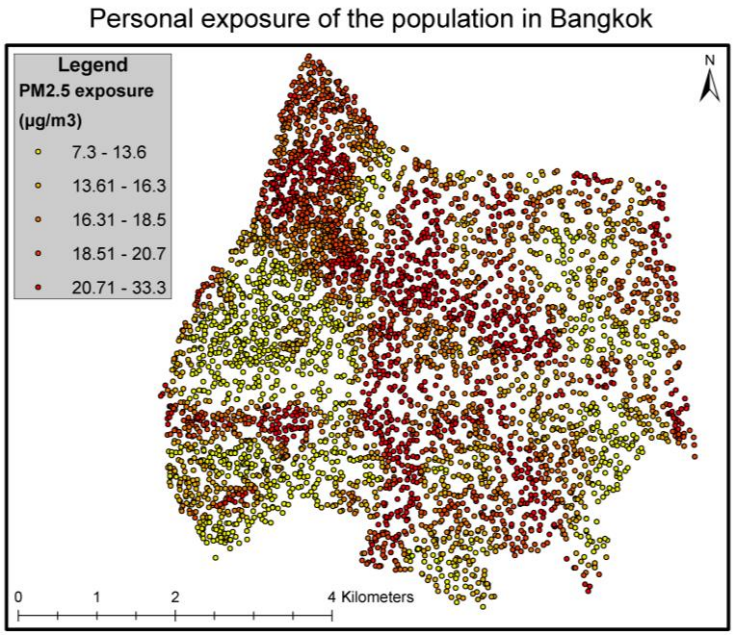


Figure 25. The personal exposure of the population in Bangkok. There a five classes which are based on a quantile classification. For a better visibility just 1% of the population is shown on this map.

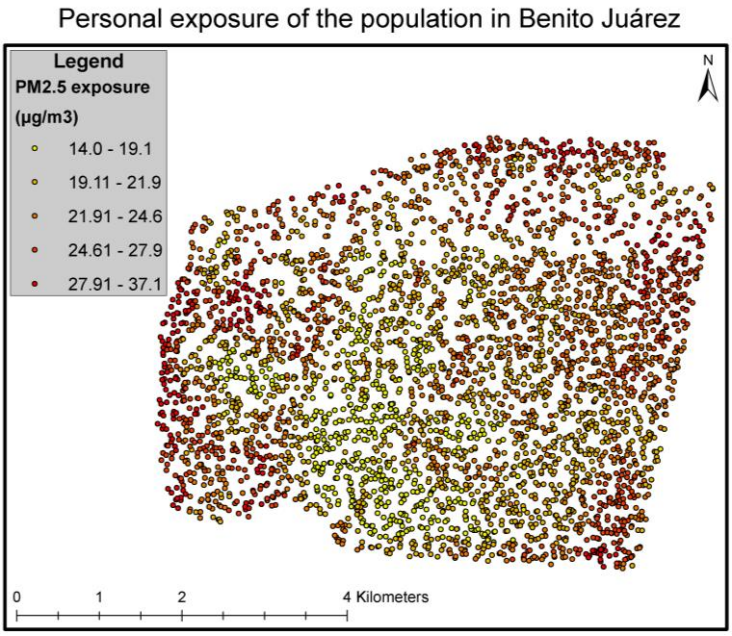


Figure 26. The personal exposure of the population in Benito Juárez. There a five classes which are based on a quantile classification. For a better visibility just 1% of the population is shown on this map.

Table 9 shows the exposure of the population to $PM_{2.5}$. The results show that just a small part of the population of the 6 modelled districts in Bangkok is exposed to $PM_{2.5}$ values below the threshold value of the Air Quality Guideline (10 µg/m^3). In Benito Juárez, the borough in Mexico City the whole population is exposed to $PM_{2.5}$ values above 10 microgram per cubic meter. The table shows that for Bangkok and Mexico City the largest part of the population (resp. 65% and 76%) is exposed to PM2.5 values between 15.1 and 25. This is partly due to the size of this range in comparison with the range of interim target 3. For Bangkok just a small part of the population belongs to the interim target 1 class (4.5%). People in this class have a 15% higher long-term mortality risk in comparison with the lowest class (Air Quality Guideline) according to the WHO report (World Health Organization, 2006). In the district of Benito Juárez a considerable number of people are exposed to values above 25 µg/m^3. It is remarkable that almost the whole population of Benito Juárez is exposed to values above 15 microgram per cubic meter. The explanations for the large differences between the two cities can be found in section 4.3.2. Besides the distribution of the roads, another explanation could be the distribution of the population, which is discussed below. Both modelled areas have 3 more people in the statistics than the population figures mentioned in section 3.2.3. This is probably a small error in the 'create random points' tool in ArcGIS.

	$PM_{2.5}$ ($\mu g/m^3$)	Bangkok districts (number of people)	%	Benito Juárez (number of people)	%
Interim target 1	> 25	21896	*4.5*	90386	*23.4*
Interim target 2	15.1-25	320043	*65.3*	293506	*76.1*
Interim target 3	10.1-15	132031	*27.0*	1550	*0.4*
Air quality guideline	0-10	15926	*3.3*	0	*0.0*
Total		**489896**	***100***	**385442**	***100***

*** This class ranges from 25.1 to 35 in the WHO report, but the model outputs contain values higher than 35.**

Table 9. The exposure of the population to $PM_{2.5}$. The first and second column provide the air pollution classes as defined by the WHO (2006). The other columns provide information on the number of people exposed to certain values of air pollution.

