TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

KARI PIETILÄ

LAND-USE REGRESSION MODEL FOR ASSESSING LONG-TERM EXPOSURE TO ULTRAFINE PARTICLES IN AMSTERDAM

Master's Thesis

Examiners: prof. Leena Korpinen &
prof. Risto Raiko
Examiners and topic approved by
the Council of the Faculty of Natural
Sciences on November 4th, 2015

ABSTRACT

KARI PIETILÄ: Land-Use Regression Model for Assessing Long-Term Exposure to Ultrafine Particles in Amsterdam
Tampere University of Technology
Master of Science Thesis, 61 pages, 8 Appendix pages
November 2015
Master's Degree Programme in Environmental and Energy Technology
Major: Environmental Health
Examiners: Professor Leena Korpinen and Professor Risto Raiko

Keywords: ultrafine particles, land-use regression, exposure assessment

Geographic Information Systems (GIS) and statistics based land-use regression (LUR) models are widely used for modeling small-scale spatial variation of ambient air pollution. These models have successfully been utilized in cohort studies where individual exposure of participants needs to be estimated. LUR has been used to model gases and particles alike, but to date, there are no published studies that would have utilized LUR in assessing cohort members' long-term exposure to ultrafine particles.

Using measurement and GIS data from previous studies, two land-use regression models for UFPs were developed for the city of Amsterdam, the Netherlands. With two slightly different models it could be assessed whether model performance was sensitive to observations that had been assigned unrealistic traffic intensities. The models were validated using the holdout method. In holdout validation (HV) the original datasets were divided into several sample pairs, new models were developed with partial set of data and these models then validated with unused data. Finally, developed LUR models were utilized in a cohort of 4,986 people to estimate participants' long-term exposure to UFPs.

The two land-use regression models performed almost equally, both explaining approximately 44% of the variability in measured particle number count, which was used as a proxy for ultrafine particles. The models incorporated inverse distance to the nearest major road as the most important predictor variable, reflecting the importance of transportation as a source of UFPs. Validation indicated that both models were stable.

Exposure estimates from applying the LUR models were fairly similar and reasonable. The correlation between the estimates from the two models was 0.76. However, the estimates should be used with caution because of the limited explanatory power of the LUR models and inherent limitations of geographic data. Further to this, exposure assessment did not account for the different exposure levels that individuals may experience when they move around the city during their days.

As a way forward, developing more robust land-use regression models is important. In general, GIS and traffic data improve on a fast basis, which in turn should improve LUR models. Even then, there is a strong need to validate assigned exposures with personal monitors.

TIIVISTELMÄ

KARI PIETILÄ: Paikkatietojärjestelmiin ja lineaariseen regressiomalliin pohjautuva arvio pitkäaikaisesta altistumisesta ultrapienille hiukkasille Amsterdamin alueella
Diplomityö, 61 sivua, 8 liitesivua
Marraskuu 2015
Ympäristö- ja energiatekniikan diplomi-insinöörin tutkinto-ohjelma
Pääaine: Ympäristöterveys
Tarkastajat: professori Leena Korpinen ja professori Risto Raiko

Avainsanat: ultrapienet hiukkaset, lineaarinen regressio, altistuksen arviointi

Pienmittakaavaista ilmansaastepitoisuuksien vaihtelua mallinnetaan yleisesti paikkatietojärjestelmiin ja lineaariseen regressioon pohjautuvien mallien avulla. Näitä malleja on käytetty onnistuneesti kohorttitutkimuksissa, joissa yksilöiden altistus saasteelle pitää arvioida. Lineaarisia regressiomalleja on käytetty niin kaasujen kuin hiukkastenkin mallintamiseen, mutta toistaiseksi ei ole julkaistu tutkimuksia, joissa regressiomallien avulla olisi selvitetty pitkäaikaista altistusta ultrapienille hiukkasille.

Tässä työssä on kehitetty kaksi lineaarista regressiomallia ultrapienille hiukkasille hollantilaista Amsterdamin kaupunkia varten. Hieman erilaisten mallien avulla voitiin arvioida missä määrin eräiden havaintopisteiden epärealistiset liikennemäärät vaikuttivat lopputulokseen. Molemmat mallit validoitiin käyttäen holdout-menetelmää. Alkuperäisestä aineistosta otettiin ensin useita erillisiä otospareja ja toisen otoksen perusteella kehitettiin uusia regressiomalleja, jotka sitten validoitiin käyttämättömällä vastinotoksella. Tämän jälkeen alun perin kehitettyjä malleja käytettiin kohortin jäsenten altistuksen arviointiin. Kohorttiin kuului 4 986 ihmistä.

Molemmat lineaariset regressiomallit selittivät noin 44% havaitusta hiukkasten lukumäärän ja siten myös ultrapienten hiukkasten vaihtelusta. Molemmissa malleissa ensisijaisena muuttujana oli etäisyys lähimpään päätiehen käänteisenä, mikä heijastaa liikenteen merkitystä ultrapienten hiukkasten lähteenä. Validointi osoitti, että molemmat mallit olivat vakaita.

Malleista saadut arviot kohortin jäsenten altistuksesta olivat melko hyvät ja yhtäläiset, sillä mallien antamien altistusten korrelaatiokerroin oli 0.76. Tuloksia pitää kuitenkin käyttää varoen, johtuen regressiomallien osittaisesta selitysvoimasta ja paikkatiedon rajoituksista. On myös hyvä huomioida, että altistuksen arvioinnissa ei otettu huomioon niitä konsentraatioita, joille ihmiset altistuvat kotiensa ulkopuolella.

Tulevaisuudessa parempien mallien kehittäminen on tärkeää. Yleisesti ottaen paikkatiedon tarkkuus ja liikennemallit kehittyvät nopeasti, minkä pitäisi parantaa myös regressiomallien selitysvoimaa. Tästä huolimatta määritetyt altistustasot pitäisi myös pyrkiä validoimaan henkilökohtaisilla mittareilla.

PREFACE

This thesis marks the end of my Master's studies in Environmental Health. The research for my topic was carried out at the Institute of Risk Assessment Sciences (IRAS) of Utrecht University in the Netherlands. At IRAS, my tasks involved not only this thesis but also the application of the results in a cohort study in order to assess whether long-term exposure to ultrafine particles is associated with mortality. Findings from epidemiological studies are intended for publication in peer-reviewed journal articles, and therefore the results of the cohort study are not disclosed here.

Guidance from IRAS was instrumental in creating this thesis and completing the cohort study. I wish to express my gratitude toward Professor Bert Brunekreef for offering me this topic as well as instructing and supporting me throughout my time at the institute. It is also my pleasure to thank Associate Professor Gerard Hoek for guidance and mentoring as well as everyone else who contributed to my work in any way.

I would also like to express my appreciation for Professor Leena Korpinen and Professor Risto Raiko in Finland. Thank you for providing swift feedback and comments on this thesis and for the support in finalizing my degree. Finally, I wish to thank everyone in my family, group of friends, and others close to me who have supported me in any way during my years at the university.

Helsinki, November 27th 2015

Kari Pietilä

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

CI	Confidence Interval
Cook's D	Cook's distance
CORINE	Coordination of Information on the Environment
CPC	Condensation particle counter
DISTINVMAJORC1	Inverse distance to the nearest major road
DMPS	Differential mobility particle sizer
EEA	European Environment Agency
EEA_5000	Population density within a buffer of 5000 meters
EPIC	European Prospective Investigation into Cancer and Nutrition
EPIC MORGEN	Cohort compiled as part of EPIC and MORGEN
ESCAPE	European Study of Cohorts for Air Pollution Effects
GIS	Geographic Information Systems
HR	Hazard Ratio
HV	Holdout validation
INTARESE	Integrated Assessment of Health Risk of Environmental Stressors in Europe
IRAS	Institute for Risk Assessment Sciences
LOOCV	Leave-one-out cross-validation
LUR	Land-Use Regression
MORGEN	Monitoring Project on Risk Factors for Chronic Diseases
NWB	Nationale Wegen Bestand (Dutch national road network)
PM	Particulate matter
$PM_{0.1}$	Particles with a diameter of less than 0.1μm; ultrafine particles
$PM_{2.5}$	Particles with a diameter of less than 2.5μm; fine particles
PM_{10}	Particles with a diameter of less than 10μm; coarse particles
OLS	Ordinary Least Squares
PNC	Particle number concentration
PORT_5000	Port area within a buffer of 5000 meters
SD	Standard Deviation
SEM	Standard Error of the Mean
SMPS	Scanning mobility particle sizer
TRAFNEAR	Traffic on nearest road
UFP	Ultrafine particles
URBGREEN_5000	Urban green area within a buffer of 5000 meters
VIF	Variance Inflation Factor

β	regression coefficient
C_{UFP}	concentration of ultrafine particles
P10	10th percentile
P90	90th percentile
$Pr > \|x\|$	p-value associated with statistic x
R^2	coefficient of determination (R-Squared)
X	predictor variable
Z	Z score

1 INTRODUCTION

There is strong evidence that exposure to ambient particulate matter (PM) is associated with adverse health effects including cardiovascular and all-cause mortality (Hoek et al. 2013; Brook et al. 2010). Most published studies have focused on fine particles and coarse particles, with respective diameters of less than 2.5 and 10 micrometers ($PM_{2.5}$, PM_{10}). Since research efforts aim to identify the most hazardous characteristics of air pollution, focus of interest has recently shifted towards a smaller fraction of PM, i.e. ambient ultrafine particles (UFPs, $PM_{0.1}$).

Ultrafine particles are a mixture of solid particles and liquid droplets with a diameter of 0.1 micrometers or less. Due to their vast number, small diameter and high surface area, UFPs are potentially more harmful to human health than larger particles (HEI 2013). Indeed, there is growing evidence of independent health effects associated with short-term exposure to UFPs but more research is still needed (Rückerl et al. 2011). To date, no studies have been published about long-term UFP exposure and its impact on health mainly due to difficulty in assessing annual exposures for various study groups.

The aim of this thesis is to build upon earlier research on ultrafine particles and health. Utilizing geographic information systems (GIS) and statistics based land-use regression (LUR) models, long-term exposure to ultrafine particles is assessed for the members of a retrospective cohort living in the city of Amsterdam, the Netherlands. LUR is an established method in modeling intraurban concentrations of various air pollutants, especially in cases where high spatial anomalies in observed concentrations have typically been a challenge (Hoek et al. 2008a). Exposure estimates from various LUR models have been applied to several epidemiological analyses but there are no published studies that would have utilized land-use regression in assessing cohort members' long-term exposure to UFPs.

In order to lay ground for the utilized methods, important background information on ultrafine particles is presented in the first couple of chapters of this thesis. In Chapter 2, typical characteristics of ultrafine particles are presented with regard to aspects that make them unique from larger particle size fractions. Chapter 3 conveys the most important considerations regarding UFP exposure so as to link UFP characteristics with plausible health impacts as well as to motivate the use of chosen study methods. Then, Chapter 4 presents a theoretical review of the research methods, namely land-use regression and exposure assessment.

In the latter part of the thesis, research materials, the application of the study methods, and results are presented. The materials include measurement data, geographic information and cohort addresses, which all are described in Chapter 5. The development of land-use regression models is presented in Chapter 6. The results from land-use regression are source material for exposure assessment, which in turn is described in Chapter 7. Finally, summary results are presented in Chapter 8 and their importance is discussed in Chapter 9. The conclusion of this thesis is available in Chapter 10.

2 CHARASTERISTICS OF ULTRAFINE PARTICLES

Ultrafine particles have a number of distinct features that distinguish the size range from larger particles. The most important characteristics of UFPs are presented in this chapter and any meaningful differences to larger particles are mentioned where appropriate.

2.1 Key Characteristics

Ultrafine particles are a fraction of airborne particulate matter, which is a mixture of solid particles and liquid droplets (Martins et al. 2010). With a diameter of 100 nanometers or less, UFPs are the smallest in the entire spectrum of particulate matter. This compares to the size of poliovirus that is 30nm in diameter (Oberdörster et al. 2005).

UFPs contribute little to the particulate mass but they are dominant as to the total number of airborne particles (HEI 2013). Due to this, UFPs have high surface area per unit of mass as compared with larger particle sizes. These aspects may also be observed in Figure 1, which depicts various concentration metrics as a function of particle diameter.

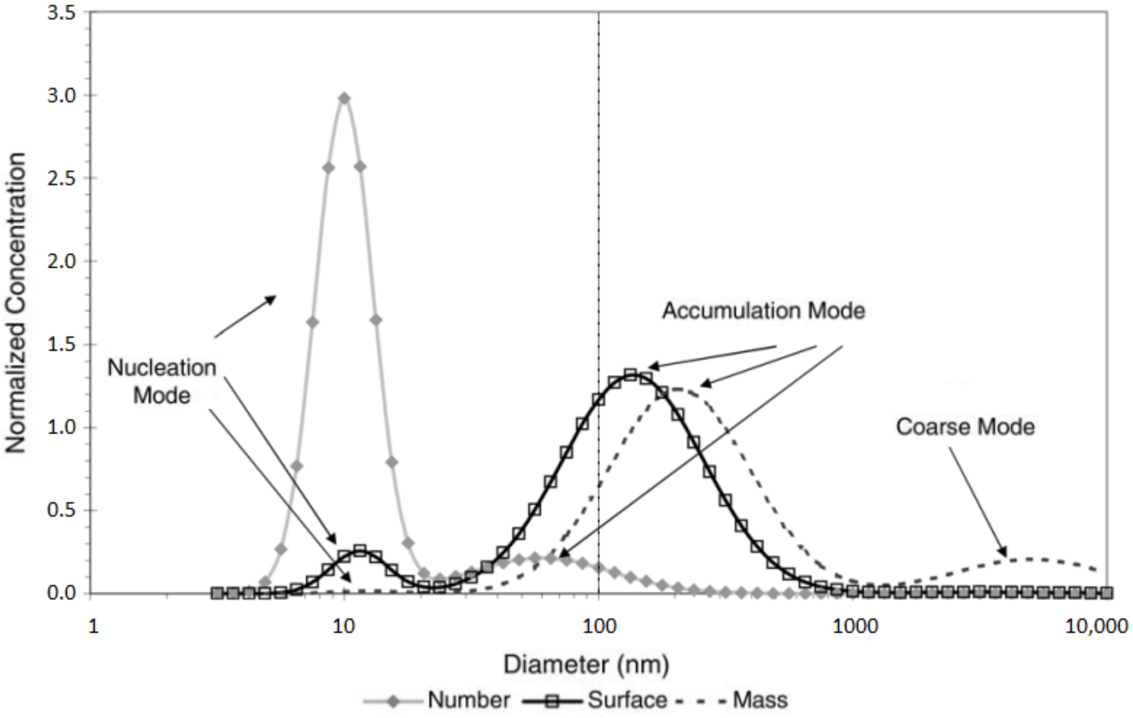


Figure 1. *Normalized particle size distributions of typical roadway aerosol (HEI 2013)*

In Figure 1, concentration-diameter functions are weighted by number, volume and mass. As it may be observed, particles between 8-20nm in diameter contribute the most to particle number while particles closer to the limit of 100nm are the main cause for observed mass and surface concentrations.

While ultrafine particles are pivotal to the total number of airborne particles, UFP concentrations show sharp spatial anomalies and substantial variation across a single city. The highest concentrations are generally observed near combustion sources but dilution is fast with increasing distance from the source (HEI 2013). In 2010, Karner and colleagues published a meta-analysis of 41 studies about traffic-related pollutants, and showed that the concentration of above 3-nm particles declines 60% when the distance from the road edge reaches 100 meters. When the distance reaches 200 meters, concentration is not distinguishable from the urban background. In contrast to UFPs, larger particle size fractions show much less spatial variability (HEI 2013).

In addition to spatial variation, ultrafine particle concentrations show high temporal variation due to diurnal and seasonal patterns. For instance, seasonal 10-fold variability in hourly UFP concentration has been reported in Los Angeles (HEI 2013). However, ultrafine particle concentrations have been shown to fluctuate similarly between different intraurban sites (Hoek et al. 2008b).

2.2 Formation and Scavenging Mechanisms

There are several mechanisms via which UFPs may form. UFPs may be emitted directly or they can form from the nucleation of supersaturated vapors as exhaust cools down. Particles formed directly are called primary particles whereas particles nucleated in the atmosphere are secondary. Third recognized formation process is associated with spontaneous chemical reactions in the atmosphere. Chemical reactions of various compounds tend to produce regionally dispersed UFPs whereas combustion-related processes lead to more localized anomalies. (Sioutas et al. 2005; Morawska et al. 2008)

As UFPs are released into the air or formed in the atmosphere they grow by the means of condensation and coagulation. Initially, the smallest particles are below 10nm in diameter but over the timescales of a few hours, coagulated UFPs may become over 100nm in diameter. Thus these particles no longer belong to the ultrafine size range. This atmospheric scavenging mechanism concerns primary UFPs. Secondary UFPs are removed when nucleated particles evaporate after continued dilution of the exhaust plume. Evaporation may also lead to the shrinkage of particles so that only the solid core remains. (Morawska et al. 2008; HEI 2013)

2.3 Sources

The sources of ultrafine particles can be categorized in several ways. Firstly, ultrafine particles can be of anthropogenic or natural origin. Secondly, anthropogenic sources include both unintentionally and intentionally produced particles. Intentionally produced particles in the ultrafine range are referred to as engineered nanoparticles. These nanoparticles are increasingly used in nanotechnology and medicine. (Oberdörster et al. 2005)

Research on ultrafine particles and engineered nanoparticles are somewhat distinctive fields due to differences in their formation and properties, such as the presence of adsorbents (Oberdörster et al. 2005). This thesis acknowledges the distinction and thus makes no further reference to engineered nanoparticles. The following two subchapters describe the sources of ultrafine particles in detail, following the categorization into natural and anthropogenic sources.

2.3.1 Natural sources

Natural sources of ultrafine particles constitute the background concentration that is experienced everywhere at different levels. Typically 30-50% of measured UFP concentration is from natural sources. (Morawska et al. 2008)

The natural sources of ultrafine particles include temporal forest fires and volcano eruptions as well as continuous occurrence of sea spray and, most importantly, various gas-to-particle conversions (Oberdörster et al. 2005). The nucleation of low-volatile gas-phase compounds into particles and their subsequent growth has been observed in forests and coastal areas. The process involves e.g. monoterpenes ($C_{10}H_{16}$) emitted by forest trees as well as sulphuric acid (H_2SO_4), ammonia (NH_3) and water (H_2O). (Morawska et al. 2008; Kulmala et al. 2000)

2.3.2 Anthropogenic sources

The major anthropogenic sources of ultrafine particles are largely identified with the help of emission inventories and source apportionment (HEI 2013). Different studies suggest similar source categories although the relative importance of a particular source varies with the location.

As presented by the Health Effect Institute (2013), road and non-road transportation, particularly diesel engines are traced as major contributors to UFP emissions in urban areas. Gasoline engines and motor oil are also important sources. Together these sources may account for up to 90% of the total UFP emissions right next to busy roads (HEI 2013).

The importance of traffic as a source of particulate matter is echoed by the meta-analysis by Morawska et al. (2008), in which PNC was calculated at eight different environments ranging from rural surroundings to urban ones. The authors utilized 71 measurements from several independent studies, and reported that the mean and median concentrations were higher for traffic environments as compared to other types of sites. These results may also be observed in Figure 2 below.

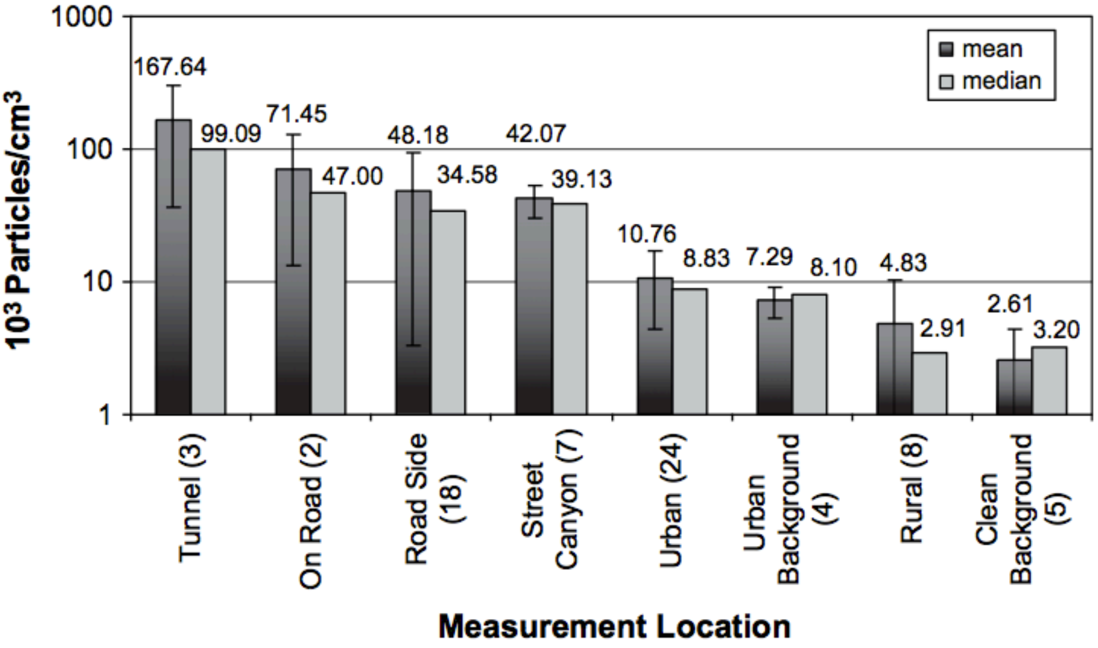


Figure 2. *Particle number concentration for various environments (Morawska et al. 2008)*

As it can be seen from Figure 2, particle number concentration can be over 4 times higher at road environments and along street canyons as compared to urban background sites. The PNC is especially high in tunnels where dilution with ambient air is limited. In rural areas, observed PNC can be close to that of clean background as suggested by the meta-analysis.

