ESCAPE

Exposure assessment manual

Version July 2010

Foreword

This manual provides procedures for air pollution exposure assessment within ESCAPE that are not covered in the overall study manual. The overall study manual specifies the monitoring program including site selection. The current manual starts form there and adds modeling issues (GIS data, modeling procedures, validation, extrapolation of exposure estimates back in time).
This version replaces the August 2009 version that was used to guide the collection of GIS data. A track change version is available indicating the differences between this version and the August 2009 version. The main changes are in chapter 2-5. Experiences obtained in the first year GIS workshop in April 2010 were incorporated in this updated manual.

Table of contents

1. Introduction

The purpose of this ESCAPE exposure assessment manual is to provide a manual for all the steps from air pollution monitoring to exposure assessment for addresses of study participants. The monitoring methods and site selection has been specified in the overall study manual. The steps covered in the current text include:

1. Calculation of annual average air pollution concentrations (section 2)
2. Collection of Geographic Information System (GIS) data (Section 3)
3. GIS analyses for the coordinates of the monitoring sites and the addresses of study participants (Section 4)
4. Development of land use regression and potentially other models (Section 5)
5. Assessment of air pollution exposure at addresses of study participants (Section 6) (including geocoding of addresses which is described in a separate ESCAPE geocoding procedure, appendix IV)
6. Exposure estimation back in time (Section 7).

For some study areas and studies there may be additional exposure issues, e.g. both ESCAPE data and previous local exposure data may be available, or no local exposure data at all may be available (Section 8). Section 9 describes the coordination activities to harmonize exposure assessment which is mostly performed locally. Section 10 shows a time planning.

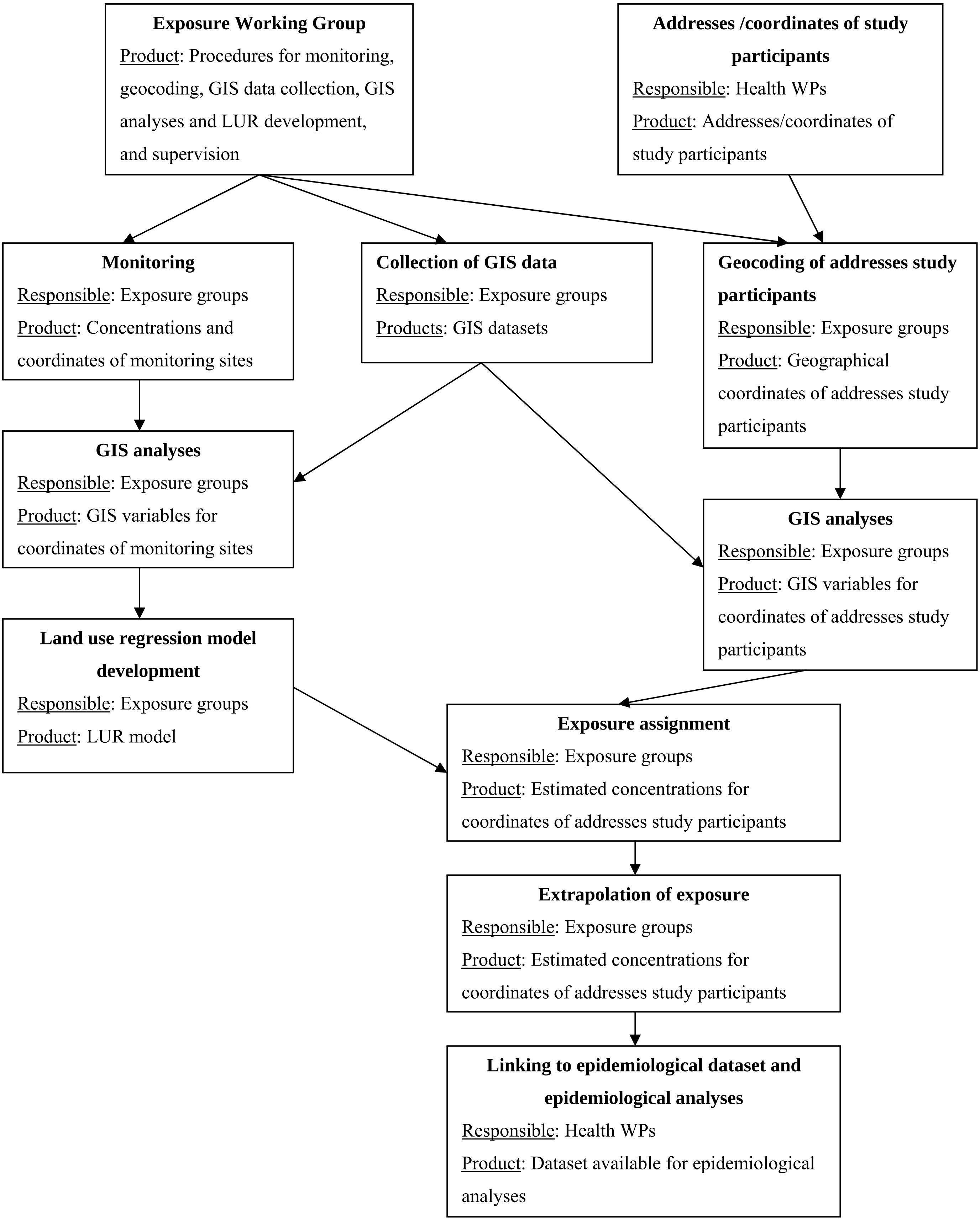
Overview of exposure assessment process

For the geographical coordinates of the ESCAPE monitoring sites and the addresses of cohort members GIS data will be collected. These data will be used as potential predictor variables in exposure models to predict air pollution concentrations at addresses of cohort members, i.e. using land use regression (LUR) models.

The GIS analyses, exposure assessment model development and exposure assessment for the addresses of study participants will be conducted by the groups who do the measurements within the study areas. If this is not feasible for logistic or scientific reasons, analyses will be performed centrally. The exposure groups who conduct the monitoring campaigns have budget to conduct the GIS analyses, to develop exposure assessment models and to estimate exposure for the study participants.

Figure 1 below shows the data structure of the exposure assessment process and who is responsible for which step (Health WPs or Exposure groups, i.e. persons in a group that conduct the measurements, conduct geocoding, conduct GIS-analyses, develop LUR models and estimate exposure for cohort addresses). The ESCAPE Exposure Working Group will provide procedures for monitoring, geocoding (which is described in a separate Geocoding manual), GIS data collection, GIS analyses and LUR development, and will provide supervision of these steps. Transfer of data has to take place when the addresses of study participants and/or geographical coordinates of these addresses will go from the Health WPs to the exposure groups for geocoding and exposure assessment. Depending on the privacy regulations for each study, this may introduce a privacy issue. Please discuss with the people from the Health WPs that are responsible for the study data whether there is a privacy issue. After the exposure assessment has been conducted for the geographical coordinates, the exposure estimates will be transferred to the Health WPs for epidemiological analyses. Because of the large number of study areas and studies and the diversity of these study areas and studies, the described procedures in this ESCAPE exposure assessment manual may not always exactly apply to each study (area). Questions about the procedures in the ESCAPE exposure assessment manual can therefore be sent to IRAS (Rob Beelen (r.m.j.beelen@uu.nl) and / or Gerard Hoek (g.hoek@uu.nl)).

Figure 1: Data structure of the exposure assessment process, and who is responsible for which step


Exposure Working Group
Product: Procedures for monitoring, geocoding, GIS data collection, GIS analyses and LUR development, and supervision
Addresses /coordinates of study participants
Responsible: Health WPs
Product: Addresses/coordinates of study participants
Monitoring
Responsible: Exposure groups
Product: Concentrations and coordinates of monitoring sites
Collection of GIS data
Responsible: Exposure groups
Products: GIS datasets
Geocoding of addresses study participants
Responsible: Exposure groups
Product: Geographical coordinates of addresses study participants
GIS analyses
Responsible: Exposure groups
Product: GIS variables for coordinates of monitoring sites
GIS analyses
Responsible: Exposure groups
Product: GIS variables for coordinates of addresses study participants
Land use regression model development
Responsible: Exposure groups
Product: LUR model
Exposure assignment
Responsible: Exposure groups
Product: Estimated concentrations for coordinates of addresses study participants
Extrapolation of exposure
Responsible: Exposure groups
Product: Estimated concentrations for coordinates of addresses study participants
Linking to epidemiological dataset and epidemiological analyses
Responsible: Health WPs
Product: Dataset available for epidemiological analyses

2. Calculation of annual average concentrations for and geocoding of the monitoring sites

The calculation of annual average concentrations proceeds in four steps after field work and laboratory analysis have been completed:

1. Reporting of analysis results (filter weights etc) by IRAS to the centers
2. Data entry of field forms and inclusion of analysis results in Excel files prepared by IRAS. This is the responsibility of the local centers doing the field work
3. Consistency checks by IRAS, resulting in a final dataset of measurements, including definition of average blank levels, limits of detection and precision.
4. Calculation of annual averages for all monitoring sites. The procedures are slightly different for the NO_x only and the PM+NO_x because in the NO_x only centers measurements are performed simultaneously for all sites. Calculation of annual average concentrations should be conducted by the Exposure groups based on the measurement data.

Measurements at the ESCAPE monitoring sites will not be performed simultaneously. For example, PM measurements in one season are spread over 4 x 2 weeks, and for both PM and NOx additional measurements will be conducted if there are missing data due to loss of samplers etc. Therefore differences may occur due to temporal variation, which have to be adjusted before calculation of the annual average concentration. The adjustment has to be conducted using data from the ESCAPE reference site or a continuous monitoring site.

In the calculations use the following rules to include individual samples:

- They should fulfill SOP requirements for sampling time (> 67%), for flow for PM measurements (start and end > 8 l/min) and for analysis
- Do not change measurements below the detection limit (do not set them to a fixed value)
- Do not replace (small) negative values with zero or another fixed value
- Do not include duplicates, as they are not available for all measurements

2.1 NO_x only sites

The simple arithmetic mean of the available measurements per site is taken after adjustment for temporal variation using the difference between the sampling period and the annual average of the study period (see below). The annual average is calculated from (urban or rural) background station routine monitoring data for the ESCAPE study period. This adjustment is applied to all samples. For sites with three valid measurements, the adjustment is just scaling, the difference between sites is not affected. For sites with less than three measurements the adjustment also limits bias related to missing samples in a high or low pollution sampling period.

The adjustment will be similar to the procedure for the PM+NO_x sites, with the difference that no ESCAPE reference site has been designated. For adjustment, data from an urban or rural background NO_x monitor will be used. If multiple monitors are available, use all monitors that have at least 75% data capture, i.e the annual average should consist of monitoring data from at least 75% of 365 days.

The adjustment procedure is as follows for NO_x, NO_2 and NO separately:

- calculate the annual average for the continuous monitoring site(s) for the 12 months period from the start date of the ESCAPE measurements: $c_c(avg)$
- calculate for the continuous measurement site the difference of the measurement for each 2-week period t in which ESCAPE measurements have been conducted (t=1 to 3 or more) from the annual average : $dc_c(t) = c_c(t) – c_c(avg)$, with : c_c being the concentration at the routine continuous monitoring site(s)
- subtract the difference for period t from the measurement at site i (i=1 to 40) in period t: $c_{i\text{ , adjusted}}(t) = c_i (t) - dc_c(t)$
- Calculate the arithmetic mean of the adjusted concentrations and the standard error of the mean to document how well the mean is established
- Calculate the arithmetic mean of the original concentration measurements and the standard error of the mean and prepare a scatterplot of these unadjusted average concentrations versus the adjusted concentrations including the R^2 and the linear regression equation linking them. This is performed to document the impact of adjustment.

For areas where no routine background monitoring data are available, another approach is necessary to avoid bias for sites with less than three measurements. Calculate the arithmetic mean concentration for each of the three sampling periods using only sites with three valid measurements. The average of these three value is then used as the $c_c(avg)$ in the procedure above. This calculation only removes bias for sites with less than three observations, it does not scale to true annual averages.

2.2 PM+NO_x only sites

The simple arithmetic mean of the available measurements per site is taken after adjustment for temporal variation using the difference between the sampling period and the annual average of the study period (see below). The annual average is calculated from the ESCAPE reference site for the ESCAPE study period. This adjustment is applied to all samples.

The adjustment procedure is as follows for PM_{10}, $PM_{2.5}$ and absorbance separately:

- calculate the annual average for the ESCAPE reference site for the 12 months period from the start data of the ESCAPE measurements: $c_r(avg)$,
- calculate for the ESCAPE reference site the difference of the measurement for each 2-week period t in which ESCAPE measurements have been conducted (t=1 to 12 or more) from the annual average $dc_r(t) = c_r(t) - c_r(avg)$
- subtract the difference for period t from the measurement at site i (i=1 to 20 / 40) in period t: $c_{i\,,\,adjusted}(t) = c_i\,(t) - dc_r(t)$
- Calculate the arithmetic mean of these adjusted concentrations and the standard error of the mean to document how well the mean is established
- Calculate the arithmetic mean and the standard error of the mean of the original concentration measurements and prepare a scatterplot of these concentrations versus the adjusted concentrations including the R^2 and the linear regression equation linking them. This is performed to document the impact of adjustment.

In case of missing data at the reference site, the data have to be imputed from data from a routine network, provided that the correlation is sufficiently high. It is considered a waste of resources if the data from a complete measurement period have to be excluded. Imputation will be performed using regression analysis. Communicate with IRAS if you need to do this (g.hoek@uu.nl). A high correlation is necessary (e.g R larger than 0.70) to do this. For components that are not measured in the network (such as absorbance), one can include other pollutants such as Black Smoke or NO_x.

2.3 Calculated concentrations

The following annual average concentrations will then be calculated for each ESCAPE monitoring site using the unadjusted and adjusted concentrations:

- PM2.5
- PM10
- PM_{coarse} (PM10-PM2.5)
- Absorption coefficient PM2.5
- NOx, NO2 and NO

The adjusted concentrations will be used in the exposure assessment modeling. The standard error of the mean may be used in further modeling.

We will further have data on elemental composition of PM2.5. These will become available in a later stage. We will use a selection of elements taking into account toxicity, indicator for sources and data quality.

2.4 Geocoding of monitoring sites

Rather than averaging GPS coordinates recorded at several site visits, the x and y coordinates of the monitoring locations will be extracted from accurate, preferably digital, maps (e.g. cadastral or topographical survey maps of 1:10,000 or better). Preliminary investigations into the GPS readings of the first year groups showed unacceptable variation at some sites of repeated GPS readings.

Each study area will use the appropriate local coordinate system / national grid. Re-projection might be necessary to convert all spatial data sets (monitoring sites, GIS data sets and cohort address locations) into this coordinate system. For assistance please contact Kees de Hoogh at Imperial College (c.dehoogh@imperial.ac.uk).

3. Collection of GIS datasets

Potential predictor variables for exposure modeling will be derived from GIS datasets. Table 1 lists some of the potential predictor variables that can be applied in modelling spatial variations in air pollution concentrations. Important predictor variables include various traffic representations, population density, land use, physical geography (e.g. altitude) and climate, but there might be other relevant area-specific predictor variables (Hoek et al., 2008).