The differences between the average model output and the average personal exposure can be found in table 10. For both cities the average personal exposure is higher than the average model output. This indicates that the population is more concentrated at locations with higher air pollution values. One of the explanations for this is the distribution method of the population. This distribution is done with the World Population Estimate (WPE). The WPE uses several factors to determine where people are located, like mentioned in the theoretical framework. Two of these factors are the location of roads and road intersections. According to the methodology of the WPE the presence of roads and road intersections indicates a higher population density. For this study it means that more people were assigned to locations in the proximity of roads. These are the locations with higher air pollution values and this is an explanation for the difference between the average model output and the average personal exposure.

	Average model output	Average personal exposure
Bangkok districts	16.3	17.4
Benito Juárez	22.0	22.6

Table 10. The differences between the average model output and the average personal exposure of the population. The average model output is the version without values on major roads and a buffer of 10 meters around these major roads.

Also the differences between exposure values at 50 metres and 200 metres from a major road were calculated. This was done to find out the influence of a location on the exposure of the population. The locations which were analyzed are the same as the locations used for the sensitivity analysis. Table 11 shows that the absolute differences between the locations at 50 metres of a major road and 200 meters of a major road range from 1.6 to 4.4 $\mu g/m^3$. The table also shows that the relative differences range from 8.9 to 31.9 percent. Three comparisons between the 50 and 200 meter points show a decrease of around ten percent and one comparison shows a considerable larger decrease of more than 30%.

City	Location of measurement	Model output (µg/m^3)	*Difference*	*Difference (%)*
Bangkok 50m (1)	100°32'28.864"E 13°43'41.352"N	13.8	***4.4***	***31.9 %***
Bangkok 200m (1)	100°32'30.982"E 13°43'45.716"N	9.4		
Bangkok 50m (2)	100°30'48.318"E 13°43'6.319"N	13.9	***1.6***	***11.5 %***
Bangkok 200m (2)	100°30'46.676"E 13°43'1.733"N	12.3		
Benito Juárez 50m (3)	99°8'48.643"W 19°23'33.206"N	19.1	***1.7***	***8.9 %***
Benito Juárez 200 m (3)	99°8'53.486"W 19°23'33.549"N	17.4		
Benito Juárez 50m (4)	99°8'42.602"W 19°23'54.976"N	22.5	***2.4***	***10.7 %***
Benito Juárez 200m (4)	99°8'37.802"W 19°23'54.617"N	20.1		

Table 11. Model outputs measured at 50 and 200 meters of a major road. The first two columns show the location of the points where the output values were measured. The numbers indicate which pairs belong together. The third column shows the $PM_{2.5}$ values. The last two columns show the difference between the 50 and 200 meter locations.

5. Discussion, limitations and recommendations

This chapter discusses the results of this research. This is done to give an answer to the research questions provided in chapter 1. The main activity of this research was to apply the London LUR model in Bangkok and Mexico City in order to research the possibilities of applying land use regression models in cities which they were not developed for. Sub-activities were discussing the availability of the data, validating the model and data, sensitivity analysis and modelling the exposure of the population. This chapter starts with discussing and interpreting the results in order to find an answer to the research questions. The results are also compared with previous findings from the literature. The last two sections of this chapter deal with the limitations of this research and recommendations for future research.

5.1 Discussion of the results

This section discusses the sub-questions in order to find an answer to the main question. The answer to the main question can be found in the next chapter. The main question of this research is: 'To what extent is it possible to model air pollution concentrations in cities outside Europe using the London ESCAPE LUR model, how valid is the model and to what extent can the personal exposure of the population to air pollution be modelled?'

5.1.1 Input data

The first sub-question that is dealt with is: 'Which input data can be used to compare different cities outside Europe using the model?' The results of sub-question 1, based on a search on the internet and in scientific literature, show that using global datasets can sometimes be problematic. Especially the land cover dataset and population density dataset have a number of disadvantages in comparison with the datasets used in the ESCAPE project. This means that additional work is needed in order to use them with the land use regression models. The most problematic issue of the land cover dataset is that this dataset does not make a distinction between high and low density residential land use. This distinction is however sometimes required in the land use regression models. A solution could be to make this distinction with additional data, like satellite imagery of city lights. The main disadvantage of the World Population Estimate, the global population density layer, is that the spatial resolution is lower (250m) compared to the dataset used in the ESCAPE project (100m). A solution to obtain a more accurate dataset is provided in section 5.1.3. The elevation dataset (SRTM) has a global coverage and was also used in the ESCAPE project, which means that this dataset can be used without any problems. The spatial resolution which is available now is even higher than the spatial resolution of the dataset used in the ESCAPE project. OpenStreetMap can be used as the road network dataset, but the quality can differ per area. Also the classification of the roads is not always uniform, which can partly be explained by the fact that anyone can edit and add information to OSM (C. Liu et al., 2015). Obtaining a traffic intensity dataset can be very problematic. It is not available with a global coverage and even on a national or regional scale a dataset was not available for Bangkok and Mexico City. This means that additional work needs to be done, like making assumptions on the traffic intensity or monitoring the traffic.

5.1.2 Model errors

The second sub-question is: 'To what extent do model errors occur when the model is applied in a study area outside Europe and what causes these errors?' The validation of the input data showed a weak positive correlation between the official ESCAPE data and the data used in this research project. In comparison with the ESCAPE results, the data used in this project underestimates the $PM_{2.5}$ values in Rotterdam. The differences between the RMSE and the MAE indicate that there were not many large errors in the predictions. This indicates that the data used in this research constantly underestimates the air pollution in Rotterdam.

Besides the validation of the input data also the model was validated against independent observations of air pollution. This is a validation of the average model output. Validation at a higher spatial resolution was not possible, due to a lack of data from ground monitoring sites. Validation of the average model output means that the within-city variability of air pollution values could not be validated. The validation of the model showed that the London model predicted the $PM_{2.5}$ values better for the Bangkok area than Benito Juárez. This could be caused by the differences in geographical structure of the cities, especially the total length of roads and the distribution of the roads across the surface. It is possible that the modelled areas of Bangkok have a similar spatial structure as London, which is one of the conditions for a successful model transfer (Poplawski et al., 2009). Also Jerrett et al. (2005) mention that a similar geographic structure is important to achieve a successful model transfer. Although it is hard to say whether the Bangkok area has a similar geographic structure as London, it is clear that the geographic structure of Bangkok and Benito Juárez has a considerable impact on the output results of the model. On the other hand, the spatial resolution of the remote sensing data is limited (approximately 10 kilometres at the equator), which was also stated in the theoretical framework (Hystad et al., 2011). This means that validation with remote sensing data gives a rather rough estimation of the $PM_{2.5}$ values in an area. The difference between the Bangkok area and Benito Juárez supports the argument that local calibration can improve the predictive capability of land use regression models (Allen et al., 2011). By using local calibration characteristics of cities can be taken into account which can improve the predictive capability of the models. However, when local calibration is not done, the 'city-specificity' of a land use regression model should always be kept in mind. The reason is that for any LUR model there is a consideration between specific variables which may or may not be taken into account and the transferability of a model (Jerrett et al., 2007). It is likely that the London model is created specifically for London, without having transferability of this model in mind, because that was not one of the objectives of the ESCAPE project. The consequence is that applying this model to other cities will lead to worse predictions of the air pollution.

The validation of the model and the validation of the input data show that both of the errors mentioned by Vienneau et al. (2010) occurred to a certain extent when the model of London was transferred to Bangkok and Mexico City. The validation of the model shows that, especially for Benito Juárez, other factors also influence the air pollution which are not taken into account in the London regression equation. The validation of the input data shows that the input data is to such an extent different that it leads to an underestimation of the air pollution in Rotterdam.

The sensitivity analysis showed that the traffic intensity variable ('*i*') has less influence on the model output than the variable which measures the total road length within a buffer of 500 meters. This means that errors in the traffic intensity dataset have less influence on the model output than errors in the road network dataset. However, this finding applies to the regression equation of London and can be different for other regression equations of the ESCAPE project. The fact that the traffic intensity variable has less influence on the model output than the roads variable is in this research project advantageous for the predictive capacity in Bangkok and Mexico City. The reason is that the quality of the input data also influences the model output (Ryan & LeMasters, 2007). Because the traffic intensity data is based on rather basic assumptions, it is positive that this variable does not have much influence on the output of the model. The fact that the '*l*' variable has quite some influence on the model output means that it is important to use correct road network datasets. This will lead to a more accurate estimation of the air pollution.

5.1.3 Exposure modelling

The third sub-question is: 'To what extent is it possible to model the personal exposure of the population to air pollution?' The results of the third sub-question have shown that $PM_{2.5}$ values for any location in the modelled areas can be estimated, however the within-city variability could not be validated, which was discussed in the previous section. The model outputs of Bangkok and Mexico

City were combined with the randomly distributed population in order to estimate the exposure for each person. The population was classified according to the classification of the WHO (2006). This means that it is possible to estimate the exposure of the population to air pollution, however there were some uncertainties in the distribution of the population. The random distribution did not take household size into account and also the population was not excluded from locations, like rivers or parks. This means that there is quite a difference between the actual distribution of the population and the distribution used in this research project.

The degree of uncertainty is hard to estimate, but the actual distribution of the population is probably more concentrated, due to people living in apartments and because the inhabitants do not live on locations, like roads and rivers. The results have shown that the location of the population certainly matters in determining the personal exposure. The difference between a location at 50 meters of a major road and a location at 200 meters of a major road can cause differences in air pollution values up to 4.4 $\mu g/m^3$. This means that an inaccurate distribution of the population causes differences in the (average) personal exposure of the population. To get a more accurate distribution of the population the population numbers should be combined with building footprints, building volume and number of building floors. The methodology of Lwin and Murayama (2009) would be a suitable approach to distribute the population in a more accurate way. But, the ESCAPE project combined health data of individuals with air pollution concentrations measured at their home addresses (Cyrys et al., 2012; Eeftens et al., 2012). This should theoretically be possible with the data and results of this research project, because the models can be expanded to entire cities.

Besides the uncertainty of the population distribution there are also uncertainties related to the air pollution map. This uncertainty could be caused by the fact that the $PM_{2.5}$ model was originally developed for London. City-specific factors of Bangkok and Mexico City which explain $PM_{2.5}$ variations can be lacking in this model. Local calibration is a suitable methodology to solve this issue (Allen et al., 2011). Moreover, the uncertainty could not be examined because ground monitoring stations data was not available for these areas. Besides, the sensitivity analysis has shown that the model is especially sensitive for differences in the *'I'* variable, which is the total road length of all roads within 500 meters. An increase of 20% of this variable led to an increase of the model outputs (both for the average model output and the selected spatial locations in both cities) of more than 10%, which is up to 3 $\mu g/m^3$. With this information it is still hard to say what the degree of uncertainty is, but it shows that it is important to have a complete and up-to-date dataset. The results have in any case shown that both, the distribution of the population and the air pollution maps, can cause wrong estimates of the personal exposure of the population. This can be up to 4.4 microgram per cubic meter for a difference in location of 150 meters and 3 microgram per cubic meter when the '*I*' variable is raised with 20%.

5.2 Research limitations

One major limitation of this research is some of the data which was not available. This applies in particular to the traffic intensity data and the assumption made about this data is very basic. Besides, the data used to estimate the number of vehicles within the cities was based on data from different years.

Another limitation is that the models are not dynamic, because they only provide an annual average of the $PM_{2.5}$ values in a city. This means that influences of the weather, the time of the year and the time of the day are not taken into account. However, this limitation does not make the models useless for the objective of estimating air pollution values at home addresses of individuals. But it makes the models less suitable for objectives which require air pollution values at a certain time of the year or day.

The validation of the model is also a limitation in this research. The lack of monitoring sites within both areas (Bangkok and Benito Juárez) made it impossible to validate the model on specific locations, for example in the proximity of certain roads. Therefore a validation with remote sensing data was done which was only possible with the averages of the model outputs.

A limitation of the exposure modelling is the distribution of the population. The distribution of the population was not very accurate because a random distribution was used. Also the population was not excluded from unsuitable residential locations, like roads and rivers. This leads to a more scattered distribution than the actual population distribution. Another weak factor of this methodology is that households were not taken into account. The consequence is that some personal exposure values are not correct, for instance for people which were assigned a location on a major road. Section 5.1.3 provides more information on how the distribution methodology could be improved by taking building footprints into account. This methodology can be used for further research to get a more accurate population distribution.

5.3 Recommendations

Based on the limitations of this research, mentioned in the previous section, this section will provide recommendations for further research. The first recommendation is that in further research improved methodologies should be used to estimate traffic intensity data. Another option is to use existing data on traffic intensity. This would make the air pollution estimations more accurate. However, the amount of time that needs to be spend on using other methodologies to estimate traffic intensity should be weighted against the influence of the traffic intensity on the output of the regression equation.

Another recommendation is to use a more extensive validation approach in further research. One of the options could be to apply a land use regression model in an area or city with several air pollution ground monitoring sites. This makes it possible to validate the LUR model more accurately than done in this research. With an improved validation methodology it is possible to identify differences between cities on street level. This gives valuable information on how to improve the transferability of land use regression models.

One recommendation, which is not based on a limitation mentioned in the research limitations section, is to apply the models to entire cities. The areas modelled in this research project are rather small in comparison with the entirety of Bangkok and Mexico City. Because the London regression equation is developed for an entire city, it would be more suitable to apply the model to another entire city, instead of a small part of it.

For future research it would also be useful to research what is needed to calibrate the models to local conditions. It would be interesting to know how much time and effort this would cost and what the results are of calibrating the models. This information could be used to weigh up locally calibrated models against models without local calibration.

6. Conclusion

This study was done to explore the possibilities of applying land use regression models to cities outside Europe to estimate air pollution concentrations and examine the exposure of the population. The reason to study this is that using one LUR model with ubiquitous datasets for several cities makes it easy to compare cities with each other and gives the possibility to estimate air pollution in a quick way. This should make it possible to estimate air pollution concentrations with few costs and without having to install air pollution samplers, which can be time consuming when cities abroad have to be modelled. This study examined the following question, based on the issues described in this paragraph: 'To what extent is it possible to model air pollution concentrations in cities outside Europe using the London ESCAPE LUR model, how valid is the model and to what extent can the personal exposure of the population to air pollution be modelled?'

The main question is divided in three sub-questions. The first sub-question deals with the input data and is formulated as follows: which input data can be used to compare different cities outside Europe using the model? The second sub-question deals with the model errors which are explored by validation of the model, validation of the input data and a sensitivity analysis. The second sub-question is formulated as follows: to what extent do model errors occur when the model is applied in a study area outside Europe and what causes these errors? The last sub-question is meant to research to which extent the personal exposure of the population can be modelled. The last sub-question is: to what extent is it possible to model the personal exposure of the population to air pollution?

The findings of the first sub-question show that it is dependent on the input requirements of the regression equation whether the equation can be calculated. Regression equations which require elevation data and road data can be calculated without much extra pre-processing, with respectively the Shuttle Rader Topography Mission dataset and the OpenStreetMap dataset. On the other hand, regression equations which require population density data and land cover data require additional activities to make them suitable for the LUR models. The World Population Estimate can be used for the global population density data, but a more accurate dataset should be created. This can for instance be done by combining this data with building footprints, building volumes and number of floors per building. The GlobeLand30 dataset can be used for the global land cover data. However, this dataset does not make a distinction between high and low residential land use, which requires additional work to obtain this data. The traffic intensity dataset is the most problematic dataset to obtain. This dataset was not available on a global or national level. If this dataset is not available on a local level either, assumptions need to be made about the number of cars on the roads.

The findings of the second sub-question show that model errors can partly be explained by the use of the input data. The validation of the input data showed a weak positive correlation between the ESCAPE output and the outputs of this research project. However, the model outputs of this research project underestimated the $PM_{2.5}$ values in Rotterdam. The absolute errors showed that there were not many large errors in the predictions. This implies that there is a constant underestimation of $PM_{2.5}$ values in Rotterdam.