Since traffic is less prominent in rural areas, the relative importance of sources not affiliated with transportation becomes greater. In these areas industry, residential and commercial heating, as well as cooking are important factors to consider. Further to this, some studies suggest that large proportion of rural particulate matter comes from an unknown source. (HEI 2013)

According to Health Effects Institute (2013), transportation along with other before-mentioned source categories account for approximately 90% of all anthropogenic UFP emissions. The institute reports that the rest, approximately 10% of emissions, originate from agriculture, waste disposal and other miscellaneous sources. The miscellaneous sources include e.g. several indoor activities not related to cooking, of which burning of pure wax candles was recognized as one of the most important contributors to indoor particle number concentration by Afshari et al. (2005).

2.4 Chemical Composition

When it comes to the chemical composition of ultrafine particles, comprehensive information is not available. One challenge is that composition of UFPs may change seasonally (Morawska et al. 2008). However, an indication of general composition is given e.g. by Cass et al. (2000), who measured UFP composition in seven Southern Californian cities over period 1995-1997.

On average, Cass and colleagues' study found out that UFPs were composed of 50% organic compounds; 14% trace metal oxides; between 5-10% elemental carbon, sulphates and nitrates; almost 4% ammonium; as well as approximately 0.5% sodium and chloride. The most abundant catalytic metals were iron, titanium, chlorine and zinc. Although measurements were carried out in seven cities, these all were located in Southern California. Therefore these results cannot be generalized to e.g. European cities. (Cass et al. 2000)

2.5 Measurement of Ultrafine Particles

Particulate matter can be measured in several different ways. Fine and coarse particles are typically measured by their mass but ultrafine particles are most often measured by their number. Measuring UFP mass is not practical as commercial balances are not accurate enough. Further to this, sample contamination with larger particles can alter the results significantly. (HEI 2013)

The number of ultrafine particles is typically measured with condensation particle counters (CPC) in which particles are counted as they pass through a laser beam. CPC alone counts particles of all sizes, i.e. total particle number concentration (PNC) per unit volume of air. Even so, PNC is often used as a proxy for ultrafine particles. There is support to this approximation as several studies show that about 90% of the total PNC is within the ultrafine range. (HEI 2013; Morawska et al. 2008)

Using CPC in combination with particle sizers, number concentration within a certain size range can be obtained. Such technologies include differential and scanning mobility particle sizers (DMPS/SMPS). As UFPs are defined by their diameter, categorizing particles with non-spherical shapes can sometimes be ambiguous and depend on the measurement technology. (Morawska et al. 2008; Sioutas et al. 2005)

CPCs can count particles as small as 2nm in diameter. When using particle sizers, this detection limit is often set up higher than what is technically possible. Compromising on the lowest possible detection limit permits a larger measurement range. Still, CPCs without particle sizers generally show significantly higher concentration results than DMPS/SMPS do. (Morawska et al. 2008)

3 IMPLICATIONS FROM EXPOSURE TO AMBIENT ULTRAFINE PARTICLES

There are several considerations with regard to exposure to ultrafine particles that are important to contemplate. These include circumstances that affect exposure and human factors that indicate what health effects are plausibly associated with exposure. These considerations are reviewed in the text that follows.

3.1 Exposure Characterization

In general, exposure means the cumulative concentration experienced in several microenvironments over a period of time (Morawska et al. 2008). Since UFP concentrations show high spatial anomalies, assessing population-level exposure cannot rely on a central monitoring site (Hoek et al. 2008a). Development of regional dispersion models and land-use regression models is an attempt to address this issue. With the help of these, UFP concentrations may be assessed at different locations of a city. The models are reviewed more closely in Chapter 4.

Due to the fact that people spend a considerable amount of their time indoors, both at home and in work, personal exposure to ultrafine particles is largely determined by the indoor concentration of UFPs. Indoor exposure in turn is dependent on the infiltration from outdoors and indoor sources. Exposure to indoor sources is often temporary in nature and includes events such as cooking, use of heaters and candle burning (Afshari et al. 2005). On the other hand, infiltration from outdoors occurs continuously and thus constitutes the indoor background concentration.

While indoor sources do contribute to the daily exposure to UFPs, outdoor sources are more relevant consideration when it comes to assessing the health impacts from exposure to ambient air pollution. In a study by Wallace and colleagues (2010), it was reported that 36% of the daily UFP exposure of a suburban nonsmoker was due to outdoor sources. Exposure in vehicles was reported separately, and it was 17% of the total daily exposure. Together these total over 50%. As the study was carried out in a suburban environment the authors argued that the share of outdoor sources to the daily exposure should be higher than this in urban environments and lower in rural environments.

Due to the need to study the impact of outdoor UFPs on human health, it is important to assess the extent to which ambient ultrafine particles penetrate indoors. As reviewed by the Health Effects Institute (2013), infiltration is affected by several factors, such as ventilation rates within buildings, presence of local outdoor sources, wind speed and season. Consequently, particle number counts are generally less indoors as compared to outdoors. In a study by Zhu et al. (2005), outdoor particle number concentrations outside of four apartments in Los Angeles were approximately 1.5-2 times higher than the concentration indoors.

Since indoor penetration of outdoor UFPs does occur, an important question is whether infiltration rates correlate with the variation of outdoor concentrations. Zhu et al. (2004) reported some evidence to this, i.e. that there is an association between the diurnal variability of outdoor and indoor particle number concentrations. Similar findings were reported by Hoek et al. (2008b), who studied 152 homes across 4 European cities. As can be seen from Figure 3 below, the average daily variability of PNC indoors tracked closely that of outdoors, only at a lower level. The concentrations were the lowest during night-time and peaked during the morning rush hour. Concentrations then stayed elevated up until evening when they started to gradually decline.

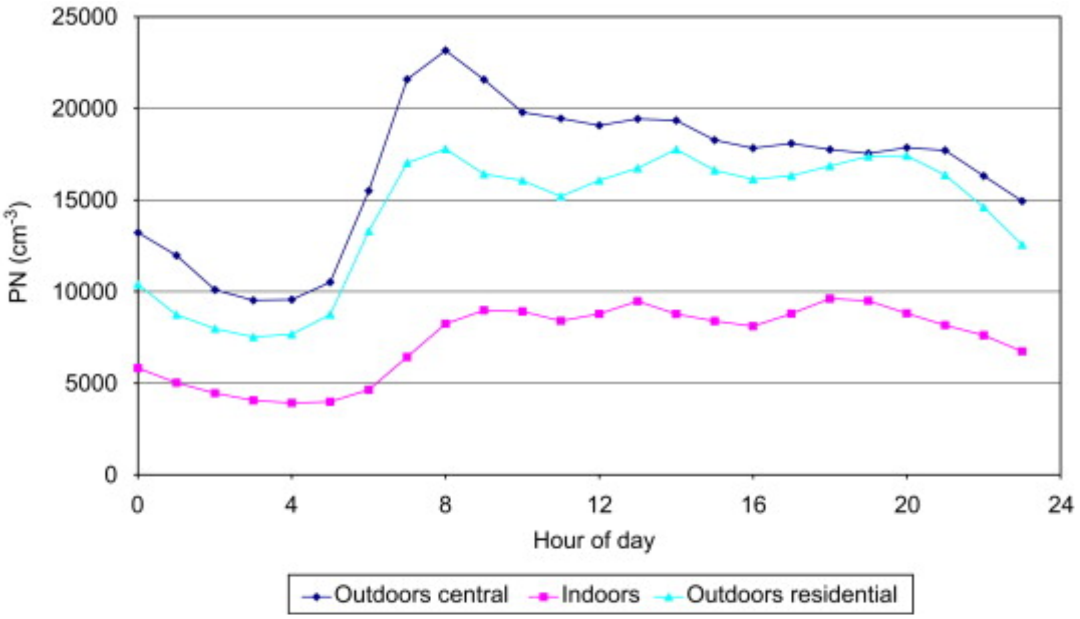


***Figure 3**. The average daily variability of outdoor and indoor PNC (Hoek et al. 2008b)*

Hoek et al. (2008b) also calculated how well indoor and outdoor concentrations correlated with each other. The Pearson correlation coefficient between indoor and outdoor particle number count was 0.58 in Amsterdam, which was one of the cities where measurements were done. In other cities, the coefficient ranged from 0.41-0.80. These results suggest that concentration data from outdoors may be used as somewhat reasonable proxy for traffic-related UFP exposure, keeping in mind that the level of exposure is not at the level of the outdoor concentration. This is an important result for epidemiological analyses.

With regard to the size-dependent indoor penetration of ultrafine particles, there is some evidence of different infiltration rates. Zhu et al. (2004) reported lowest indoor/outdoor ratios (0.1-0.4) for particles between 10-20nm whereas the highest ratios (0.6-0.9) were reported for particles in the 70-100-nm range. However, the authors noted that they had less statistical confidence in data below 20 nm. Of note are also the findings that the composition of particles may change during infiltration. Especially volatile particles may change or be lost completely during indoor penetration (Sioutas et al. 2005).

3.2 Considerations Regarding Exposure

In order to fully understand what health impacts UFP exposure may cause, physiological considerations need to be factored in. Whereas exposure routes determine which organs and body systems are most susceptible to ultrafine particles, physiological defense mechanisms limit the dose experienced by the target organs. Plausible health impacts in turn are derivative from these two factors.

3.2.1 Exposure Routes and the Human Respiratory System

Ultrafine particles can become in contact with the human body via respiratory system, skin or gastrointestinal tract. Very little uptake has been documented by either the gastrointestinal tract or skin, albeit translocation to the lymphatic system does occur from areas of broken or flexed skin. The major exposure route is the respiratory system, which is also what most in vivo toxicity studies have focused on. (Oberdörster et al. 2005)

The respiratory deposition of particles is dependent on a variety of physiological factors such as the level of physical activity, posture, sex, and breathing mode as well as wind condition and particle characteristics. Further to this, particles of different sizes deposit in somewhat different regions within the respiratory system. These regions may be classified as the extrathoracic region, tracheobronchial region and alveolar region as presented in Figure 4 on the following page. (Bartley et al. 2011)

The anatomy of the human respiratory system is well known, and in this context a brief description of the before-mentioned regions suffice. Firstly, air and inhaled particles enter the extrathoracic region via mouth or nostrils. The anatomy of the region involves also nasal and oral cavities as well as different parts of pharynx, and the larynx. (Marieb 2011)

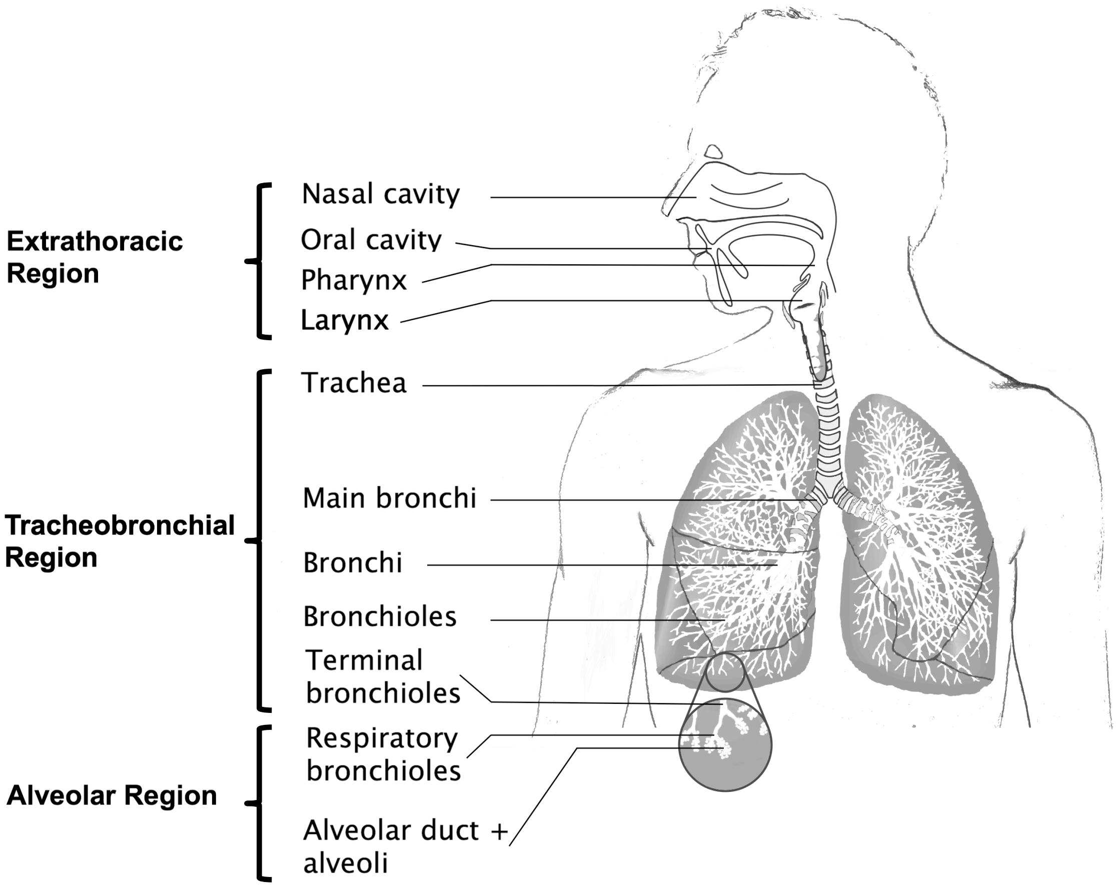


***Figure 4**. Human respiratory system, adapted from Bartley et al. (2011) & Marieb (2011)*

After passing the extrathoracic region, air and particles proceed to the tracheobronchial region, which comprises of trachea, bronchi and terminal bronchioles. Lastly, the alveolar region is synonymous with the respiratory zone, which is where gas-exchange occurs. As seen in Figure 4, the region consists of respiratory bronchioles, alveolar ducts and alveoli. (Oberdörster et al. 2005; Bartley et al. 2011; Marieb 2011)

3.2.2 Deposition in the Human Respiratory Tract

As mentioned, the respiratory tract deposition of particles is dependent on various physiological and environmental factors. Therefore absolute deposition rates do not exist but they vary with e.g. physical activity. However, various deposition rate functions do look similar (Bartley et al. 2011), which is why several patterns may be observed from them.

One of the most important remarks is that ultrafine particles deposit more readily in the alveolar region in contrast to larger particle sizes (HEI 2013). This may be observed from Figure 5 where the respiratory tract deposition of particles is plotted as a function of particle size for a healthy male human subject at rest. The function represents a meta-analysis of several journal articles and is widely accepted in the scientific community (Geiser et al. 2010). The figure is available on the following page.

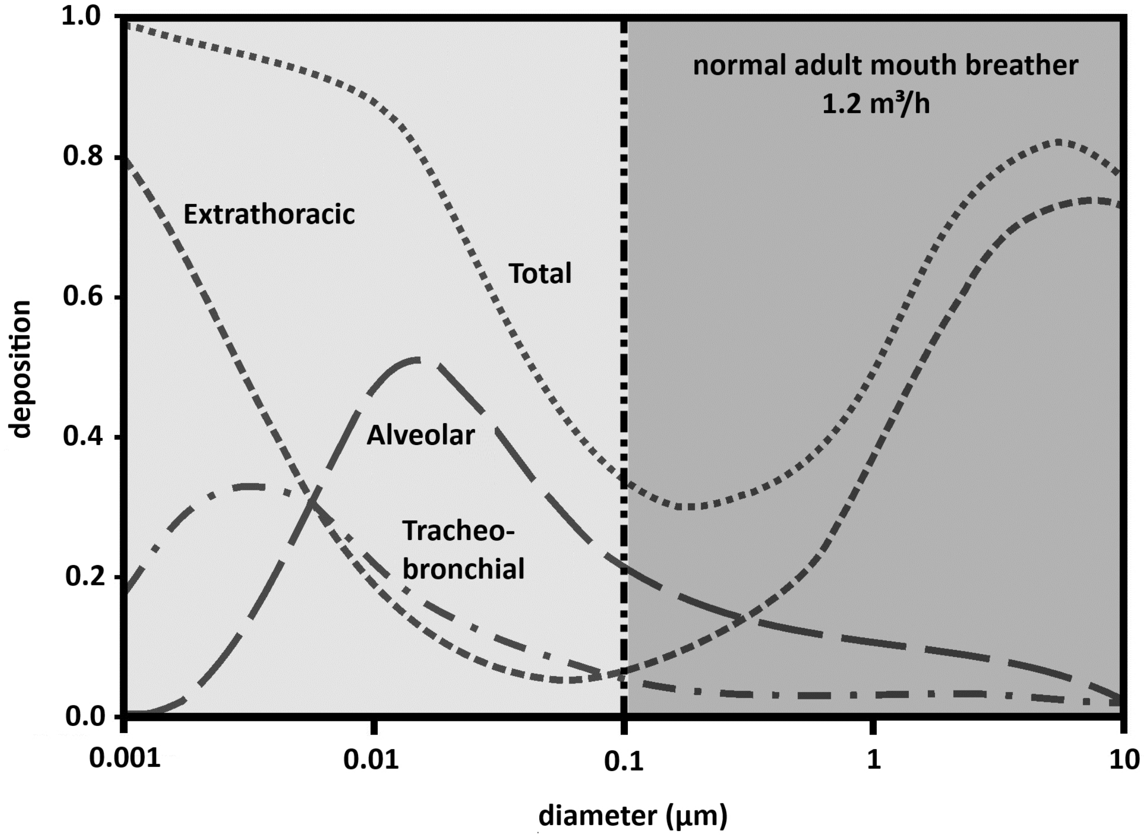


***Figure 5**. The respiratory tract deposition of particles for a male human subject as a function of particle size, adapted from Geiser et al. (2010)*

As can be seen from Figure 5, almost all 1-nm particles deposit in the respiratory tract, while approximately 80% of them deposit already in the extrathoracic region. In comparison, 7-nm particles are deposited about equally in the extrathoracic, tracheobronchial and alveolar areas. Ultrafine particles of over 7nm in diameter are most likely to deposit in the alveoli. This coincides with the 8-20nm range that contributes the most to particle number concentration in ambient air as presented in Figure 1 in Chapter 2.

In spite of varying deposition patterns, ultrafine particles deposit more homogenously as compared to larger particles. This is due to the fact that ultrafine particles have the ability to move via diffusion (Kreyling et al. 2006).

3.2.3 Clearance and Translocation of Ultrafine Particles

All deposited particles are not retained in the respiratory system as there are clearance and excretion mechanisms that remove foreign debris and pathogens. On one hand, biosoluble particles and particle components may be dissolved chemically. Solutes are then absorbed, diffused, bound to subcellular structures or cleared into blood and lymphatic circulation. On the other hand, insoluble particles are cleared with the help of physical mechanisms. (Oberdörster et al. 2005)

There are a couple of different physical clearance mechanisms. Firstly, cilia present in the extrathoracic and tracheobronchial areas move suspended particles toward pharynx from where contaminated mucus is swallowed into the stomach for digestion and excretion. Secondly, macrophages phagocytize particles in the alveolar region where there are no cilia. Macrophages with internalized particles then move toward the mucociliary escalator, which in turn moves macrophages toward pharynx. (Marieb 2011; Oberdörster et al. 2005).

There is some evidence that ultrafine particles are cleared slower and less completely from the lungs as compared to particles of larger size. This may lead to particle accumulation and translocation within the body. For instance, it has been reported that it may take up to 700 days in humans for macrophages to reach the mucociliary escalator. Further to this, studies with rats have shown that ultrafine particles are not effectively phagocytized by alveolar macrophages as opposed to larger particles. (HEI 2013; Oberdörster et al. 2005)

Ineffective clearance mechanisms lead to retention and accumulation of ultrafine particles, which increase their interaction with lung cells. There is evidence from studies with animals that UFPs may move across the lung epithelium into interstitial spaces. Mechanisms for translocation are not well understood but some studies show that UFPs may move through endocytosis and exocytosis. Factors affecting translocation include particle size, surface chemistry and probably charge. (HEI 2013; Oberdörster et al. 2005)

Once UFPs have reached pulmonary interstitial spaces, they may be further transported into cardiovascular and lymphatic systems. With blood, UFPs may be distributed into organs, such as liver, spleen, heart and kidneys. Neuronal uptake and translocation to the brain may also occur through olfactory nerves. However, the importance of neuronal uptake in humans has been questioned. (Oberdörster et al. 2005)

3.3 Plausible Health Effects from Exposure

Since health impacts from the exposure to ultrafine particles are not well-known, UFP characteristics and physiological considerations largely determine what type of health effects are most likely. In their 2013 report, Health Effects Institute listed three types of plausible health impacts: 1) effects on the respiratory system, 2) effects on the cardiovascular system, and 3) effects on the neurological system. There are several mechanisms via which the health impacts are hypothesized to occur. These are summarized in Figure 6 on the following page.