Table 1. Potential predictor variables that can be applied in land use regression modelling

Variable	**Specification**	**Spatial scale / buffer size**
Traffic intensity nearest street	Motor vehicles per day, if possible separated into light, medium-heavy and heavy vehicles	NA
Distance to nearest street	Typically distance of object to centre of the road	NA
Traffic intensity buffers	Motor vehicles per day, if possible separated into light, medium-heavy and heavy vehicles in buffers around the sampling point	Circles with radii of e.g. 100m, 300m, and 500m around the sampling point
Height	Height above ground	NA
Distance to nearby major road	Typically distance of object to centre of the nearest major road	Within 500 meter
Traffic intensity on nearest major road	Motor vehicles per day, if possible separated into light, medium-heavy and heavy vehicles in buffers around the sampling point	NA
Population density	Population density in buffers around the sampling point	Circles with radii of 300m, 1000m and 3000m around the sampling point
Household density	Household density in buffers around the sampling point	Circles with radii of 300m, 1000m and 3000m around the sampling point
Land use	Land use in buffers around the sampling point (e.g. residential land, industry, urban green)	Circles with radii of 300m, 1000m and 3000m around the sampling point
Altitude	If important altitude differences exist	NA
Meteorology	If important meteorological differences exist (e.g. wind speed, wind direction, temperature)	NA
Distance to features	E.g. distance to sea, major lake or specific source area, border crossing	
Coordinate variables	If the study area is large there might be regional variation in air pollution concentrations. These regional variations may not be taken into account by the other variables.	NA

NA = not applicable

European wide GIS data will be centrally obtained and re-projected by Imperial College for each study area and provided at the April 2010 and February 2011 GIS/LUR workshops. This will be done in order to ensure consistency, facilitate licensing and avoid duplication of effort. Locally available GIS data should also be collected by each centre as not all predictor variables are centrally available. Local GIS data may also be better than that available in the central GIS data. Both the central and local GIS predictor variables will be evaluated similarly for exposure model development.

3.1 Central GIS datasets

The following data sets will be made available to partners.
In Table 2 it is described which datasets are "Default" or "Backup" datasets. "Default" datasets are the central GIS datasets that can be used regardless whether similar local GIS data are available. "Backup" datasets are the central GIS datasets that should only be used when similar local GIS datasets are not available.

Table 2: Available central GIS datasets.

Central GIS dataset	**Default / backup**
Digital road data	Backup / Default
Land use data	Default
Population	Backup
Altitude	Backup

Below some more detailed information about the central available GIS datasets.

1. *Digital road data.*
   These data are essential inputs to land use regression and dispersion models.
   High resolution road data was obtained by Imperial College for all countries represented in ESCAPE. Eurostreets version 3.1 is a 1:10,000 digital road network which is based on the TeleAtlas MultiNet TM. Attributes include name of street, road classification, route number, speed and length. The FRC road classification classes are described in Appendix I. Imperial College will re-project the EuroStreets data for each study area. It is important to note that no traffic intensity is attached to this data set.
   Checking for availability of local GIS road network data with traffic intensities attached has therefore a high priority.

2. *Land use data.*

   These data are key inputs to the land use regression models. The data will also be used as the geographic base for all GIS data sets.

   CORINE land cover 2000 (CLC2000) is available from the EEA as a 1:100,000 seamless vector database. It comprises 44 land cover classes, and has a spatial minimum mapping resolution of 25 hectares. CORINE data are not available for Norway and Switzerland. Imperial College will re-project the CORINE data for each study area into the relevant coordinate system.

   We will use the land use categories as used in a recent LUR paper in the UK and Netherlands (Vienneau et al, 2010). The original CORINE categories will be used/regrouped by summation as indicated in Table 4. Appendix II describes the CORINE classes that will be used.

3. *Population density data*

   High resolution modelled population data for ca. 2001 will be distributed by Imperial College (available from the INTARESE project).

4. *Altitude*

   Central altitude data have not been collected but the SRTM 90m Digital Elevation Data is available for download from the following website: http://srtm.csi.cgiar.org/. Please contact Imperial College if assistance is needed in re-projecting these data for your study.

3.2 Local GIS datasets

Local GIS data should be collected by the local centers that do the exposure assessment modeling. Local data should be collected with a reasonable effort. Table 3 describes all local GIS datasets for which it should be evaluated whether these data are available in your study area (in decreasing order of priority, together with the required resolution/accuracy). Please indicate also if data are not available and why these data are not available (e.g. not existing, costs too high, not accurate enough). The Exposure Working Group will evaluate this for each study area. Below (page 18) guidelines are given which information the Exposure Working Group would like to receive from each dataset. Please collect this information and send that to the Exposure Working Group.

If available, data for different years, i.e. for the current situation (because monitoring takes place for this time period) and for the relevant time window of exposure, should be collected. Retrospective information on changes in land use, road networks, traffic flows etc can be used to reconstruct historical trends.

Table 3: Local GIS datasets that should be collected in each study area, in decreasing order of priority, and with required accuracy / resolution

Local GIS dataset	**Maximum resolution / accuracy**
Local road network with linked traffic intensities or road type	10 meter
Population and/or household density	100 meter
Altitude	100 meter
Information about height of buildings, canyon street and other street configuration	10 meter
Study area specific local data: for example, information about wood smoke, distance to sea/lake, distance to major air pollution source.*	**
Land use	100 meter
Emission data	1000 meter
Satellite data	100 meter

* If applicable to your study area

** Depending on the variable

The local GIS are described in more detail below.

1. *Local digital road network with linked traffic intensities or road type*

   As described above an Europe-wide digital road network will be centrally available. Although this road network is accurate, it includes only a road classification but no traffic intensities (for road classifications in the central road network, see Appendix I). Local GIS road network data with traffic intensities attached or with road type information should be collected. The required accuracy for the local digital road network should be 10 meter. This should apply to road sections between intersections as well (for route finding databases this is not critical). A further issue is the completeness of the database, realistically especially for major roads.

   In some areas traffic intensities might already be linked to a local digital road network, while in other areas traffic intensity data are available but not linked yet to a road network. The collection and linkage of traffic intensities to a digital road network should be done locally. Collecting of the traffic intensity data and linking to the road network may be time consuming! Collection of traffic intensity data, especially those for municipal roads, is often problematic as they are only available for a small number of streets, and mainly on major roads, in many cities. If possible, traffic intensity data should be collected for different years, i.e. current data, but also years in the relevant time window of exposure. Preferably, traffic intensity data should be collected for different traffic types (light-duty, medium-duty and heavy-duty traffic). However, if this is not available total traffic intensity can be used. Linkage of traffic intensity data to a digital road network can be done based on street section, street name or street number.

If traffic intensity data are not available, the length of specific road types / classifications can be used as potential predictor variables. E.g in the UK, roads are classified as motorways, A-roads, B-roads. Several land use regression applications have successfully explored the use of the length of specific road types without traffic intensity data. For this the European-wide digital road network will be used because this road network is accurate and has a road classification (Table 4 and Appendix I).

2. *Population and household density*

   Local population density data should be collected because the central available data are modeled and not validated. In addition to local population density data, household density data are also interesting to collect. Some studies successfully used household density as potential predictor variable for air pollution concentrations. The correlation between population density and household density is normally high. If possible, data should be collected for different years, i.e. for the current situation but also for the relevant time window of exposure. The required accuracy for population and household density data is 100 meter.

3. *Altitude*

   Altitude data are needed for study areas with relevant altitude differences. If readily available and of better spatial resolution, local altitude data may be used. Otherwise the SRTM 90m Digital Elevation Data may be used: http://srtm.csi.cgiar.org/. The required accuracy for altitude data is 100 meter.

4. *Information about address height, height of surrounding buildings and street configuration such as canyon type*

   Evaluate whether GIS databases on height of surrounding buildings and location of street canyons or other information about street configuration (e.g. information about traffic speed) are available. Because air pollution concentrations have a vertical gradient with lower concentrations at increasing height, but with higher concentrations in canyon streets, this information might be useful. This will not be easy to obtain, but a reasonable effort should be made to obtain data. For smaller studies (for example < ~1000 subjects, e.g. some birth cohorts), data could be obtained manually, using the form used for monitoring site characterization. The importance of these variables has recently been illustrated in a few land use regression studies. Inclusion of these variables would be a contribution from ESCAPE.

   If these data are not available in a GIS, for each monitoring site it could be evaluated what the height is and whether it is located in a street canyon. A limitation may however be that all traffic sites are on ground level or first floor, so the effect of height for traffic sites might be difficult to evaluate. This information can then be added to the developed final land use regression model and it can be evaluated whether the model improves by adding these data. If not, data about height and street canyon are not needed for the addresses of the study participants. If height and location of street canyon improve the model but these data are not available in a GIS, alternatives might be used for estimating exposure at addresses. Information about height is sometimes available from study questionnaires or the floor can sometimes be deduced from the address information. Further, dispersion models could be used to assess the effect of height on air pollution concentrations. This 'height' factor could be applied to the land use regression estimates.

5. *Study area specific local GIS data*

   This may include information about wood smoke, distance to sea/lake, distance to major air pollution source, distance to harbor, etc. For each area it has to be evaluated whether such study area specific local data are needed and if these data are available in a GIS. If available in GIS, these data should be evaluated when developing the LUR model.

6. *Land use data*

   The central available CORINE land use data have been shown to be predictors in intra-urban land use regression models. Raw CORINE land cover data have however a resolution of ~25 ha, but can be used in grids of 100m. The quality of CORINE data and classification varies may not be entirely consistent between countries. Further, CORINE data may not have incorporated specific land use for a specific area or country. It is thus recommend that each centre collects local GIS land use data. We anticipate that only modest gains will be obtained by getting local data, hence collection of local land use data has modest priority. If available, data for different years, i.e. for the current situation as well as for the relevant time window of exposure, should be collected. The required accuracy for local land use datasets is 100 meter.

7. *Emission data*

   The central available emission data are only available for 50km grids and can therefore not be used for modeling at smaller scales. Availability of more detailed local emission data should be checked (preferably for different years, i.e. for the current situation and for the relevant time window of exposure). The required accuracy for local emission data is 1000 meter. Source-specific data should be obtained.

8. *Satellite data*

   We will check the possibilities to include satellite data into the assessment. Satellite data will be used as potential predictor variables in the exposure models, for example satellite data of land use. There are no central satellite data available, so the availability and relevance of satellite data should be evaluated locally. Please evaluate for your study area which satellite data are available and whether they can be used as predictor variables in the exposure assessment model. The required accuracy for the satellite data is 100 meter. Su has recently demonstrated the usefulness of using satellite data for characterizing street configuration, though in a geographically simple setting (Su, 2008).

The availability of all the local GIS datasets as described in Table 3 should be checked. For each of the evaluated local GIS dataset the following information is then needed, please collect this information for each local GIS dataset:

- Whether the local GIS dataset as described in Table 3 is available. And if not, why the dataset is not available
- Name of GIS dataset
- Description of GIS dataset
- Type of data
  Are the data vector data or raster data? And if the data are raster data, how large are the grids?
- Accuracy/resolution of the dataset
  What is the resolution of the dataset? Because of the large spatial variation in air pollution concentrations close to roads and in urban areas, geographic precision of GIS databases of potential predictor variables is important and should be evaluated and documented. Further, evaluate whether there are differences in accuracy within a study area.
- Completeness
  Does it cover the whole study area? Which area does the dataset cover (e.g. national dataset)?
- Coordinate system
  What is the coordinate system of the GIS dataset (local, national, latitude/longitude coordinate system etc)? For the GIS analyses, all data have to be in the same coordinate system. Preferably conversion to different coordinate systems has to be restricted to a minimum, because this may result in loss of accuracy.
- Year(s) of data availability
  If available, data for different years, i.e. for the current situations as well as for the relevant time window of exposure, should be collected.
- (Potential) Costs of the dataset

Send this information to Rob Beelen (r.m.j.beelen@uu.nl) AND Gerard Hoek (g.hoek@uu.nl), so the data can be evaluated in the Exposure Working Group, and an overview can be made of which GIS data are available in the different study areas.

3.3 Area-level potential confounders for the epidemiological analysis

Area-level potential confounder data will not be used as potential predictor variable for the exposure model. It is listed here because it is efficient that these data are collected by the same persons collecting data on GIS predictor variables.

Several studies have shown that apart from individual socio-economic status (SES)-level, the SES level of for example a neighbourhood may also be an important confounder. Appendix III gives an overview of used area-level confounder variables in other studies and discusses which area-level confounder variables may be relevant within the ESCAPE study. This will be further discussed with the Health WP leaders. Please evaluate for your study area which of the described possible area-level confounder variables are relevant for your study area (or maybe there might be other potential data relevant), whether these data are available and for which spatial scale and which time period these data are available. As these data will be used in the epidemiological analyses, it should also be discussed with the people responsible for the epidemiological analyses for that study area which data are needed and relevant.

4. GIS analyses

With the collected GIS datasets GIS analyses will be conducted to derive the values for the predictor variables for the coordinates of the monitoring sites and the addresses of the study participants (section 4.1). For mapping purposes, also the values for the centroids of 100 m grids over the study area should be obtained (see chapter 5, this is only necessary for the predictor variables that are included in the final LUR models).
In addition, GIS analyses will also be conducted to collect traffic variables that will be used as independent exposure variables in the epidemiological analyses (Section 4.2). Some of the traffic variables will also be used as predictor variable in the LUR models. Section 4.3 gives a short explanation of some of the key GIS-analyses that may be conducted.

It is important to note that for the monitoring sites the geographical coordinates will be already available after the monitoring campaign has finished and GIS analyses can then already be conducted for the coordinates of the sites regardless whether the coordinates for the address of study participants are already available. Because for some studies the addresses of study participants first have to be geocoded, which may cost some time, the GIS analyses for the coordinates of the monitoring sites and the development of land use regression models should already start. The GIS analyses for the coordinates of study participants' addresses can then start after the addresses have been geocoded.

4.1 Predictor variables for land use regression development

Before starting with the GIS analyses for the predictor variables, please make first a table with the *a priori* selected predictor variables for you study area (see also Table 1 for potential relevant predictor variables) by updating Table 4. Table 4 shows the buffers and directions of effect that are defined and should be used within the ESCAPE project. Do not use other definitions.
The coordinates will only be offered if a model has been developed to test if the model with more explicit variables can be improved with coordinates (describing slow trends in background).
Please also use the variable names as described in Table 4 because a combined dataset with all data from the different study areas will be made.

Table 4. Predictor variables with predefined variable names, units, defined buffer sizes, transformations of the predictor variables and directions of effect.

GIS dataset	Predictor variable	Name variable[1]	Unit	Buffer size (radius of buffer in meter)	Transformation	Direction of effect
Background						
-	Coordinate variables[2]	XCOORD, YCOORD	m	NA	Local decision	NA
CORINE	High density residential land [3]	HDRES	m^2	100, 300, 500, 1000, 5000	-	+
CORINE	Low density residential land [3]	LDRES	m^2	100, 300, 500, 1000, 5000	-	+
CORINE	Industry [3]	INDUSTRY	m^2	100, 300, 500, 1000, 5000	-	+
CORINE	Port [3]	PORT	m^2	100, 300, 500, 1000, 5000	-	+
CORINE	Urban green [3, 4]	URBGREEN	m^2	100, 300, 500, 1000, 5000	-	-
CORINE	Semi-natural and forested areas [3, 5]	NATURAL	m^2	100, 300, 500, 1000, 5000	-	-
Local land use			m^2	100, 300, 500, 1000, 5000	-	Following CORINE
Population density	Number of inhabitants [3]	POP	N(umber)	100, 300, 500, 1000, 5000	-	+
Household density	Number of households	HHOLD	N(umber)	100, 300, 500, 1000, 5000	-	+
Altitude	Altitude	SQRALT	m	NA	square root	-
Traffic [6]						
Local road network	Traffic intensity [6] on nearest road	TRAFNEAR	$Veh.day^{-1}$	NA	-	+
Local road network	Distance to the nearest road	DISTINVNEAR1 DISTINVNEAR2	m^{-1}, m^{-2}	NA	Inverse distance and inverse distance squared	+
Local road network	Product of traffic intensity on nearest road and inverse of distance to the	INTINVDIST INTINVDIST2	$Veh.day^{-1}m^{-1}$ $Veh.day^{-1}m^{-2}$	NA	-	+

	nearest road and distance squared					
Local road network	Traffic intensity on nearest major road [7]	TRAFMAJOR	Veh.day^{-1}	NA	-	+
Local road network	Distance to the nearest major road [7]	DISTINVMAJOR1 DISTINVMAJOR2	m^{-1}, m^{-2}	NA	Inverse distance and inverse distance squared	+
Local road network	Product of traffic intensity on nearest major road and inverse of distance to the nearest major road and distance squared [7]	INTMAJORINVDIST INTMAJORINVDIST2	Veh.day^{-1}m^{-1} Veh.day^{-1}m^{-2}	NA	-	+
Local road network	Total traffic load of major roads in a buffer (sum of (traffic intensity * length of all segments)) [7]	TRAFMAJORLOAD	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000	-	+
Local road network	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	TRAFLOAD	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000	-	+
Local road network	Heavy-duty traffic intensity on nearest road	HEAVYTRAFNEAR*	Veh.day^{-1}	NA	-	+
Local road network	Product of Heavy-duty traffic intensity on nearest road and inverse of distance to the nearest road and distance squared	HEAVYINTINVDIST HEAVYINTINVDIST2	Veh.day^{-1}m^{-1} Veh.day^{-1}m^{-2}	NA	-	+
Local road network	Heavy-duty traffic intensity on nearest major road	HEAVYTRAFMAJOR	Veh.day^{-1}	NA	-	+
Local road network	Total heavy-duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	HEAVYTRAFMAJORLOAD	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000	-	+
Local road network	Total heavy-duty traffic load of all roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	HEAVYTRAFLOAD	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000	-	+
Central road network	Road length of all roads in a buffer	ROADLENGTH	m	25, 50, 100, 300, 500, 1000	-	+
Central road	Road length of major roads in a buffer [8]	MAJORROADLENGTH	m	25, 50, 100, 300, 500, 1000	-	+

network						
Central road network	Distance to the nearest road	DISTINVNEARC1 DISTINVNEARC2	m^{-1}, m^{-2}	NA	Inverse distance and inverse distance squared	+
Central road network	Distance to the nearest major road [8]	DISTINVMAJORC1 DISTINVMAJORC2	m^{-1}, m^{-2}	NA	Inverse distance and inverse distance squared	+
	Aspect ratio (sum height buildings both side of road divided by road width) [9]	CANYON [9]	m/m	NA		

[1] Variable name: Combining name and buffer size, for HDRES: HDRES_100, HDRES_300, HDRES_500,HDRES_1000, HDRES_5000

[2] The coordinates will only be offered if a model has been developed to test if the model with more explicit variables can be improved with coordinates (describing slow trends in background).

