Environment International 92–93 (2016) 202–209

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Back-extrapolated and year-specific NO_2 land use regression models for Great Britain - Do they yield different exposure assessment?

John Gulliver [a,*], Kees de Hoogh [b,c], Gerard Hoek [d], Danielle Vienneau [b,c], Daniela Fecht [a], Anna Hansell [a]

[a] UK Small Area Health Statistics Unit (SAHSU), MRC-PHE Centre for Environment & Health, Imperial College London, Norfolk Place, W2 1PG London, UK
[b] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland
[c] University of Basel, Petersplatz 1, 4003 Basel, Switzerland
[d] Institute of Risk Assessment Sciences, University of Utrecht, Yalelaan 2, 3584 CM Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 14 January 2016
Received in revised form 21 March 2016
Accepted 29 March 2016
Available online xxxx

Keywords:
Air pollution modelling
Back-extrapolation
Exposure assessment
GIS
Land use regression
Nitrogen dioxide

ABSTRACT

Robust methods to estimate historic population air pollution exposures are important tools for epidemiological studies evaluating long-term health effects. We developed land use regression (LUR) models for NO_2 exposure in Great Britain for 1991 and explored whether the choice of year-specific or back-extrapolated LUR yields 1) similar LUR variables and model performance, and 2) similar national and regional address-level and small-area concentrations. We constructed two LUR models for 1991using NO_2 concentrations from the diffusion tube monitoring network, one using 75% of all available measurement sites (that over-represent industrial areas), and the other using 75% of a subset of sites proportionate to population by region to study the effects of monitoring site selection bias. We compared, using the remaining (hold-out) 25% of monitoring sites, the performance of the two 1991 models with back-extrapolation of a previously published 2009 model, developed using NO_2 concentrations from automatic chemiluminescence monitoring sites and predictor variables from 2006/2007. The 2009 model was back-extrapolated to 1991 using the same predictors (1990 & 1995) used to develop 1991 models. The 1991 models included industrial land use variables, not present for 2009. The hold-out performance of 1991 models (mean-squared-error-based-R^2: 0.62–0.64) was up to 8% higher and ~1 μg/m^3 lower in root mean squared error than the back-extrapolated 2009 model, with best performance from the subset of sites representing population exposures. Year-specific and back-extrapolated exposures for residential addresses ($n = 1.338{,}399$) and small areas ($n = 10.518$) were very highly linearly correlated for Great Britain ($r > 0.83$). This study suggests that year-specific model for 1991 and back-extrapolation of the 2009 LUR yield similar exposure assessment.

1. Introduction

Land use regression (LUR) modelling has been widely used to estimate exposures for a range of air pollution metrics (Adam-Poupart et al., 2014; Aguilera et al., 2008; Briggs et al., 2000; de Hoogh et al., 2014; Montagne et al., 2015; Yang et al., 2015) over different spatial and temporal scales (Hystad et al., 2011; Liu et al., 2012; Tang et al., 2013; Vienneau et al., 2013; Wang et al., 2014). One of the potential uses of LUR is historical exposure assessment to meet the needs of cohort studies with recruitment in the past and life-course epidemiology. Limiting factors in the application of LUR to earlier years are the availability of historical air pollution measurements (i.e. required to train and evaluate models) that adequately represent population exposures and, to a lesser extent, data on historical patterns of land use, transportation and industrial activity. Transferring current or recent LUR models to earlier years (i.e. back-extrapolation) has been shown to have potential in estimating air pollution exposures for historical periods (Chen et al., 2010; Cesaroni et al., 2012; Eeftens et al., 2011; Gulliver et al., 2013; Levy et al., 2015). The main premise for back-extrapolation is that air pollution sources are similar and the spatial structure of concentration surfaces remains stable over time.

NO_2 is one of the main pollutants of health concern, which can be viewed as a proxy for transport-related exposure (HEI, 2010). We previously evaluated back-extrapolation of LUR models with a spatial resolution of 200 m for Great Britain (GB) for the year 2009 to predict concentrations of NO_2 in 1991 at 451 diffusion tube monitoring sites (Gulliver et al., 2013). The 2009 models were developed using predictor variable data from that period (2006/2007). Back-extrapolation of the 2009 models was performed using predictor variables from nearest years (1990 & 1995) as possible to the target year (1991). We showed that back-extrapolation for the whole of GB was valid for up to

* Corresponding author at: UK Small Area Health Statistics Unit, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, St Mary's campus, Norfolk Place, London W2 1PG, UK.
E-mail address: j.gulliver@imperial.ac.uk (J. Gulliver).

http://dx.doi.org/10.1016/j.envint.2016.03.037
0160-4120/

18 years earlier (mean-squared-error-based-R^2 (MSE-R^2) ~0.55), but models performed poorly in some regions, especially the regions Midlands and North of England (North): MSE-R^2: 0.01–0.24.

In this paper, we use the 451 NO_2 monitoring sites in 1991 to produce GB LUR models for 1991 for comparison with back-extrapolated models. We used historical predictor variables from the nearest years (1990 & 1995), both to develop new models for 1991 and apply the back-extrapolated 2009 model. We addressed two specific questions to inform ongoing and future epidemiological studies:

1) How different are NO_2 LUR models developed specifically for 1991 in terms of variables and performance compared to an available model for 2009 back-extrapolated to 1991?
2) Does the choice of a year specific (1991) or back-extrapolated (2009) model yield different national and regional address-level and small-area exposure assessments, given known regional differences in air pollution levels and number of monitoring sites?