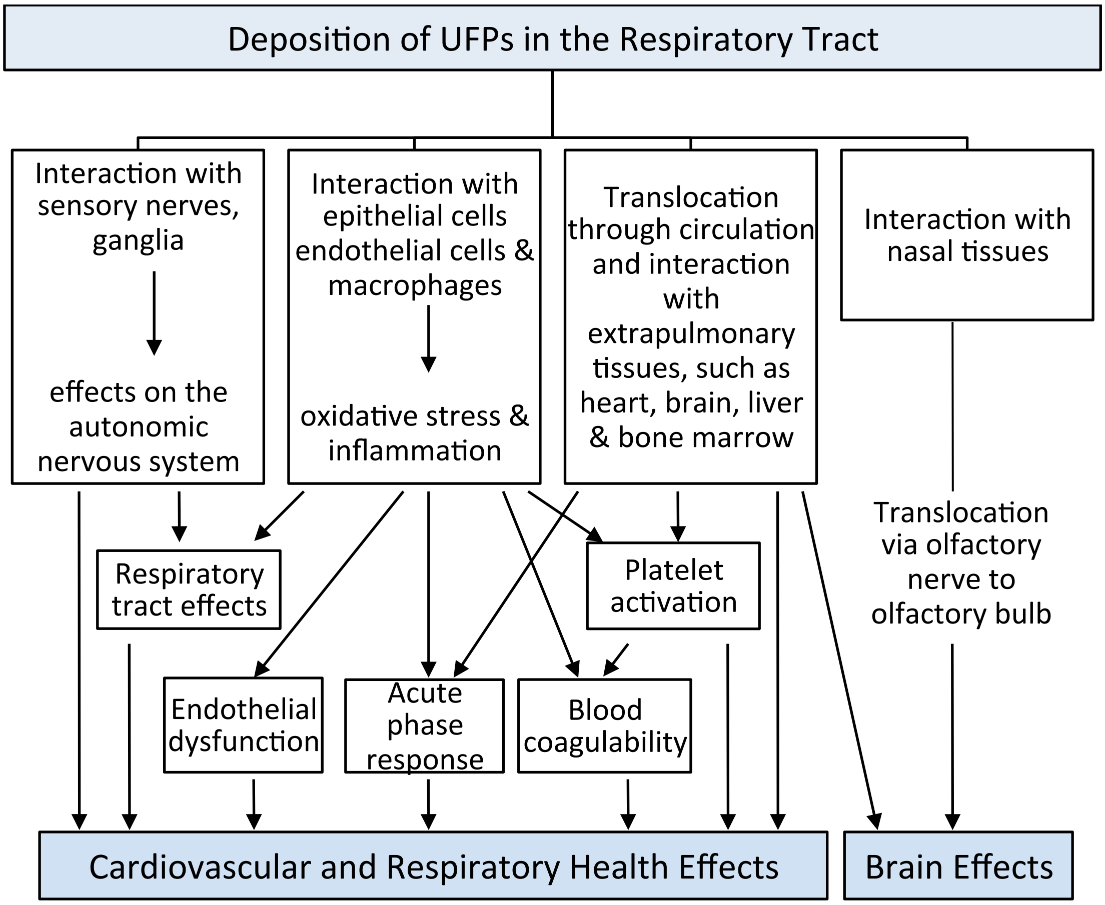

Deposition of UFPs in the Respiratory Tract
Interaction with sensory nerves, ganglia
effects on the autonomic nervous system
Interaction with epithelial cells endothelial cells & macrophages
oxidative stress & inflammation
Translocation through circulation and interaction with extrapulmonary tissues, such as heart, brain, liver & bone marrow
Interaction with nasal tissues
Translocation via olfactory nerve to olfactory bulb
Respiratory tract effects
Platelet activation
Endothelial dysfunction
Acute phase response
Blood coagulability
Cardiovascular and Respiratory Health Effects
Brain Effects

Figure 6. *Hypothesized pathways via which UFPs may cause health effects, adapted from HEI (2013)*

As can be seen from Figure 6, it is hypothesized that oxidative stress, inflammation, particle translocation, respiratory reflexes and increased blood coagulability are among the mechanisms that may be responsible for negative health effects associated with UFP exposure. Some health effects may be caused by series of processes while others are linked to a certain mechanism, as is the case with particle translocation to the olfactory bulb, which may cause neurological effects.

The conceivable health effects are studied with the help of controlled animal studies, experimental studies with humans and epidemiological studies. Studies with animals suggest that UFP exposure induces airway inflammation at very high concentrations, although maybe not at commonly experienced levels. UFP exposure may also enhance allergic responses and provide for the progression of atherosclerosis. Inflammatory responses in the brain of some animals have also been observed. However, simultaneous exposure to fine particulate matter, different responses in different species and the general limitations of laboratory studies complicate the interpretation of the results. (HEI 2013)

Experimental studies with humans show inconsistent findings. Some studies show reductions in lung function and increase in airway inflammation while others do not show any pulmonary effects at all. Similarly, cardiovascular responses vary between studies, which have explored in particular vascular function, heart rate variability, cardiac repolarization, and blood coagulation. The short duration of exposure, small sample size and other limitations may mirror the diversity of the findings. (HEI 2013)

Like experimental studies, epidemiological studies have been likewise inconsistent in their findings regarding health impacts from exposure to ultrafine particles. Nonetheless, there is suggestive evidence that short-term exposures to ambient UFPs may increase acute mortality, i.e. mostly cardiovascular mortality, as well as morbidity from respiratory and cardiovascular diseases. As of 2013, no epidemiological studies of long-term exposure to ambient UFPs had been conducted. (HEI 2013)

As an attempt to synthesize accumulated, yet contradicting knowledge on the health impacts of UFP exposure, an expert elicitation was formed in 2009. This group of twelve European epidemiologists, toxicologists and clinicians rated how likely they regarded the existence of an independent causal relationship between increased short-term UFP exposure and any given hypothesized health endpoint. All-cause mortality, hospital admissions for cardiovascular and respiratory diseases, the aggravation of asthma, and decrease in lung function received medium to high ratings by most experts. (Knol et al. 2009)

When it comes to long-term exposure to UFPs, the likelihood of a causal relationship with all-cause mortality, cardiovascular and respiratory morbidity and lung cancer were rated mostly medium by the expert elicitation (Knol et al. 2009). Since these types of health effects are possible, it is important to assess the association between long-term exposure to UFPs and various health endpoints. To date, such studies have not been published mainly due to difficulty in assessing annual exposures for various study groups. Therefore exposure assessment is an important step forward. Methods to assign long-term exposures to the participants of cohorts are presented in the following chapter.

4 THEORETICAL FRAMEWORK FOR MODELING AND EXPOSURE ASSESSMENT

The concentration of ultrafine particles may be estimated by several methods, which are reviewed in this chapter. After reviewing modeling alternatives, land-use regression is described in detail since it is the utilized method. Further to this, exposure assessment is explained in this chapter in order to shed light on all the methods that are utilized in this thesis.

4.1 Introduction to Modeling

As described in Chapter 3, there are several considerations with regard to the assessment of UFP exposure. Firstly, exposure means the cumulative concentration experienced in several microenvironments over a period of time. These microenvironments may have very different UFP concentration levels due to the fact that particle numbers vary temporally and spatially, even within one city. Secondly, while it is of interest to assess the health effects from exposure to ambient particulate matter, people spend a considerable amount of their time indoors. Due to infiltration however, concentration data from outdoors may be used as somewhat reasonable proxy for traffic-related UFP exposure.

In order to assess exposure to ultrafine particles, personal monitors might be utilized. Their benefit is the ability to measure exposure in different microenvironments but they are not feasible in epidemiological studies where cohorts may consist of thousands of people. Instead, exposure must be assessed indirectly. Central monitoring sites are used for pollutants, which are dispersed somewhat homogenously over a city but that approach is not realistic for ultrafine particles (Hoek et al. 2008a). Instead, concentrations at different locations of a city may be assessed with the help of geostatistical methods, regional dispersion models or land-use regression models, as described in the text that follows.

Geostastical methods include various interpolation methods such as kriging, triangulation and inverse distance weighing. These models require a set of monitoring sites that are used in predicting concentrations at unsampled sites. Kriging is the most common method used in air pollution research as it has the advantage of producing not only predicted values but also their standard errors. Other interpolation methods do not produce estimates on statistical errors. (Jerrett et al. 2004)

Dispersion models for particulate matter are numerous and include e.g. Gaussian models, Langrangian/Eulerian models, as well as models utilizing computational fluid mechanisms or aerosol dynamics. They utilize data on emissions, meteorological conditions and topography. Conservation of mass is typically assumed at each time step. Consequently, dispersion models are most useful in predicting mass concentrations. If particle number concentration is to be modeled, specific care with regard to particle chemistry and atmospheric dynamics must be taken. Efforts to incorporate these parameters have thus far produced models that have not been able to accurately predict particle number concentrations. (HEI 2013; Jerrett et al. 2004; Holmes et al. 2006)

Land-use regression (LUR) is another attempt to model particle number concentrations within a city. LUR utilizes a spatially dense network of measured air pollution data and variables derived from geographic information systems (GIS). In LUR, statistical modeling is used so as to determine what type of geographic information correlates with the measured concentrations. Concentrations outside of measurement sites are then predicted with the help of site-specific geographic characteristics. (Eeftens et al. 2012)

Land-use regression has been shown to generally outperform geostatistical methods, while comparisons with dispersion modeling suggest approximately equal performance (Hoek et al. 2008a). Considering this and the need to model number concentrations, land-use regression was applied in this thesis. The method is described in detail in the following subchapters.

4.2 Land-Use Regression

As described, land-use regression models utilize a spatially dense network of measured concentration data and variables derived from geographic information systems (GIS). These variables are also called predictor variables. The measurement of ultrafine particles was reviewed in Chapter 2.5, whereas the calculation and utilization of predictor variables are described in the text that follows.

4.2.1 Geographic Information Systems in Land-Use regression

There are several books about GIS that describe how it can be utilized in various analyses. Briefly, GIS is a computer system for managing spatial data. The data is often restricted to two spatial dimensions and mapped with the help of geographic coordinates. The functional capabilities of GIS include e.g. data manipulation, combination, transformation, visualization, query, analysis and modeling. In land-use regression, only data combination and query are utilized, whereas modeling is done separate from GIS. (Bonham-Carter 2014)

Land-use regression can involve the utilization of several datasets as long as they may be merged together using coordinates as a link. Predictor variables are computed from this data, typically as buffers of various radii around each measurement site. The selection of buffer size should ideally reflect known dispersion patterns. Further to this, buffer size is instrumental in determining the performance of the LUR model. (Hoek et al. 2008a)

The assortment of predictor variables is dependent on the availability of data and the features of the study area. Various land-use regression studies have incorporated between 55-140 different predictor variables. Typical predictor variables have included e.g. population density, land-use and several traffic-related variables as well as sometimes meteorology and altitude. In addition to buffers of various sizes, some predictor variables may express the distance to the nearest air pollution source. (Hoek et al. 2008a; Eeftens et al. 2012)

In their meta-analysis, Hoek et al. (2008a) pointed out that various land-use regression models have been developed with little attention to problems associated with geographic datasets. Some of the identified issues include accessibility, completeness and precision as well as varying data compilation periods. The latter was said to be a potential issue in retrospective exposure assessment.

4.2.2 Model Development

Mathematically, land-use regression is an application of linear regression, which is a well-known modeling method. There are several books about linear regression, which describe the method in detail. Briefly, land-use regression for ultrafine particles compares to multiple linear regression model, where the concentration of ultrafine particles (C_{UFP}), as the dependent variable, is regressed against predictor variables, denoted by X in the equation below:

$$C_{UFP} = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \beta_0. \qquad (1)$$

The intercept term is β_0 and all other betas express the rate of change in concentration for a unit change in the respective predictor variable. (Chatterjee et al. 2013)

The regression model is fitted using ordinary least-squares (OLS) method, which minimizes the residual sum of squares while estimating the true regression line. The object of regression is to find a set of variables that best explain the variability of measured concentrations. This is measured with the adjusted coefficient of determination (adjusted R^2). Like R-squared, it measures the goodness of fit but also adjusts for the number of variables in the model as not to inflate the explanatory power of the model. (Chatterjee et al. 2013)

After the model with the highest adjusted R^2 is found, the results should be evaluated with respect to significance, multicollinearity and influential observations. In addition, it should be assessed whether the final model complies with the OLS regression assumptions. These include the normality of regression residuals; constant variability of the residuals, i.e. homoscedasticity; expected value of 0 for all residuals; and lack of spatial autocorrelation among the residuals. Compliance with the first three is assessed with the help of several residual plots, which present the error term, i.e. difference between true PNC and estimated PNC, in various ways. The last is assessed with Global Moran's I, which evaluates whether spatial patterns are clustered, dispersed, or random. (Anselin et al. 1991; Eeftens et al. 2012; Chatterjee et al. 2013)

Studies by Hoek et al. (2010) and Eeftens et al. (2012) have successfully utilized land-use regression in estimating concentrations of various pollutants. Hoek and colleagues were also first to utilize LUR for ultrafine particles although they did not utilize the model for any cohort. This thesis follows the procedures developed in these previous studies. Namely, supervised stepwise regression is used to develop the models, as described and applied in Chapter 6.

4.2.3 Validation of Land-Use Regression Models

Land-use regression models are not only tested against OLS regression assumptions but they must also be validated with regard to their ability to predict concentrations at unmeasured sites. This can be done with the help of data that was not used in developing the model. However, such data often does not exist and the use of other validation methods comes into question. These include leave-one-out cross validation, K-fold cross validation and holdout validation.

The most commonly utilized method in land-use regression is leave-one-out cross-validation (LOOCV) where a new model is developed with n-1 sites and the predicted concentration at the left-out site is compared with measured concentration at that site. This procedure is repeated n times. To measure the performance of the model, overall goodness of fit, i.e. R-squared, is calculated. Usually the structure of the models remains constant, i.e. predictor variables do not change from model to model. (Hoek et al. 2008a)

In holdout validation (HV) the approach is to divide the original dataset into two so as to create a new model based on one subset and validate it with the other. These datasets are also called training dataset and test dataset, respectively. Evaluations based on holdout validation may rely heavily on how the subsets are formed. (Hoek et al. 2008a; Schneider et al. 1997)

In K-fold cross validation original dataset is partitioned into k subsets, and the holdout method is repeated k times. Each time, one subset is used as the test set while the others are put together as a training dataset. Thus the method is a combination of leave-one-out cross validation and holdout validation. (Schneider et al. 1997)

Since LOOCV is an overly optimistic validation method in LUR models that are developed with a small number of observations (Wang et al. 2012), this thesis opts for holdout validation. That is, the original dataset is divided into two a number of times and new models are developed with partial data. These new models are then validated with unused data. This is done by the means of predicting PNC at unused sites and then regressing these predictions against measured PNC. Utilization of the described method is available in Chapter 6.

4.2.4 Notions about Land-Use Regression

Although the design and execution of a measurement campaign was not a part of this thesis, some important notions about measurement campaigns should be made on the grounds of completeness of the theory as well as interpretation and usefulness of the LUR models. Bearing this in mind, the most important design issues in a measurement campaign are the number and distribution of measurement sites as well as the number and allocation of measurement days.

When it comes to the number and distribution of measurement sites, there is no definite methodology that should be followed. Typically, researchers aim to maximize the contrast in predictor variables, e.g. by measuring concentrations near and far away from pollution sources (Hoek et al. 2008a). For instance, Hoek and colleagues (2011) utilized data where 50 measurement sites were divided into traffic and background sets.

Sufficient number of measurement sites is affected by local geography and the size of the city (Hoek et al. 2008a). While results from the Spanish city of Girona suggest that LUR models should be based on over 80 sites (Basagaña et al. 2012), studies conducted in Oslo and Toronto did not find significant differences between models of 40 and 65 sites as compared to those with 80 and 94 sites, respectively (Hoek et al. 2008a). Previous LUR-models for ultrafine particles have been developed with 46-80 sites (Hoek et al. 2010; Abernethy et al. 2013). LUR models in this thesis were developed with 46 and 43 sites as shown in Chapter 6.

As pointed out by Hoek and colleagues (2008), another consideration in the applicability of a land-use regression model is the number of measurement days during the monitoring campaign. Atypical weather conditions may distort the results even if measurements have been carried out periodically over four seasons. However, 60 days is considered a sufficient number for measuring PM_{10} for regulatory purposes in the USA (EPA 2014). This thesis utilized data on measurements that were performed non-simultaneously in fifty locations, each of which was measured for 7 days.

4.3 Exposure Assessment

The foundation for exposure assessment has been established in the previous text. Briefly, particle number concentration at cohort members' home addresses can be predicted using land-use regression models. Predicted concentrations may then be used as a proxy for personal exposure to ultrafine particles. However, it should be kept in mind that some particles are lost during indoor penetration and that the actual exposure is not at the level of the outdoor concentration. In addition, predicting individual exposures based on concentration at home does not reflect the fact that people move around the city during their days. However, this is a problem for all exposure assessment methods except personal monitoring or biomonitoring (Hoek et al. 2008a).

Application of the land-user regression models is straightforward. When predictor variables are known at the addresses of interest, these variables may be inserted into regression functions in order to obtain a prediction of PNC at that site. Some variable values may have to be truncated in case they are more extreme than the values used in creating land-use regression models. This is to ascertain that relationship between model variables stays linear. (Wang et al. 2014)

After personal exposures are assessed for the participants of a cohort, the results may be utilized in a medical study so as to assess whether exposure is associated with adverse health effects such as cardiovascular mortality. Since findings from epidemiological studies are intended for publication in peer-reviewed journal articles, such analysis is not presented in this thesis. Nonetheless, application of the described exposure assessment method is presented in Chapter 7.

5 MATERIALS

The development of land-use regression models requires geographic information and data on measured concentrations. Exposure assessment in its turn requires information on cohort addresses. This chapter presents these materials along with how they were obtained.

5.1 Measurement Data

The annual mean particle number concentrations were acquired from a 2011 study, which was conducted by Hoek and his colleagues. They in turn relied on measurement data that was collected by Puustinen et al. (2007). The data was available for 50 sites within the city of Amsterdam in the Netherlands.

Details about the measurement campaign have been published before. Briefly, Puustinen et al. (2007) measured particle number among other pollutants directly outside of 50 homes in Amsterdam between October 2002 and February 2004. The sites were divided into 22 traffic sites and 28 background sites. At all sites, the aim was to measure 24-hour average concentration within a period of one week. The measurements were not done simultaneously in different locations due to the limited availability of equipment. However, measurements were continuous at an urban background site.

Particle number measurements were done using TSI’s condensation particle counter model CPC 3022A following standard operating procedures. According to the manufacturer’s spec sheet (TSI 1999), the utilized particle counter is run with supersaturated butanol that condenses onto sample particles in order to produce larger and more easily detectable droplets. These droplets are then counted with an optical detector. When the concentration is below 10,000 particles per cubic centimeter, the detector counts individual pulses produced by passing particles. Higher concentrations are measured by detecting the intensity of scattered light. The particle counter detects particles down to 7nm in diameter.

Using the data that Puustinen and colleagues (2007) collected, Hoek et al. (2011) calculated site-specific annual mean concentrations of measured pollutants. First they subtracted measured 24-hour concentrations from the simultaneously measured concentration at the urban background site. In case Puustinen and colleagues succeeded in their measurements every day, there were seven 24-hour measurements per site. However, the number of successful measurement days varied.

After subtraction, the arithmetic differences of concentrations between the two measurement sites were averaged, i.e. differences were summed up and divided by the number of successful measurement days. The overall annual mean concentration at the urban background was then added to the average difference to obtain an estimate for annual mean concentration at the measurement site. This resulted in concentrations that ranged from approximately 12,200 to 87,000particles/cm^3.

Utilizing Esri's ArcGIS software, the 50 measurement sites may be plotted on a map. This results in a visual representation of the measurement sites that can be seen in Figure 7 below. Red dots symbolize the 50 locations where the measurements were carried out. The slightly bigger black dot represents the urban background site.

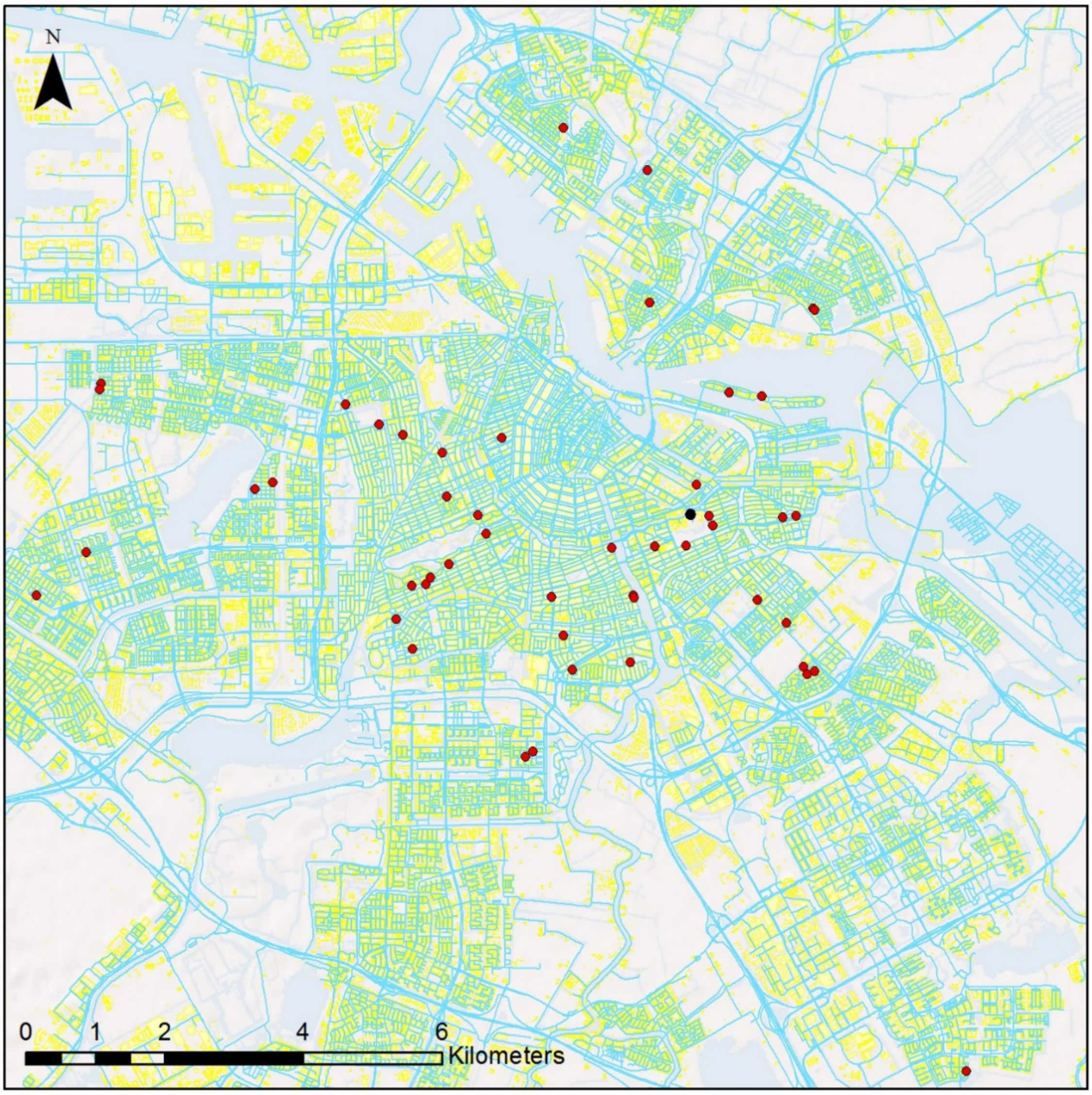


Figure 7. *Distribution of the measurement sites within the city of Amsterdam*

As can be seen from Figure 7, measurement sites encircled the Amsterdam city centre, where housing is less prominent. Generally, measurement locations were chosen so as to cover a large amount of different types of sites. For instance, sites were located near different types of geographical features, such as the River Amstel, canals, parks and various kinds of streets.

