EFFECT OF THE NUMBER OF MEASUREMENT SITES ON LAND USE REGRESSION (LUR) MODELS OF AIR POLLUTANTS

Xavier Basagaña, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
David Agis, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
Inma Aguilera, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
Marcela Rivera, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
Maria Foraster, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
Audrey de Nazelle, *Centre for Research in Environmental Epidemiology (CREAL), Spain*
Nino Künzli, *Swiss Tropical and Public Health Institute Basel, Switzerland*

Background and Aims: Land use regression (LUR) models used in epidemiologic studies on air pollution usually rely on air pollution measurements at a few locations across a city. The number of locations commonly varies from 20 to 100. Our aim is to explore the effect of the number of locations (N) on LUR model performance.
Methods: The REGICOR-AIR study measured NO_2 in 159 locations in the two adjacent cities of Girona and Salt (Spain). A total of 102 potential predictors like traffic, road type and land cover entered a supervised forward selection procedure to derive the final model. Training datasets of 20, 50, 80, 110 and 140 points were obtained by randomly sampling the original dataset. The remaining points were used as validation datasets. The process was repeated 200 times for each N. Adjusted R^2 (R^2a) and leave-one-out cross-validation R^2 (R^2cv) were calculated in the training datasets, and validation R^2 was obtained from the validation datasets. The median value over the 200 iterations was reported.
Results: With 20 points, the median R^2a and R^2cv were 79% and 71% but the validation R^2 was negative, indicating that the regression equation resulted in a worse prediction than assigning the overall mean to each point of the validation sample. Both R^2a and R^2cv decreased with increasing N to reach R^2a=54% and R^2cv=52% for N=140. Conversely, the validation R^2 increased with increasing N, from negative values (N=20) to 44% (N=140). The most frequent predictive variable entered in only 32% of the models for N=20, while for N=140 four variables appeared in more than 90% of the final models.
Conclusions: Increasing the number of measurement points increases both the predictability and the stability of the models. Leave-one-out cross-validation R^2 overestimates the predictive ability of the model. This bias is more important at small sample sizes.
References: