Procedure for exposure assessment for cohort addresses

Version April 6 2010

Preface

Based on the first experiences with exposure assessment in some areas, we provide a more extended description of the procedure for exposure assessment for cohort addresses in order to clarify and harmonize the exposure assessment in all ESCAPE study areas. Specifically, a few extreme outliers were found, that are likely unrealistic. We need an explicit decision on how to truncate extreme values of predictor variables. Use this procedure when estimating concentrations for cohort addresses in your study area. Deviation from this procedure is possible after contact with WP2 coordinators. This text does not replace chapter 6 of the exposure manual but extends it, please therefore also read chapter 6 of the exposure manual.

Introduction

After approval of your LUR models by the WP2 coordinators, the final LUR models will be used to estimate outdoor air pollution concentrations at the addresses of study participants. 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. 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. When the values for the predictor variables are available for the coordinates of the addresses, outdoor air pollution concentrations for the coordinates of the cohort addresses are estimated by filling in the developed exposure model.

Procedure for exposure assessment

Although we aimed to cover the complete range of predictor variable values with the locations of our monitoring sites, values for predictor variables at the cohort addresses may fall out of this range, also because we have only a limited number of monitoring sites in each study area. In addition, the coordinate of an address could also be located by data inaccuracies (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.

When using predictor values that are outside the range as estimated for the monitoring sites this could result in over-predictions, because we cannot guarantee that the relationship between the concentration and the predictor variable remains linear outside the range of values at the monitoring sites.

Therefore, please first conduct descriptive analyses of the predictor variables for both:
1) the monitoring sites, and
2) the cohort addresses,
before estimating air pollution concentrations using the LUR model. The descriptive analyses for the monitoring sites should be conducted separately for the PM (N=20) and the NOx (N=40) sites (for the PM/NOx areas), and separately for all sites and background sites, because the distribution of predictor variable values between these sites can be different.

In the two German study areas, it was observed that for a limited number of cohort addresses predictor variables indeed were more extreme than observed at the monitoring sites, resulting in a few extreme concentrations, which would very likely create problems in the epidemiological analysis. This occurred especially in the small-scale traffic variables, e.g. very small distances to roads (e.g. 0.25 m to center of the road), resulting in extreme values for the inverse distance and inverse distance * intensity variables. Also, the traffic load in 50 and 100m buffer was more extreme in a few cases than at the monitoring sites (e.g. because a freeway situation was present for the cohort addresses and the monitoring sites were more intra-urban).

Because we cannot document that the associations increase linearly well outside the range that we observed at the sites and extreme values are problematic in epidemiological analysis, we think it is safer to truncate these predictor values to the maximum that was observed at the monitoring sites. This despite the fact that it is very likely that more extreme values of predictor variables realistically occur at cohort addresses. We considered allowing some extrapolation outside the observed range (e.g. 50%), but believe any number to select is difficult to defend. With this procedure we accept that we may underestimate the highest exposures somewhat, but they remain the highest. This procedure seems reasonable if there is a limited number of values that fall outside the range at the monitoring sites. It is therefore critical to document the percentage of observations that was truncated. In the two German studies this procedure resulted in more realistic predictions.

<u>The proposed procedure is then that we truncate the values for predictor values at the cohort addresses and these will be given the highest value which occurs at one of the monitoring sites for that specific predictor variable</u> (NB. for PM/NOx areas this highest value depends on the pollutant that will be estimated, as well on whether it is all sites or background model). This will be done for all values above the highest (or lowest) value at a monitoring site, regardless of how much higher this value is.

For documentation purposes, please estimate then concentrations for the addresses using both the original predictor values (i.e. the untruncated values) and using the truncated values.
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.
Further checks that need to take place are that the 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).
After these checks have been conducted, please send the following description to IRAS (Gerard Hoek; g.hoek@uu.nl and Rob Beelen; r.m.j.beelen@uu.nl):

- Description of the predictor variables at the monitoring sites (for PM/NOx areas: separately for the PM and NOx sites; and separately for all sites and background sites) (only necessary for predictor variables that enter one or more LUR models): Provide the following information: N, Min, P1, P5, P10, P25, P50, mean, P75, P90, P95, P99 and Max
- Description of the predictor variables at the addresses (only necessary for predictor variables that enter one or more LUR models):
  1) For the *untruncated* predictor variables: N, Min, P1, P5, P10, P25, P50, mean, P75, P90, P95, P99 and Max
  2) For the *truncated* predictor variables: N, Min, P1, P5, P10, P25, P50, mean, P75, P90, P95, P99 and Max, and for each variable the number (and percentage) that has been truncated. Please also indicate for which pollutant each variable will be used in the model (NB. for PM/NOx areas and for all sites vs background sites the number of truncated records could differ based for which pollutant the variable has been used).
- Description of the estimated concentrations at the addresses:
  1) Based on the *untruncated* predictor variables: N, Min, P1, P5, P10, P25, P50, mean, P75, P90, P95, P99 and Max
  2) Based on the *truncated* predictor variables: N, Min, P1, P5, P10, P25, P50, mean, P75, P90, P95, P99 and Max
  3) Correlation and scatterplot between the estimated concentrations based on the *untruncated* and *truncated* predictor variables for the different pollutants.

If you have any questions, please send these to us together with these descriptions.

We will evaluate these descriptions and after approval from us you can continue with the final steps (see below). If you want to deviate from the procedure please discuss this with us. In that way we harmonize the exposure assessment over the different study areas.

Final steps and transferring of the estimated concentrations to the health studies

After approval of the estimated concentrations for the addresses, the exposure estimates for all addresses can be transferred to the groups that will conduct the epidemiological analyses (NB First extrapolation over time should be applied – See chapter 7 of the exposure manual).

In addition, the R^2 value and RMSE value of the *cross-validation* results as well as the R^2 value and RMSE value 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.

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, if the amount of cells is feasible. For large study areas a cruder grid could be used. The same procedures as for estimating for the cohort addresses should be followed.

A further step is to extrapolate the modelled concentrations back in time. We will also provide an extended procedure for this.