The model errors are also caused by the model itself and not just by the input data. The validation of the models showed that there is a reasonably good match between the remote sensing data grid cells and the average model outputs. However, the differences between Bangkok and Mexico City also show that local calibration would be a suitable solution to take into account city-specific characteristics, like the distribution of the roads. This will lead to smaller prediction errors.

The sensitivity analysis showed that in the London regression equation especially the total road kilometres within a buffer of 500 meters has a substantial influence on the model output. The traffic intensity has just little influence on the model output. This means that it is particularly important to have a road network dataset which is accurately and unambiguously mapped in order to get a representation which is as realistic as possible. On the other hand, it is less important to have accurate traffic intensity data. Especially by using a correct road network dataset it is possible to obtain smaller prediction errors.

The results of the third sub-question have shown that the personal exposure of the population in Bangkok and Mexico City can be estimated with the regression equation. However, the weighting layer used for the random distribution still has a rather low spatial resolution (250 meters). On the contrary, the results show that it is possible to estimate the $PM_{2.5}$ value for any location in a research area. This makes it possible to estimate the air pollution at home addresses of individuals. This data could be used in epidemiological research.

The research has shown that it is possible to model air pollution concentrations in cities outside Europe with a spatial resolution of 10 metres. The validity of the models seems to depend a lot on the distribution of roads in a city. The distribution of the population used in this research can be suitable for a quick comparison between cities, but could be improved by taking more factors into account, like the locations of residential buildings. The main conclusion is that a model can be created in a rather quick way, with acceptable results, without the need of measuring air pollution with ground monitoring stations.

7. Bibliography

Allen, R. W., Amram, O., Wheeler, A. J., & Brauer, M. (2011). The transferability of NO and NO2 land use regression models between cities and pollutants. *Atmospheric Environment*, *45*(2), 369–378. doi:10.1016/j.atmosenv.2010.10.002

Arino, O., Gross, D., Ranera, F., Bourg, L., Leroy, M., Bicheron, P., … Weber, J.-L. (2007). GlobCover: ESA service for Global Land Cover from MERIS. *Geoscience and Remote Sensing Symposium*, 2412–2415. doi:10.1109/IGARSS.2007.4423328

Beckerman, B. S., Jerrett, M., Serre, M., Martin, R. V, Lee, S.-J., van Donkelaar, A., … Burnett, R. T. (2013). A hybrid approach to estimating national scale spatiotemporal variability of PM2.5 in the contiguous United States. *Environmental Science & Technology*, *47*, 7233–7241. doi:10.1021/es400039u

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., … 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*(2), 10–23. doi:10.1016/j.atmosenv.2013.02.037

Brauer, M., Hoek, G., Vliet, P. Van, Meliefste, K., Fischer, P., Heinrich, J., … Brunekreef, B. (2003). Estimating Long-Term Average Particulate Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems. *Epidemiology*, *14*(2), 228–239.

Briggs, D. J., De Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., & Smallbone, K. (2000). A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Science of the Total Environment*, *253*(1-3), 151–167. doi:10.1016/S0048-9697(00)00429-0

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *Lancet*, *360*(9341), 1233–1242. doi:10.1016/S0140-6736(02)11274-8

CBS, Central Bureau of Statistics (2015). Wijk- en buurtkaart 2014. http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2015/wijk-en-buurtkaart-2014-art.htm. Accessed on 17-2-2016.

Citypopulation (2010). Krung Thep Mahanakhon (Bangkok). http://www.citypopulation.de/php/thailand-admin.php?adm1id=10. Accessed on 1-2-2016.

Cyrys, J., Eeftens, M., Heinrich, J., Ampe, C., Armengaud, A., Beelen, R., … Hoek, G. (2012). Variation of NO2 and NOx concentrations between and within 36 European study areas: Results from the ESCAPE study. *Atmospheric Environment*, *62*(2), 374–390. doi:10.1016/j.atmosenv.2012.07.080

Donkelaar, A. van, Martin, R. V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., & Villeneuve, P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environmental Health Perspectives*, *118*(6), 847–855. doi:10.1289/ehp.0901623

EEA, European Environment Agency (2009). Population density disaggregated with Corine land cover 2000. http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-2. Accessed on 20-2-2016.

EEA, European Environment Agency (2010). Corine land cover 2006 raster data. http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster. Accessed on 16-2-2016.

EIONET, European Topic Centre on Spatial Information and Analysis (2000). Corine Land Cover

classes. http://sia.eionet.europa.eu/CLC2000/classes/index_html. Accessed on 22-1-2016.

Eeftens, M., Beelen, R., Hoogh, K. De, Bellander, T., Cesaroni, G., Cirach, M., … Hoek, G. (2012). Development of Land Use Regression Models for PM 2 . 5 , PM 2 . 5 Absorbance , PM 10 and PM coarse in 20 European Study Areas ; Results of the ESCAPE Project. *Environmental Science and Technology*, *46*(20), 11195–11205.

ESCAPE (2013). ESCAPE Manuals. http://www.escapeproject.eu/manuals/index.php. Accessed on 21-10-2015

ESOC. Empirical Studies of Conflict (2010). Administrative boundaries of Mexico. https://esoc.princeton.edu/file-type/gis-data#Mexico. Accessed on 1-2-2016.

ESRI (2015a). Bangkok population. http://www.arcgis.com/home/item.html?id=acc1dd91b0b64030a0f9fabced8a92a9. Accessed on 1-2-2016.

ESRI (2015b). World Population Estimate. http://www.arcgis.com/home/item.html?id=ac0401d78fa24a10a9151ffe50f35afe. Accessed on 15-2-2016.