5.2 Geographic Information

Geographic information was obtained from the same datasets, which were successfully utilized in the European Study of Cohorts for Air Pollution Effects (ESCAPE) project. The datasets included European Environment Agency's (EEA) Corine land cover 2000, Eurostreets version 3.1 road network data, as well as Dutch national road database (Nationale Wegen Bestand, NWB). Details about these datasets are presented in the Table 1 below and in the following text.

Table 1. *Sources of geographic information*

Dataset	Description	Positional Accuracy	Year of compilation
CORINE 2000	Land cover data	Better than 100m	2000
Eurostreets v3.1	Central road network	5-12m	2008
NWB	National road network with linked traffic intensities	~10m	2008
Population	Population density	N.A.	2001

As can be seen from Table 1, the accuracy and the compilation year of different datasets varied. The positional accuracy of CORINE 2000 is less than that of road network data and the worst possible accuracy of about 100 meters means that land cover data should not be used with small buffer zones. On contrary, road network data is quite accurate in both datasets. The years of compilation are acceptable for this study, since the city plan of Amsterdam did not change considerably during the first decade of the 2000s.

In order to give background information on available data, the datasets are next described in detail. Firstly, CORINE (Coordination of information on the environment) is a program run by the European Commission in order to provide information on land use for policy makers and other interested stakeholders. The European Environment Agency (EEA) maintains Corine land cover (CLC) database, which distinguishes 44 different land cover classes. The classes are grouped in a three-level hierarchy, where the main classes are 1) artificial surfaces, 2) agriculture areas, 3) forests and semi-natural areas, 4) wetlands, and 5) water bodies. (EEA 2002)

Secondly, Eurostreets version 3.1 is based on a commercial TeleAtlas MultiNet TM dataset. Tele Atlas is a wholly owned subsidiary of the Dutch automotive navigation system manufacturer TomTom. The attributes of Eurostreets include the name of the street, functional road classification, route number, speed limits and length. (Eeftens et al. 2012; Spatial Insights 2014)

Next, the Dutch national road database is a network that consists of intersections connected by road sections. NWB integrates several types of different data such as traffic intensities and road crashes. In addition to regular roads, also all separate footpaths, bicycle tracks and unsurfaced roads are included in the database in case they have a street name. (SWOV 2014)

Finally, population density data was available from the Integrated Assessment of Health Risk of Environmental Stressors in Europe (INTARESE) Project. In this dataset population density – available from the EEA – is modeled in 100m grids across different European countries. (IEHIAS 2010)

5.3 Cohort

This thesis utilized the Monitoring Project on Risk Factors for Chronic Diseases (MORGEN) cohort, which is a Dutch contribution to the European Prospective Investigation into Cancer and Nutrition (EPIC). In short the cohort is referred to as EPIC MORGEN. The cohort was compiled by the Dutch National Institute for Public Health and the Environment (Rijksinstituut voor Volksgezondheid en Milieu) from 1993 to 1997. (Beulens et al. 2009)

EPIC MORGEN consists of a general population sample from the Dutch towns of Amsterdam, Doetinchem and Maastricht. A total of 50,766 people aged 20-59 years were invited to participate, while 22,769 people completed questionnaires and medical check-up that were prerequisites for inclusion in the cohort. Other details about the cohort have been published in the EPIC-NL cohort profile. (Beulens et al. 2009)

Since EPIC MORGEN cohort consists of participants from three Dutch towns, the cohort was restricted to those living in Amsterdam for the purpose of this study. There were 4,986 such cases. Further details about the utilization of the cohort in exposure assessment are presented in Chapter 7.

6 LAND-USE REGRESSION

In this chapter, available data is utilized in land-use regression modeling, which consists of several steps. Firstly, predictor variables are calculated and assigned to all measurement sites with the help of GIS software. This information is then reviewed with respect to accuracy. Secondly, land-use regression models are developed with the available data. Lastly, developed models are validated using the holdout method.

6.1 Assigning Predictor Variables to Measurement Sites

Assigning geographic information, i.e. calculating predictor variables at each measurement site was done with the help of Esri's ArcGIS software and Python scripts. First, all 50 measurement sites were plotted on an empty map in ArcGIS using X- and Y-coordinates based on the Dutch RD coordinate system. All sites were given unique identification numbers so that they could be called in different programs.

Then, plotted points and their metadata were imported in a geodatabase, which is a common data storage and management framework for ArcGIS. The information was imported in the vector format, which is provides for a more precise basis for calculating predictor variables as opposed to the raster format.

Next, geodatabase and all land-use datasets were processed in Python in order to assign predictor variables to each measurement site. Python scripts were developed previously as part of the ESCAPE project and therefore this step did not require any new programming. The scripts in question calculated predictor variables out of the baseline data, including distances to air pollution sources, such as distance to nearby roads, as well as various values of land-use data in a buffer, e.g. area of industrial land in a buffer of 100m. In contrast to studies published under ESCAPE, the 25-meter buffer for several traffic-related variables was rendered useless. This is due to the high uncertainty of geographic precision within that buffer. All other calculated predictor variables are presented in Appendix 1.

6.2 Adjustments to the Assigned Data

After geographic information was assigned to each site, the resulting dataset was combined with the concentration data using site identification number as a link between the two. This was to ascertain that the datasets were combined correctly. Then the resulting dataset was examined with respect to coverage and accuracy of data.

In case particular geographic characteristics were non-existent in over 20 sites, i.e. a predictor variable was given a value zero, the variable was removed from further analysis. This is due to the fact that LUR models may be developed only with such predictor variables that are widespread enough. The procedure led to the removal of high-density residential housing in every buffer since only four sites were assigned values in the category. This reflects the fact that housing in Amsterdam is predominantly low density. Also industrial, port and semi-natural areas within the buffers of 100, 300, 500 and 1000 meters were removed, as well as urban green area within the buffers of 100, 300 and 500 meters.

From traffic-related variables, traffic load on major roads within the buffer of 50 meters was removed as it had been assigned only 26 non-zero values. Two variables describing the road length of major roads in the buffers of 50 and 100 meters were neither included. Only 19 and 23 sites had been assigned values in these buffers respectively. All discarded predictor variables can be seen in Appendix 1.

Finally, for all sites, the value of traffic intensity on the nearest road was categorically replaced with the value of traffic intensity on the nearest major road in case these two roads were within 25 meters from each other. This was done because the major road is more important source of traffic-related pollutants than the quieter street. The same transformation was done for heavy-duty traffic intensity. The cut-off value of 25 meters was chosen because it is unlikely that adjacent streets within that distance would be located on different sides of a building block. From the sample of 50 sites, there were 10 sites where this transformation was done. At all of these sites, the roads passed the building block from the same or adjacent side as judging from Google Earth.

6.3 Descriptive Analysis

In order to find potential outliers and errors in the data, the predictor variables were analyzed further. Particularly, variables that describe the same or similar features were compared with each other. As only some traffic variables were directly comparable, the main interest was to assess whether similar conclusions could be drawn from them.

Assigned traffic variables were based on two different datasets, i.e. the Dutch national road network and the Eurostreets network as described in Chapter 5. In the Dutch national road network all roads with a daily traffic intensity of 5,000 vehicles or more were classified as major roads. In comparison, functional road classes were the basis for classifying major roads in the Eurostreets dataset. In this thesis, the road classes that were regarded as major roads included 0) motorways, 1) main roads of major importance, and 2) other main roads. The classification was the same as in ESCAPE.

The accuracy of the road network data may be assessed with the help of distance variables. NWB and Eurostreets correlated well with each other when it comes to the distance to the nearest road, as measured in meters. Accounting all 50 sites, Pearson's correlation coefficient was 0.89. In contrast, the coefficient was 0.57 with regard to the distance to the nearest major road. This demonstrates how different ways of classifying major roads has an effect on overall layout of the major road network.

Due to the fact that measurement sites were classified into traffic and background sites, higher traffic counts were to be expected on traffic sites than on background sites. This premise was examined by forming boxplots that represent traffic intensity for the two site types. These boxplots can be seen below.

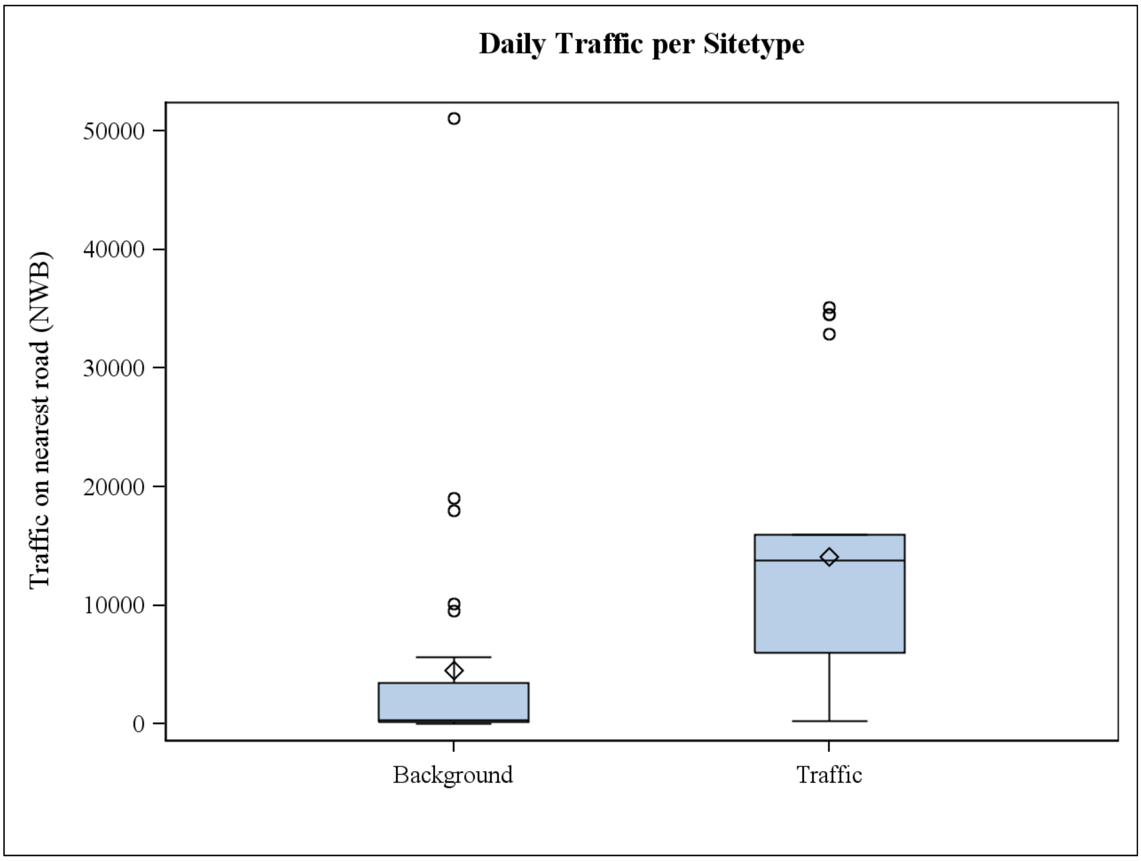


Figure 8. *Estimated traffic intensities at the measurement sites by site type*

The boxplots reveal several potential outliers in the data, i.e. points outside of whiskers. The maximum traffic intensity should be 5,000 vehicles per day at all urban background sites but in Figure 6 it is not the case. All points that are above this limit are considered erroneous, i.e. either site classification or traffic intensity is wrong. All traffic sites with traffic intensity of below 5,000 vehicles per day are similarly problematic. However, when it comes to traffic sites, several points well above the interquartile range cannot be considered outliers as the values conform to the definition of a traffic site, i.e. traffic intensity of over 5,000 vehicles per day without an upper limit.

All sites were further examined on ArcGIS and Google Earth in order to visually inspect if the assigned traffic intensities were reasonable given the layout of roads at the site. This inspection revealed several points where assigned traffic intensity was not a true representation of reality. Likely reason for erroneous data was the fact that some footpaths and bicycle tracks were included in NWB. This then confused modeled traffic intensity in parts of the city.

6.4 Data Selection for Land-use Regression Models

Similar to Hoek et al. (2011) only such sites were selected for further analysis where there were three or more successful measurement days for PNC or $PM_{2.5}$ within the one-week measurement campaign. Out of 50 sites, 46 fulfilled this criterion. These were therefore the sites with what land-use regression modeling could be started.

In addition to excluding sites based on the number of successful measurement days, some sites were removed because of the discrepancies between different traffic variables. Since high traffic volumes are imperative to air pollution, attention was paid to traffic on major roads.

Although the classification of major roads was different in Eurostreets compared with NWB, some similarities between the variables were to be expected. Mainly, the highest traffic intensities were anticipated on major roads due to their greater capacity as compared to regular city streets. It was assumed that only 0-2 class roads (Eurostreets major roads) could carry over 15,000 vehicles per day. In case NWB predicted over 15,000 vehicles per day on the nearest road and there was no such high-capacity major road nearby (a distance of less than 25m), the site was excluded from modeling.

The described procedure removed three more sites from further analysis. These sites were concurrently the three background sites with unrealistically high traffic intensities in Figure 6. Removed sites were finally inspected in Google Maps so as to confirm that assigned traffic intensities in these sites were indeed incorrect.

Land-use regression based on the exclusion of sites due to lack of successful measurement days constitutes Model A, whereas the exclusion of three more sites due to discrepancies in traffic data concerns only Model B. Developing these two models provided a form of sensitivity analysis with respect to how the number of measurement sites and unrealistic traffic intensities affect model performance. Details about the site-specific particle number concentration of these two models are presented in the table on the next page.

Table 2. *Particle number concentration per model and site type*

Model	Site	N	min	P10	median	P90	max
A	Traffic	20	22064	26436	40353	70543	86902
	Background	26	12248	13289	22359	32179	46633
B	Traffic	20	22064	26436	40353	70543	86902
	Background	23	12248	13289	21832	31655	46633

As can be seen from Table 2 above, the removal of three sites from Model A decreased the median and 90th percentile (P90) particle number concentrations but the extreme values and 10th percentile (P10) remained the same. Since traffic sites in both models were exactly the same, the difference between the two models is limited to the distribution of PNC at the background sites.

6.5 Development of Land-use Regression Models

As described in Chapter 4, land-use regression models were developed using supervised stepwise regression. In the first step, univariate regression analyses were conducted for all potential predictor variables (see Appendix 1 for the complete list), i.e. each predictor variable was regressed against measured particle number concentrations. The predictor variable that explained the largest percentage of variability of measured concentrations (as indicated by adjusted R^2) was recorded. It was then entered as the first variable in the model.

Next, more variables were introduced into the model in the order of the highest additional increase in adjusted R^2. This was done only if the increase in adjusted R^2 was more than 1%. Further to this, the variable had to conform to the direction of effect, i.e. the regression slope, defined a priori. Urban green and semi-natural areas were regarded as not a source of UFPs and therefore they were assigned with a minus sign a priori. All other variables were assigned with a plus sign a priori. The a priori direction of effect for variables already included in the model was not supposed to change when new variables were introduced into the model.

New variables were added until the increase in adjusted R^2 was below the 1-% threshold. After that, model variables were examined with respect to their significance and multicollinearity. Eeftens et al. (2012) established limits for respective p-values and variance inflation factors (VIFs), which were followed in this thesis. That is, in case a variable p-value was over 0.10 or VIF was over 3, the variable in question was removed from the model.

Then the model was scrutinized with respect to influential observations and no Cook's distance (Cook's D) of over 1 was allowed. Finally, the analysis of residuals was performed in order to assess whether the model complied with the OLS regression assumptions, which were presented in Chapter 4.

The models were developed in SAS Institute's SAS software with the help of REG procedure where PNC was entered as the dependent variable and different predictor variables were entered as independent variables. The selection method was based on adjusted R^2. Details about the model development are presented in the following two sections.

6.5.1 Model A

In total of 46 observations with their predictor variables were used to develop the land-use regression Model A. A short SAS script utilizing the regression procedure of the software was developed for this purpose, and an excerpt of the script is described below:

```
PROC REG;
MODEL PNC = <predictors> / START=1 STOP=1 SELECTION=ADJRSQ;
RUN;
```

The script describes the first step of regression. SAS commands are in bold, and to save space, the whole list of predictor variables is substituted with <predictors>. In options, START and STOP together designate that only one variable is to be entered into the model in the first step. The selection method is based on adjusted R^2 as described before.

After running the script, the most predictive single predictor variable with adjusted R^2 of 0.3638 was judged to be inverse distance to the nearest major road (*DISTINVMAJORC1*) as calculated from the Eurostreets data. Its univariate parameter estimate was positive 360087, which conformed to the a priori defined direction of effect. The variable was therefore entered as the model's first.

After the first predictor variable was found, modeling continued onto multivariate modeling. This was done by keeping the first variable constant and testing which new variables increased model performance the most. The following excerpt describes the next step in SAS:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 <predictors> /
BEST=5 INCLUDE=1 START=1 STOP=2 SELECTION=ADJRSQ;
RUN;
```

In step two <predictors> comprised of all predictor variables except *DISTINVMAJORC1*, which came out as the first variable in the initial step. The variable was therefore included in all potential models with INCLUDE=1. In the above script STOP denotes that model may comprise two variables and BEST is there to simply limit the amount of regression outcomes to five best.

Table 3 shows the first five variables that could be added into the model based on the additional increase in adjusted R^2. First on the list is urban green area in the buffer of 5000 meters (*URBGREEN_5000*). It increases the explained adjusted R^2 by more than 10% to 0.4006. The parameter estimate of -0.00187 conforms to the a priori direction of effect and the sign of the regression slope for the first predictor variable does not change. Therefore, the variable passed the criteria and it was added into the model.

***Table 3**. Potential predictor variables for Model A in step 2*

Variable Count	Adjusted R-Square	Last Variable in the Model	Parameter Estimate	A priori direction of effect
1	0.3638	*DISTINVMAJORC1*		
2	0.4006	*URBGREEN_5000*	-0.00187	-
2	0.3910	*HEAVYTRAFLOAD_300*	-0.00429	+
2	0.3898	*HEAVYTRAFLOAD_500*	-0.00144	+
2	0.3885	*HEAVYTRAFMAJORLOAD_300*	-0.00423	+
2	0.3876	*MAJORROADLENGTH_300*	-5.25968	+

After adding *URBGREEN_5000* into the model, particle number count was regressed against urban green in the buffer of 1000 meters. This was done so as to examine whether a smaller buffer size of the included variable could provide additional value. The SAS script for this trial can be written as below:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 URBGREEN_5000 URBGREEN_1000;
RUN;
```

Adding urban green in 1000-m buffer penalized the model's adjusted R^2 by almost 1 percentage point. Therefore modeling was continued without the variable. Next step in SAS followed the script as follows:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 URBGREEN_5000 <predictors> /
BEST=5 INCLUDE=2 START=2 STOP=3 SELECTION=ADJRSQ;
RUN;
```

In step three <predictors> included all predictor variables except *DISTINVMAJORC1* and *URBGREEN_5000*, which were obtained in the first two steps. These variables were incorporated into all potential models with INCLUDE=2 as defined in the script options above. STOP denotes that model could consist of three variables.

The next five candidate variables are presented in Table 4 below. First on the list is population density within the 5000-meter buffer (*EEA_5000).* The variable increases the model's adjusted R^2 by some 9% to 0.4367. Further to this, the parameter estimate of 0.01654 conforms to the predefined direction of effect. As the directions of effect of preceding variables do not change, the variable *EEA_5000* was added into the model.

Table 4*. List of potential predictor variables for Model A in step 3*

Variable Count	Adjusted R-Square	Last Variable(s) in the Model	Parameter Estimate	A priori direction of effect
2	0.4006	*DISTINVMAJORC1, URBGREEN_5000*		
3	0.4367	*EEA_5000*	0.04882	+
3	0.4345	*HEAVYTRAFLOAD_500*	-0.00151	+
3	0.4315	*HEAVYTRAFMAJORLOAD_500*	-0.00429	+
3	0.4301	*LDRES_5000*	0.00045	+
3	0.4293	*HEAVYTRAFLOAD_300*	-0.00429	+

As in the second step, smaller buffers of the newly included variable were presented into the model in case they were to provide additional value. However, none of the smaller buffers increased the model's adjusted R^2 and therefore they were not included in the model. Modeling proceeded to step four, the SAS script of which is presented below:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 URBGREEN_5000 EEA_5000 <predictors> /
BEST=5 INCLUDE=3 START=3 STOP=4 SELECTION=ADJRSQ;
RUN;
```

Like in the earlier steps, previously included variables were excluded from the <predictors>. Further to this, corresponding adjustments were made into the model options. Namely, the number of variables was controlled with INCLUDE, START and STOP.

