Environmental Modelling & Software 105 (2018) 17–23

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment

David W. Morley, John Gulliver*

MRC-PHE Centre for Environment & Health, Department of Epidemiology & Biostatistics, Faculty of Medicine, Imperial College London, W2 1PG, London, UK

ARTICLE INFO

Article history:
Received 3 March 2017
Received in revised form
25 January 2018
Accepted 30 March 2018

Keywords:
Land use regression
R
GIS
Air pollution
Epidemiology
Exposure assessment

ABSTRACT

Land use regression (LUR) is commonly used to estimate air pollution exposures for epidemiological studies. By statistically relating a set of geolocated measured pollutant values with explanatory variables defining sources and modifiers of air pollution patterns, such as land cover characteristics, traffic flow and intensity, it is possible to predict pollution levels at unsampled locations. LUR utilises simple linear regression, but the generation of predictor variables, application of the model and the supervised iterative approach to model development means an analyst must be a competent user of both GIS and statistical packages. Here we present an application to simplify the LUR modelling process for exposure scientists and environmental epidemiologists. RLUR is a user-friendly application built using the statistical and GIS capabilities of the R programming language. The main aim of this software is to provide an introduction to the LUR process without the need for specific GIS or statistical expertise.

1. Introduction

Epidemiological cohort studies attempting to statistically link the long-term effects of air pollution to specific diseases require a representative exposure estimate for a defined time period, usually for many thousands of individuals. Recent examples include the BioSHaRE (Cai et al., 2017) and ESCAPE (Jacquemin et al., 2015) projects to study, for example, asthma incidence in relation to a range of air pollutant metrics (e.g. $PM_{2.5}$, PM_{10}, NO_2, NO_X) in several European countries. The use of portable personal monitoring devices would often be preferred over exposure modelling, but this has the limitation of being restrictively expensive and time consuming for larger studies of long-term health effects. As a result, modelling methods may be used to predict exposure at desired locations (e.g. home address). Although studies in the past relied on exposure proxies such as distance to the nearest road (Hoek et al., 2002), most recent and current studies use either dispersion modelling (deterministic) or statistical techniques (e.g. spatial interpolation, land use regression (LUR)). Given a detailed inventory of a point or multiple point sources, a dispersion model (such as ADMS-Urban, Cambridge Environmental Research Consultants (2008)) takes into account the spread of a pollutant factoring in meteorology, atmospheric circulation, land cover and terrain. This has the potential to give accurate exposure data, but is reliant on obtaining detailed input data on source emissions and may also be computationally restrictive over very large areas (e.g. countries or continents). A second set of empirical modelling methods utilises monitoring of pollutants at fixed geographic locations. At the simplest level these can be transformed to an areal prediction using spatial interpolation methods or geostatistical techniques such as Kriging (Briggs et al., 2000). However, depending on the density of sampling locations, interpolation may not express the inherent variability due to local-scale land cover (Gulliver et al., 2011). For example, Karner et al. (2010) show a rapid decrease in a number of traffic derived pollutants with distance from the road edge, resulting in often a halving of concentration within 150m. LUR is able to capture this local variability by relating the measured pollutant levels to a set of GIS derived land cover variables in a linear regression that are directly associated to the pollution such as distance to the road edge and relative traffic intensities. LUR has the advantage of being computationally undemanding when compared to dispersion models and efficient to apply to large cohorts often on national scales. The main limitation of LUR, however, is that it requires a network of fixed-site air pollution monitors to derive a model, or calibrate an existing model transferred from another area.

* Corresponding author.
E-mail address: j.gulliver@imperial.ac.uk (J. Gulliver).

https://doi.org/10.1016/j.envsoft.2018.03.030
1364-8152/

Although conceptually quite simple (based on linear regression), the generation of LUR variables, model construction, and application to unsampled locations (e.g. cohort addresses) requires a combination of specialist GIS and statistical software applications. Here, we present a tool called RLUR to allow a user to follow a workflow to develop a LUR model from extracting GIS variables, model creation and validation, to generation of long-term (i.e. annual) exposure estimates. RLUR is designed to be user-friendly to allow users not skilled in GIS or statistics to generate LUR predictions from their own data. Using GIS data (e.g. land cover, road network, population), RLUR automatically generates potential predictor variables in proximity to air pollution monitoring sites (e.g. the training dataset) using distance and buffering routines. This removes the need to manually, or via bespoke scripts, apply spatial operations in a GIS. There is no compromise to the spatial accuracy of variable generation by using RLUR instead of using commercially available GIS. The basic requirements are vector data (shapefiles) to include land cover with a unique land cover type per area, a road network attributed with annual average daily traffic flows (separately into 'all vehicles' and 'heavy goods vehicles'), and a point data set representing the number of households and/or population at each location. These types of data are readily available in these formats so users inexperienced in GIS should be able to load data, develop, and evaluate a LUR model. Where data are not available in an appropriate format then some basic GIS training may be required to edit attribute tables (e.g. to summarise traffic as a single annual average value for each road link in the road network attribute table).