[3] Area of that land use in the buffer (m^2)

[4] Corine Urban green is the sum of classes 141 and 142

[5] Corine semi-natural is the sum of classes 311, 312, 313, 321, 322, 323, 324, 331, 332, 333, 334, 335, 411, 412, 421, 422, 423, 512, 521, 522 and 523

[6] Traffic intensities are traffic intensities per 24h

[7] Definition of major road for local road network: road with traffic intensity > 5,000 mvh/24h

[8] Definition of major road for central road network: classes 0, 1, and 2 (+ classes 3 and 4 based on local knowledge and decision)

[9] To be decided later how to be exactly defined and how to be used

NA is not available

This paragraph describes some background information about the size of the buffer area for a predictor variable. Predictor variables in land use regression models are usually computed as circular zones around each monitoring site, using buffer functions available in GIS. This is done to evaluate the impact of predictor variables at different spatial scales. The selection of buffer size is crucial in determining the performance of the model, and the spatial resolution of the estimates. Ideally, buffer sizes should be selected to take account of known dispersion patterns. Various monitoring studies have shown that the impact of a major road on concentrations of traffic-related air pollutants declines exponentially with distance to the road. Beyond about 100 m from a major urban road, or 500 m from a major freeway, variability is limited. In inner-city areas, however, buildings may cause marked departures from this simple distance-decay pattern (e.g. for street canyons). There is also evidence to suggest that air pollution concentrations fall virtually to background levels behind a row of uninterrupted buildings. Especially in the compact European urban areas, much of the variation in traffic-related air pollution is therefore extremely local. The use of the larger buffer sizes (300, 500, 1000 m) is included to reflect the sum of emissions in a larger area, not specifically the nearest roads. In urban areas the buffer size should however be even smaller (for example a radius of 100 meters). For variables such as population or address density, and land use the buffer size may however be larger (see Table 4).

Traffic intensity on the nearest street is an important variable. In some local road networks, minor roads were present but without traffic intensity attached. Several local road networks with traffic intensity data, only contained the major roads. Minor roads were completely excluded from the network. In order to use the variable traffic intensity and distance to nearest street, a low value should be assigned to these minor roads. We will all assign the value 500 to these streets. In the first scenario this is sufficient as the distance to the nearest street is valid. In the second scenario (incomplete network), also the distance to the nearest street is incorrect. The following steps should be made:

- Calculate distance to nearest street (DISTNEAR) with the local network and central road network (DISTNEARC)
- If DISTNEAR – DISTNEARC < 10 m, then keep DISTNEAR and TRAFNEAR
- If DISTNEAR – DISTNEARC > 10 m, then replace DISTNEAR with DISTNEARC. If additionally, the central road has class 3 or higher the standard low value is assigned. If it has class 0, 1, 2 we will assign the average traffic intensity of major roads, provided this occurred in a small number of cases.

4.2 Traffic exposure variables

GIS analyses will also be conducted to collect traffic variables for the addresses of study participants that could be used as independent exposure variables in the epidemiological analyses.

The following traffic variables have to be collected (if data are available). The first four are also predictors in exposure modeling, the MAJORROAD variable is an additional variable.

- Traffic intensity on the nearest road (TRAFNEAR) and inverse distance to the nearest road (DISTINVNEAR1), based upon local road network
- Traffic intensity on the nearest major road (TRAFMAJOR) defined as a road with traffic intensity > 5,000 vehicles / day and inverse distance to the nearest major road (DISTINVMAJOR1), based upon a local road network
- Total traffic load (intensity*length) on major roads in a 100m buffer (TRAFMAJORLOAD), based upon local road network
- Indicator variable indicating whether a coordinate is within 50m of a class 1 or 2 type road and/or within 100m of a class 0 road (=motorway), based upon central road network. This needs to be calculated and named MAJORROAD.
- Sum of road length of major roads defined as class 0, 1 or 2 (and possibly classes 3 or 4 based upon local knowledge) from the central road network within a 100m buffer (MAJORROADLENGTH).

These traffic indicators will be included in the epidemiological analyses together with a modelled background concentration. For this we will use NO_2 as this pollutant is available for all study areas and is measured at more locations than PM, which is important when we will model background only sites (typically about 50% of all sites). The modelling of this background concentration will be described in section 5.2.2 (background concentrations land use regression modelling).

4.3 Explanation GIS analyses

The GIS analyses will be conducted using ArcGIS. A multi-day training workshop will be organized about GIS analyses (and land use regression modelling) in the beginning of 2010 for the groups that were in measurement year 1 (and early 2011 for the groups that were in measurement year 2).

The main analyses that will be conducted are calculating distances from coordinates to air pollution sources (for example nearby roads) and calculating the value of a predictor variable in a buffer around a coordinate (for example area of industry in a buffer). After finishing these analyses, conduct some checks. For example select the observations with the lowest and highest traffic intensities and the observations with the highest percentage of industry and check whether this is correct for these coordinates using for example Google Earth or other route maps.

The following text briefly describes commands that can be used for the main GIS analyses. More details about these commands can be found in the help-function of ArcGIS.

1. Creating multiple buffers around monitoring sites.

In ArcToolBox go to Analysis Tools → Proximity → Multiple Ring Buffer. Enter relevant buffer distances (see Table 4). Set the 'Dissolve Option' to 'NONE' then press 'OK'.

2. Intersect

The multiple buffers around each monitoring site are intersected with GIS data (e.g. road data, landcover data) In ArcToolBox go to Analysis Tools → Overlay → Intersect. In the Intersect window select the input features (e.g. roads and the buffer shape files) and add them to the Feature list. In the Output Feature Class enter a name for the new shapefile. Click 'OK'.

3. Update Geometry

The intersect tool cuts out features (lines or areas) which falls inside the buffer. ArcGIS does not automatically recalculate the length (or area) for these intersected features in shape files. The geometry of the resulting shape file, therefore, needs updating. To do this open the attribute table of your newly created intersect shape file and right-click the Length (or Area) field and choose Calculate Geometry from the drop down menu. In the Calculate Geometry dialog, check that 'Property', Coordinate system and 'Units' are correct (e.g. length = m, area = m2) and click 'OK'. If there is no length (or area) field in the shape file first add it as a new field (type = Double) to the attribute table and then calculate the geometry.

4. Calculating Proximity (to roads)

A commonly used indicator in epidemiological investigations of traffic related air pollution and health is the distance between subject's home address and the nearest road. In ArcInfo this can be computed by the NEAR command. In ArcMap this is done by spatial join, by right clicking on the monitoring site and from the drop-down menu selecting Join and Relates → Join. In the Join Data dialog you can append data from another layer to monitoring sites file. Choose the 'spatial location' option. In the Join Data dialog, under 1) choose the layer which you want to join to the monitoring sites (e.g. the road shape file). Under 2) tick the second button so that each monitoring site will be given all the attributes of the nearest road segment. Under 3) you need to give a name to the new layer which will be created. The resulting shape file will have a new field 'distance' which is the distance to the nearest road segment.

5. Area-weighting

Area weighting is used to redistribute data from one geography (source map unit, e.g. census areas) onto a different, non-overlapping geography (target map unit, e.g. buffers). It is often used to redistribute census totals to calculate small-area population estimates. The total population from the census remains unchanged after area weighting.

$$P_t = \sum_{s=1}^{S} \frac{P_s \times A_{ts}}{A_s}$$

Where: P_t is the population in target map unit t; P_s is the population in source map unit s; A_s is the area of source map unit s; and A_{ts} is the area of target map unit t overlapping source map units.

The steps are as follows:

- Intersect buffer with census area (with population data attached)
- Rename the area field of the census shape file as 'area_tot'
- Add a new item to the intersect field called 'area_int'
- Recalculate the geometry of 'area_int'
- Add a new field 'pop_weighted'
- Use the field calculator to compute the area-weighted population in the 'pop_weighted' field: ('area_int' /'area_tot') x 'census_pop'

The sum of the weighted population within each buffer can be computed using for example PIVOT table in Excel.

DEALING WITH RASTER DATA

Until now we have only described analyses with vector based GIS data (roads, landcover). However, GIS data is also available in raster format (e.g. altitude and population) and is handled in a different way.

6. Extract values to points (for Altitude)

The altitude variable is not calculated in buffers. Instead this variable is simply the altitude (in m) at the point location (x,y coordinates) of the monitoring site and/or cohort address. The tool to extract values from a raster to x,y coordinates is found under Spatial Analyst Tools → Extraction → Extract Values to Points. Add the input point features (monitoring sites / cohort addresses) and the input raster (altitude) to create a new shape file with the altitude value attached.

7. Raster to poly (for Population)

To calculate buffer variables from raster GIS data the raster must first be converted to vector data (polygons) using the following steps:

- Raster to poly (Spatial Analyst → Convert → Raster to Features)
- Add new field 'area_new'
- Calculate geography
- Add new field 'multiplier'
- Calculate 'multiplier = area_new / area_cell'
- Add new field 'new_pop'
- Calculate 'new_pop = muliplier x gridcode'

Instead of creating a polygon feature for each cell in the raster, adjacent raster cells with the same value (e.g. 50) are merged into one larger polygon feature with that value (50). This needs to be corrected so we do not underestimate the real population in our polygon shape file. The multiplier in the steps above is used to indicate merged cells so this correction can be applied. The area_cell is constant and refers to the original cell resolution of the raster (length x width of a cell in m^2).

Now step 5 (area-weighting) can be used to compute the buffers for the population variable.

As described above, a multi-day training workshop will be organized about GIS analyses (and land use regression modelling) in the beginning of 2010 for the groups that are in measurement year 1 (and in the winter of 2011 for the groups that are in measurement year 2).

5. Development and validation of land use regression and other exposure models

5.1 Introduction land use regression modeling

After the GIS analyses have been conducted, for each monitoring location the following information is available:

- Geographical coordinate
- Annual average air pollution concentration (measured in the period 2008-2010)
- Values for potential predictor variables

The average concentrations and values for potential predictor variables will be used to develop prediction models using stochastic modelling techniques. Stochastic modelling techniques involve developing statistical associations between potential 'predictor variables' and measured pollutant concentrations. Regression techniques are often used for this purpose and the technique is often called land use regression (LUR) modelling. The developed regression equations are then used to predict concentrations at unsampled sites (i.e. the coordinates of addresses of study participants). This technique was successfully developed in a number of studies (Hoek et al 2008). These studies have shown that regression models can explain a large part of the spatial variations in air pollution concentrations.

The sections below describe the procedures for land use regression development (Section 5.2) and checks and procedures for validation of the land use regression models (Sections 5.3 and 5.4). Section 5.5 lists the other modelling techniques we will evaluate in a selected number of study areas.

5.2 Land use regression model development

Modelling is performed mainly by the local group who conducted also the measurements. Central supervision and training will be provided for the study teams. A multi-day training workshop will be organized to harmonize development of land use regression models (in combination with explanation of how to conduct GIS analyses). This training workshop was organized in April 2010 for the groups that are in measurement year 1 (and early 2011 for the groups that are in measurement year 2).

The default is that land use regression (LUR) models will be developed for each study area separately. Data from different areas will also be pooled to develop a model based on data from a larger number of monitoring sites. Study areas could be combined for modelling when areas are comparable (factors of interest are for example distance to sea, mountains, urbanization). A 'dispersion'-potential is useful to combine data from different areas. "Dispersion"-potential describes the degree of dispersion of a certain area related to geography and meteorological conditions.

The default is that models will be developed for each air pollutant separately. This implies that models will be made for PM_{10}, $PM_{2.5}$, PM_{coarse} (PM10-PM2.5), absorbance of $PM_{2.5}$, NO_x and NO_2. We added coarse PM to this because of recent interest in the health effects of this PM fraction. We will drop NO because it is measured with lower quality than NO2 and NOx. Moreover NOX is a better marker of primary emissions than NO. The number of sites available for PM models is smaller than for NOx, so in section 5 we evaluate whether PM models can be made using data from NO_x. In addition, we will have data on elemental composition. Elements will be grouped in meaningful source-related groups instead of using each individual element.

Untransformed concentrations will be used as these are more readily interpretable. Usually the use of untransformed concentrations will only modestly violate normality of the distributions, but this will be carefully evaluated. A further advantage of using untransformed concentrations is that concentration data that are used in dispersion models are also untransformed concentrations.
Some studies used the untransformed concentrations whereas other studies use the logarithm of the concentration (Gilbert et al 2005; Henderson et al. 2007; Ryan et al. 2007; Moore et al. 2007; Jerrett et al. 2007) in an attempt to better approximate a normal distribution of the residuals. When a log transformation is used, the interpretation of the model changes from an absolute contribution of variables in the model to a relative change.

The concentration data from the ESCAPE reference site in PM + NOx study area should not be used for LUR model development, because these data have been used for the time adjustment.

Before starting with model development first conduct descriptive analyses of the annual average air pollution concentrations and the potential predictor variables. This includes boxplots of concentration data and predictor variables, a.o. to evaluate whether there are outliers. Further, evaluate also the variation in predictor variables over the monitoring sites, for example check whether there are many monitoring sites with value zero for a predictor variable. Then the correlation between predictor variables will be calculated to assess the potential for co-linearity in developing models. Scatterplots including a linear regression equation and R^2 of the concentrations of the different pollutants will be prepared. The correlation is useful because it shows the potential we have to disentangle independent effects of different pollutants and the potential we have to use NO2 to predict PM. The relationship between NO_2 and NO_x is of interest to check especially the linearity of the relationship. In dispersion models NOx is modelled and then transformed into NO2, because the relationship is non-linear and dependent on the ozone concentration (Beelen et al. 2010). Standard linear regression will be used to develop a LUR model that best predicts the measured concentrations, i.e. a model that maximizes the percentage explained variability (R^2) and minimizes the error (RMSE – Root Mean Square Error).