ESRI (2015c). Map Gives New Insights into Global Population. http://www.esri.com/esri-news/arcnews/summer15articles/map-gives-new-insights-into-global-population. Accessed on 15-2-2016.

FIMEVIC, Fideicomiso para el Mejoramiento de las Vías de Comunicación del Distrito Federal (2001). Diagnosis of mobility of people in Mexico City. http://www.fimevic.df.gob.mx/problemas/1diagnostico.htm. Accessed on 13-2-2016.

Gallego, F. J. (2010). A population density grid of the European Union. *Population and Environment*, *31*(6), 460–473. doi:10.1007/s11111-010-0108-y

Geofabrik (2015). OpenStreetMap data extracts. http://download.geofabrik.de/. Accessed on 2-2-2016.

Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montréal, Canada, with a land-use regression model. *Journal of the Air & Waste Management Association (1995)*, *55*(March 2015), 1059–1063. doi:10.1080/10473289.2005.10464708

GlobeLand30 (2010). Global land cover. http://www.globallandcover.com/GLC30Download/index.aspx. Accessed on 16-2-2016.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. doi:10.1068/b35097

Henderson, S. B., Beckerman, B., Jerrett, M., & Brauer, M. (2007). Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science and Technology*, *41*(7), 2422–2428. doi:10.1021/es0606780

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, *42*(33), 7561–7578. doi:10.1016/j.atmosenv.2008.05.057

Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., … Demers, P. (2011). Creating national air pollution models for population exposure assessment in Canada. *Environmental Health Perspectives*, *119*(8), 1123–1129. doi:10.1289/ehp.1002976

IBM (2016). IBM SPSS software. http://www-01.ibm.com/software/nl/analytics/spss/. Accessed on 22-2-2016.

INEGI (2010). Número de habitantes. http://www.cuentame.inegi.org.mx/monografias/informacion/df/poblacion/default.aspx?tema=me&e=09. Accessed on 1-2-2016.

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., … Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*, *15*(2), 185–204. doi:10.1038/sj.jea.7500388

Jerrett, M., Arain, M. A., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N. L., … Finkelstein, M. M. (2007). Modeling the Intraurban Variability of Ambient Traffic Pollution in Toronto, Canada. *Journal of Toxicology and Environmental Health, Part A*, *70*(3-4), 200–212. doi:10.1080/15287390600883018

Johnson, M., Isakov, V., Touma, J. S., Mukerjee, S., & Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, *44*(30), 3660–3668. doi:10.1016/j.atmosenv.2010.06.041

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE*, *7*(2). doi:10.1371/journal.pone.0031743

Liu, C., Xiong, L., Hu, X., & Shan, J. (2015). A Progressive Buffering Method for Road Map Update Using OpenStreetMap Data. *ISPRS International Journal of Geo-Information*, *4*(3), 1246–1264. doi:10.3390/ijgi4031246

Liu, Y., Paciorek, C. J., & Koutrakis, P. (2009). Estimating Regional Spatial and Temporal Variability of PM2.5 Concentrations Using Satellite Data, Meteorology, and Land Use Information. *Environmental Health Perspectives*, *117*(6), 886–892. doi:10.1289/ehp.0800123

Luchtmeetnet (2016). Luchtmeetnet uitleg. http://luchtmeetnet.nl/uitleg. Accessed on 20-1-2016.

Lwin, K., & Murayama, Y. (2009). A GIS approach to estimation of building population for micro-spatial analysis. *Transactions in GIS*, *13*(4), 401–414. doi:10.1111/j.1467-9671.2009.01171.x

Marshall, J. D., Nethery, E., & Brauer, M. (2008). Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*, *42*(6), 1359–1369. doi:10.1016/j.atmosenv.2007.08.012

NASA (2014). U.S. Releases Enhanced Shuttle Land Elevation Data. http://www2.jpl.nasa.gov/srtm/. Accessed on 22-1-2016.

OSM, OpenStreetMap (2015). OpenStreetMap. http://www.openstreetmap.org/. Accessed on 9-2-2016.

PCRaster (2015). PCRASTER. http://pcraster.geo.uu.nl/. Accessed on 23-10-2015.

Pope, C. A., & Dockery, D. W. (2006). Health Effects of Fine Particulate Air Pollution: Lines that Connect. *Journal of the Air & Waste Management Association*, *56*(6), 709–742. doi:10.1080/10473289.2006.10464485

Poplawski, K., Gould, T., Setton, E., Allen, R., Su, J., Larson, T., … Buzzelli, M. (2009). Intercity transferability of land use regression models for estimating ambient concentrations of nitrogen dioxide. *Journal of Exposure Science & Environmental Epidemiology*, *19*(1), 107–117. doi:10.1038/jes.2008.15

Pozzi, F., Small, C., & Yetman, G. (2002). Modeling the Distribution of Human Population With Night-

Time Satellite Imagery and Gridded Population of the World. *Pecora 15/Land Satellite Information IV/ISPRS Commission I/FIEOS 2002 Conference Proceedings*, 9.