2. Materials and methods

2.1. Monitored concentrations of NO_2

We acquired data from 451 diffusion tube sites for 1991 and data from 187 routine monitoring sites that form part of the Automatic Urban and Rural Network (AURN) in the UK for 2009 as detailed in Gulliver et al. (2013). We chose 1991 for developing year-specific LUR models as that year had a relatively large number of monitoring sites where all diffusion tubes had been analysed at the same laboratory (Bush et al., 2001). Average measured concentrations were broadly similar in 1991 (39.33 μg/m^3) and 2009 (36.69 μg/m^3) but are not directly comparable due to different locations of monitoring sites and monitoring methods.

2.2. Development of variables

Data on monitored NO_2 concentrations for 1991 diffusion tube sites, site name, site type, and geographic coordinates (six-figure British National Grid), were integrated into the geographic information system ArcGIS version 10 (ESRI, Redlands, CA). Characteristics of 1991 sites differ compared to the 2009 Automated Urban and Rural Network (AURN). Firstly, in 1991 geographic coordinates were recorded with a ground precision of 100 m whereas for 2009 sites have a precision of 1 m. We therefore restricted the 1991 models to a minimum buffer distance of 200 m and mapping of concentrations for both the 1991 and back-extrapolated 2009 models to a 200 m grid. Secondly, the classification of site types differs between the two years (for 1991 there are 26 classes of site type related to land cover classes and emissions sources but no information on proximity to road sources whereas for 2009, site type is classified as: industrial, roadside, urban centre, urban background, suburban, or rural).

Table S1 in Supplementary material summarises the data and variables used in LUR models. Data on land cover, road geography and altitude were integrated into ArcGIS. Land cover variables were derived from a combination of CORINE land cover for Europe (CLC) and the Land Cover Map of Great Britain (LCM) as in Gulliver et al. (2013). Both CLC, comprising 44 categories of land cover on a 100 m grid, and LCM, comprising 25 categories of land cover on a 25 m grid are available for the year 1990. LCM contains only two classes of urban land (i.e. urban and suburban). Industry, ports, airports, construction sites, green urban areas and sport and leisure facilities from CLC were, therefore, substituted into the 25 m LCM where they were intersected with ‘urban land’. This enhanced land cover data set was then aggregated into seven groups for use in modelling: high density urban, low density urban, industry and commercial, urban green areas, agriculture, semi-natural land, and water.

Data on road geography came from the Ordnance Survey via the agreement for UK educational establishments (www.digimap.co.uk). Meridian-1 from 1995 was applied to 1991 models. Roads are separated into four classes: motorways, A-roads (major trunk roads, dual carriageways and arterial roads), B-roads (roads with significant traffic flows but not in ‘A’ class), and minor roads (urban side-roads and country lanes). Data on altitude were obtained from the PANORAMA™ digital terrain model from Ordnance Survey via Digimap.

Each dataset was, in turn, intersected with a common 25 m grid covering the whole of GB and summed within each grid cell. “Buffering”, using the ArcGIS FOCALSUM tool with the circle option (subsequently referred to as buffers), was used to sum the contribution (i.e., area or length) of each land cover and road variable around each monitoring site. For land cover, buffers were created for 0.2, 0.3, 0.4, 0.5, 0.75, 1, 2, 3, 5, 10 and 20 km. For road length 0.2, 0.3, 0.4, 0.5, 0.75 and 1 km buffers were used. Both zero-centred and ring buffers were constructed. Altitude (and log-transformed altitude) was obtained for each monitoring site using a point-in-grid function. Coordinates of each monitoring site were recorded for applying either second or third order trend surfaces.

Monitoring sites were pooled into model ‘training’ and ‘evaluation’ (i.e. hold-out validation (HOV)) groups, using a 75:25 split. To control for potential geographical bias in this selection process, monitoring sites were stratified by five regions (South East, South West & Wales, Midlands, North, and Scotland) and site type. The balance of site types by region was checked for the training and evaluation groups and the LUR variables attributed to the monitoring sites were then imported into SPSS (version 20) for development of the regression models.

2.3. Modelling strategy for year-specific (1991) models

Historic monitoring networks were designed to target areas of industrial pollution in order to control air quality at source, hence 49% (221) of the 451 NO_2 monitoring sites in 1991 were located in the North region, with a concentration of heavy industry but with only 26% of the population of Great Britain in 1991. In order to study the potential effects of monitoring site selection on exposure estimates, we developed two NO_2 LUR models for 1991: 1) using all 451 available monitoring sites (model 1991A), and 2) using a reduced set of 186 monitoring sites (model 1991B) proportionate to the population in each region, relative to the region with the highest population (South East). Models were developed using a supervised forward approach that we, and others (Beelen et al., 2010; Gulliver et al., 2013; Vienneau et al., 2013), have previously used. In this study we included a rule that each of the main sources of NO_2 variability must be represented in the model (i.e. road traffic and at least one class of urban land) based on information on national source emissions contributions (Carslaw et al., 2011).

Following the supervised forward approach, LUR models were developed such that: 1) each variable in the final model had a significant correlation with the monitored concentration ($p < 0.05$), 2) the direction of effect met predetermined expectations, 3) the direction of effect of predictors already in the model did not change as subsequent predictors were added, and 4) variables adding < 1% to the explained variation in monitored concentrations were excluded.

2.4. Back-extrapolation

We used the best performing 2009 model on HOV samples of monitoring sites as described in Gulliver et al. (2013) to back-extrapolate NO_2 concentrations to 1991 for comparison with year-specific models. We applied the 2009 model using the relevant predictor variables from the same data sets (i.e. 1990 & 1995) used to develop year-specific models. Using annual average NO_2 concentrations from all five available concurrent background monitoring sites for 1991 and 2009 from the AURN, we calculated the average difference in concentrations of NO_2 between 1991 and 2009. We then added the resulting average