A supervised forward stepwise procedure will be used. Predictor variables have to be a priori defined (see Section 4.1 and Table 4). The regression models will then be developed using these a priori defined predictor variables. There is no restriction regarding the number of predictor variables that is used in the final models. The model should be developed while stating in advance the sign of the regression slope of a specific predictor in the model (see Table 4). For example we know that traffic emissions will increase the concentration. Regression models that include a negative slope for traffic intensity variables will therefore not be accepted as the final model. We therefore do not use automatic selection procedures, implemented in statistical packages. Both SAS, STATA, SPSS or another statistical software package (but not Excel) can be used to develop regression models.

We will develop two models. The first model is based upon all sites and all predictors (overall concentration including background and local traffic impacts – see section 5.2.1). The second model will only include background sites and predictors (background model – see section 5.2.2). The second model is added because the background concentration can be used in conjunction with traffic intensity variables as an alternative exposure model (see Section 4.2). The centrally and locally available GIS predictor variables will be evaluated at the same time. There is no a priori priority for one of the datasets.

5.2.1 Overall concentrations (full dataset)

In step 1, univariate regression analyses will be conducted for all possible predictor variables so that each predictor variable is regressed against monitored concentrations. The concentrations of all sites will be used. The model with the highest adjusted explained variance (adjusted R^2) is regarded as the 'start model'. To this 'start model' the remaining variables will be added separately, and the effect on the adjusted R^2 recorded. The predictor variable with the highest additional increase in adjusted R^2 will be maintained in the model if three criteria are satisfied: (1) the increase in adjusted R^2 is greater than 1%, (2) the coefficient conforms to the pre-specified direction, and (3) the direction of effect for predictors already included in the model does not change. This ensures that models involving counter-intuitive associations be avoided, even if they give a stronger basis for prediction as indicated by adjusted R^2 value.

When a variable is included, other buffer sizes of the same variable can be offered to the model, both smaller and larger buffers. In model development the original sizes can be offered to judge whether they provide additional explained variability. Due to co-linearity, slopes may be instable but the predicted values and R^2 are valid. In the final model, we will rewrite the model using 'outer or inner rings' of buffers. For example, if urban green with a 1000m buffer is included in the model first (see Table 4) and then urban green within a 100m buffer with both negative signs, the final model will be written as urban green in a 100m buffer and urban green in 1000m buffer minus urban green in 100m buffer. This will result in more interpretable regression slopes.

The addition of variables in this supervised stepwise process will be repeated until there are no remaining predictor variables that add more than 1% to the adjusted R^2 of the previous regression model, which results in a 'intermediary model'.

The last step is to evaluate the significance of the variables in the model. In using adjusted R^2 as an inclusion criterion, some variables may become highly non-significant as additional variables are included in the model. As a final step, therefore, variables with p value >0.10 will be sequentially removed from the model, starting with the least significant, until all predictor variables in the 'final model' have a $p \leq 0.10$.

5.2.2 Background concentrations

In this scenario regression models will be separately developed for background concentrations only. The background model will be used in epidemiological analyses including traffic indicators (section 4.2). The model for the background scale will be developed using only background sites (both regional and urban background sites). For developing the background model only background predictor variables will be evaluated (Table 4). The model development procedure is the same as described above for the 'Overall concentration' scenario.

5.3 Checks for regression analysis including spatial autocorrelation

Standard diagnostic tests for ordinary least squares regression will be applied to the final models:

- Influential observations

  The influence of each observation on the estimates will be measured. Influential observations are those that have a large influence on the parameter estimates. Cook's D, which is a measure of influence, will be used to evaluate whether there are influential observations. Cook's D measures the change to the estimates that results from deleting each observation. If there are influential observations, it will be evaluated whether the parameter estimates in the regression model change when this influential observation is excluded from the analyses. Default is however that no data will be excluded.
- Heteroscedasticity of the residuals

  A plot will be made of the monitored concentrations and the residuals to evaluate whether there is heteroscedasticity.
- Plots of and tests for normality of the residuals

  A test for normality of residuals will be conducted (this test is however not that important).

The most important is the test for influential observations.

In addition, ordinary kriging will be conducted on the residuals of the final regression models to evaluate spatial autocorrelation in the residuals. For this the coordinates and the residuals for each monitoring site are needed. In addition, Moran's I will be calculated, as this statistic provides a significance test of spatial autocorrelation which is not available for kriging. If there is spatial autocorrelation of the residuals the assumption of independence for the residuals is violated. In most land use regression studies, it was observed that residuals of regression models did not exhibit spatial autocorrelation anymore, suggesting that ordinary linear regression is appropriate. Semi-variograms will be made for the residuals which will be used to evaluate whether there is a pattern of spatial autocorrelation in the residuals. If needed, ordinary kriging can be centrally conducted by IRAS / Imperial College. R will be used to conduct kriging.

If the tests for ordinary least squares regression are good, and if there is no spatial autocorrelation of the residuals, then the linear regression model can be used to predict concentrations at the cohort addresses. If there is significant spatial autocorrelation of the residuals, we will use universal kriging methods instead of ordinary least squares regression modelling. The resulting universal kriging model will then be validated, and will then be used to estimate concentrations at cohort addresses. R will be used to conduct universal kriging.

5.4 Model validation

Model validation is a crucial part of applying land use regression methods. Various approaches have been taken with respect to validation.

Within ESCAPE we will use the leave-one-out cross-validation, in which a model is developed for n-1 sites and the predicted concentrations are compared with the actually measured concentrations at the left-out site. This procedure is repeated n times and the overall level of fit between the predicted and observed concentrations, across all sites, then computed as a measure of model performance. The structure of the model remains constant for each estimate (Brauer et al. 2003; Hochadel et al. 2006). The performance measures that will be evaluated are the correlation between measured and estimated concentration, a scatter plot between measured and estimated concentration, and the R^2 and RMSE value of the regression equation.

Other studies have also used other validation methods. An approach is to sub-divide the monitoring sites into a training dataset for model development and a smaller group of sites for model validation (Briggs et al. 1997). This approach requires less intensive computer processing, but may be disadvantaged by the a priori division of sites (e.g. concentrations measured at the training and validation sites may differ, especially when the total number of sites is small). However, within ESCAPE there is only a limited number of monitoring sites.

5.5 Summary of LUR output

In summary, for each of the following pollutants we will have available the model used for predicting concentrations using all predictors and the cross-validation R^2 and RMSE. We will also have available a background model used for predicting background concentrations using only background predictors and the cross-validation R^2 and RMSE. The pollutants we will model include:

- PM2.5
- PM10
- PM_{coarse} (PM10-PM2.5)
- Absorption coefficient PM2.5
- NOx and NO2
- “Source-specific” elements from XRF

5.6 Central ESCAPE database

Further, the datasets from each study area including concentration data, coordinates, site IDs and predictor variables that have been used for developing these models will be combined in one large centrally available dataset. Therefore it is also important to use the variable names as described in Table 4. The exact format of the database is included in appendix V.

As has become clear during the plenary ESCAPE meeting in 2010, the comparison of our measurements with routine measurements is important for the EU policy makers. Therefore also submit these comparisons (which were included in the study manual already) as separate Excel files.

5.7 Documentation of LUR models

To harmonize development of LUR models by the individual centers further, we will also prepare a standard documentation of the models including all the steps listed in the previous section, that is both descriptive analyses and the final model and cross-validation. This documentation will be reviewed by the ESCAPE Exposure WG. Submit the documentation to IRAS, r.m.j.beelen@uu.nl

Document:

1. Distribution of adjusted annual average concentrations (min, 10th percentile, 25th percentile, 50th percentile, 75th percentile, 90th percentile, max and arthitmetic mean)
2. Scatterplots with correlation of adjusted averages of all pollutants
3. Scatterplots of unadjusted versus adjusted with correlation for each pollutant
4. Distribution of predictor variables (see 1, use names of Table 4)
5. Correlation between predictor variables
6. Final developed model with
    a. Variables included in the model
    b. Regression coefficients, standard error and p-value of all variables including intercept
    c. R^2, adjusted R^2 and RMSE of model development
    d. R^2, adjusted R^2 and RMSE of cross validation

5.8 Other exposure assessment methods

Other exposure assessment methods that will be evaluated in specific areas within ESCAPE are:

- Dispersion models

  The use of dispersion models for exposure assessment will be evaluated in a selection of the study areas in a later stadium of the project. We will not develop new dispersion models, but use already existing dispersion models. Dispersion models have been used in Stockholm, Oslo, Copenhagen and the SAPALDIA study in Switzerland. Some of the variables that will be used as potential predictor variables in the LUR models are the same as for dispersion models, and might be used.

  An attractive option is to use the output of a dispersion model as one of the inputs of a land use regression model (in addition to the other potential predictors mentioned before). In areas where dispersion models are available, this can be a sensitivity analysis.
- Bayesian Maximum Entropy (BME) models

  A proposal about BME with a description on how to compare BME with LUR modeling will be written by Audrey De Nazelle. If possible, BME will be evaluated in 3 study areas (Barcelona, UK and the Netherlands).
- Latent variable approach which can also be applied in Bayesian Maximum Entropy models. This is a technique which makes use of the information on different variables measured at a variable number for sites. In ESCAPE, models for traffic particles could be developed based on the 20 PM sites and 40 NOx sites, using the relationship between NO2 and PM at the 20 sites.
- Co-kriging is a geo-statistical technique which uses information and information on other pollutants to estimate the concentration of PM at unmeasured sites. One option is to estimate the PM concentrations at the 20 sites NOx sites where no PM measurements have been conducted and then apply the regular LUR procedures of section 5.2
- Focalsum

  Focalsum procedures to estimate concentrations may be applied to a selection of the study areas. Emission and monitoring data are needed for the focalsum procedures. Imperial College will be able to conduct the focalsum procedures.

6. Estimation of air pollution exposure at addresses of study participants

The final LUR model will be used to estimate outdoor air pollution concentrations at the addresses of study participants. The model has been developed using monitored concentrations in the period 2008-2010. These concentrations will thus be assigned to the addresses of study participants.

Exposure for the study participants will be estimated based on the geographical coordinates of the addresses of study participants. These geographical coordinates will be made available by the health work packages when the addresses have been geocoded. If the addresses have not been geocoded yet, the groups that will do the GIS analyses and exposure assessment will geocode the addresses of cohort members which will be made available by the health work packages. The procedure for geocoding is described in the ESCAPE geocoding procedure (Appendix IV). The geographical coordinates of the addresses or the addresses of study participants will be transferred from the health work packages to the groups who will do the GIS analyses and exposure modeling in the respective study areas. Because of this data transfer this may introduce a privacy issue.

Either GIS analyses have been conducted parallel to the GIS analyses for the coordinates of the monitoring sites, or the GIS analyses for the coordinates of the addresses will be conducted separately. For the coordinates of the addresses the same GIS analyses will be conducted and the same potential predictor variables will be collected compared with the coordinates for the monitoring sites (see Sections 3 and 4). When the values for the predictor variables are available for the coordinates of the addresses, it is relatively straightforward to estimate outdoor air pollution concentrations for the coordinates of the cohort addresses by filling in the developed exposure model.

Exposure estimation will be conducted for all available addresses with coordinates. For example residential histories might be available, and in some cohorts information may be available for home and work / school / day care address, and also for these addresses predictions can be made.

The distribution of the estimated concentrations at the addresses will be explored and compared with the distribution of the measured concentrations at the monitoring sites in order to assess the predicted concentrations and to evaluate whether there are any extreme predictions. If there are extreme predictions it will be checked why there are extreme predictions (e.g. extreme values for predictor variables). If necessary, such extreme values for predictor variables could be truncated and given the highest value which occurs at one of the monitoring sites for that specific predictor variable. Other reasons might be that the

coordinate of an address is located by coincidence (e.g. because GIS datasets and/or geocoding are not accurate) on an air pollution source (for example a major road) resulting in high estimated concentrations.
Further checks that will take place are that the coordinates / addresses with the highest and lowest estimated concentrations will be checked with Google Earth/Maps or a topographical map. It will be evaluated whether the coordinates / addresses with the highest concentration are located close to an air pollution source (for example a major road), and whether the coordinates / addresses with the lowest concentrations are located in an area without air pollution sources (e.g. rural areas).
In addition, concentration maps will be made for the whole study area. This is for presentation purposes only. This means that for centroids of 100m grid cells concentrations will be estimated. This will also mean that GIS analyses should be conducted to collected data for the GIS predictor variables for each of these centroids. Because the model is already developed, only values for predictor variables included in the final LUR model have to be collected, as is done for the cohort addresses.

The exposure estimates for all addresses can then be transferred to the groups that will conduct the epidemiological analyses (NB First extrapolation over time should be applied – See Section 7). In addition, the R^2 value and RMSE value of the cross-validation results of the developed final LUR model should also be provided. These values indicate how good the model is and indicates the amount of exposure misclassification. Sensitivity epidemiological analyses can then be conducted by excluding the areas for which the LUR models have more measurement error. As our key interest is in assessing spatial variability well, the R^2 obtained from cross-validation will be used to characterize potential measurement error.

7. Exposure estimation over time

Monitoring data and developed LUR models are for the time period 2008-2010. A challenge within the ESCAPE project is that for several of the included studies, concentrations need to be estimated for the past.
The information about which time period is the most relevant for each health outcome and corresponding study area has to come from the different Health WPs.
It will be study area specific how far back in time estimations can be made. The local people from the exposure groups should evaluate temporal aspects in their study area.

We will address temporal aspects by:

1. Documenting the stability of the spatial contrast generated by the ESCAPE LUR model by collecting previous spatially resolved monitoring data (if available)
2. Documenting the temporal trend of concentrations using previous monitoring data (also possible if not spatially resolved)
3. Adjusting the ESCAPE LUR model with the observed temporal trend

The ESCAPE LUR model will be used as the main exposure variable in the epidemiological analyses. An advantage of this is that no external data (which can differ between study areas) are needed for estimating a trend over time. Time-trend adjusted variables will be included as a sensitivity analysis.

Adjustment for changes in spatial patterns is generally not possible, because of lack of data. Adjustment for changes in concentrations over time is however possible. An absolute difference between different time periods will be estimated, however this does not imply that the time trend needs to be linear. Compared to using the original data, such an adjustment for temporal trends may have no effect on the estimated relative risks in the epidemiological analyses, but it has an effect on graphs of exposure-response relationships. In addition, it may have an effect if in a study recruitment occurs over multiple years and if a study is comprised of multiple areas which have a different time trend. Furthermore the scaling is relevant if residential history is taken into account.

In summary, the exposure data for addresses of study participants that will be available for the epidemiological analyses are:

1. Concentrations based on the exposure model that has been developed based on the 2008-2010 ESCAPE air pollution data (= main exposure variable)

2. Concentrations that have been estimated after applying a temporal trend (= exposure variable for sensitivity analyses)

Procedures for evaluation of trends and backward extrapolation of annual air pollution concentrations are described in Section 7.1. Birth cohort studies require more detailed temporal resolution, e.g. estimates for each trimester during pregnancy (Section 7.2). For some studies residential history information or other addresses than only the home address might be available (e.g. work address, school addresses, etc). If these addresses are available, exposure will also be estimated for these addresses (Section 7.3).

7.1 Evaluation of changes in spatial patterns and adjustment for temporal trends

7.1.1 Evaluation of changes in spatial patterns and temporal trends

Monitoring data

In some ESCAPE study areas previous study-specific spatially dense monitoring has taken place. We re-sampled many of the sites used in 1999 in the TRAPCA study (Munich, Stockholm and the Netherlands), so that we will have direct evidence of the agreement between spatially distributed measurements of PM and NO_2 in these three European cities/areas obtained several years apart. Further information might come from the SAPALDIA study and ECRHS Spain. These data will also be used to estimate trends over time, also intra-urban. Trends over time for all site types together and for the different site types will be estimated using mixed modeling. This trend can then also be applied to the estimated LUR concentrations.