As can be seen from Table 5 in the next page, the first candidate to be added was the heavy-duty traffic load of major roads in a 300m buffer. It fulfills the criteria of increasing the model's adjusted R^2 over 1% but falls short on conforming to the a priori direction of effect. Thus it cannot be introduced into the model. The same applies to all the following heavy-duty traffic variables. However, the variable describing traffic intensity on the nearest road (*TRAFNEAR*) does fulfill all criteria and it was therefore added into the model.

Table 5. *List of potential predictor variables for Model A in step 4*

Variable Count	Adjusted R-Square	Last Variable(s) in the Model	Parameter Estimate	A priori direction of effect
3	0.4367	*DISTINVMAJORC1, URBGREEN_5000 EEA_5000*		
4	0.5140	*HEAVYTRAFMAJORLOAD_300*	-0.00668	+
4	0.5102	*HEAVYTRAFLOAD_300*	-0.00636	+
4	0.5040	*HEAVYTRAFMAJORLOAD_500*	-0.00208	+
4	0.5036	*HEAVYTRAFLOAD_500*	-0.00204	+
4	0.4502	*TRAFNEAR*	0.23239	+

After adding *TRAFNEAR*, there were in total 18 candidate variables in step 5 that would have increased the model's adjusted R^2 over 1% but none of these conformed to the a priori direction of effect. Therefore no additional variables entered the model A.

After all predictor variables were found, the p-value and variance inflation factor of each were calculated and examined. Since the p-value of TRAFNEAR was 0.1617 and over the cut-off limit of 0.10 the variable was removed from the final model. No variable was removed because of a high VIF value (over the cut-off of 3). There were also no influential observations in the model as all Cook's D values were less than 0.30 -- well below the threshold of 1 -- as can be seen from the model diagnostics in Appendix 2. Details about the final model are presented in Table 6 below:

Table 6. *Parameter Estimates for Model A*

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	33447	9549	3.50	0.0011	0
DISTINVMAJORC1	1	270214	76096	3.55	0.0010	1.34825
URBGREEN_5000	1	-0.00287	0.00107	-2.67	0.0107	1.32021
EEA_5000	1	0.04882	0.02520	1.94	0.0595	1.55413

As presented, Model A predicted particle number concentration as a function of inverse distance to the nearest major road (Eurostreets), urban green area in the 5000-meter buffer, and population density within a buffer of 5000 meters. The model explained approximately 44% of the variability in measured PNC. The equation, i.e. Model A, for predicting PNC is the following:

$$PNC = 270214\, DISTINVMAJOR_{C1} - 0.00287\, URBGREEN_{5000} + 0.04882\, EEA_{5000} + 33447 \quad (2)$$

The calculated coefficients express the rate of change in PNC for a unit change in respective variable when all other variables are kept constant. In case all predictor variables were zero or close to zero, estimated PNC would be about 33,000/cm^3. It is slightly higher than the overall median PNC of 27,000/cm^3 of the Model A sites. This reflects the importance of urban green areas in decreasing particle number concentrations.

In addition to tests done above, the model's compliance with OLS regression assumptions was checked as described in Chapter 4. Firstly, the normality of the residuals was verified from QQ-plot and histogram, which are presented in Appendix 2. Then the variability of the residuals was inspected so as to assess whether the model was based on homoscedastic error. This was to ascertain that standard errors and consequent significance tests were not biased. The heteroscedasticity test was done visually with the help of a residual plot generated in SAS, as presented below.

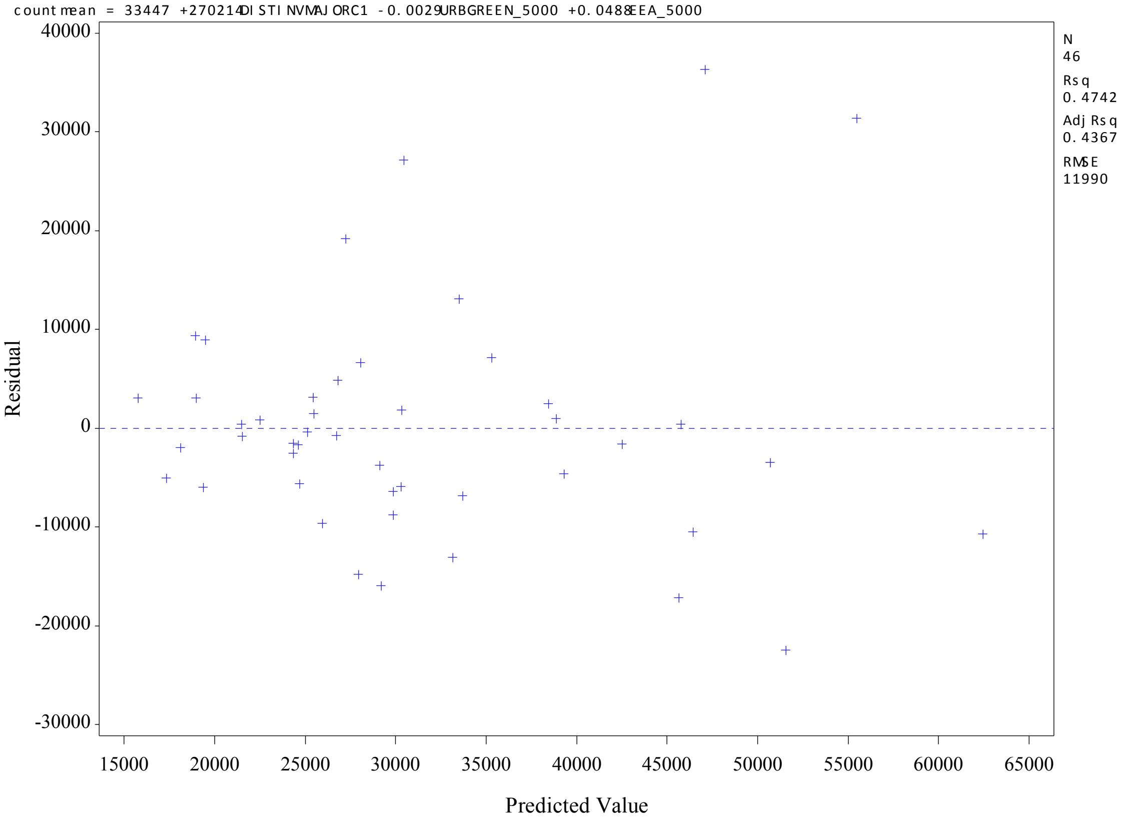


***Figure 9**. Residual plot of Model A*

As can be seen from Figure 9, the variability of residuals does not depart from the expected value 0 in any systematic manner. This implies that the regression function is computed correctly and is indeed linear. Another remark in Figure 9 is a slight fanning effect, i.e. the random variation of residuals seems to increase with increasing values on x-axis, indicating some heteroscedasticity. However, predicted values flocked towards the lower end of the scale and studentized residuals (available in Appendix 2) suggested that the four most extreme values of residuals in Figure 7 might have been outliers (|RStudent|>2). One of these outliers was related to an observation absent in Model B. Based on these findings, the model was considered acceptable.

As a final step in the analysis of residuals, Global Moran's I was calculated in order to assess spatial autocorrelation among regression residuals so as to validate the test of heteroscedasticity and the estimated value of adjusted R^2. Moran's I was calculated in SAS, where the following script was utilized:

```
PROC VARIOGRAM;
  COMPUTE NOVARIOGRAM AUTOCORRELATION (WEIGHTS=DISTANCE);
  COORDINATES XC=x YC=y;
  VAR RESIDUAL;
RUN;
```

The variogram procedure is used in analyzing spatial data. Autocorrelation statistics are requested under normality assumption in the compute statement with AUTOCORRELATION. Every measurement site is linked with all the other sites using distance as a weight (WEIGHTS=*DISTANCE*). Due to this, there is no need to assign lag distances or maximum number of lags. Therefore NOVARIOGRAM is also written in the compute statement. Coordinates are given as they are in the source file and *RESIDUAL* is inserted as the variable being analyzed. The resulting autocorrelation statistics (Moran's I) are presented in Table 7 below.

***Table 7.** Autocorrelation Statistics for Model A*

Assumption	Coefficient	Observed	Expected	SD	Z	Pr > |Z|
Normality	Moran's I	0.0897	-0.0222	0.392	0.286	0.7751

From Table 7 above Moran's I of about 0.09 can be observed. It is fairly close to zero, which is the value when there is no spatial clustering present. Associated Z-score and its p-value likewise suggest that there is no spatial autocorrelation among regression residuals (small Z-score, large p-value). Considering this and the other tests presented before, Model A was considered acceptable in terms established land-use regression methodologies. Further validation of the model is presented in subchapter 6.6.

6.5.2 Model B

In total of 43 observations with their predictor variables were used to develop the land-use regression Model B. As with Model A, SAS scripts were developed to find the predictor variables that best explain the variability in measured PNC. The first excerpt of the script can be seen below:

```
PROC REG;
MODEL PNC = <predictors> / START=1 STOP=1 SELECTION=ADJRSQ;
RUN;
```

The most predictive single predictor variable with adjusted R^2 of 0.3659 was found to be – like in Model A – inverse distance to the nearest major road as calculated with Eurostreets data. Its positive parameter estimate conformed to the a priori direction of effect. The variable was therefore entered as the model's first. Modeling continued as follows:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 <predictors> /
BEST = 5 INCLUDE = 1 START =1 STOP= 2 SELECTION = ADJRSQ;
RUN;
```

The script did not differ from that of Model A's step 2. As described earlier, *DISTINVMAJORC1* was excluded from <predictors> and adjustments were made into the model options. Table 8 below shows the five best candidate variables based on running the script.

***Table 8**. List of potential predictor variables for Model B in step 2*

Variable Count	Adjusted R-Square	Last Variable in the Model	Parameter Estimate	A priori direction of effect
1	0.3659	*DISTINVMAJORC1*		
2	0.4024	*TRAFNEAR*	0.44452	+
2	0.4016	*URBGREEN_5000*	-0.00194	-
2	0.3920	*HEAVYTRAFLOAD_300*	-0.00443	+
2	0.3906	*PORT_5000*	0.00108	+
2	0.3905	*HEAVYTRAFMAJORLOAD_300*	-0.00442	+

As can be seen from Table 8, the first candidate variable on the list is traffic intensity on the nearest road (*TRAFNEAR*). It increases the adjusted R^2 by approximately 10% to 0.4024. Also, the parameter estimate of 0.44452 conforms to the a priori direction of effect. As the direction of effect of the preceding variable does not change, the variable *TRAFNEAR* was added into the model.

In step three, modeling continued in the established manner. The SAS script took into account previously included variables in both the model statement and options. Finding the third variable followed the script as presented below:

```
PROC REG;
MODEL PNC = DISTINVMAJORC1 TRAFNEAR <predictors> /
BEST=15 INCLUDE=2 START=2 STOP=3 SELECTION=ADJRSQ;
RUN;
```

The output of the script above shows 15 best variables in terms of increase in adjusted R^2 as it was only the 15th candidate that made the cut, i.e. increased the adjusted R^2 by more than 1% and conformed to the a priori direction of effect. The variable in question is the sum of port areas within a buffer of 5000 meters (*PORT_5000*) as can be seen from Table 9 below. Adding *PORT_5000* increased the adjusted R^2 by some 9%.

***Table 9**. List of potential predictor variables for Model B in step 3*

Variable Count	Adjusted R-Square	Last Variable(s) in the Model	Parameter Estimate	A priori direction of effect
2	0.4024	*DISTINVMAJORC1, TRAFNEAR*		
3	0.5499	*HEAVYINTINVDIST*	-190.82774	+
3	0.5495	*HEAVYTRAFLOAD_100*	-0.06958	+
3	0.5402	*HEAVYTRAFMAJORLOAD_100*	-0.06735	+
3	0.5400	*HEAVYTRAFNEAR*	-18.63367	+
3	0.5347	*HEAVYINTINVDIST2*	-2249.11954	+
3	0.5307	*HEAVYTRAFLOAD_50*	-0.16571	+
3	0.5266	*HEAVYTRAFMAJORLOAD_50*	-0.16350	+
3	0.5130	*INTMAJORINVDIST2*	-198.64159	+
3	0.5039	*INTINVDIST2*	-157.54672	+
3	0.5014	*INTINVDIST*	-157.54672	+
3	0.5011	*HEAVYTRAFMAJOR*	-12.69289	+
3	0.4705	*INTMAJORINVDIST*	-16.96443	+
3	0.4443	*HEAVYTRAFLOAD_300*	-0.00516	+
3	0.4408	*HEAVYTRAFMAJORLOAD_300*	-0.00506	+
3	0.4391	*PORT_5000*	0.00122	+

As there were no smaller buffers of the port-variable to introduce, modeling proceeded directly to next step. In step 4, there were 16 candidate variables but none of these conformed to the a priori direction of effect. Therefore no additional variables entered the model. As seen in Table 10 below, all variables had a p-value below 0.10 and VIF below 3. Therefore none of the variables were removed from the model. Furthermore, there were no influential observations with a Cook's D of over 1 as can be seen in Appendix 3.

Table 10. *Parameter Estimates for Model B*

Variable	**Degrees of Freedom**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	15944	4185	3.81	0.0005	0
DISTINVMAJORC1	1	222384	88744	2.51	0.0165	1.69531
TRAFNEAR	1	0.48811	0.23127	2.11	0.0413	1.61848
PORT_5000	1	0.00122	0.00063	1.90	0.0646	1.06154

As presented, Model B predicted particle number concentration as a function of inverse distance to the nearest major road (Eurostreets), traffic on the nearest road and port area within a buffer of 5000 meters. All parameters were significantly different from zero and the predictors were not highly correlated with each other. The Model B, like Model A, explained about 44% of the variability in measured PNC. The equation, i.e. Model B, for predicting particle number concentration is the following:

$$PNC = 222384\, DISTINVMAJOR_{C1} + 0.48811\, TRAFNEAR + 0.00122\, PORT_{5000} + 15944 \qquad (3)$$

The calculated coefficients express the rate of change in PNC for a unit change in respective variable when all other variables are kept constant. In case all predictor variables were zero or close to zero, estimated PNC would be close to 16,000/cm^3. It is well below the median PNC of about 27,000/cm^3 of the Model B sites. This reflects the importance of traffic and port in increasing particle number concentrations.

As with the Model A, an analysis of residuals was performed for Model B. Firstly, the normality of the residuals was verified from QQ-plot and residual histogram, which are presented in Appendix 3. Then the variability of the residuals was examined so as to assess whether the model was based on homoscedastic error. This was to ascertain that standard errors and consequent significance tests were not biased. The heteroscedasticity test was done visually with the help of a residual plot, which is presented on the following page.

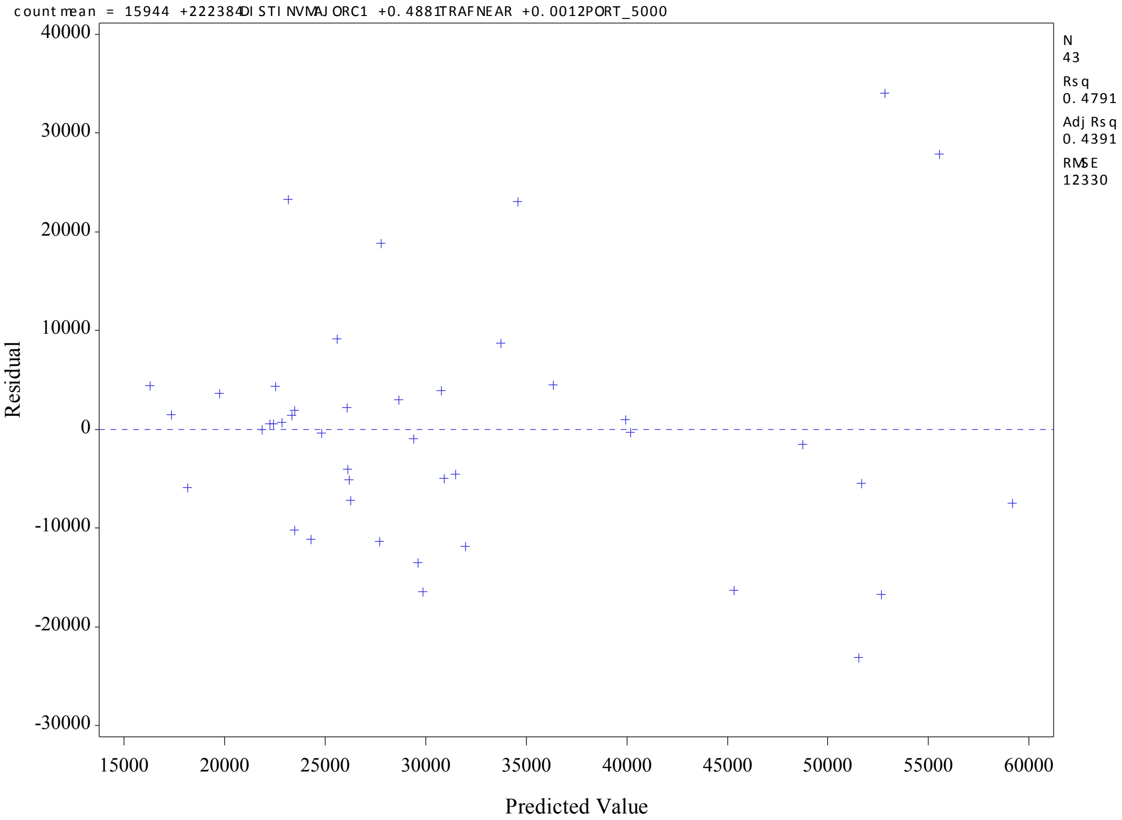


***Figure 10**. Residual plot of Model B*

As can be seen from Figure 10, the variability of the residuals does not depart from the mean value 0 in any systematic manner. This implies that the regression function is computed correctly and is indeed linear. Further to this, there isn't any systematic pattern to residuals in the plot. Based on these findings, the model was considered appropriate.

As a final step in the analysis of residuals, Moran's I was calculated in order to check spatial autocorrelation among regression residuals so as to validate the test of heteroscedasticity and the estimated value of adjusted R^2. Moran's I was calculated in SAS with a similar script as presented in Chapter 6.5.1. The resulting autocorrelation statistics (Moran's I) under normality assumption can be seen from Table 11, which is presented below.

***Table 11.** Autocorrelation Statistics of Model B*

Assumption	Coefficient	Observed	Expected	SD	Z	Pr > |Z|
Normality	Moran's I	0.00557	-0.0238	0.419	0.0701	0.9441

The observed Moran's I of less than 0.01 as well as the low Z-score and high p-value all suggest that there is no autocorrelation among regression residuals. Considering this and the other tests presented before, Model B was considered adequate. Further validation of the model is presented in the following chapter.

6.6 Validation of the Models

The performance of the land-use regression models was evaluated using holdout validation, a procedure described in Chapter 4. The original datasets were divided into training and test datasets 20 times so that each validation result contributed 5% to the overall result. As a form of sensitivity analysis, the original datasets were also split in two separate ways, as can be seen in Figure 11 below.

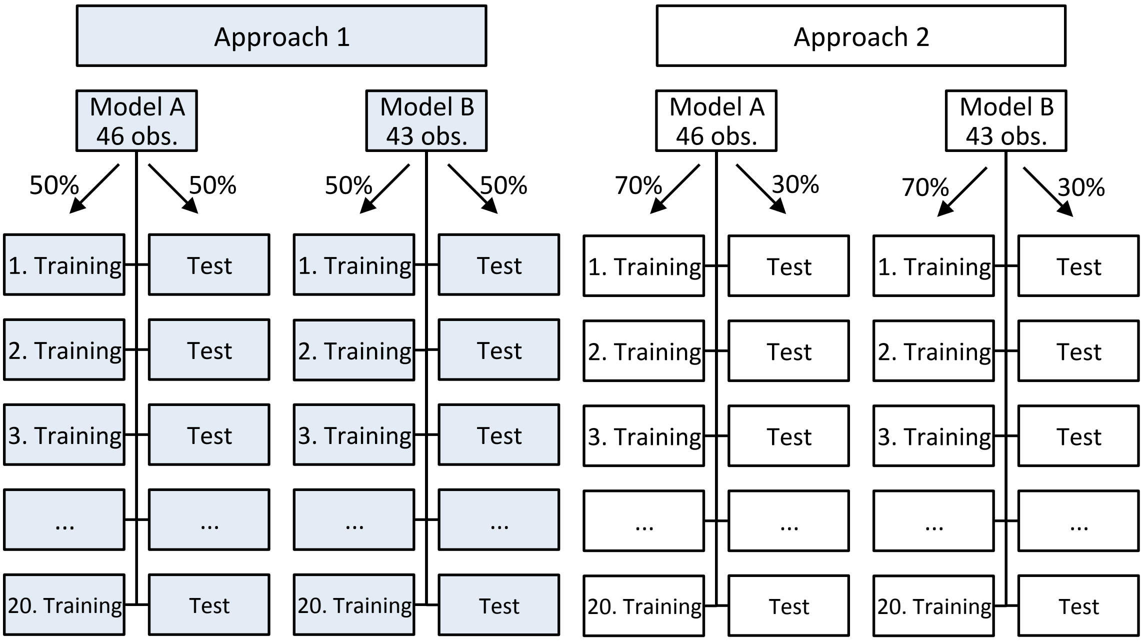


***Figure 11**. Validation of the land-use regression models*

As can be seen in Figure 11, the first approach was to split the original datasets in two so that half of the dataset could be used to validate a LUR model based on the other half. These two datasets constitute test and training dataset respectively. In Approach 2, 70% of the observations were assigned to the training dataset and 30% to the test dataset. These two approaches were both employed 20 times.

All datasets were generated using Surveyselect procedure on SAS. In order to include different kinds of sites in both types of datasets, and thus increase the range of values for predictor variables, the original dataset was divided with respect to sitetype. That is, 50/70% of the traffic sites were included in the training dataset and the rest in the test dataset. The same was done for the background sites. The selection procedure in SAS is described in detail on the following page.