We anticipate that potential users will include exposure scientists wishing to undertake an initial assessment of the potential of LUR in a particular study, epidemiologists who may wish to gain understanding of LUR to help the interpretation of findings from subsequent epidemiological studies, and students studying in the areas of exposure science, epidemiology, and public health. The underlying R code is made available as open source to allow users to customise the framework in terms of included explanatory variables and modelling methods specific to particular problems.

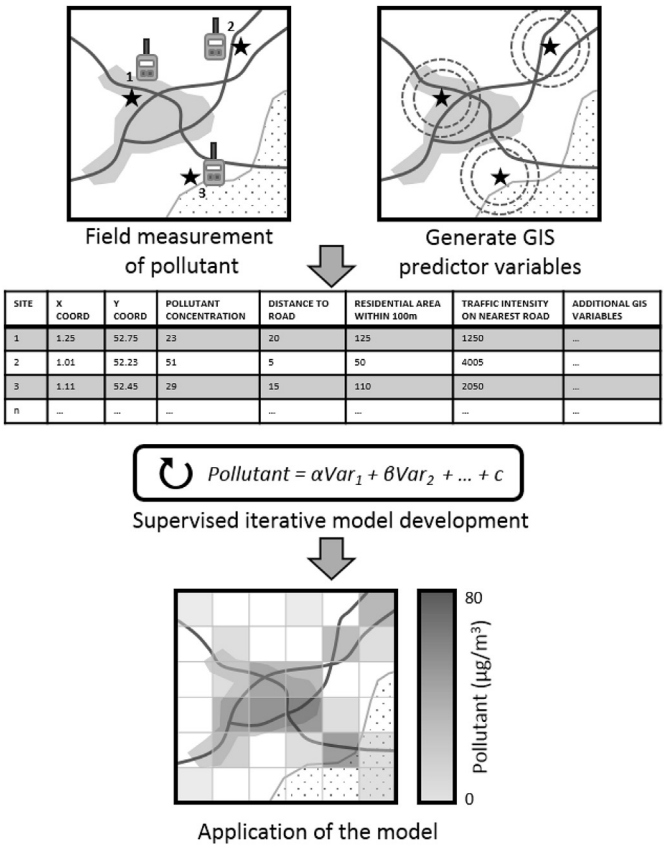

Fig. 1. Workflow for generating and applying an LUR. The pollutant of interest is measured at a set number of locations in the study area. GIS predictor variables are also generated at these points to create the training dataset. Model development (linear regression) follows a supervised approach of including and dropping variables (shown in arbitrary units for illustration only). The final model can be applied to estimate exposure over a regular grid to create a continuous surface or at known address point locations.

2. LUR methodology

LUR is an established method, first introduced to environmental epidemiology by Briggs et al. (1997), then known as 'regression mapping', and has been reviewed by Ryan and LeMasters (2007), Hoek et al. (2008), Hoek et al. (2015) and Gulliver and de Hoogh (2015). The basic workflow of developing and applying an LUR model is given in Fig. 1.

The initial step is to obtain a set of georeferenced air pollutant concentration measurements. These may be derived from a bespoke measurement campaign or sourced from an existing monitoring network such as the UK's Automatic Urban and Rural Network[1] (AURN). Importantly, the distribution of sites should be as representative as possible of conditions over the study area and characterise both high (e.g. road traffic, industrial) and low (e.g. residential, suburban) pollution environments. The number of sample sites needed will vary depending on the scale of the study, but LUR models are commonly developed with between 20 and 100 training locations (Hoek et al., 2015).

Secondly, GIS-derived spatial predictor variables are created for each monitoring location to empirically relate to measured pollution values. GIS data sets need to be gathered for the entire study area and also a surrounding area to account for the influence on external sources of pollution that will affect measurement stations and address locations that are near to the perimeter. Candidate variables should include significant emissions sources that could explain the spatial variability in measured air pollutant concentrations, and subsequently human exposures. Commonly, data sets used include a representation of the road network with associated traffic flow data of light (cars) and heavy (lorries, buses) vehicles, land cover classifications (residential, urban parkland, industrial, farmland, woodland etc.) population density and altitude in areas of complex terrain. Variables are derived from these using basic GIS operations such as circular buffering, to calculate the total area of land cover of a particular category or the total road length (with varying radii usually between 25m and 5000m) around a monitoring site, or the distance (often log or inverse transformed) to the nearest major or minor road optionally weighted by the corresponding traffic flow of light or heavy vehicles. Variables generated by RLUR follow those derived for the ESCAPE project (Eeftens et al., 2012). The ESCAPE variables cover a large number of predictors related to sources (e.g. roads) and sinks (natural landscape) and have been used extensively in cohort studies in Europe (http://www.escapeproject.eu/publications.php); other LUR studies have used similar types of variables but with different naming conventions (Ryan and LeMasters, 2007). Since the ESCAPE project, other studies outside Europe, e.g. in Australia (Dirgawati et al., 2015), China (Meng et al., 2015) and South Africa (Muttoo et al., 2018), have adopted the ESCAPE variable list, thus it has global application especially in urban areas. Table S1 in the supporting information for

[1] https://uk-air.defra.gov.uk/networks/network-info?view=aurn.