In the large majority of the study areas, monitoring of air pollution by routine continuous network sites has been in operation for many years so that trends over time can be addressed. Concentration data have to be collected for all types of sites (regional, urban, traffic) and as far back in time as possible. We will first evaluate a European database based upon AIRBASE, in which an assessment has been made of changes in sites and monitoring methods by the European Topic Center on Air pollution (dr Frank de Leeuw).
AIRBASE does not have all sites and is only complete after 2000. Local centers should therefore try to collect reliable routine monitoring data. An annual mean should consist of concentrations of at least 75% of the days in a year (i.e. 75% data capture for annual mean). If the study area contains a sufficient number of sites (e.g. > 5), some indication of the spatial stability can be obtained by comparing routine monitoring data over time and the ESCAPE predictions versus the routine monitoring data. If a small number of sites exist, trends over

time for all site types together and for the different site types separately will be estimated using mixed modelling.
Correlations between concentrations of different years will be calculated. Previous studies have shown that the correlation between air pollution concentrations of different years is high, even over a period of more than 10 years (Beelen et al, 2007). Whether a trend can be estimated over a time period longer than 10 years is area-specific, and depends whether there have been large changes in road network, emission sources, land use, etc.

Disadvantages of using previous monitoring data may be that the data do not exist far enough back in time for some ESCAPE study areas, and the composition of for example PM10 has been changed over time which cannot be taken into account with these data. A careful evaluation of changes in the network needs to be performed, as monitoring techniques may have changed, sites may have changed or (traffic) conditions around specific sites may have changed. Further, different time trends within one study area cannot usually be modeled, so the assumption is that the pattern over space is similar in an area.

We will further evaluate the possibility to assess historic time trends using satellite data.

Modelling data

- Background pollution maps
  Models to estimate background concentrations of PM and NOx in previous years exist in some ESCAPE study areas, typically on fairly large spatial scales of no less than 5 x 5 km (for example in the Netherlands, UK and Sweden, but possibly also other countries). These maps have been developed using a combination of network data and dispersion modelling and are updated each year. We will evaluate the usefulness of these data after comparing with our monitoring data. These models could possibly also be used to evaluate spatial patterns in air pollution concentrations for different time periods. Similar to information from previous monitoring data, a disadvantage may be that the data do not exist far enough in time.
- Historical emission data and dispersion models
  Historical emission databases of sufficient resolution and quality exist in only a few locations (e.g. Stockholm), so that the usability of dispersion modelling is limited. Where available, dispersion models will be used to asses trends in (spatial pattern) time. It will however give too optimistic results, because the same model is used for each year, only input data will be different for different years. Historical emission data and dispersion models will therefore be less useful to estimate trends in air pollution concentrations, but could be used for assessing changes in spatial patterns over time.

Historical GIS data

Historical data on land use, road networks etc. exist as well which allow us to judge whether the spatial ranking of current ambient concentrations has been stable. These historical data should thus be collected (see Section 3).

Correlations between values (for example for land use) between different years will be evaluated and trends over time will be evaluated. For data like road networks, it will be evaluated whether the road network changed over the years, and which changes occurred. Further, the correlation between traffic intensities of different years can be calculated. A recent study (Beelen et al, 2007) showed that major roads remained likely in place, and the major roads are the roads that are most important for air pollution exposure. Spatial pollution patterns probably tend to be fairly stable across large urban areas as for example land use and the road network of major roads often do not change quickly or abruptly over time.

IRAS will provide a questionnaire about previous monitoring data and about previous modeling which will be distributed to each of the ESCAPE study areas. This information can then be used to estimate historical air pollution concentrations and evaluate spatial patterns.

7.1.2 Backward extrapolation of air pollution concentrations

Although documentation of changes is the main response towards time trends, we will perform sensitivity analyses using backward extrapolation of annual air pollution concentrations to test the sensitivity of the epidemiological findings to temporal trends. The default is to use one absolute time trend correction difference, calculated as the absolute difference between 2009 and the year for which an estimate was desired (e.g. 1999). If solid data from multiple sites exist, we may use different factors within the study area, e.g. differentiating between rural and urban or background and traffic sites. Typically, the number of sites will not allow that.

7.2 Exposure estimation for birth cohort studies

While modelled annual average concentrations are sufficient for most study areas within the ESCAPE project, pregnancy outcome studies in WP3 require more detailed temporal resolution. In pregnancy outcome studies, it is common to express exposure as the average concentration per month or trimester of a specific pregnancy (Slama et al. 2007). The required exposure thus needs to contain a spatial and temporal component. To date this has not been systematically evaluated.

One simple option is to develop LUR models using annual average concentrations and then use continuous routine monitoring data to produce a temporally varying component. This approach makes the assumption that the spatial pattern is constant in time.

This was for example used in the TRAPCA study. The spatial exposure estimates were yearly averages that did not allow testing for a higher susceptibility to atmospheric pollutants during a given trimester of pregnancy. To seasonalize the exposure model (i.e. include a temporal component depending on the conception and delivery dates), the temporal observations observed in one background station in Munich operated by the Bavarian Environmental Protection Agency were applied to the exposure. For NO_2, this was done by averaging the NO_2 daily mean levels over the pregnancy of each woman (of continuous sampling site), by dividing this average by the average NO_2 level during the TRAPCA measurement campaign from 1999-2000, and multiplying the corresponding coefficient by the NO_2 estimate from the TRAPCA II spatial model. Within ESCAPE, we will use the absolute difference instead of the ratio, as ratios may be problematic for low concentrations.

Using the same approach, trimester-specific exposure variables were estimated. The assumption was that temporal variations in the considered atmospheric pollutants were similar across the metropolitan area. Although reasonable, this assumption was likely to have induced exposure misclassification, which was believed to be minor compared with that which would exist when temporal variations in air pollution had been ignored.

A similar approach will be used for the pregnancy outcome studies within WP3 of the ESCAPE study.

Daily air pollution data from one or more background continuous monitoring sites are therefore needed. Only sites with more than 75% data capture will be used to estimate pregnancy-specific or trimester-specific exposure estimates.

7.3 Residential history and other addresses

For some studies residential history information or other addresses than only the home address might be available (e.g. work address, school addresses, etc). Data on residential histories of study participants allow us to back calculate exposures as well. If these addresses are available, exposure will also be estimated for these addresses.

A questionnaire about address information in each of the cohorts has been distributed to each of the Health WPs.

8. Additional exposure issues

The sections above describe the ‘standard’ modeling procedures for ESCAPE study areas in which monitoring will be conducted. There may however be some additional exposure issues in some study areas or within some studies. In some study areas both ESCAPE monitoring may have taken place as well as previous spatially resolved models may be available that could be used for exposure assessment (Section 8.1). For some large studies, there may be study areas for which no ESCAPE measurements will be conducted, but which have spatially resolved models (e.g. study areas within SAPALDIA or ECRHS) (Section 8.2). For some multi-center studies, there may be study areas for which no ESCAPE measurements will be conducted *and* for which no spatially resolved models are available (study areas in France and United Kingdom), but for which national (governmental) monitoring data are available (Section 8.3).

Please contact the Health WPs to ask which areas also could be included for the exposure assessment besides the areas for which ESCAPE monitoring will take place. If for more study areas exposure could be assessed, this would result in a larger number of study observations available for epidemiological analyses.

If one or more of these additional exposure issues apply to a study area or a study, please inform the exposure assessment working group (Rob Beelen: r.m.j.beelen@uu.nl and Gerard Hoek: g.hoek@uu.nl), and it will be evaluated for each study area how to use all available data in the best way, and how exposure can be estimated.

If procedures as described in Section 8 have been used for exposure assessment this should also be informed to the Health WPs. Sensitivity analyses could then be conducted with and without these study areas.

8.1 Both ESCAPE data and previous monitoring and / or modeling data available

For some study areas both ESCAPE exposure data and spatially resolved models are available. These previous exposure data may for example be for a time period which is closer to or in the relevant time window of exposure for a specific health outcome.

Exposure will be estimated in 2 ways for these study areas: (1) using the ESCAPE procedures, and (2) using the previous modeling data. Both previous LUR and/or dispersion exposure assessment models can be used. The agreement between the two estimates will be calculated at the residential addresses, which includes two sources of variation: temporal trend and method. If the model can be applied again, agreement at the ESCAPE and previous study monitoring sites (in case of a previous LUR study) can also be calculated.

The ESCAPE monitoring data will then be used for validation. The previous exposure assessment model will be used to estimate the concentrations at the ESCAPE monitoring sites. The correlation and scatterplot between measured and estimated concentration, mean difference (and range/SD) between measured and estimated concentration, and R^2 value and RMSE-value of regression analysis between measured and estimated concentration will then be calculated to evaluate the validity of the model on the ESCAPE monitoring sites. See also section 7.1.

Because of harmonization and comparison with other ESCAPE study areas, in all study areas exposure will be estimated using the ESCAPE procedures and epidemiological analyses will be conducted with the ESCAPE exposure estimates. As a sensitivity analysis, previous modelling exposure estimates will be used.

8.2 For specific study area no ESCAPE data available, but spatially resolved models available

For some multi-center or national studies that are included in ESCAPE there may be study areas for which no ESCAPE monitoring will be conducted, but for which previous local exposure data are available. For example, ESCAPE monitoring will not take place in all SAPALDIA and ECRHS areas, while previous local exposure assessment may be available for these areas.

First, it should be evaluated whether the ESCAPE models developed for other areas can reasonably be applied in the study area despite the fact that no monitoring takes place. Recalibration of models may be an option.

Second, if that is too uncertain, exposure for the participants in these study areas will be estimated using these previous exposure assessment models. For these areas, the previous exposure models can however not be validated using ESCAPE monitoring data (as described in Section 8.1). If the models have been shown to agree well with ESCAPE monitoring in other (SAPALDIA or ECRHS) areas, this may not be a serious limitation.

Further we suggest to the Health WPs that because for the comparison and harmonization within the ESCAPE project, the data from these study areas will only be included in the epidemiological analyses as a sensitivity analysis (so the default is to conduct the epidemiological analyses without the data from these areas).

8.3 No ESCAPE data and no spatially resolved models available

For some multi-center or national studies in ESCAPE, there may be study areas within such studies for which no ESCAPE measurements will be conducted *and* for which no local monitoring and/or modeling data are available, i.e. study areas in France and the United Kingdom where study participants live in the whole country but ESCAPE measurements will only be conducted in selected monitoring areas. Although no local exposure data are available, monitoring data from fixed (governmental) monitoring sites are available in these study areas.

For France, measurements will be conducted in a limited number of monitoring areas (Paris, Grenoble, Lyon, and Marseille), and only in Paris PM and NOx will be measured while in the other areas only NOx will be measured. The areas are distributed in different regions of France. Because PM is measured only in Paris and because of the geographical and other differences over France, for PM only a model for Paris will be made. However, for NOx a model will be made with which also exposure can be estimated for addresses located outside the monitoring areas. Because France has a national NOx network, this network will be used to estimate the regional background concentration for the whole country (using interpolation/kriging). Within the monitoring areas a NOx LUR model will be made to estimate the urban background component and the local traffic component, using a combined model for the different study areas. Further, for each study area separately a study specific NOx LUR model will be made using only the ESCAPE monitoring data. A comparison will be made how well the combined model performs in the monitoring areas compared to the study-area specific model.

A few of the sites for each monitoring area have been located around the city in background locations and preferably a few sites will be collocated with sites of the national monitoring network to evaluate (systematical) differences between the concentrations measured by the fixed monitoring sites and by the ESCAPE monitoring sites.

For the epidemiological analyses this will mean:

1. Analyses with all addresses.
   Exposure will be estimated using data from the regional monitoring network and a combined LUR model for the different study areas.
2. Analyses with only addresses in areas where ESCAPE monitoring has been conducted.
   Exposure will be estimated for each study area separately using a study area specific LUR model.

For the UK, Imperial College will provide a proposal for monitoring and modeling.

Monitoring will be conducted in Manchester (PM and NOx) and Bradford (NOx and limited number of PM measurements). Further, it was planned to conduct a monitoring campaign in Oxford (PM and NOx). The "Oxford" cohort is however spread over the whole country, and it is therefore not worthwhile to conduct a specific monitoring campaign in Oxford. Instead of in Oxford, the monitoring campaign will be conducted in the Thames valley (London, Oxford).

Similar to France it will not be possible to develop a nationwide PM exposure assessment model. The UK has also a nationwide NOx measurement program, which can be used to estimate the regional NOx background concentration, and within the monitoring areas a combined LUR model will be developed to estimate the urban background and local traffic component. For each study area separately a study specific NOx model will be made using only the ESCAPE monitoring data (same procedure as in France – see above).

For the epidemiological analyses this will mean:

1. Analyses with all addresses.

   Exposure will be estimated using data from the regional monitoring network and a combined LUR model for the different study areas.
2. Analyses with only addresses in areas where ESCAPE monitoring has been conducted.

   Exposure will be estimated for each study area separately using a study area specific LUR model.

9. Coordination

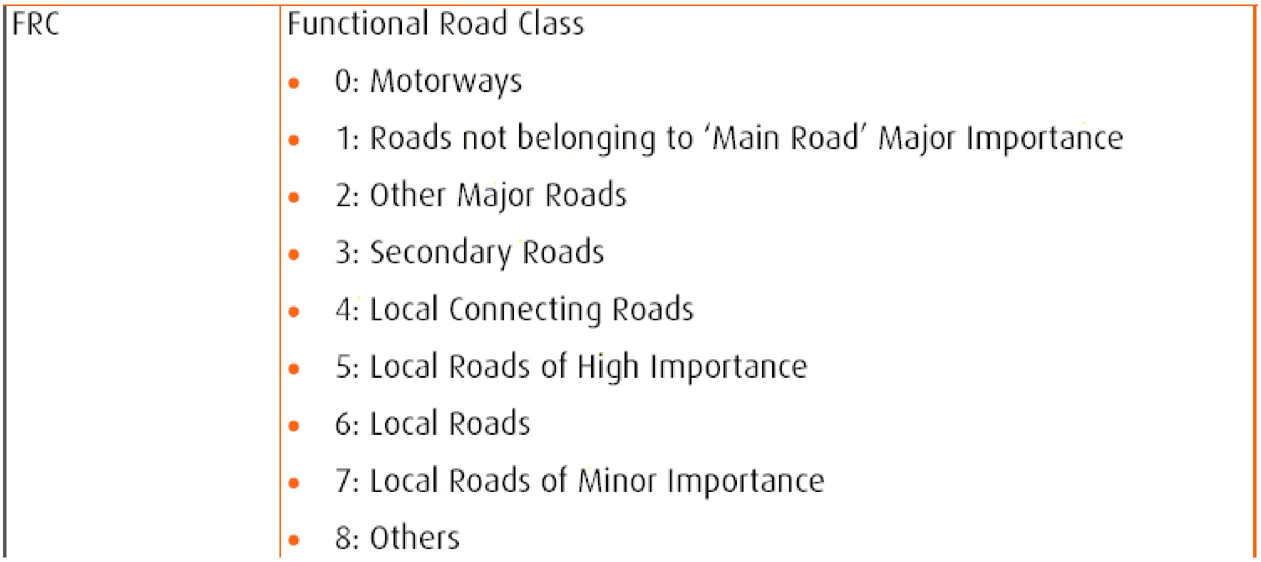
The exposure assessment procedure will be coordinated and supervised by the WP leaders of WP2 and the Exposure Working Group. To ensure that all groups use the same procedures for estimating exposure and to harmonize exposure assessment which is mostly performed locally, the procedures for exposure assessment are described in this ESCAPE exposure assessment manual. Further, we will organize a workshop in which the procedures will also be explained and training will be provided how GIS analyses, LUR model development and exposure assessment should be conducted. The Exposure WG will also supervise the development of models in various stages of the project, by reviewing locally developed models.

10. Time planning

The table below describes the time planning and the date when each step in the exposure assessment should be finished. Please note that some steps will cost a considerable amount of time and may need some work in advance. For example, the collection and evaluation of GIS data may cost some time. Further, the geocoding of addresses of study participants may be time consuming (a Geocoding manual is available). And the extrapolation of air pollution estimates over time may cost some time in advance, e.g. the collection of historical air pollution data, dispersion/emission data and GIS data. Please make sure to start in advance with all these steps.