Ran, Y., & Li, X. (2015). First comprehensive fine-resolution global land cover map in the world from China—Comments on global land cover map at 30-m resolution. *Science China Earth Sciences*, *58*(9), 1677–1678. doi:10.1007/s11430-015-5132-4

Ross, Z., English, P. B., Scalf, R., Gunier, R., Smorodinsky, S., Wall, S., & Jerrett, M. (2006). Nitrogen dioxide prediction in Southern California using land use regression modeling: potential for environmental health analyses. *Journal of Exposure Science and Environmental Epidemiology*, *16*(2), 106–114. doi:10.1038/sj.jea.7500442

Ryan, P. H., & LeMasters, G. K. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology*, *19*(s1), 127–133. doi:10.1080/08958370701495998

Schneider, A., Friedl, M. a., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on "urban ecoregions." *Remote Sensing of Environment*, *114*(8), 1733–1746. doi:10.1016/j.rse.2010.03.003

SEDAC, Socioeconomic Data and Applications Center (2000). Global Rural-Urban Mapping Project (GRUMP), v1. http://sedac.ciesin.columbia.edu/data/collection/grump-v1. Accessed on 6-11-2015.

SEDAC. Socioeconomic Data and Applications Center (2010). Global Roads. http://sedac.ciesin.columbia.edu/data/collection/groads. Accessed on 22-1-2016.

SEDAC. Socioeconomic Data and Applications Center (2012). Global Annual PM2.5 Grids. http://sedac.ciesin.columbia.edu/data/set/sdei-global-annual-avg-pm2-5-modis-misr-seawifs-aod-1998-2012. Accessed on 9-2-2016.

Slama, R., Morgenstern, V., Cyrys, J., Zutavern, A., Herbarth, O., Wichmann, H.-E., … The LISA Study Group. (2007). Traffic-Related Atmospheric Pollutants Levels during Pregnancy and Offspring's Term Birth Weight: A Study Relying on a Land-Use Regression Exposure Model. *Environmental Health Perspectives*, *115*(9), 1283–1292. doi:10.1289/ehp.

UN-Habitat (2013). Private motorized transport in Bangkok, Thailand. http://unhabitat.org/wp-content/uploads/2013/06/GRHS.2013.Case_.Study_.Bangkok.Thailand.pdf. Accessed on 13-2-2016.

USGS (U.S. Geological Survey) (2015a). Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global. https://lta.cr.usgs.gov/SRTM1Arc. Accessed on 10-11-2015.

USGS (U.S. Geological Survey) (2015b). Shuttle Radar Topography Mission (SRTM) Void Filled. https://lta.cr.usgs.gov/SRTMVF. Accessed on 16-2-2016.

Vienneau, D., de Hoogh, K., Beelen, R., Fischer, P., Hoek, G., & Briggs, D. (2010). Comparison of land-use regression models between Great Britain and the Netherlands. *Atmospheric Environment*, *44*(5), 688–696. doi:10.1016/j.atmosenv.2009.11.016

Wang, M., Beelen, R., Bellander, T., Birk, M., Cesaroni, G., Cirach, M., … brune. (2014). Performance of Multi-City Land Use Regression Models for Nitrogen Dioxide and Fine Particles. *Environmental Health Perspectives*, *122*(8), 843–850. doi:10.1289/ehp.122-A322

Wikipedia (2015). List of cities proper by population. https://en.wikipedia.org/wiki/List_of_cities_proper_by_population. Accessed on 21-10-2015

World Health Organization. (2006). *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Summary of risk assessment*.

http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf?ua=1. Accessed on 22-2-2016.

8. Appendices

8.1 Appendix 1: reclassification of the road network data

The table below shows the reclassification of the roads, based on the reclassification scheme of Beelen et al. (2013).

Road type	Reclassification
0 Motorways	Major road
1 Main road – major importance	Major road
2 Other major roads	Major road
3 Secondary road	Major road / road*
4 Local connecting roads	Major road / road*
5 Local roads of high importance	Road
6 Local roads	Road
7 Local roads of minor importance	Road
8 Others	Road

*** Depends on local knowledge and depends on local knowledge and decision making**

Table 12. Reclassification scheme of the roads, based on the article of Beelen et al. (2013).

Table 13 shows the reclassification used in this research, based on the reclassification of Beelen et al. (2013).

OpenStreetMap (Bangkok & Mexico City)	Reclassification
Abandoned	Delete
Bridleway	Delete
Bus_guideway	Road
Bus_stop	Delete
Construction	Road
Corridor	Delete
Cycleway	Delete
Elevator	Delete
Escape	Delete
Footway	Delete
Ford	Road
Junction	Road
Living_street	Road
Motorway	Major road
Motorway_link	Major road
Path	Delete
Pedestrian	Delete
Planned	Delete
Platform	Delete
Primary	Major road
Primary_link	Major road
Proposed	Delete
Raceway	Delete
Residential	Road
Rest_area	Delete
Road	Road
Rural	Road

Secondary	Road
Secondary_link	Road
Service	Road
Services	Delete
Steps	Delete
Tertiary	Road
Tertiary_link	Road
Track	Road
Trunk	Major road
Trunk_link	Major road
Unclassified	Road
Unsurfaced	Road
Yes	Delete

Table 13. Reclassification of the OpenStreetMap dataset. The first column shows the OpenStreetMap road names. The second column shows how the roads are classified. 3 options were available: the roads could be deleted (for example when it was a bicycle path), or they were classified as road or major road.

8.2 Appendix 2: pre-processing & calculation of the variables

This appendix shows how the variables were calculated. For each variable a short description will be provided.

8.2.1 Variables of the London $PM_{2.5}$ regression equation

The regression equation of London consists of two variables: '*i*' and '*l*'. This equation is applied to Bangkok and Mexico City.