```
PROC SURVEYSELECT DATA=full_dataset
OUT=training_n RATE=(A A) SEED=n;
STRATA Sitetype;
RUN;
```

The script utilizes stratified selection on the basis of sitetype as defined by the STRATA command. Observations from both site types are selected by number *A* that is defined in RATE. Since selected observations are assigned to the training dataset, the value of *A* was 0.5 in Approach 1 and 0.7 in Approach 2. The SEED (n) is a positive integer that is needed to partition the dataset randomly. In order to generate 20 different datasets, this integer ran from 1 to 20. Next, test datasets were created as written below:

```
PROC SORT DATA=full_dataset BY id;
RUN;
PROC SORT DATA=training_n by id;
RUN;

DATA validation_n;
MERGE full_dataset training_n (IN=a);
BY id;
IF a NE 1;
RUN;
```

The script sorts the full dataset as well as the training dataset in ascending order, using identification number as criterion. Then these two datasets are merged using such criteria that there is no match between the id:s. Since mismatch is associated with observations left out from the training dataset, the script generates the test dataset.

After generating the datasets, new LUR models were developed with the training data using the same criteria as when developing Models A and B. These new models constitute “partial models” since they were not developed with the full set of observations. The partial models were then validated with respective test data, i.e. predictions of PNC at test sites were compared with what was measured. In SAS, validation was done with the help of the regression procedure as described below. The given formula is for the first partial model developed under Approach 1.

```
DATA validation_1;
Predicted_PNC=338736*DISTINVMAJORC1-0.00215*URBGREEN_5000+45935;
RUN;