	Month of completion	
Activity	*1st year group*	*2nd year group*
Collection of GIS data	February 2010	December 2010
GIS analyses	May 2010	February 2011
LUR model development	September 2010	May 2011
Geocoding addresses study participants (if needed) and GIS analyses for addresses	October 2010	August 2011
Evaluation of and collection of (SES) area level confounders	October 2010	August 2011
Collection of traffic indicator variables for coordinates of addresses of study participants	November 2010	September 2011
Assessment of air pollution estimates for addresses based on LUR model	November 2010	September 2011
Extrapolation air pollution estimates over time	November 2010	September 2011
Additional issues	December 2010	October 2011

Appendix I: Central road network classes



FRC	Functional Road Class • 0: Motorways • 1: Roads not belonging to 'Main Road' Major Importance • 2: Other Major Roads • 3: Secondary Roads • 4: Local Connecting Roads • 5: Local Roads of High Importance • 6: Local Roads • 7: Local Roads of Minor Importance • 8: Others

Appendix II: CORINE 2000 classes

CLC_CODE	LABEL1	LABEL2	LABEL3	RGB	HD res	LD res	Ind	Port	Urb Green	Semi Nat
111	Artificial surfaces	Urban fabric	Continuous urban fabric	230-000-077	x					
112	Artificial surfaces	Urban fabric	Discontinuous urban fabric	255-000-000		x				
121	Artificial surfaces	Industrial, commercial and transport units	Industrial or commercial units	204-077-242			x			
122	Artificial surfaces	Industrial, commercial and transport units	Road and rail networks and associated land	204-000-000						
123	Artificial surfaces	Industrial, commercial and transport units	Port areas	230-204-204				x		
124	Artificial surfaces	Industrial, commercial and transport units	Airports	230-204-230						
131	Artificial surfaces	Mine, dump and construction sites	Mineral extraction sites	166-000-204						
132	Artificial surfaces	Mine, dump and construction sites	Dump sites	166-077-000						
133	Artificial surfaces	Mine, dump and construction sites	Construction sites	255-077-255						
141	Artificial surfaces	Artificial, non-agricultural vegetated areas	Green urban areas	255-166-255					x	
142	Artificial	Artificial, non-	Sport and	255-					x	

	surfaces	agricultural vegetated areas	leisure facilities	230-255						
211	Agricultural areas	Arable land	Non-irrigated arable land	255-255-168						
212	Agricultural areas	Arable land	Permanently irrigated land	255-255-000						
213	Agricultural areas	Arable land	Rice fields	230-230-000						
221	Agricultural areas	Permanent crops	Vineyards	230-128-000						
222	Agricultural areas	Permanent crops	Fruit trees and berry plantations	242-166-077						
223	Agricultural areas	Permanent crops	Olive groves	230-166-000						
231	Agricultural areas	Pastures	Pastures	230-230-077						
241	Agricultural areas	Heterogeneous agricultural areas	Annual crops associated with permanent crops	255-230-166						
242	Agricultural areas	Heterogeneous agricultural areas	Complex cultivation patterns	255-230-077						
243	Agricultural areas	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation	230-204-077						
244	Agricultural areas	Heterogeneous agricultural areas	Agro-forestry areas	242-204-166						
311	Forest and semi natural areas	Forests	Broad-leaved forest	128-255-000						x

312	Forest and semi natural areas	Forests	Coniferous forest	000-166-000						x
313	Forest and semi natural areas	Forests	Mixed forest	077-255-000						x
321	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Natural grasslands	204-242-077						x
322	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Moors and heathland	166-255-128						x
323	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Sclerophyllous vegetation	166-230-077						x
324	Forest and semi natural areas	Scrub and/or herbaceous vegetation associations	Transitional woodland-shrub	166-242-000						x
331	Forest and semi natural areas	Open spaces with little or no vegetation	Beaches, dunes, sands	230-230-230						x
332	Forest and semi natural areas	Open spaces with little or no vegetation	Bare rocks	204-204-204						x
333	Forest and semi natural areas	Open spaces with little or no vegetation	Sparsely vegetated areas	204-255-204						x
334	Forest and semi natural areas	Open spaces with little or no vegetation	Burnt areas	000-000-000						x
335	Forest and semi natural areas	Open spaces with little or no vegetation	Glaciers and perpetual snow	166-230-204						x

411	Wetlands	Inland wetlands	Inland marshes	166-166-255						x
412	Wetlands	Inland wetlands	Peat bogs	077-077-255						x
421	Wetlands	Maritime wetlands	Salt marshes	204-204-255						x
422	Wetlands	Maritime wetlands	Salines	230-230-255						x
423	Wetlands	Maritime wetlands	Intertidal flats	166-166-230						x
511	Water bodies	Inland waters	Water courses	000-204-242						
512	Water bodies	Inland waters	Water bodies	128-242-230						x
521	Water bodies	Marine waters	Coastal lagoons	000-255-166						x
522	Water bodies	Marine waters	Estuaries	166-255-230						x
523	Water bodies	Marine waters	Sea and ocean	230-242-255						x
999	NODATA	NODATA	NODATA							
990	UNCLASSIFIED	UNCLASSIFIED LAND SURFACE	UNCLASSIFIED LAND SURFACE							
995	UNCLASSIFIED	UNCLASSIFIED WATER BODIES	UNCLASSIFIED WATER BODIES	230-242-255						

Appendix III: Possible area-level confounders for the ESCAPE project

1. Summary

The area-level confouders that are possibly relevant for the ESCAPE project are described in the Table below. Please evaluate for your study area which of the described possible ecologic covariate (or maybe there might be other potential data) are relevant for your study area, whether these data are available and for which spatial scale and which time period these data are available. The relevant spatial scale can differ for covariates and smaller (e.g. neighborhood) and larger (e.g. city-wide) spatial scales might be important.

Please provide an overview with what is available for your own study area (see Section 7 for a more extensive explanation of the to be evaluated ecologic covariates).

Table: Possible ecologic confounders identified in the literature, which might be relevant for ESCAPE

Demographic and socioeconomic environment
Income
Income inequality
Proportion households with income below minimum
Housing values
Ratio between owner-occupied houses and rented houses
Unemployment rate
Proportion of people who depend on benefits
Proportion people with severe financial problems
Proportion of single mothers
Infrastructure deprivation
Residential stability/population change
Urbanization
Region/province
Climate and physical environment
Water hardness

Health services
Availability and accessibility of health services
General practitioner deprivation score

2. Background – Evaluated ecologic covariates in the HEI reanalysis of the Six Cities and ACS study

The proposed ecologic covariates in the HEI reanalysis of the Six Cities and the ACS study were classified in 3 categories: demographic and social environment, climate and physical environment, and health services. Finally, in the reanalysis 20 ecologic covariates were used from a longer list of 30 potential ecologic covariates (Table 1). A more detailed description of the ecologic covariates that were included in the reanalysis can be found on page 178 of the reanalysis report.(Krewski, Burnett et al. 2000)

Briefly, eight measures of the social environment were considered: population change, percentage of white residents, percentage of black residents, mean income of residents in 1979, poverty level in 1979, income disparity as measured by the Gini coefficient, unemployment in 1979, and percentage of residents age 25 or older who had completed high school. In terms of the physical environment, altitude, water hardness, and climate (average maximum temperature, average monthly variation in maximum temperature, average daily relative humidity, and average monthly variation in daily relative humidity). Four gaseous co-pollutants were also used in the reanalysis: CO, NO_2, O_3, and SO_2. Two measures of the provision of health care services were used: number of physicians per 100,000 residents and number of hospital beds per 100,000 residents. However, it was not possible to obtain data on certain ecologic covariates for some of the cities included in the ACS study.(Krewski, Burnett et al. 2000)

Reasons for not including some proposed ecologic covariates in the analyses ranged from no data available (barometric pressure) and no reliable data available (crime rate and other airborne toxic substances) to that the data do not necessarily represent a useful group-level variable (race, serum iron levels, and air conditioning). A full description of the proposed ecologic covariates that are not used is given in Appendix E of the HEI reanalysis report.(Krewski, Burnett et al. 2000)

Table 1: Proposed ecologic covariates in the HEI reanalysis

Demographic and socioeconomic environment

Used in the reanalysis	*Not used in the reanalysis*
Income	Crime rate
Education	Race
Income disparity	Serum iron levels
Population change	
Poverty	
Unemployment	
Percent whites	
Percent blacks	

Climate and physical environment

Used in the reanalysis	*Not used in the reanalysis*
Temperature	Barometric pressure
Temperature variation	Air conditioning
Relative humidity	Geographic position: latitude and longitude
Relative humidity variation	Other airborne toxic substances
Altitude	Radon gas
Water hardness	
Gaseous copollutants: CO, NO_2, O_3, SO_2	

Health services

Used in the reanalysis
Number of physicians
Number of hospital beds

3. Relevance of the proposed ecologic covariates in the HEI reanalysis for study areas within the ESCAPE project

The demographic variables that were considered in the HEI reanalysis may all be relevant for the different study areas in the ESCAPE project. Also some of the climate and physical environment variables may be relevant for some of the study areas within the ESCAPE project depending also on the size of the study area. Gaseous co-pollutants are certainly relevant, but will be treated as exposures instead of ecologic covariates.

It has to be evaluated which of the proposed ecologic covariates in the HEI reanalysis are relevant for each of the different study areas within the ESCAPE project. Furthermore, the HEI reanalysis used ecologic covariates on a metropolitan scale. This is probably not the most relevant scale. ‘Neighborhood’ or ‘postal code areas’ are probably better as the spatial scales for ecologic covariates in the ESCAPE study areas since large differences in demographic variables exist in urban areas. In addition, larger scale than city-level may be relevant, for example region (see also Table 3 below).

4. Possible ecologic covariates identified in the literature

Similar to the search for possible ecologic covariates in the reanalysis of the ACS study, we also made use of 3 categories (i.e. demographic and socioeconomic environment, climate and physical environment, and health services) to conduct our own literature search for possible ecologic covariates. Both a Medline search and reference-tracking was conducted to identify possible relevant ecologic covariates. The possible ecologic covariates found in the literature are shown in Table 2. It is however not sure whether all these data are available on an area level or whether these data are available on the preferred area level and for the preferred time period in all study areas within the ESCAPE project. It has also to be evaluated which of the identified possible ecologic covariates may be relevant for each of the study areas within the ESCAPE project.

Some of these ecologic covariates were identified from studies investigating the effects on mortality, other ecologic covariates were identified from studies which investigated the effects on other health outcomes. Whether an ecologic covariate is relevant in a study area depends thus also on the health outcome under study.

Table 2: Possible ecologic confounders identified in the literature

Demographic and socioeconomic environment
Income
Income inequality
Proportion households with income below minimum
Housing values
Ratio between owner-occupied houses and rented houses
Unemployment rate
Proportion of people who depend on benefits
Proportion people with severe financial problems
Proportion of single mothers
Infrastructure deprivation
Residential stability/population change
Urbanization
Region/province
Climate and physical environment
Water hardness
Health services
Availability and accessibility of health services
General practitioner deprivation score

5. Discussion

It is important to conceptualize the causal pathways by which ecologic covariates can affect health.(Pickett and Pearl 2001). Differences in health between areas may be caused by two mechanisms:

1. Socioeconomic differences or differences in area-bound factors such as for example pollution differences between areas;
2. Selective migration of “healthy” people out of “non-healthy” areas (for example cities).(Lucht and Verkleij 2001)

For ecologic covariates, both the average and the spread of ecologic covariates may be worthy of examination (Pickett and Pearl 2001), for example mean income versus income range in a neighborhood. Ecologic covariates may also be inter-correlated and therefore the correlation between different ecologic covariates has to be evaluated.

Most of the possible ecologic covariates found in the literature are social economic status (SES) related covariates. To investigate the effect of these ecologic SES covariates, individual level SES covariates have to be accounted for. Furthermore, different alternative measures of SES may be considered jointly because each measure may represent different aspects of social status and may be associated with different intermediate risk factors in the relation between SES and health (Backlund, Sorlie et al. 1999). Therefore, in the analyses, account was taken for individual level SES covariates and different alternative area level measures of SES.

However, results from studies indicate that the choice of area level variables may be less critical than ensuring correct control for individual level social economic status. (Fiscella and Franks 1997; Lucht and Verkleij 2001; Pickett and Pearl 2001; Osler, Prescott et al. 2002; Steenland, Henley et al. 2004) For example, the results of a study with analysis of pooled data from two cohort studies (13,710 women and 12,018 men) indicated that area based income inequality did not affect all cause mortality after adjustment for individual income and other risk factors. The authors concluded that Denmark's welfare system, that is based on a Nordic model, may even out the effect of area inequality.(Osler, Prescott et al. 2002) In addition, a study by Fiscella and Franks had almost the same conclusion, i.e. family income, but not community income inequality independently predicted mortality. This study was a longitudinal study in the US where 14,407 people (aged 25-74 years) were followed from 1971-5 until 1987.(Fiscella and Franks 1997) A longitudinal study in Canada where 2,116 people (age 18-75 years) were followed from 1990 through December 1999, found that neighborhood socioeconomic characteristics (neighborhood income, educational level, unemployment rate) were not significantly associated with mortality. However, within advantaged neighborhoods, the importance of individual socio-economic characteristics for mortality is increased relative to disadvantaged neighborhoods. This has also been demonstrated in the US and thus also, although less pronounced, in Canada, in a setting with universal access to basic health and social services.(Veugelers, Yip et al. 2001) A recent cohort study by Steenland *et al.* on individual- and area-level socioeconomic status variables as predictors of mortality with 179,383 study participants suggested that the predictive value of area-level socioeconomic status variables varies by cause of death but is less important than individual-level socioeconomic status variables.(Steenland, Henley et al. 2004)

Studies in the Netherlands also showed that after correction for more individual level SES factors smaller health differences between neighborhoods were found.(Lucht and Verkleij 2001) However, the findings of a longitudinal study (GLOBE-study) in the Netherlands (baseline: 1991, with a follow-up time of 6 years) in 8,506 men and women aged 15-74 years, indicated that particular indicators of neighborhood SES were related to all cause mortality of men and women in an urban setting. After the stringent control for individual SES, the neighborhood percentage of unemployed or disabled persons, and the percentage who reported severe financial problems continued to affect mortality risks. The educational and occupational indicators of neighborhood SES were also related to mortality, but less strongly, and their effects were no longer statistically significant after control for individual-level socioeconomic indicators.(Bosma, van de Mheen et al. 2001)

Although ecologic covariates may be less important than individual-level variables, both individual and ecologic variables will be evaluated as potential confounders in the ESCAPE project.

6. Relevant spatial scale for ecologic covariates

A study by Reijneveld et al. (Reijneveld, Verheij et al. 2000) in the Netherlands examined the impact of geographical classification on the clustering of poor health (as measured by 4 indicators in a interview: self rated health, physical symptoms, mental symptoms, and long term physical limitations) per area and on the size of the differences in health by area deprivation. Three classifications were used:

1. *Neighborhoods* are areas with a similar type of buildings, often delineated by natural boundaries. Because of this, they are socioculturally rather homogenous and therefore relate to “real” communities, but their population size varies a great deal.
2. *Postcode sectors* (postal code areas) have a logistic origin, adequate post delivery, and were designed at a national level. They had to comprise similar numbers of addresses and therefore, their average population size varies less. Postcode sectors do not further have a (emotional) meaning to most of their residents.
3. *Boroughs* concern aggregates of socioeconomically comparable neighborhoods; they mostly exist in urban areas. In some of the bigger cities of the Netherlands they have their own public administration.

Regarding homogeneity, ecologic, area-bound factors may have a greater impact on health if an area relates to a socioculturally homogeneous, “real”, community. Therefore, the independent area effect on the clustering of poor health was largest for neighborhoods. A large part of these area effects can be explained by differences between areas in the socioeconomic composition of their population.

However and more importantly, the choice of the geographical classification had hardly any impact on the size of the health differences by area deprivation in this study.(Reijneveld, Verheij et al. 2000)

Because this study has been conducted in Amsterdam, the conclusions cannot automatically be applied to other cities/towns and countries.