The '*i*' variable is a combination of the traffic intensity on the nearest major road and inverse of distance to the nearest major road. The first step was to reclassify the roads of both cities, this is based on the reclassification scheme in appendix 8.1. Then the total number of vehicles in both cities were spread out over the road network: the major roads got 75% and the roads 25%. The length of each road segment within each class (major road or road) determined how much vehicles each road segment would get, which is the traffic intensity of the road segment. Then the distance to nearest major road was calculated for each grid cell centre point in the areas (Bangkok & Mexico City) was calculated. Subsequently for each grid cell centre point the following calculation was done:
(Distance to nearest major road^{-1} * traffic intensity on this major road)

The '*l*' variable calculates the total length of all roads within a buffer of 500 meters. The first step was to create a fishnet (shapefile raster) with a spatial resolution of 10 meters. The next step was to intersect this fishnet with the road network. Then the geometry of each intersected road segment was calculated. Then the total length of all roads within a fishnet cell was calculated by summarizing the length of the roads per fishnet cell ID. Subsequently this was converted to a map-file in order to do the 500 meter buffer calculations in PCRaster (PCRaster, 2015). An example of how the total road length within a 500 meter buffer was calculated with PCRaster is:

```
from pcraster import *

setglobaloption("unitcell")
Expr = readmap("Fishnet500_10.map")
Ben_500mBuf_10m = windowtotal( Expr, 50)
report(Ben_500mBuf_10m, "Ben_500mBuf_10m.map")
aguila("Clone.map", Ben_500mBuf_10m, Expr)
```

The ‘windowtotal’ operator uses a square window to sum all the values within the defined distance (in this case 50 cells of 10 meters). The last step was to clip both variables (‘*i*’ & ‘*l*’) to the right extent and to calculate the regression equations. The PCRaster scripts can be found in appendix 1.

8.2.2 Variables of the Rotterdam $PM_{2.5}$ regression equation

The regression equation of Rotterdam consists of three variables: ‘*r*’, ‘*m*’ and ‘*t*’.

The ‘*r*’ variable is the variable with the $PM_{2.5}$ regional background values. The first step was to collect the data of the (city) background monitoring sites from Luchtmeetnet.nl. The next step was to calculate the average $PM_{2.5}$ value for 2015 for all monitoring sites. With Inverse Distance Weighting an interpolation was made of all the monitoring sites. Two regional estimate maps were created: one with all the (city) background monitoring sites and one with two of these monitoring sites left out. The reason is that these two monitoring sites were in and very near the modelled area in Rotterdam. This could have too much influence on the variable and the regression equation.

The ‘*m*’ variable consists of the total length of all major roads within a 50 meter buffer. This variable was calculated in the same way as the ‘*l*’ variable of the London regression equation, but then for major roads and another buffer size.

The ‘*t*’ variable is the total traffic load of major roads in a buffer of 1000 meters. The first step is to create a fishnet (shapefile raster) of 10 meters. Then an intersection needs to be done with the major roads. The next step is to calculate the new length (geometry) of the intersected major roads. Then the traffic load of these intersected major roads can be calculated by multiplying the number of vehicles on the road with the length of the major road segment. Subsequently the total traffic load of major roads per fishnet cell can be calculated by summarizing the traffic load of the major roads per fishnet cell ID. Subsequently this was converted to a map-file in order to do the 1000 meter calculations in PCRaster with the window total command (which is similar to the command used for the ‘*l*’ variable).

The last step was to clip the variables to the right extent and to calculate the regression equations. The PCRaster scripts can be found in appendix 1.

8.3 Appendix 3: $PM_{2.5}$ python scripts

The code below shows the regression equation used to model the $PM_{2.5}$ for Benito Juárez.

```
from pcraster import *

# reading inputs
INTMAJORINVDIST = readmap("intmajorrast.map")
ROADLENGTH_500 = readmap("benclip_500m.map")

# PM2.5 London regression equation
res = 7.19 + 1.38 * 10E-03 * INTMAJORINVDIST + 2.65 * 10E-04 * ROADLENGTH_500

report(res, "Final_pm_25_Ben.map")
```

The code below shows the regression equation used to model the $PM_{2.5}$ for the six districts in Bangkok.

```
from pcraster import *
```

```
# reading inputs
INTMAJORINVDIST = readmap("Bk_trafload.map")
ROADLENGTH_500 = readmap("Bk_500mBufOutput.map")

# PM2.5 London regression equation
res = 7.19 + 1.38 * 10E-03 * INTMAJORINVDIST + 2.65 * 10E-04 * ROADLENGTH_500

report(res, "Final_pm_25_Bk.map")
```

The code below shows the regression equation used to model the $PM_{2.5}$ for Rotterdam.

```
from pcraster import *

# reading inputs
REGIONALESTIMATE = readmap("IDWCL.map")
MAJORROADLENGTH_50 = readmap("Maj50BufCL.map")
TRAFMAJORLOAD_1000 = readmap("TrafMaj1000CL.map")

# PM2.5 regression equation for the Netherlands
res = 9.46 + 0.42 * REGIONALESTIMATE + 0.01 * MAJORROADLENGTH_50 +2.28 * 10E-09 *
TRAFMAJORLOAD_1000

report(res, "Final_pm_25_Rd_MyOutput.map")
```

The code below shows an example of how all the major roads within a 50 meter buffer were calculated.

```
from pcraster import *

setglobaloption("unitcell")
Expr = readmap("fishnetraster.map")
Rdam_Major50mBuf = windowtotal( Expr, 5)
report(Rdam_Major50mBuf, "Rdam_Major50mBuf.map")
```