PROC REG DATA=validation_1 OUTEST=hv ADJRSQ;
MODEL PNC=Predicted_PNC;
PLOT Predicted_PNC*PNC;
RUN;
```

First, a partial model based on the training dataset is applied to test data, i.e. PNC is predicted at unmeasured sites by utilizing the values of predictor variables at those sites.

Then predicted PNC is regressed against measured PNC so as to estimate how well these two types of data fit.

In case variables in the test dataset were out of range as defined by the minimum and maximum values of those same variables in the training dataset, they were truncated to the closest range limit. This was done since the linearity of the models can be guaranteed only within the defined range. The results from the application of the holdout method are presented in the graphs below and in Appendices 4-7 in more detail.

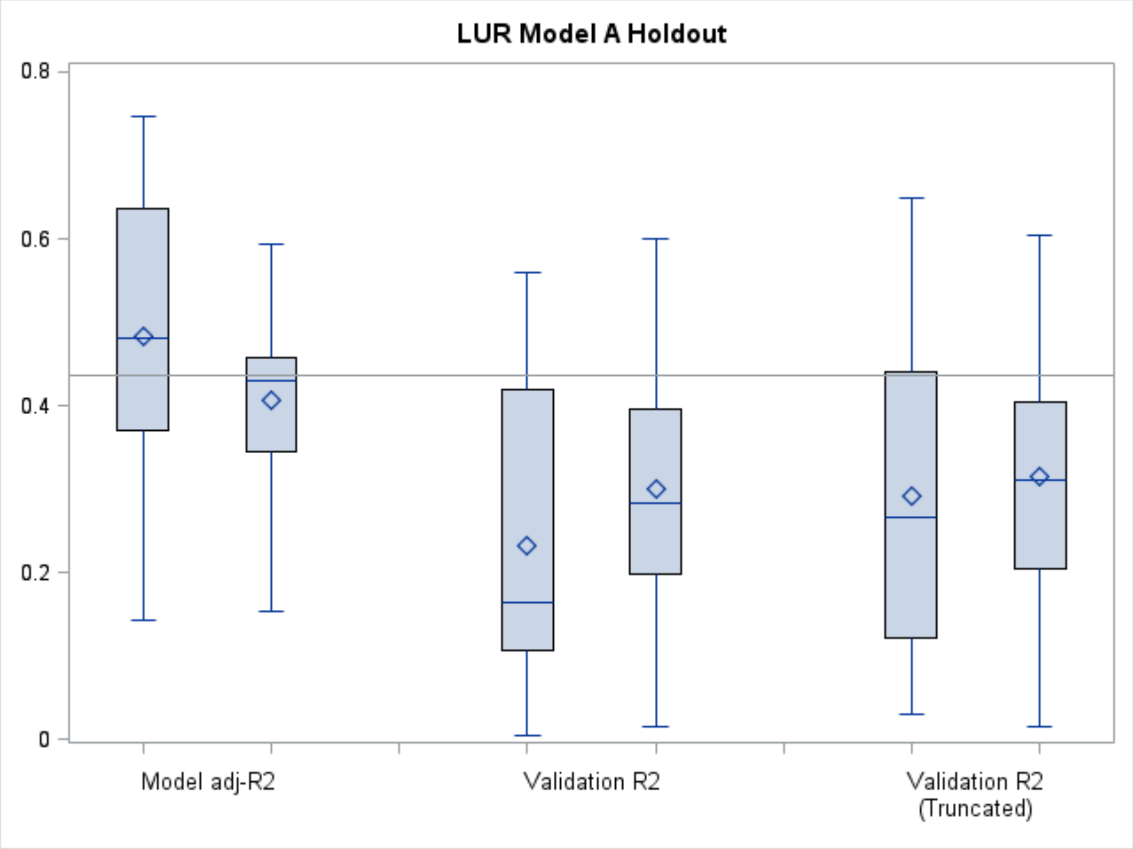


***Figure 12**. Validation of LUR Model A with approach 1 and 2*

As can be seen from Figure 12, partial model adjusted R-squares were generally higher than validation R-squares. The partial model adjusted R-squares also decreased and got more precise when they were developed with a larger number of measurement sites, and got closer to the Model A adjusted R^2 of 0.4367 as represented by the horizontal line. On contrary, validation R-squares generally increased when training datasets were larger.

When considering partial models based on smaller training datasets (Approach 1), the mean validation R^2 was 0.2320±0.0405 (SEM), or 48% of the mean partial model adjusted R^2. After truncation, however, its value increased to 0.2912±0.0416, i.e. 60% of the mean partial model adjusted R^2. Large difference between truncated and non-truncated validation R-squares indicates that the performance of the models is fairly sensitive to some of the observations. When considering partial models based on larger training datasets (Approach 2), respective percentages were 73% and 78% as can be calculated from the data in Appendix 5. These results indicate that Model A is sufficiently stable.

When it comes to the predictor variables entering the models, there were on average 1.95 variables in models developed with smaller training sets and 2.35 variables in models developed with larger training sets. This compares to three variables entering Model A. Out of Model A variables, *DISTINVMAJORC1* entered 65% of the models utilizing Approach 1 and 75% of the models utilizing Approach 2. Respective percentages were 35% and 55% for *URBGREEN_5000* and 15% and 25% for *EEA_5000*. More detailed description of variables entering the models is available in annexes 4-5. Next, validation results for Model B are presented.

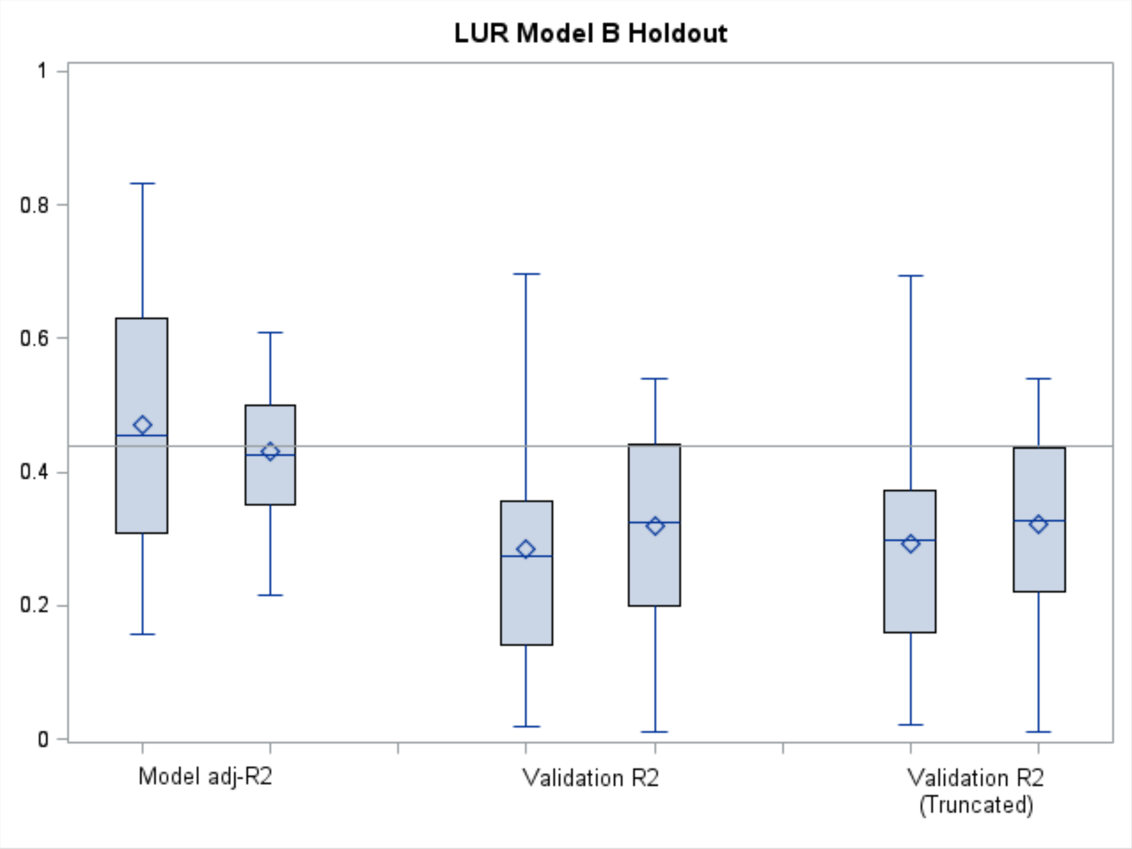


***Figure 13**. Validation of LUR Model B with approach 1 and 2*

When it comes to the validation of Model B, trends were similar but the variability of the results smaller. The means of partial model adjusted R-squares were close to the adjusted R^2 of Model B (represented by the horizontal line in Figure 13), and the difference between those was only 7% utilizing Approach 1 and 2% utilizing Approach 2.

When considering models based on smaller training datasets (Approach 1), the mean validation R^2 of 0.2853±0.0391 was 61% of the mean partial model adjusted R^2. As can be seen from Appendix 6, truncation increased it only to an extent, i.e. to 0.2929±0.0395 (62% of the mean partial model adjusted R^2). When considering models based on larger training datasets (Approach 2), respective percentages were both 74%. Small difference between truncated and non-truncated validation R-squares indicates that the performance of the models was not dependent on individual observations. This and the difference of 26% between the mean model adjusted R^2 and non-truncated validation R^2 indicate that Model B is sufficiently robust and performs slightly better than Model A.

When it comes to predictor variables entering the models, there were on average 2.15 variables in models developed with both smaller and larger training datasets. Out of Model B variables, *DISTINVMAJORC1* entered 45% of the models utilizing Approach 1 and 65% of the models utilizing Approach 2. Respective percentages were 25% and 30% for *TRAFNEAR* and 10% and 35% for *PORT_5000*. More detailed descriptions of variables entering the models are available in appendices 6-7.

7 EXPOSURE ASSESSMENT

Exposure assessment is a vital step for conducting a cohort study on air pollution and health. The assessment can be done with the help of land-use regression models as presented in this chapter.

7.1 Framework for Exposure Assessment

The framework for exposure assessment was established in Chapters 3 and 4. Briefly, outdoor particle number concentration may be used as a somewhat reasonable proxy for predicting the exposure to ultrafine particles, even if subjects reside indoors. However, it should be kept in mind that some particles are lost during indoor penetration and that the actual exposure is not at the level of the outdoor concentration. In addition, predicting individual exposures based on concentration at home does not reflect the fact that people move around the city during their days. However, this is a problem for all exposure assessment methods except personal monitoring or biomonitoring (Hoek et al. 2008a).

Taking note of the limitations of the methodology, cohort members' exposure to UFPs was predicted as analogous to outdoor PNC. These estimates were obtained from using the previously developed land-use regression models A and B. Before applying the models, predictor variables had to be assigned for cohort addresses and adjusted as discussed next.

7.2 Adjustments to Predictor Variables

Since predictor variables had already been calculated for cohort addresses as part of ESCAPE, no additional work in ArcGIS was needed. However, similarly to the adjustments done in developing the LUR models, the value of traffic intensity on the nearest road (*TRAFNEAR*) was replaced with the value of traffic intensity on the nearest major road (*TRAFMAJOR*) in case these two roads were within 25 meters from each other. This was done since *TRAFNEAR* was one of the variables in Model B.

Another adjustment had to do with the observed ranges of predictor variables. An important remark about land-use regression is that linear relationship of the model variables can be guaranteed only within the range of values that are observed at the measurement sites. That is, linearity cannot be guaranteed below the minimum values or over the maximum values, as presented under *Measurement Sites* in Table 12.

***Table 12**. Extreme values of applied predictor variables*

	Measurement Sites			Cohort		
Model A	**N**	**min**	**max**	**N**	**min**	**max**
DISTINVMAJORC1	46	0.00100957	0.11682275	4986	0.000264678	0.165966387
URBGREEN_5000	46	5275804	13100483	4986	505024	15443570
EEA_5000	46	137883	465981	4986	11068.6	470909.8
Model B						
DISTINVMAJORC1	43	0.00100957	0.11682275	4986	0.000264678	0.165966387
TRAFNEAR	43	16	35151	4986	0	133626
PORT_5000	43	0	11776748	4986	0	15059396

As can be observed from Table 12, there were out-of-range values for all relevant predictor variables pertaining to the cohort addresses. Since some of the values were out of range, they needed to be truncated to the closest range limit. The overall effect of this kind of range restriction can be assessed with the number of truncated values, as presented in Table 13 below, or with the altered distribution of values as presented in Appendix 8.

***Table 13**. Truncation, number and percentage truncated*

Predictor Variable	Truncated to lower limit (N)	%	Truncated to higher limit (N)	%	Total (N)	%
DISTINVMAJORC1	113	2.27	5	0.08	118	2.35
URBGREEN_5000	243	4.87	340	6.80	583	11.67
EEA_5000	106	2.13	71	1.40	177	3.53
TRAFNEAR	137	2.75	69	1.36	206	4.11
PORT_5000	0	0.00	449	8.99	449	8.99

As can be seen from Table 13, all but *PORT_5000* were truncated on both sides of the valid range. That and *URBGREEN_5000* were truncated mostly downwards. The remaining three variables were truncated mostly upwards.

7.3 Application of Land-Use Regression Models

After predictor variables were adjusted for valid ranges, cohort members' exposure to UFPs could be predicted. The method was to apply equations (2) and (3) with the predictor variables that were calculated for cohort addresses. Since range restriction affects predicted exposures, the effect of truncation was first examined. The results are presented in Table 14 on the following page.

Table 14. *Extreme values of predicted exposures*

	Measured Concentrations (particles/cm^3)		Predicted Exposures (particles/cm^3)			
	Lowest	**Highest**	**Lowest, Non-Truncated**	**Highest, Non-Truncated**	**Lowest, Truncated**	**Highest, Truncated**
Model A	12248	86902	6941	80295	8637	69605
Model B	12248	86902	16005	86465	16176	72974

As can be seen in Table 14, truncation led to tighter ranges of predicted exposures, i.e. the lowest predicted exposures increased with the range restriction whereas the highest predicted exposures decreased when the predictor variables were truncated. However, the effect of the range restriction was minimal when it comes to the lowest predicted exposure based on Model B.

Due to the factors discussed before, exposure assessment was based on the truncated values of predictor variables. The correlation between the predictions from Model A and Model B was good, 0.76. However, the distributions of predicted exposures were slightly different from each other, as may be observed in Figure 14 below.

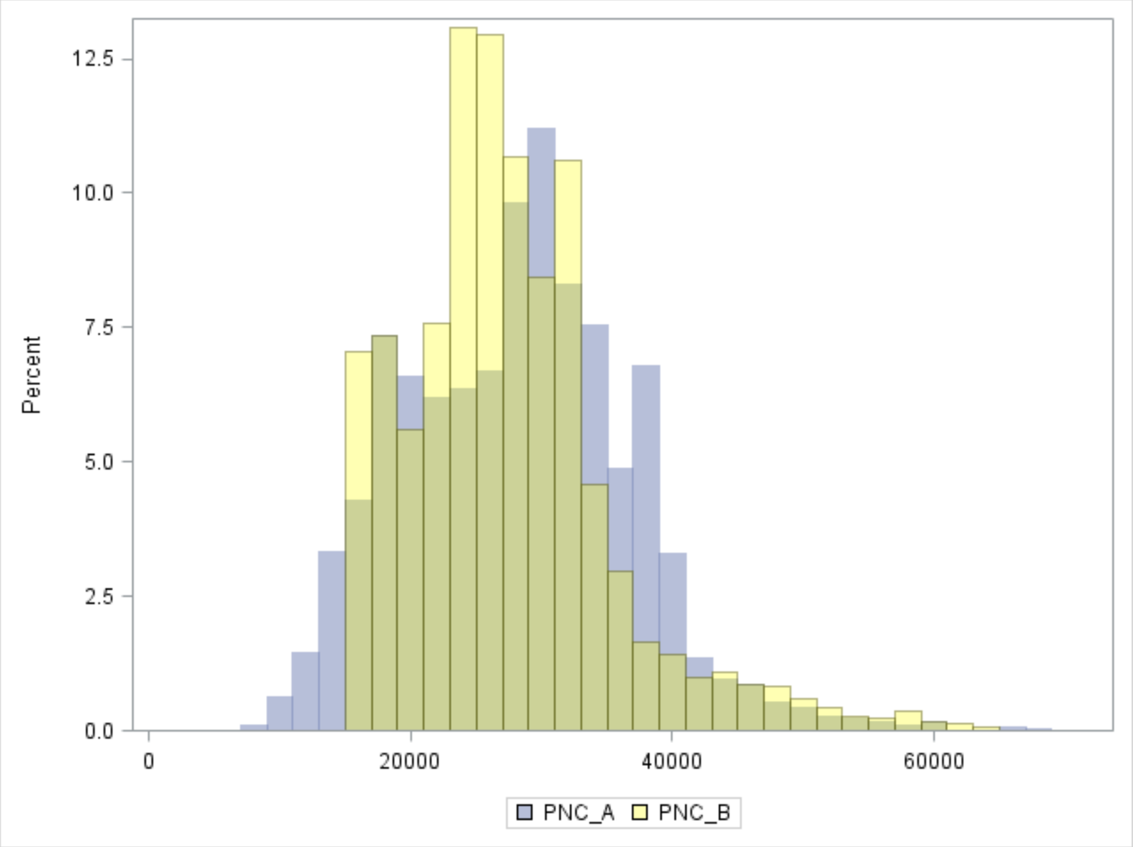


Figure 14*. Distribution of predicted UFP exposures based on Model A and Model B*

In Figure 14, one bin, i.e. an interval, represents 2,000 particles. As can be observed, predictions based on Model A were fairly symmetric and unimodal. Predictions based on Model B were skewed right.

On average, exposure estimates from Model A were slightly higher than those from Model B. Another remark is that over 5% of the predictions based on Model A were lower than the lowest predicted exposure based on Model B. However, neither corresponds to the lowest measured concentration of about 12,000 particles/cm^3 that was actually observed at the measurement sites. Also the highest predicted exposures were lower than the highest observed concentrations at the measurement sites, as may be seen in Table 14.

8 SUMMARY RESULTS

In summary, land-use regression Model A (N=46) predicted PNC as a function of inverse distance to the nearest major road, urban green area in a buffer of 5000 meters, and population density within the buffer of 5000 meters. Variables in Model B (N=43) were inverse distance to the nearest major road, traffic on nearest road and port area within the buffer of 5000 meters. Even though included variables were not the same, the F-test on both models rejected the null hypothesis, which claims that all predictor variables are 0 (p-value < 0.0001). In fact, both models explained about 44% of the variability in measured particle number concentrations.

Since predictor variables have different units of measurement, the coefficients in the models may not be compared with each other directly. The commensuration of predictor variables may be achieved by the means of calculating PNC for the commonly reported 10th-90th percentile range. The values of regression terms may then be compared with each other, as written in Table 15 below.

***Table 15**. Commensuration of Predictor Variables*

Model A	**10th-90th Range**	**Regression Term**	**Standard Error**
DISTINVMAJORC1	0.06085663	16444	4631
URBGREEN_5000	5013489	-14389	5364
EEA_5000	242015	11815	6099
Intercept	-	33447	9549
Model B			
DISTINVMAJORC1	0.06085663	13534	5401
TRAFNEAR	15897.3	7760	3677
PORT_5000	8850063	10797	5657
Intercept	-	15944	4186

Regression term is the 10th-90th range multiplied by respective regression coefficient

As can be seen from Table 15, the intercept term affects PNC the most. It lacks a clear interpretation but in the case of Model B, it represents the concentration not accountable to predictor variables. From the predictor variables, inverse distance to the nearest major road made the highest contribution to PNC in both models, i.e. shorter distance to a major road was associated with elevated concentrations. As described before, this variable alone explained about 36-37% of the variability in observed PNC. When it comes to the other predictors, all but urban green areas in the buffer of 5km were positively correlated with PNC.

With regard to testing the assumptions of linear regression, no clear violations were found. That is, residuals were approximately based on normal distribution and appropriately homoscedastic. Further to this, there was no spatial autocorrelation among regression residuals. These interpretations were stronger with Model B but sufficiently strong also with Model A.

When it comes to the validation of the land-use regression models, several observations were made. Firstly, the mean adjusted R^2 of the models developed with partial set of observations was close to the adjusted R^2 of the original models A and B. In the case of Model A, the difference was 3-4 percentage points, and in the case of Model B, it was only 1-3 percentage points. In both, the difference decreased with increasing number of observations in the training datasets.

These results indicate that validating Model A and Model B indirectly with 20 stratified random samples was an appropriate method. Predictor variables included in the partial models did change from one model to another but increasing the number of sites in the training datasets led to models that were more similar to the original models A and B.

When the number of sites was increased in the training datasets, model adjusted R-squares got smaller. At the same time, holdout validation R-squares increased. In both cases, estimations got more precise. This effect was considerable with the partial models that were developed to validate Model A. The presence of outliers absent from Model B may have had an influence in the observed trend. Truncation provided some evidence to this, as the mean of validation R^2 hardly changed with truncation when Model B was considered, but truncation considerably improved the validation R^2 associated with Model A. Further to this, the impact of truncation diminished when training datasets were larger as the outliers were more likely to be distributed into the training dataset, leading thus to greater valid ranges for predictor variables. See table 16 on the next page for details.

Table 16*. Holdout Validation of Models A and B*

						Variable present (% of times)				
			Mean (SD)							
	Training Sites (N)	Test Sites (N)	Model adj. R^2	HV R^2	Truncated HV R^2	*DISTINMAJORC1*	*URBGREEN_5000*	*EEA_5000*	*TRAFNEAR*	*PORT_5000*
A	23	23	0. 48 (0.17)	0.23 (0.18)	0.29 (0.19)	65	35	15	-	-
	32-33	13-14	0.41 (0.11)	0.30 (0.15)	0.32 (0.15)	75	55	25	-	-
B	21-22	21-22	0.47 (0.20)	0.29 (0.17)	0.29 (0.18)	45	-	-	25	10
	30-31	12-13	0.43 (0.09)	0.32 (0.14)	0.32 (0.14)	65	-	-	30	35

As can be seen from Table 16, holdout validation R-squares were modestly lower than the average adjusted R^2 of respective partial models. The difference between the mean model adjusted R^2 and the mean holdout validation R^2 was 11 percentage points when the models were developed with larger training datasets, indicating stable models. All partial models included 1-4 predictor variables, of which *DISTINMAJORC1* was found most often (65-75% of times).

When applying land-use regression models to cohort addresses, individual exposure estimates were obtained. The correlation between predictions from Model A and Model B was good, 0.76. As may be observed from Table 17 below, the mean and median predictions were slightly higher with Model A whereas the lowest and the highest predictions were higher with Model B.

Table 17*. Predicted UFP exposures in the EPIC MORGEN Amsterdam subcohort*

Model	N	Mean	min	5%	25%	50%	75%	95%	max	5th-95th Range
A	4986	28152	8637	14623	21413	28556	33673	41330	69605	26708
B	4986	27435	16176	16727	22469	26487	31382	43122	72974	26395

It can be further seen in Table 17 that the predicted exposures ranged from below 10,000 to about 70,000 particles/cm^3. However, 90% of the predicted exposures were within a range of approximately 26,000 particles/cm^3. These results compare to the 12,200-87,000 range observed at the measurement sites.

9 DISCUSSION

For the first time, land-use regression modeling has been utilized in predicting cohort members' long-term exposure to ultrafine particles. New LUR models were developed as part of the thesis, and thus this study adds to a small but growing number of land-regression models that have been developed for ultrafine particles.

The developed land-use regression models comprise Model A and Model B, with stricter inclusion terms for observations in Model B. In Model A, four sites out of fifty were excluded due to lack of successful measurement days. On top of this, three observations with highly overestimated traffic intensities were left out from Model B. However, the results show that LUR modeling was not overly sensitive to incorrect traffic estimates. In fact, both models explained about 44% of the variability in measured particle number concentrations.

Predictor variables included in the models were similar to those in previously published LUR models for ultrafine particles. With the exception of urban green as a mitigating variable, all other predictors have been included in previous LUR models in some form or another (Abernethy et al. 2013, Hoek et al. 2010, Rivera et al. 2012). Consistent with previous findings, the most important predictor variable was related to traffic in both Model A and Model B. The predictor variable in question is inverse distance to the nearest major road. Since transportation is a well-known source of PM emissions, proximity to major roads is indeed an important consideration. Model B also incorporated traffic intensity, which is a more direct predictor of PNC. From the remaining predictor variables, inclusion of port areas, population density and urban green in the buffer of 5000 meters reflects shipping emissions, commuting intensity and increased dispersion of PNC, respectively.

The LUR models A and B performed similarly to previous LUR models for ultrafine particles. In their 2011 study, Hoek and colleagues developed the first LUR model for ultrafine particles, also for the city of Amsterdam. With partly similar data they were able to explain 44% of the variability in measured particle number concentrations when site observations not derived from GIS were removed from their model. Utilizing different types of criteria in model development, Abernethy and colleagues (2013) achieved the adjusted R^2 values of 0.37-0.53. Meanwhile Rivera and colleagues (2012) reported an adjusted R^2 of 0.50 for the city of Girona in Spain.

Previous LUR models have typically reported leave-one-out cross-validation (LOOCV) R^2 values for model validation. Eeftens et al. (2012) reported differences of mostly less than 15 percentage points between LOOCV R^2 and model R^2 for several LUR models that were developed for $PM_{2.5}$. With regard to LUR models for ultrafine particles, Abernethy and colleagues (2013) reported differences in the range of 10-16 percentage points, whereas the difference was 3 percentage points for the model for Girona by Rivera et al. (2012).

Since LOOCV is an overly optimistic validation method in LUR models that are developed with a small number of observations (Wang et al. 2012), this thesis opted for holdout validation. When validating partial LUR models against unused test data, the lowest difference between mean model R^2 and non-truncated HV R^2 was 11 percentage points for both models A and B. This may be compared to the range of 8-29 percentage points that Wang et al. (2012) reported for LUR models developed for NO_2. Similarly to what Wang and colleagues reported, it was found out that the number of sites in the training datasets affected the results. Namely, model adjusted R^2 got smaller and holdout validation R^2 increased with an increasing number of sites in the training dataset.

With regard to assessing cohort members' long-term exposure to UFPs, the mean and median predictions were slightly higher with Model A whereas the lowest and the highest predictions were higher with Model B. In both cases, 90% of predicted exposures were within the range of approximately 26,000 particles/cm^3. Since this is the first time when land-use regression has been used to predict long-term exposure to ultrafine particles, these results cannot be compared with previous studies. Nonetheless, over 95% of predicted exposures were within the range that was actually measured at the measurement sites. However, since the highest predicted exposures were lower than the highest measured concentrations, it is likely that the models could not accurately predict all the highest exposures.

There are several considerations that should be kept in mind when interpreting the results. Firstly, the LUR models A and B were able to explain variability in PNC only to an extent. This is partly reflected by the less than desirable quality of some GIS data. For instance, bicycle/pedestrian paths in NWB come with motor vehicle traffic data when the paths have been given a street name. In some cases, a named pedestrian bridge over a canal may therefore distort modeled traffic data within large areas. Likewise, a named pedestrian street over a large urban green or semi-natural area may give false traffic estimates in the proximity of such land.

Another limitation with GIS data has to do with the contradictory nature of some geographic features. For instance, decreased UFP concentrations can be expected near rivers since water bodies provide for a microclimate where air pollution becomes diluted. However, intense shipping on a river increases emissions. If these two effects cannot be separated, treatment of rivers as a predictor variable can be problematic. In this study, rivers were absent from the predictor variable *semi-natural areas*, which otherwise included water bodies.