Van der Lucht and Verkleij (Lucht and Verkleij 2001) found that mortality differences between neighborhoods in cities are greater than mortality differences between cities and non-urban areas. But mortality differences between neighborhoods do not only exist in cities, but can also been found between neighborhoods in non-urban areas. It has been estimated that all cause mortality in the "poorest" neighborhoods is 13 percent higher compared with all cause mortality in the "wealthiest" neighborhoods. However, mortality differences between neighborhoods are especially great among men and women younger than 65 years. At older ages, these mortality differences between neighborhoods are small, especially among women.(Lucht and Verkleij 2001)

The exact geographical classification of areas is thus possibly of less importance for studying health differences between areas. Furthermore, the relevant spatial scale can differ for covariates.

7. Available and to be used ecologic covariates within the ESCAPE project

In the previous sections it has been described which ecologic covariates may be relevant for the ESCAPE project, and which spatial scale may be relevant. However, what is possible also depends on the information that is available.

Please evaluate for your study area which of the described possible ecologic covariate (or maybe there might be other potential data) are relevant for your study area, whether these data are available and for which spatial scale and which time period these data are available. Also evaluate the number and percentage of missing values in each dataset, because not all information may be available for the whole study area. If there are many missing values, these data should not be used because it would result in a decrease in the number of observations available for epidemiological analysis.

As an example, below an overview is given which GIS-data might be relevant and are available in the Netherlands. Using the geographical coordinates of the addresses of the study participants for each study participant a value for an ecologic covariate can be generated using GIS analyses. In Table 3 an overview is shown which potential ecologic covariates, and at which scale and for which year(s), are available for the Netherlands. All these covariates can be evaluated as potential ecologic covariates in the epidemiological analyses.

Please provide also such an overview for your own study area and discuss this also with the people responsible for the epidemiological analyses in that WP.

Table 3: Available potential ecologic covariates in the Netherlands.

Name GIS coverage	***Potential predictor variable***	***Spatial scale***	***Year***
“Wijk en buurt”	Average income per inhabitant	District/quarter/borough Neighbourhood (‘wijk’)	1995, 1997, 2001, 2003, 2005
	Percentage persons with low income (below the 40th percentile of the Dutch income distribution)		
	Percentage persons with high income (above the 80th percentile of the Dutch income distribution)		
	Percentage persons who depend on benefits		
“COROP”	Percentage persons with low income (below the 40th percentile of the Dutch income distribution)	COROP area: consist of a central point (e.g. a city) and the surrounding economic and social region (the Netherlands is divided in 40 COROP areas)	1995, 1997, 2001, 2003, 2005
	Percentage persons with high income (above the 80th percentile of the Dutch income distribution)		

Appendix IV: ESCAPE geocoding procedure

ESCAPE

Geocoding procedure

Version August 2009

Geocoding

Each of the addresses of study participants should be geocoded, i.e. a geographical coordinate for each address should be determined. If available, other addresses than baseline home address, such as work addresses, residential history or day care/school addresses can be geocoded. With these coordinates Geographic Information System (GIS) analyses can be conducted and the values for potential air pollution predictor variables and area level covariates (socioeconomic status) can be determined for each of the addresses, followed by exposure assessment for the addresses.

The validity of epidemiologic research using geocoded addresses for exposure assessment depends on the percentage of addresses that can be geocoded and the positional accuracy of locations of geocoded addresses.

Geocoding can be characterized in terms of its fundamental components: the *input data*, *output data*, *geocoding procedure*, and *reference dataset* (Goldberg, Wilson et al. 2008). The input data are the addresses that have to be geocoded and which contain attributes capable of being linked to some datum that has been previously geographically coded. The geocoding procedure determines the appropriate geographical coordinate to return for an address based on the values of its attributes and the values of attributes in the reference dataset. This is by far the most complicated portion of the geocoding process. The reference dataset consists of the geographically coded information that can be used to derive the appropriate geographic code for an input. The output data are the geocoded addresses determined by the geocoding procedure to geocode the input.

Geocoding is an important step in the exposure assessment process, and should be given considerable attention. Further, geocoding is no trivial task and may have costs.

The following sections describe the different components of the geocoding process that should be used to geocode addresses of epidemiologic studies within ESCAPE. It is however not possible to describe all the different problems and possibilities within the geocoding process. If there are any questions, please contact IRAS or Imperial College.

Geocoding: already done, yes or no?

In general, there are three options: (1) geocoding is already done; (2) geocoding has not been done yet, but a geocoding reference dataset is available; and (3) geocoding has not been done yet *and* there is no geocoding reference dataset available.

If addresses have already been geocoded, the accuracy and validation should be described. An explanation about what is needed for this description can be found below in the “Geocoding procedure” and “Validation” sections.

For the study areas for which the addresses have not been geocoded yet, the procedures below can be used to geocode the addresses. If addresses of study participants have not been geocoded yet, the most logical option is that it is done by the local centers, i.e. the groups who do the local air pollution measurements and GIS and LUR modelling. Geocoding should be done locally because no EU-wide address geocoding database is available. Each local center should evaluate which geocoding reference datasets are available. If no geocoding reference dataset is available for a specific study area a solution has to be found together with IRAS, Imperial College and the exposure assessment working group.

The input data

The input data consist of an address record table containing all addresses to be geocoded. Required attributes in the address record table include often street address, house number, postal code, and city name.

The default is that baseline home addresses of study participants will be geocoded. If available, other addresses than baseline home address, such as work addresses, residential history or day care/school addresses should also be geocoded.

The exact addresses of addresses should be geocoded, i.e. street name/postal code plus house number. If address information on a less detailed level - for example only postal codes, on street level, or addresses without house number - would be used then the geographical coordinate for that home address would be inaccurate. Because air pollution close to busy roads varies within tens of meters, it is important to use the exact address of study participants for geocoding.

Reference dataset

The reference dataset should contain both information about addresses and the corresponding geographical coordinates.

It is important that the percentage correctly geocoded addresses is as high as possible. Evaluate therefore the completeness of the reference dataset: i.e. does the reference dataset cover the complete study area where the study participants’ addresses are located?

Because GIS-analyses will be (mostly) conducted in the local or national centers, the coordinates in the reference dataset should have the same coordinate system compared with the to be used GIS-data. If the coordinate systems are different it is often possible to convert one of the coordinate systems into the other coordinate system. This may however result in some loss in accuracy, so preferably the coordinate systems should be the same.

The required accuracy of geocoding should be at least 5-10 meters (see also Validation section). It is thus also important to evaluate what kind of coordinates are available in the reference dataset (e.g. no excessive rounding of the coordinates). Reference databases can usually consist of coordinates of the centroids of buildings (homes) or parcels. Preferably coordinates of centroids of buildings should be used (sometimes these can also be front door coordinates), because if parcels are large then the centroid of the parcel may be away from the building (this applies especially for buildings in rural areas).

Preferably the reference dataset should be for the same year as the study's baseline year to avoid assigning incorrect coordinates to addresses because of postal code or house number changes over time. If no reference dataset is available for the study's baseline year, a reference dataset should be used which is as close as possible to the baseline year. However, also evaluate whether there are changes in precision in datasets from different years. If more recent reference datasets are more accurate this may favour the use of these more recent datasets.

The percentage of geocoded addresses could be improved using multiple reference datasets or methods. However, this may introduce error because datasets or methods may have different accuracy and/or completeness. We therefore recommend using only a single reference dataset, i.e. the best reference dataset.

Also evaluate the costs of the different datasets. If there is more than one reference dataset for a study area evaluate the coordinate system, available years, accuracy, completeness and costs to decide which reference dataset is the best for a specific study area.

An example reference dataset is the Address Coordinates Netherlands database in The Netherlands, which consists of all registered addresses by the Dutch postal service. The accuracy of this reference dataset is high with 95.5% of all coordinates located at the centroid of the correct building, 6.0% located at the centroid of the correct parcel, and only 0.5% not located in the correct building or parcel.

Geocoding methods

In general there two methods for linking a geographical coordinate to an address: building/parcel matching or interpolation.

With the building/parcel matching method, each building/parcel has an address and these addresses are linked to a geographic file that contains both addresses and corresponding geographical coordinates. Linkage based on address information (postal code, street name, house number, city name) can be conducted for example in SAS or Access. A further advantage of this method is that geocoding is done automatically. Field and/or manual methods (e.g. using aerial photography or Google maps) should therefore not be used.
The interpolation method attempts to match each address to an address-ranged street segment georeferenced within a streetline database and then interpolates the position of the address along that segment (Zimmerman, Fang et al. 2007). This is based on the proportional distance between the address on a record and the address range for a street segment. This method is also used in most car navigation systems. Small positional errors may occur because this method assumes a homogenous distribution of addresses along a street segment. Such positional errors have been shown to be a function of street length and may be larger for rural streets which are typically longer than their urban counterparts (Hay, Kypri et al. 2008). Because the required accuracy of geocoding should be at least 5-10 meters, the interpolation method may not always be accurate enough. This may result in misclassification and bias of potential air pollution exposure at addresses of study participants.
The building/parcel method should therefore preferably be used.

Geocoding procedure

To summarize the sections above: preferably geocoding should be conducted using the exact addresses of study participants and with a complete and accurate reference dataset. Linking of coordinates to addresses should preferably be done using the building matching method. Before starting with the geocoding procedure (and also when geocoding has already been conducted), give a description of:

- The input data:
  - Type of addresses (exact address (postal code plus house number), or only postal code, city name etc)
  - Which addresses are available and can be geocoded; and for which year are these addresses available:
  baseline address, residential history, work address, school/day care address, etc
- The reference dataset:
  - Completeness: for example does it cover the whole study area, which percentage of the total number of addresses in the area is included, etc.
  - Accuracy and whether accuracy is similar for the whole dataset or whether there are

areas which are less accurate; This information may be known from the supplier of the data
  - Coordinate system
  - Whether building or parcel centroids have been used, or other
  - Year for which geocodes are available
- The geocoding method:
  - building matching, parcel matching or interpolation method, or other method

When geocoding, regardless of the to be used method, the following the steps can be followed. Describe the different steps, and describe the percentage of addresses that could be geocoded in each step (if applicable):

1. Manually inspect all records before geocoding and document for the to be geocoded records:
   - the total number of records
   - the number of records with (partly) missing address information, e.g. missing postal code or house number (for these addresses geocoding will not be possible if no additional information about the missing information will be available).
2. For records for which an address or address component is missing, a detailed investigation should be performed to determine an appropriate and/or corrected address to be geocoded based on other information associated with the record. Commonly used sources include websites, phone books, utility records, and various vital records. More complete addresses result in a higher percentage of geocoded addresses.
3. Carefully check all the addresses to ensure that their attributes (street address, postal code, house number, city name etc) are clearly separated and correctly formatted for automatic geocoding with the reference dataset. If necessary the address input data should be standardized. The way to standardize depends on the geocoding reference dataset. For example streetname, postal code, house number could be standardized, for example no capitals.
4. Link in an automated way coordinates to each address based on the common attributes in both address input data and reference dataset, e.g. postal code, house number, city name. This can be done in SAS, or other statistical package, or Access for example.
5. Document the number and percentage of addresses for which a geographical coordinate could be assigned.
6. Document the number of non-geocoded addresses and the reasons why it was not possible to geocode (e.g. postal box, postal code / house number missing / standardized in a wrong

way / other / no clear reason). A manual search in the reference dataset could be conducted to document this.

7. Repeat, if necessary, the steps 2-6 for the addresses that have not been geocoded yet. If available, more recent versions of the used reference dataset could be used for the addresses that could not be geocoded to check whether geographical coordinates are available in these more recent datasets.
8. Document the final number and final percentage of addresses for which a geographical coordinate could and could not be assigned
9. If possible, evaluate whether the addresses that could not be geocoded are clustered in the same area(s). This could for example be done based on the city name because that might be available for all records. If addresses in certain areas cannot be geocoded this may result in some selection effect.

Considerations about geocoding

- Because air pollution exposure cannot be estimated for records for which no geographical coordinates are available, the percentage of geocoded addresses should be as high as possible. Other studies have reported geocoding rates ranging from 20% to 100% depending on factors such as the number of problematic addresses, quality of addresses, and type of geocoding method (McElroy, Remington et al. 2003). Recent studies in The Netherlands typically had geocoding rates higher than 90 - 95%. Such geocoding rates were also common in other recent studies in other countries.
- Records for which geographical coordinates could not be linked and for which therefore no air pollution exposure could be made will drop out of the epidemiological analyses. This may result in selection bias when for example non-geocodable records are clustered in the same area(s). It is therefore important that the percentage of geocoded addresses is as high as possible and that non-geocodable addresses are not clustered.
- Error may occur where for less specific addresses (for example when the house number is missing) a coarser geocode is linked (for example centroid of a postal code area). This reduces the specificity of the address and may lead to situations where coordinates are linked that do not reflect ground truth, i.e. the actual position. This results in a mix of accuracy levels within a geocoded dataset. Therefore, do not link geographical coordinates to addresses with non-complete address information.
- It has been shown that geocoding match rates may be lower for rural areas than for urban areas (Hay, Kypri et al. 2008). The reasons for this difference in accuracy include the

following: (1) rural areas tend to be less specific, with rural delivery routes and post office boxes sometimes used instead of street addresses; (2) there is more frequent use of unofficial or colloquial place names in rural areas; (3) when using the interpolation method there are larger interpolation errors due to longer street segments; and (4) roadway reference data for rural areas are less accurate than they are for urban areas. It is therefore important to try to document whether there are areas where the accuracy is less.

- Furthermore, in rural areas parcels may be large and contain many structures, so that a residence location may be different from the centroid of the parcel. This illustrates also that preferably building centroids should be used.

Validation

Because large spatial variability of air pollution concentrations occurs within tens of meters from major roads geographical accuracy is important. As a general rule, spatial data must be much more accurate than the minimum distance used in spatial analysis for the results to be meaningful. Within ESCAPE the accuracy of the geocoding procedure should therefore be at least 5-10 meters.

The accuracy of the geocoding procedure should be documented in two ways: (1) describe the claimed accuracy by the supplier of the reference dataset and (2) the accuracy of the geocoding procedure should also be evaluated by conducting our own validation.

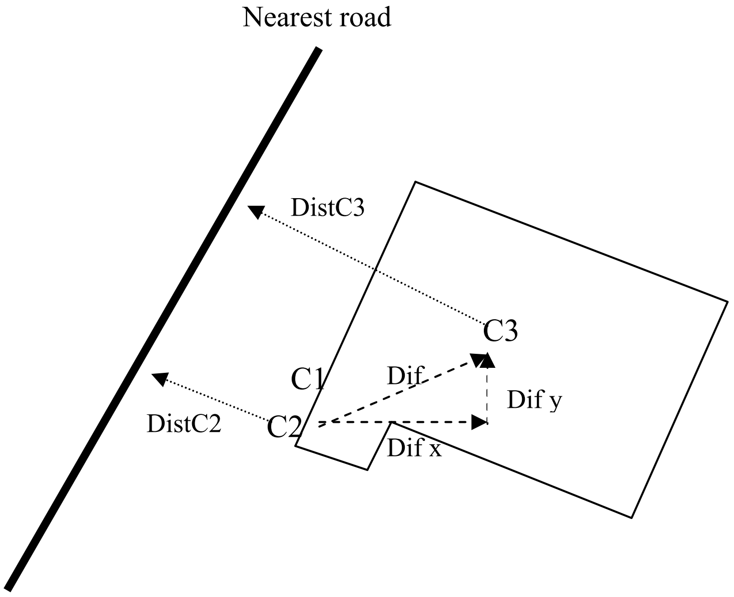
Our own validation will be conducted using the locations of ESCAPE monitoring sites. For all monitoring sites GPS readings have been made and the exact address is available when the site was located at a building. This exact address can then be geocoded using the procedures described above. The coordinates obtained by GPS readings and the coordinates obtained by the geocoding procedure can then be compared. Because the air pollution measurements and GPS readings will often be conducted at a drain pipe of the building, at a balcony or in the garden of a building, there will however be some difference between the coordinates using GPS readings and using geocoding procedures. The following methods will be used to conduct the validation:

1. For each location geocoded by the reference dataset and the GPS, the distance between each pair of geocoded addresses and its corresponding GPS surveyed location can be calculated: $d_i = [(x_i - x_{i0})^2 + (y_i - y_{i0})^2]^{1/2}$
   where d_i is positional error between the geocoded location of address i using the geocoding reference dataset and the GPS-surveyed location of address i; x_i and y_i are the x and y coordinates of the geocoded location of address i obtained from the geocoding reference dataset; x_{i0} and y_{i0} are the x and y coordinates of the GPS-surveyed location of

address i. Document the mean difference, range and standard deviation for the distance for all locations together, and separately for the different site types (regional background, urban background and traffic) to evaluate whether there are differences between urban and more rural areas. Further, document this also separately for homes and public building, e.g. hospitals, municipality health centers, schools etc, because the ground floor dimensions of such public buildings are mostly larger resulting in larger differences between GPS location of monitoring site and centroid of building for public buildings even when the geocoded address is correct.