With regard to the temporal aspect of data, land-use records were available for 2000-2008, measurements were carried out in 2002-2004, and the cohort was compiled in 1993-1997. This could be problematic but since measurement locations downtown Amsterdam have not generally been altered in recent years, the use of data originating from different years can be considered appropriate.

Other limitations of the study design have to do with the general drawbacks of exposure assessment. Even if ambient UFP concentrations at cohort addresses were modeled well, they would not reflect personal exposure perfectly. Issues like indoor infiltration rates, daily activities and time spent home affect personal exposure to a great extent. Validating predicted concentrations with personal monitors is therefore an important research need.

10 CONCLUSION

Land-use regression is by now a well-established method to model intraurban air pollution with high spatial concentration anomalies. With the help of land-use regression, pollutant concentrations may be predicted at unmeasured sites within constrained geographical areas. Therefore LUR models can be a basis for epidemiological analyses where the health impacts from ambient air pollution are a matter of interest.

Studies on ultrafine particles are part of the long history of air pollution research where the underlying interest has been to identify the most hazardous characteristics of air pollution. To date, strong evidence has been found on the association of exposure to ambient particulate matter and adverse health effects, including e.g. cardiovascular and all-cause mortality. Studies are inconsistent in their findings whether some of the observed health impacts could be attributed to ultrafine particles only, and there are no published studies on the long-term effects of UFP exposure.

This thesis built upon earlier research and assessed long-term exposure to ultrafine particles for the members of a cohort living in Amsterdam, the Netherlands. As part of the study, new land-use regression models for ultrafine particles were created. These models, developed with slightly different set of observations, performed approximately equally and similarly to those published before. The previous studies have examined whether LUR in general may be used in predicting ultrafine particle concentrations at unmeasured sites. This study is therefore the first that has actually utilized LUR in estimating individual exposures to UFPs.

Exposure estimates from applying the LUR models were fairly similar and reasonable. However, the estimates should be used with caution because of the limited explanatory power of the LUR models, the inherent limitations of GIS data, and the difficulty in assigning exposure to individuals that typically move around the city during their days.

As a way forward, developing more robust land-use regression models is important. In general, GIS data gets better and traffic models improve on a fast basis that then provide for a more accurate description of surrounding environments. Even then, there is a strong research need to validate assigned exposures with personal monitors.

REFERENCES

Abernethy, R. C., Allen, R. W., McKendry, I. G., & Brauer, M. (2013). A land use regression model for ultrafine particles in Vancouver, Canada. Environmental science & technology. Vol. 47, No. 10, pp. 5217-5225.

Afshari, A., Matson, U., & Ekberg, L. E. (2005). Characterization of indoor sources of fine and ultrafine particles: a study conducted in a full-scale chamber. Indoor air. Vol. 15, No. 2, pp. 141-150.

Anselin, Luc, and Serge Rey. (1991). "Properties of tests for spatial dependence in linear regression models." Geographical analysis. Vol. 23, No. 2, pp. 112-131.

Bartley, D. L., & Vincent, J. H. (2011). Sampling conventions for estimating ultrafine and fine aerosol particle deposition in the human respiratory tract. Annals of occupational hygiene. Vol. 55, No. 7, pp. 696-709.

Basagaña, X., Rivera, M., Aguilera, I., Agis, D., Bouso, L., Elosua, R., Foraster, M., de Nazelle, A., Nieuwenhuijsen, M., Vila, J. & Künzli, N. (2012). Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmospheric Environment. Vol. 54, pp. 634-642.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., Vineis, P., Xun, W., Katsouyanni, K., Dimakopoulou, K., Oudin, A., Forsberg, B., Modig, L., Havulinna, A., Lanki, T., Turunen, A, Oftedal, B., Nystad, W. & Nafstad, P. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The Lancet. Vol. 383, No. 9919, pp. 785-795.

Beulens, J. W., Monninkhof, E. M., Verschuren, M. W., van der Schouw, Y. T., Smit, J., Ocke, M. C., Jansen, E., van Dieren, S., Grobbee, D., Peeters, P. & Bueno-de-Mesquita, B. H. (2009). Cohort profile: the EPIC-NL study. International journal of epidemiology. dyp217.

Bonham-Carter, G. F. (2014). Geographic information systems for geoscientists: modelling with GIS (Vol. 13). Elsevier. 416 p.

Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R., Mittleman, M., Peters, A., Siscovick, D., Smith Jr, S., Whitsel, L. & Kaufman, J. D. (2010). Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the American Heart Association. Circulation. Vol. 121, No. 21, pp. 2331-2378.

Cass, G. R., Hughes, L. A., Bhave, P., Kleeman, M. J., Allen, J. O., & Salmon, L. G. (2000). The chemical composition of atmospheric ultrafine particles. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. Vol. 358, No. 1775, pp. 2581-2592.

Chatterjee, S., & Simonoff, J. S. (2013). Handbook of Regression Analysis. Vol. 5. John Wiley & Sons. 252 p.

European Environment Agency (EEA). (2002). Corine land cover update 2000: Technical guidelines. Available (accessed on 26.5.2014):
http://www.eea.europa.eu/publications/technical_report_2002_89

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., ... & Hoek, G. (2012). Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. Environmental science & technology. Vol. 46, No. 20, pp. 11195-11205.

Geiser, M., & Kreyling, W. G. (2010). Deposition and biokinetics of inhaled nanoparticles. Particle and Fibre Toxicology, Vol. 7, No. 2, pp. 1-17.

U.S. Environmental Protection Agency (EPA). (2014). About Reports – Monitor Values Report. Available (accessed on 16.11.2014): http://www.epa.gov/airdata/ad_about_reports.html

HEI Review Panel on Ultrafine Particles. (2013). Understanding the Health Effects of Ambient Ultrafine Particles. HEI Perspectives 3. Health Effects Institute, Boston, MA. 108p.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008a). A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric Environment. Vol. 42, No. 33, pp. 7561-7578.

Hoek, G., Kos, G., Harrison, R., de Hartog, J., Meliefste, K., ten Brink, H., Katsouyanni, K., Karakatsani, A., Lianou, M., Kotronarou, A., Kavouras, I., Pekkanen, J., Vallius, M., Kulmala, M., Puustinen, A., Thomas, S., Meddings, C., Ayres, J., van Wijnen, J. & Hameri, K. (2008b). Indoor–outdoor relationships of particle number and mass in four European cities. Atmospheric Environment. Vol. 42, No. 1, pp. 156-169.

Hoek, G., Beelen, R., Kos, G., Dijkema, M., Zee, S. C. V. D., Fischer, P. H., & Brunekreef, B. (2010). Land use regression model for ultrafine particles in Amsterdam. Environmental science & technology. Vol. 45, No. 2, pp. 622-628.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., & Kaufman, J. D. (2013). Long-term air pollution exposure and cardio-respiratory mortality: a review. Environmental Health. Vol. 12, No. 1, 43.

Holmes, N. S., & Morawska, L. (2006). A review of dispersion modelling and its application to the dispersion of particles: an overview of different dispersion models available. Atmospheric Environment. Vol. 40, No. 30, pp. 5902-5928.

Integrated Environmental Health Impact Assessment System (IEHIAS). (2010). EU age/sex stratified population: 100 metre grid. Available (accessed on 15.5.2014):
http://www.integrated-assessment.eu/resource_centre/eu_agesex_stratified_population_100_metre_grid

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J. & Giovis, C. (2004). A review and evaluation of intraurban air pollution exposure models. Journal of Exposure Science and Environmental Epidemiology. Vol. 15, No. 2, pp. 185-204.

Karner, A. A., Eisinger, D. S., & Niemeier, D. A. (2010). Near-roadway air quality: synthesizing the findings from real-world data. Environmental Science & Technology. Vol. 44. No. 14, pp. 5334-5344.

Knol, A. B., de Hartog, J. J., Boogaard, H., Slottje, P., van der Sluijs, J. P., Lebret, E., ... & Hoek, G. (2009). Expert elicitation on ultrafine particles: likelihood of health effects and causal pathways. Particle and Fibre Toxicology. Vol. 6, No. 1, pp. 19.

Kreyling, W. G., Semmler-Behnke, M., & Möller, W. (2006). Health implications of nanoparticles. Journal of Nanoparticle Research. Vol. 8, No. 5, pp. 543-562.

Kulmala, M., Pirjola, L., & Mäkelä, J. M. (2000). Stable sulphate clusters as a source of new atmospheric particles. Nature. Vol. 404, No. 6773, pp. 66-69.

Marieb, E. N. (2012). Essentials of human anatomy and physiology. Benjamin Cummings. 10th edition. New York. 632p.

Martins, L. D., Martins, J. A., Freitas, E. D., Mazzoli, C. R., Gonçalves, F. L. T., Ynoue, R. Y., Hallak, R., Albuquerque, T. T. & de Fatima Andrade, M. (2010). Potential health impact of ultrafine particles under clean and polluted urban atmospheric conditions: a model-based study. Air Quality, Atmosphere & Health. Vol. 3, No. 1, pp. 29-39.

Morawska, L., Ristovski, Z., Jayaratne, E. R., Keogh, D. U., & Ling, X. (2008). Ambient nano and ultrafine particles from motor vehicle emissions: characteristics, ambient processing and implications on human exposure. Atmospheric Environment. Vol. 42, No. 35, pp. 8113-8138.

Oberdörster, G., Oberdörster, E., & Oberdörster, J. (2005). Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environmental health perspectives. Vol. 113, No. 7, pp. 823-839.

Puustinen, A., Hämeri, K., Pekkanen, J., Kulmala, M., De Hartog, J., Meliefste, K., ten Brink, H., Kos, G., Katsouyanni, K., Karakatsani, A., Kotronarou, A., Kavouras, I., Meddings, C., Thomas, S., Harrison, R., Ayres, J.G., van der Zee, S. & Hoek, G. (2007). Spatial variation of particle number and mass over four European cities. Atmospheric Environment. Vol. 41, No. 31, pp. 6622-6636.

Rivera, M., Basagaña, X., Aguilera, I., Agis, D., Bouso, L., Foraster, M., Medina-RAmón M., Pey, J., Kûnzli, N. & Hoek, G. (2012). Spatial distribution of ultrafine particles in urban settings: a land use regression model. Atmospheric Environment. Vol. 54, pp. 657-666.

Rückerl, R., Schneider, A., Breitner, S., Cyrys, J., & Peters, A. (2011). Health effects of particulate air pollution: a review of epidemiological evidence. Inhalation toxicology. Vol. 23, No. 10, pp. 555-592.

Schneider, J., Moore, A.W. (1997). Cross Validation, in: A Locally Weighted Learning Tutorial using Vizier 1.0. Available (accessed on: 14.4.2014): http://www.cs.cmu.edu/~schneide/tut5/node42.html

Sioutas, C., Delfino, R. J., & Singh, M. (2005). Exposure assessment for atmospheric ultrafine particles (UFPs) and implications in epidemiologic research. Environmental Health Perspectives. Vol. 113, No. 8, pp. 947-955.

Spatial Insights. (2014). TeleAtlas MultiNet Features. Available (accessed on 15.9.2014): http://www.spatialinsights.com/catalog/product.aspx?product=95

Stichting Wetenschappelijk Onderzoek Verkeersveiligheid (SWOV). (2014). National Roads Database (NWB). Available (accessed on 17.9.2014): http://www.swov.nl/UK/Research/cijfers/Toelichting-gegevensbronnen/NWB-UK.html

TSI Incorporated (TSI). (1999). Product Information – Model 3022A Condensation Particle Counter. Available (accessed on 16.8.2014): http://www.tsi.com/uploadedFiles/_Site_Root/Products/Literature/Spec_Sheets/3022A.pdf

Wallace, L., & Ott, W. (2010). Personal exposure to ultrafine particles. Journal of Exposure Science and Environmental Epidemiology. Vol. 21, No. 1, pp. 20-30.

Wang, M., Beelen, R., Eeftens, M., Meliefste, K., Hoek, G., & Brunekreef, B. (2012). Systematic evaluation of land use regression models for NO2. Environmental science & technology. Vol. 46, No. 8, pp. 4481-4489.

Wang, M., Beelen, R., Stafoggia, M., Raaschou-Nielsen, O., Andersen, Z. J., Hoffmann, B., Fischer, P., Houthuijs, D., Nieuwenhuijsen, M., Weinmayr, G., Vineis, P., Xun, W., Dimakopoulou, K., Samoli, E., Laatikainen, T., Lanki, T., Turunen, A., Oftedal, B., Schwarze, P., Aamodt, G., Penell, J., De Faire, U., Korek, M., Leander, K., Pershagen, G., Pedersen, N.L., Östenson, C-G., Fratiglioni, L. et al. (2014). Long-term exposure to elemental constituents of particulate matter and cardiovascular mortality in 19 European cohorts: Results from the ESCAPE and TRANSPHORM projects. Environment international. Vol. 66, pp. 97-106.

Wheeler, A. J., Wallace, L. A., Kearney, J., Van Ryswyk, K., You, H., Kulka, R., Brook, J.R. & Xu, X. (2011). Personal, indoor, and outdoor concentrations of fine and ultrafine particles using continuous monitors in multiple residences. Aerosol Science and Technology. Vol. 45, No. 9, pp. 1078-1089.

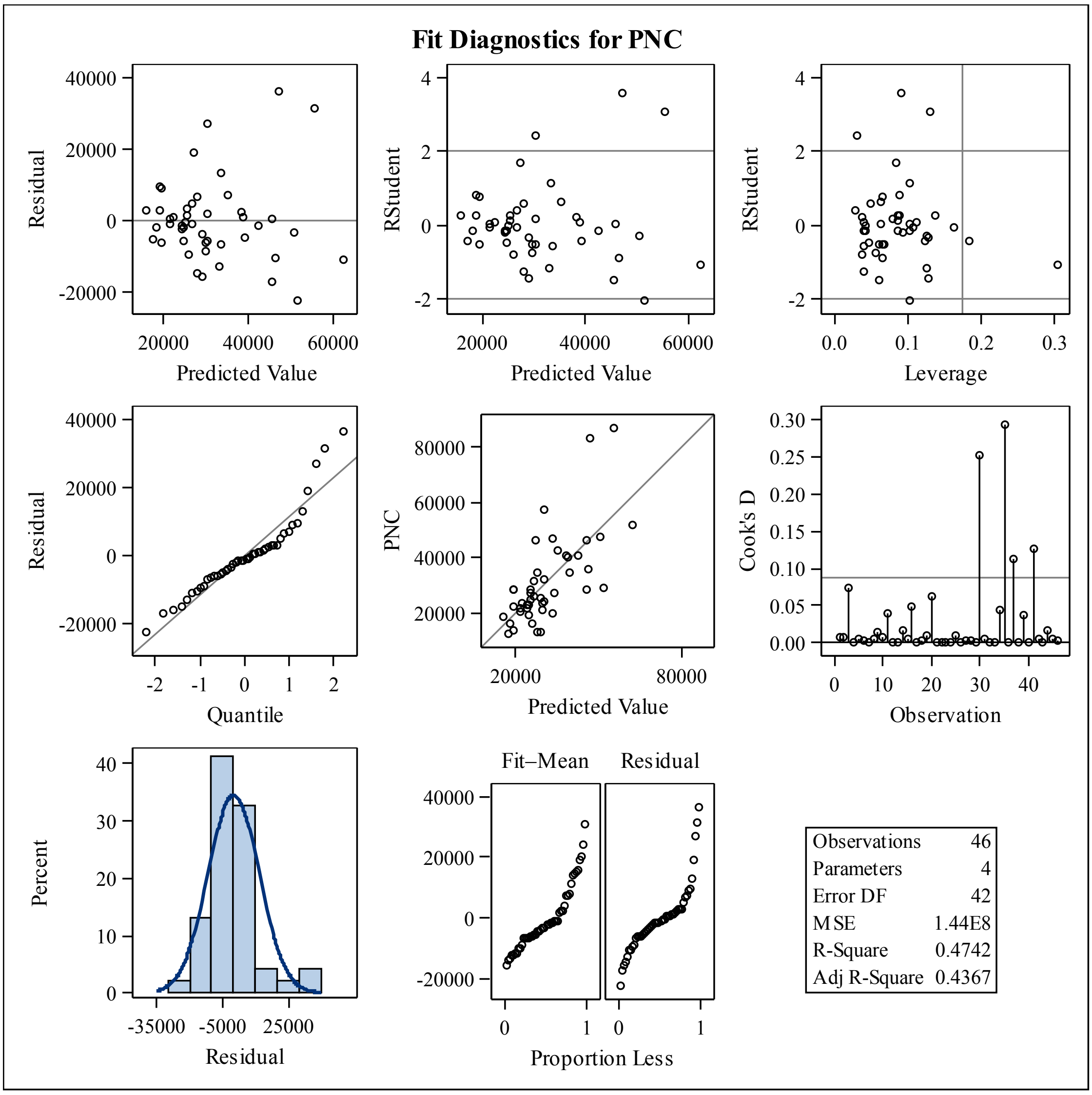
Zhu, Y., Hinds, W. C., Krudysz, M., Kuhn, T., Froines, J., & Sioutas, C. (2005). Penetration of freeway ultrafine particles into indoor environments. Journal of Aerosol Science. Vol. 36, No. 3, pp. 303-322.

APPENDIX 1: LUR PREDICTOR VARIABLES

Variable name	Description	Unit	Buffer (m)					
			50	100	300	500	1000	5000
LDRES	Low density residential land	m^2	-	X	X	X	X	X
HDRES	High density residential land	m^2	-	/	/	/	/	/
INDUSTRY	Industry	m^2	-	/	/	/	/	X
PORT	Port	m^2	-	/	/	/	/	X
URBGREEN	Urban Green	m^2	-	/	/	/	X	X
NATURAL	Semi-natural and forested areas	m^2	-	/	/	/	/	X
EEA	Population density	n	-	X	X	X	X	X
DISTINVNEAR1 DISTINVNEAR2	Distance to the nearest road, NWB	m^{-1} m^{-2}	-	-	-	-	-	-
TRAFNEAR	Traffic intensity on nearest road	vehicles/day	-	-	-	-	-	-
INTINVDIST INTINVDIST2	Product of 1) traffic intensity on nearest road, and 2) distance to the nearest road	$\frac{\text{vehicles}}{\text{day} * \text{m}}$; $\frac{\text{vehicles}}{\text{day} * \text{m}^2}$	-	-	-	-	-	-
DISTINVMAJOR1 DISTINVMAJOR2	Distance to the nearest major road, NWB	m^{-1} m^{-2}	-	-	-	-	-	-
TRAFMAJOR	Traffic intensity on nearest major road	vehicles/day	-	-	-	-	-	-
INTMAJORINVDIST INTMAJORINVDIST2	Product of 1) traffic intensity on nearest major road, and 2) distance to the nearest major road	$\frac{\text{vehicles}}{\text{day} * \text{m}}$; $\frac{\text{vehicles}}{\text{day} * \text{m}^2}$	-	-	-	-	-	-
TRAFLOAD	Total traffic load of all roads in a buffer (sum of traffic intensity * length of all segments)	$\frac{\text{vehicles}}{\text{day}} * \text{m}$	X	X	X	X	X	-
TRAFMAJORLOAD	Total traffic load of major roads in a buffer (sum of traffic intensity * length of all segments)	$\frac{\text{vehicles}}{\text{day}} * \text{m}$	/	X	X	X	X	-
HEAVYTRAFNEAR	Heavy-duty traffic intensity on nearest road	vehicles/day	-	-	-	-	-	-
HEAVYINTINVDIST HEAVYINTINVDIST2	Product of 1) Heavy-duty traffic intensity on nearest road 2) Distance to the nearest road	$\frac{\text{vehicles}}{\text{day} * \text{m}}$; $\frac{\text{vehicles}}{\text{day} * \text{m}^2}$	-	-	-	-	-	-
HEAVYTRAFMAJOR	Heavy-duty traffic intensity on nearest major road	vehicles/day	-	-	-	-	-	-
HEAVYTRAFLOAD	Total heavy-duty traffic load of all roads in a buffer (sum of heavy-duty traffic intensity * length of all segments)	$\frac{\text{vehicles}}{\text{day}} * \text{m}$	X	X	X	X	X	-
HEAVYTRAFMAJORLOAD	Total heavy-duty traffic load of major roads in a buffer (sum of heavy-duty traffic intensity * length of all segments)	$\frac{\text{vehicles}}{\text{day}} * \text{m}$	/	X	X	X	X	-
ROADLENGTH	Road length of all roads in a buffer	m	X	X	X	X	X	-
MAJORROADLENGTH	Road length of major roads in a buffer	m	/	/	X	X	X	-
DISTINVNEARC1 DISTINVNEARC2	Distance to the nearest road, Eurostreets	m^{-1} m^{-2}	-	-	-	-	-	-
DISTINVMAJORC1 DISTINVMAJORC2	Distance to the nearest major road, Eurostreets	m^{-1} m^{-2}	-	-	-	-	-	-

X = Buffer utilized in modeling; / = Buffer discarded; - = buffer not available

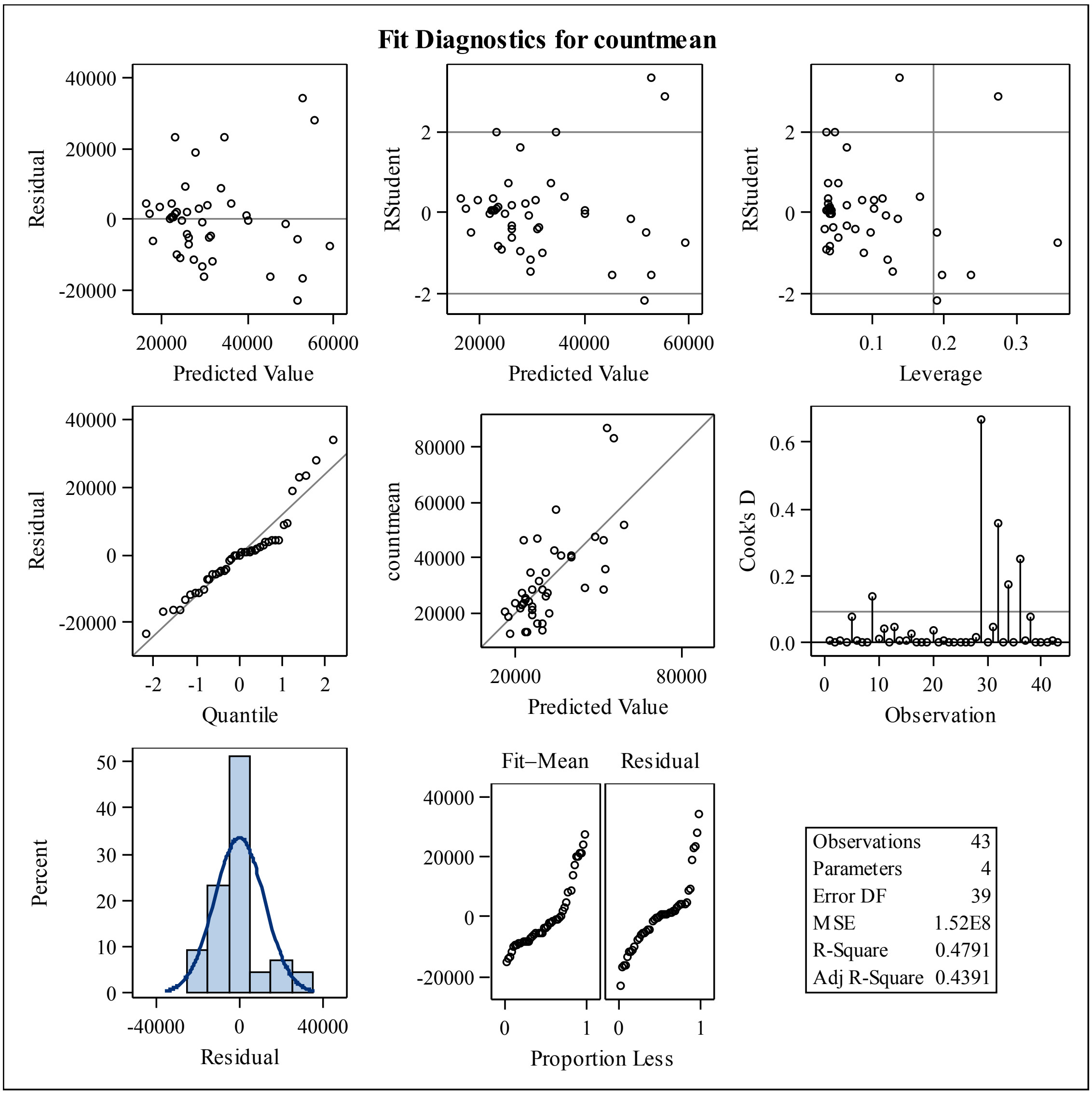
APPENDIX 2: LUR MODEL A DIAGNOSTICS



Fit Diagnostics for PNC

Observations	46
Parameters	4
Error DF	42
MSE	1.44E8
R-Square	0.4742
Adj R-Square	0.4367

APPENDIX 3: LUR MODEL B DIAGNOSTICS



APPENDIX 4: VALIDATION OF LUR MODEL A, APPROACH 1

	Model Adjusted R2	Validation R2	Validation R2 (Truncated)	Variables in the model
1	0.4714	0.2854	0.2857	DISTINVMAJORC1, URBGREEN_5000
2	0.3765	0.5116	0.5116	DISTINVMAJORC1
3	0.3684	0.1500	0.2599	TRAFLOAD_50, PORT_5000
4	0.4899	0.1536	0.1935	URBGREEN_5000, ROADLENGTH_100, EEA_5000
5	0.4986	0.4301	0.4301	DISTINVMAJORC1
6	0.2509	0.4990	0.499	DISTINVMAJORC1
7	0.6851	0.4467	0.4467	DISTINVMAJORC1
8	0.3715	0.4103	0.4102	DISTINVMAJORC1, URBGREEN_5000
9	0.3210	0.1144	0.1328	ROADLENGTH_100-1000
10	0.6381	0.0577	0.1109	DISTINVMAJORC1, HEAVYTRAFNEAR
11	0.4342	0.0065	0.1260	DISTINVMAJORC1, DISTINVNEARC2
12	0.4730	0.0441	0.0300	URBGREEN_5000, TRAFLOAD_100
13	0.6045	0.2805	0.2726	DISTINVMAJORC1, URBGREEN_5000
14	0.7469	0.1165	0.1165	MAJORROADLENGTH_1000, URBGREEN_5000, EEA_5000, PORT_5000
15	0.4926	0.1021	0.1164	DISTINVMAJORC1, PORT_5000, MAJORROADLENGTH_1000
16	0.6450	0.1831	0.4325	DISTINVMAJORC1, INTMAJORINVDIST
17	0.1438	0.5587	0.6490	DISTINVMAJORC1
18	0.3062	0.0036	0.064	INTMAJORINVDIST2
19	0.6338	01757	0.1779	PORT_5000, EEA_5000, URBGREEN_5000, ROADLENGTH_100
20	0.7187	0.1107	0.5581	INTMAJORINVDIST2, DISTINVMAJORC1
Mean	0.4835	0.2320	0.2912	
Std. error	0.0370	0.0405	0.0416	

APPENDIX 5: VALIDATION OF LUR MODEL A, APPROACH 2

	Model Adjusted R2	Validation R2	Validation R2 (Truncated)	Variables in the model
1	0.5943	0.0158	0.0152	DISTINVMAJORC1, URBGREEN_5000, MAJORROADLENGTH_1000, PORT_5000
2	0.5239	0.4743	0.4101	DISTINVMAJORC1, PORT_5000
3	0.4342	0.1720	0.1720	DISTINVMAJORC1, URBGREEN_5000
4	0.3414	0.2129	0.3505	DISTINVMAJORC1, TRAFMAJORLOAD_100
5	0.3501	0.3998	0.3998	DISTINVMAJORC1
6	0.4413	0.1247	0.1599	URBGREEN_5000, TRAFNEAR, ROADLENGTH_500
7	0.3789	0.2029	0.2365	DISTINMAJOR1, ROADLENGTH_300, URBREEN_5000, EEA_5000
8	0.4975	0.3503	0.2917	DISTINVMAJORC1, URBGREEN_5000, TRAFNEAR
9	0.3154	0.4240	0.5042	DISTINVMAJORC1
10	0.3226	0.2585	0.3319	PORT_5000, EEA_5000, ROADLENGTH_50, TRAFNEAR
11	0.3807	0.5119	0.5119	DISTINVMAJORC1, URBGREEN_5000, EEA_5000
12	0.5824	0.2072	0.2072	DISTINVMAJORC1, URBGREEN_5000, EEA_5000
13	0.4716	0.6000	0.6034	DISTINVMAJORC1, URBGREEN_5000, EEA_5000
14	0.1525	0.1271	0.1271	DISTINVMAJOR1
15	0.4250	0.3772	0.3391	DISTINVMAJORC1, URBGREEN_5000
16	0.3530	0.3865	0.4967	DISTINVMAJORC1
17	0.4376	0.2783	0.2783	DISTINVMAJORC1
18	0.4452	0.3906	0.3907	DISTINVMAJORC1, URBGREEN_5000, LDRES_5000
19	0.4429	0.1952	0.1996	DISTINVMAJORC1, PORT_5000
20	0.2472	0.2880	0.2880	ROADLENGTH_1000, URBGREEN_5000
Mean	0.4069	0.2999	0.3157	
Std. error	0.0238	0.0329	0.0332	

APPENDIX 6: VALIDATION OF LUR MODEL B, APPROACH 1

	Model R2	Validation R2	Validation R2 (Truncated)	Variables in the model
1	0.5833	0.3412	0.3815	DISTINVMAJORC1, TRAFMAJORLOAD_100, URBGREEN_5000
2	0.6248	0.3590	0.3603	DISTINVMAJORC1, URBGREEN_5000, ROADLENGTH_300
3	0.6459	0.0940	0.1122	TRAFNEAR, ROADLENGTH_100, MAJORROADLENGTH_1000, URBGREEN_5000
4	0.8324	0.3455	0.3059	INTMAJORINVDIST, DISTINVMAJORC2, URBGREEN_5000, TRAFMAJOR
5	0.5811	0.3527	0.3652	TRAFNEAR, DISTINVMAJORC1
6	0.3946	0.1197	0.1585	TRAFMAJORLOAD_100
7	0.4711	0.1622	0.1624	TRAFLOAD_50, URBGREEN_5000
8	0.7589	0.2802	0.3052	URBGREEN_5000, EEA_5000, LDRES_100, TRAFNEAR
9	0.5728	0.1029	0.1028	TRAFNEAR, PORT_5000
10	0.6383	0.2334	0.2398	TRAFNEAR, URBGREEN_5000, EEA_5000
11	0.3110	0.2279	0.2279	PORT_5000, DISTINVMAJOR1
12	0.4394	0.3090	0.3089	DISTINVMAJORC1
13	0.3047	0.0199	0.0205	INDUSTRY_5000, EEA_5000
14	0.3098	0.4923	0.4923	DISTINVMAJORC1
15	0.2808	0.4444	0.5135	DISTINVMAJORC1
16	0.3561	0.6975	0.6934	DISTINVMAJORC1
17	0.3061	0.5677	0.5677	DISTINVMAJORC1
18	0.1568	0.2693	0.2907	DISTINVMAJOR1
19	0.1670	0.0503	0.0630	ROADLENGTH_300, URBGREEN_5000
20	0.6909	0.2370	0.1866	DISTINVMAJORC1, URBGREEN_5000, ROADLENGTH_100
Mean	0.4713	0.2853	0.2929	
Std. Error	0.0439	0.0391	0.0395	

APPENDIX 7: VALIDATION OF LUR MODEL B, APPROACH 2

	Model R2	Validation R2	Validation R2 (Truncated)	Variables in the model
1	0.4989	0.2979	0.2564	DISTINVMAJORC1, PORT_5000
2	0.5336	0.2114	0.2114	DISTINVMAJORC1, URBGREEN_5000, EEA_5000, ROADLENGTH_300
3	0.4579	0.1876	0.2302	DISTINVMAJORC1, MAJORROADLENGTH_1000, URBGREEN_5000
4	0.5229	0.0113	0.0113	TRAFNEAR, PORT_5000
5	0.492	0.1895	0.2109	DISTINVMAJORC1, URBGREEN_5000, TRAFNEAR
6	0.4091	0.4475	0.4353	DISTINVMAJORC1, URBGREEN_5000
7	0.5376	0.3142	0.3147	DISTINVMAJORC1, URBGREEN_5000, EEA_5000
8	0.3542	0.4737	0.4736	DISTINVMAJORC1
9	0.4343	0.2727	0.2727	DISTINVMAJORC2, TRAFMAJORLOAD_100
10	0.6082	0.3815	0.3952	URBGREEN_5000, EEA_5000, TRAFNEAR
11	0.323	0.3083	0.3204	TRAFLOAD_50, PORT_5000
12	0.3947	0.0986	0.0986	PORT_5000, TRAFNEAR
13	0.3463	0.4815	0.4812	DISTINVMAJORC1
14	0.2166	0.4068	0.4068	DISTINVMAJOR1, PORT_5000
15	0.3487	0.4357	0.4357	DISTINVMAJORC1
16	0.4911	0.4492	0.4495	TRAFNEAR, DISTINVMAJORC1, PORT_5000
17	0.3324	0.3355	0.3355	TRAFNEAR
18	0.4998	0.1364	0.1388	DISTINVMAJORC1, PORT_5000
19	0.4091	0.5392	0.5392	DISTINVMAJORC1
20	0.4152	0.4046	0.4061	DISTINVMAJORC1, ROADLENGH_1000, URBGREEN_5000
Mean	0.4313	0.3192	0.3212	
Std. Error	0.0211	0.0320	0.0315	

APPENDIX 8: DISTRIBUTION OF PREDICTOR VARIABLES

Statistical values for predictor variables at the measurement sites

Model A	N	min	P10	median	P90	max
DISTINVMAJORC1	46	0.00100957	0.00167061	0.00783077	0.06252724	0.11682275
URBGREEN_5000	46	5275804	6511773	8574764	11525262	13100483
EEA_5000	46	137883	216244	387822	458259	465981
Model B	**N**	**min**	**P10**	**median**	**P90**	**max**
DISTINVMAJORC1	43	0.00100957	0.00167061	0.00830738	0.06252724	0.11682275
TRAFNEAR	43	16.0	90.0	735.0	15987.3	35151.0
PORT_5000	43	0	1937469	5223853	10787532	11776748

Statistical values for predictor variables at cohort addresses (non-truncated & truncated)

Predictor Variable	N	min	P10	median	P90	max
DISTINVMAJORC1	4986	0.000264678	0.002036569	0.005914848	0.023932635	0.165966387
URBGREEN_5000	4986	505024	5821783	8427855	12485104	15443570
EEA_5000	4986	11068.6	165776.5	365438.3	452673.3	470909.8
TRAFNEAR	4986	0.0	105.0	809.5	13189.2	133626.0
PORT_5000	4986	0	0	5298745	11546243	15059396
Predictor Variable	**N**	**min**	**P10**	**median**	**P90**	**max**
DISTINVMAJORC1	4986	0.00100957	0.00203657	0.00591485	0.023932640	0.11682275
URBGREEN_5000	4986	5275804	5821783	8427855	12485104	13100483
EEA_5000	4986	137883	165777	365438	452673	465981
TRAFNEAR	4986	16.0	105.0	809.5	13189.2	35151.0
PORT_5000	4986	0	0	5298745	11546243	11776748