2. Further, we will evaluate the distance in the x- and y-direction separately. The distance in x-direction can be calculated by: $d(x)_i = x_i - x_{i0}$, where $d(x)_i$ is the positional error in the x-direction between the geocoded location of address i and the GPS-surveyed location of address i. The distance in y-direction can be calculated in the same way: $d(y)_i = y_i - y_{i0}$. Document the mean difference, range and standard deviation for the x- and y-direction separately, and also separately for the different site types, and also for homes and public buildings separately. Further, make a scatterplot by plotting for each location the difference in x-direction and the difference in y-direction. This gives insight whether the geocoded addresses differ in a systematical way from the GPS-coordinates. Scatterplots can be made for all addresses together, for different site types separately, and for homes and public buildings separately.
3. Calculate for both the GPS coordinates and the geocoded addresses the distance to the nearest road, and calculate the difference between the distances to the nearest road for both coordinates. Document the mean difference, range and standard deviation for all sites together, for different site types separately, and for homes and public buildings separately.
4. Finally, plot on a map for each monitoring site:
   - the coordinate by the GPS reading
   - the coordinate by geocoding the address of the monitoring site
   - the ground floor dimensions of houses and buildings
   - the "real" location of the monitoring site. Because the exact location is known this location can also be plotted using information from the ground floor dimensions of the building and local knowledge
   - the digital road network which was used to calculate distances to the nearest road.

   Such a map gives insight in the potential different positional errors and will for example look like:



With:

C1 = ‘real’ location

C2 = coordinate from GPS reading

C3 = coordinate from geocoding the address

Dif = Distance between coordinate from GPS reading and coordinate from geocoding the address

Dif x = Distance in the x-direction between coordinate from GPS reading and coordinate from geocoding the address

Dif y = Distance in the y-direction between coordinate from GPS reading and coordinate from geocoding the address

DistC2 = Distance between coordinate from GPS reading and nearest road

DistC3 = Distance between coordinate from geocoding the address and nearest road

References

Goldberg, D. W., J. P. Wilson, et al. (2008). "An effective and efficient approach for manually improving geocoded data." Int J Health Geogr **7**(1): 60.

Hay, G., K. Kypri, et al. (2008). "Potential biases due to geocoding error in spatial analyses of official data." Health Place.

McElroy, J. A., P. L. Remington, et al. (2003). "Geocoding addresses from a large population-based study: lessons learned." Epidemiology **14**(4): 399-407.

Zimmerman, D. L., X. Fang, et al. (2007). "Modeling the probability distribution of positional errors incurred by residential address geocoding." Int J Health Geogr **6**: 1.

Appendix V: ESCAPE air pollution and GIS transfer: explanation of variables

ESCAPE air pollution and GIS data transfer:explanation of variables

Utrecht
Monday, May 31, 2010
Version 1
g.hoek@uu.nl

For documentation purposes IRAS will maintain a central database with the final air pollution and GIS data for the monitoring sites. Four sets of files will be prepared for each study area:

1. The standard Excel files including all samples, calculations and results of QA/QC
2. Individual measurement results from all monitoring sites (3 per site).
3. Annual average concentrations for all monitoring sites, used in land use regression modeling
4. Predictor data for all monitoring sites. This includes the GIS data and the monitoring site characterization (appendix 2 from the study manual).

Files will be transferred to IRAS (g.hoek@uu.nl) in Excel files. The first level is the files used for the calculations of NO2 and PM concentrations. The final results will be in the file **escape_airgis.xls**, for which this documents provides guidelines.

Excel files with calculations

Please send the final datafiles with a mail stating that this is the final data. Files have been sent back and forth with comments several times, so we want to avoid confusion.

Individual data

The individual air pollution measurements are stored in the worksheet **individual data.** An explanation of the variable names follows below. The database contains only the successful measurements used in the calculation of annual averages. Measurements failing the SOP criteria have been excluded by the centers. Blanks and duplicates are also not included. NOX-only areas can leave all data referring to PM empty. For them sampling session typically is 1, 2 or 3. Elemental composition data will be entered in a separate sheet (**elemental data)**, as we do not have the data yet and do not know yet which elements will be measured with adequate quality.

Variable	Explanation
COUNTRY	Country
STUDYAREA	Name of the study area as in ESCAPE study manual
TOWN	Town (city) in which the site is located
SITENAME	Street address, including number
SITEID	Site identification number given by each center
STARTDATE_PM	Start date of the PM sample (mm/dd/yy)
STARTDATE_NOX	Start date of the NOX sample (mm/dd/yy)
SESSION_PM	Sampling session PM, identifying sampling period (1 to ~ 12 for PM-NOX)
SESSION_NOX	Sampling session NOX, identifying sampling period (1 to ~ 12 for PM-NOX)
PM25	14-day average concentration PM2.5 (μg/m^3)
PM25ABS	14-day average absorbance PM2.5 (10^{-5} m^{-1})
PM10	14-day average concentration PM10 (μg/m^3)
PM10ABS	14-day average absorbance PM10 (10^{-5} m^{-1})
NO_2	14-day average concentration NO_2 (μg/m^3)
NO_x	14-day average concentration NO_x (μg/m^3)
NO	14-day average concentration NO (μg/m^3)
Comment	Comment such as 'imputed'

Annual average concentrations

These data are included in the worksheet 'annual average'. All unadjusted and adjusted averages are included here. Data from the ESCAPE continuous measurement site are included as well. This is intended for descriptive purposes only, not for use in the regression modeling part. A file for elemental composition will be developed later. The coarse particle concentration is added here by subtracting PM25 from PM10.

Variable	Explanation
COUNTRY	Country
STUDYAREA	Name of the study area as in ESCAPE study manual
TOWN	Town (city) in which the site is located
SITEID	Site identification number given by each center
SITETYPE	RB=regional background, UB=urban background, T=traffic
PM25UNADJ	Annual average concentration PM2.5 (μg/m^3) (unadjusted)
PM25	Annual average concentration PM2.5 (μg/m^3) (adjusted)
PM25ABSUNADJ	Annual average absorbance PM2.5 (10^{-5} m^{-1}) (unadjusted)
PM25ABS	Annual average absorbance PM2.5 (10^{-5} m^{-1}) (adjusted)
PM10UNADJ	Annual average concentration PM10 (μg/m^3) (unadjusted)
PM10	Annual average concentration PM10 (μg/m^3) (adjusted)
PM10ABSUNADJ	Annual average absorbance PM10 (10^{-5} m^{-1}) (unadjusted)
PM10ABS	Annual average absorbance PM10 (10^{-5} m^{-1}) (adjusted)
COARSE	Annual average concentration coarse particles (μg/m^3): PM10 – PM25 (adjusted)
NO2UNADJ	Annual average concentration NO2 (μg/m^3) (unadjusted)
NO2	Annual average concentration NO2 (μg/m^3) (adjusted)
NOXUNADJ	Annual average concentration NOX (μg/m^3) (unadjusted)
NOX	Annual average concentration NOX (μg/m^3) (adjusted)
NOUNADJ	Annual average concentration NO (μg/m^3) (unadjusted)
NOD	Annual average concentration NO (μg/m^3) (adjusted)
Comments	

GIS data

The files only contain basic data; thus data that can be calculated from the basic data are not included (e.g. GIS data from buffer 5000 – 1000 meter is not included since it can be calculated from the data from the 5000 and 1000 meter buffer). The common data are as described in Table 4 of the exposure assessment manual (version July 2010). If additional local data are obtained and used in modeling, add these to the file.

Name variable[1]	**Predictor variable**	**Unit**	**Buffer size**
STUDYAREA	Name study area ESCAPE manual		
SITEID	Site identification number given by each center		
XCOORD, YCOORD	Coordinate variables	m	NA
HDRES	High density residential land	m^2	100, 300, 500, 1000, 5000
LDRES	Low density residential land	m^2	100, 300, 500, 1000, 5000
INDUSTRY	Industry	m^2	100, 300, 500, 1000, 5000
PORT	Port	m^2	100, 300, 500, 1000, 5000
URBGREEN	Urban green	m^2	100, 300, 500, 1000, 5000
NATURAL	Semi-natural and forested areas	m^2	100, 300, 500, 1000, 5000
POP	Number of inhabitants	N(umber)	100, 300, 500, 1000, 5000
HHOLD	Number of households	N(umber)	100, 300, 500, 1000, 5000
SQRALT	Square root of altitude	m	NA
TRAFNEAR	Traffic intensity on nearest road	Veh.day^{-1}	NA
DISTINVNEAR1 DISTINVNEAR2	Distance to the nearest road	m^{-1}, m^{-2}	NA
INTINVDIST INTINVDIST2	Product of traffic intensity on nearest road and inverse of distance to the nearest road and distance squared	Veh.day^{-1}m^{-1} Veh.day^{-1}m^{-2}	
TRAFMAJOR	Traffic intensity on nearest major road	Veh.day^{-1}	NA
DISTINVMAJOR1 DISTINVMAJOR2	Distance to the nearest major road	m^{-1}, m^{-2}	NA
INTMAJORINVDIST INTMAJORINVDIST2	Product of traffic intensity on nearest major road and inverse of distance to the nearest major road and distance squared	Veh.day^{-1}m^{-1} Veh.day^{-1}m^{-2}	
TRAFMAJORLOAD	Total traffic load of major roads in a buffer (sum of (traffic intensity * length of all segments))	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000
TRAFLOAD	Total traffic load of all roads in a buffer (sum of (traffic intensity * length of all segments))	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000
HEAVYTRAFNEAR*	Heavy-duty traffic intensity on nearest road	Veh.day^{-1}	NA
HEAVYINTINVDIST HEAVYINTINVDIST2	Product of Heavy-duty traffic intensity on nearest road and inverse of distance to the nearest road and distance squared	Veh.day^{-1}m^{-1} Veh.day^{-1}m^{-2}	
HEAVYTRAFMAJOR	Heavy-duty traffic intensity on nearest major road	Veh.day^{-1}	NA
HEAVYTRAFMAJORLOAD	Total heavy-duty traffic load of major roads in a buffer (sum of (heavy-duty traffic intensity * length of all segments))	Veh.day^{-1}m	25, 50, 100, 300, 500, 1000
HEAVYTRAFLOAD	Total heavy-duty traffic load of all roads in a buffer	Veh.day^{-1}m	25, 50, 100,

	(sum of (heavy-duty traffic intensity * length of all segments))		300, 500, 1000
ROADLENGTH	Road length of all roads in a buffer	m	25, 50, 100, 300, 500, 1000
MAJORROADLENGTH	Road length of major roads in a buffer	m	25, 50, 100, 300, 500, 1000
DISTINVNEARC1 DISTINVNEARC2	Distance to the nearest road	m^{-1}, m^{-2}	NA
DISTINVMAJORC1 DISTINVMAJORC2	Distance to the nearest major road	m^{-1}, m^{-2}	NA
CANYON	Aspect ratio (sum height buildings both side of road divided by road width)	m/m	NA

[1] Variable name: Combining name and buffer size, for HDRES: HDRES_100, HDRES_300, HDRES_500,HDRES_1000, HDRES_5000

Site classification

This information is taken from the site characterization form, described in appendix 2 of the ESCAPE study manual. To some of the variables we added “sc” to distinguish the site characterization data from the GIS data.

Variable	**Explanation**
STUDYAREA	Study area
SITEID	Site code
SITENAME	Site address (street + street number)
XCOORD	X-coordinate[1]
YCOORD	y-coordinate[1]
LIGHTVEHFLOW	Light vehicles flow on the nearest street (cars day^{-1})
HEAVYVEHFLOW	Heavy vehicles flow on the nearest street (cars day^{-1})
DISTNEAR_SC	Distance to nearest street (m)
DISTNEARMAJOR_SC	Distance to nearest major street (m)
DISTINT_SC	Distance to nearest intersection (m)
DISTLIGHT_SC	Distance to nearest traffic light (m)
WIDTH	Width of the nearest street (m)
HEIGHT	Height of building of which home is part (m)
STREETCONFIG	1. Largely uninterrupted rows of homes on both sides of the street 2. Largely uninterrupted rows of homes on the study home side of the street, but not the other side 3. Largely uninterrupted rows of homes on the other side of the street, but not on the study home side 4. On both sides of the road interrupted rows of homes
BUILDINGINT	Buildings uninterrupted for at least 25 meter on each side (yes / no)
FLOORSAMPLE	Floor at which outdoor measurements are made (0, 1, 2, etc)
SAMPHEIGHT	Sampling height for outdoor measurements (m)
SAMPSIDE	Sampling site in backyard (B), streetside (S) or rooftop (R)
PARKING	Is there a large parking lot within 100 meter? Yes or no
INDUSTRY	Is there a small industrial plant (e.g. garage, petrol station) within 100 meter? Yes or no

[1] Indicate coordinate system used

References

Beelen, R., Hoek, G., Fischer, P., van den Brandt, P.A. & Brunekreef, B. 2007, "Estimated long-term outdoor air pollution concentrations in a cohort study", *Atmos Environ,* vol. 41, pp. 1343-1358.

Beelen, R, Voogt, M., Duyzer, J., Zandveld, P., Hoek, G. Comparison of the performances of land use regression modeling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. Submitted 2010

Gilbert, N.L., Goldberg, M.S., Beckerman, B., Brook, J.R. & Jerrett, M. 2005, "Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model", *J Air Waste Manag Assoc,* vol. 55, pp. 1059-1063.

Henderson, S., Beckerman, B., Jerrett, M. & Brauer, M. 2007, "Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter", *Environ Sci Technol,* vol. 41, pp. 2422-2428.

Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. & Briggs, D. 2008, "A review of land-use regression models to assess spatial variation of outdoor air pollution", *Atmos Environ,* vol. 42, pp. 7561-7578.

Jerrett, M., Arain, M., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N., Brook, J. & Finkelstein, N. 2007, "Modelling the intra-urban variability of ambient traffic pollution in Toronto, Canada", *Journal of Toxicology and Environmental Health, Part A,* vol. 70, pp. 200-212.

Moore, D., Jerrett, M., Mack, W. & Kuenzli, N. 2007, "A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA", *J Environ Monit,* vol. 9, pp. 246-256.

Ryan, P.H., LeMasters, G.K., Biswas, P., Levin, L., Hu, S., Lindsey, M., Bernstein, D.I., Lockey, J.E., Villareal, M., Khurana Hershey, G.K. & Grinshpun, S.A. 2007, "A comparison of proximity and land use regression traffic exposure models and wheezing in infants", *Environ Health Perspect,* vol. 115, no. 2, pp. 278-284.

Slama, R., Morgenstern, V., Cyrys, J., Zutavern, A., Herbarth, O., Wichmann, H.E. & Heinrich, J. 2007, "Traffic-related atmospheric pollutants levels during pregnancy and offspring's term birth weight: a study relying on a land-use regression exposure model", *Environ Health Perspect,* vol. 115, no. 9, pp. 1283-92.