# A scripted workflow for updating GIS regression mapping models using land use and other predictors of air pollution

*Luke D. Knibbs[1,3], Ivan C. Hanigan[2,3,4], Christy Geromboux[2,3]*

*[1] The University of Queensland, School of Public Health, [2] University Centre for Rural Health, North Coast, School of Public Health, The University of Sydney, Sydney, Australia. [3] Centre for Air pollution, energy and health Research (CAR). [4] Health Research Institute, University of Canberra, Canberra, Australia.*

## Contents



## 1 Abstract

In this study we aimed to develop a scripted workflow to estimate air pollution for points of interest. This is needed to assign exposure estimates for health studies.

## 2 Introduction

The scripted workflow allows updated geographical information system (GIS) regression mapping models that use land use and other predictors to be generated. The air pollution modelling framework is based on linear regression modelling and relies on the coefficients from a regression equation to estimate air pollution at new locations. The method is also known as land use regression (LUR) and 'GIS regression mapping', and was first introduced by Briggs *et al.* (1997). It has become a technique commonly used to create predictive spatial models (and spatiotemporal models) of air pollution levels. Such GIS regression mapping models are used extensively in epidemiological studies of health impacts of air pollution exposure (Knibbs *et al.* 2014, 2018; Hoogh *et al.* 2016; Pereira *et al.* 2017; Dirgawati *et al.* 2019; Cowie *et al.* 2019).

The updating of a GIS regression model can be conceptualised in five ways. These distinguish between updates that:

1. create predictions at new locations, using the same regression equation and predictor data
2. predict at the same locations using data for new time periods but use the same regression equation
3. use data for new time periods to develop a new regression equation but keep the same set of candidate predictors
4. use new data and new equations developed with additional candidate predictor variables
5. create predictive models for a new pollutant using the same procedures.

The scripted workflow that we developed provides a set of software tools and a set of method steps that guide a user through the development of a GIS regression mapping model. Depending on the amount of data available models can be updated relatively quickly. In addition the updated models and data are well documented and reproducible.

Primarily we focus on the type of updates that apply a previously developed model regression equation to 1) a new set of locations within the same study area and 2) a new time point. New locations can be defined by the user or can be based on a dataset containing regularly located points (i.e. a grid) or locations of the participants in a epidemiological study (i.e. exposure estimates). We aim to automate most of the process to reduce required efforts from the user.

# 3 Methods

## 3.1 Implementation of procedures from Knibbs et al. 2014

We developed a case study based on the methodology used by Knibbs *et al.* (2014). In that paper the authors developed a regression model that estimated the concentrations of the air pollutant nitrogen dioxide (NO2). We applied the methods of that study exactly so that the air pollution could be estimated for any new set of points of interest. An example is the locations of study locations for a health impact assessment.

The codes can be repurposed to facilitate new air pollution models easily by making minor modifications.

## 3.2 Software

We used the PostgreSQL / PostGIS Geographical Information Systems (GIS) database server:

- PostgreSQL www.postgresql.org
- PostGIS http://postgis.refractions.net

We implemented this on a PostGIS database (version 2.3 running PostgreSQL 9.6.15 on x86_64-pc-linux-gnu (Ubuntu 9.6.15-1.pgdg16.04+1), compiled by gcc (Ubuntu 5.4.0-6ubuntu1~16.04.11) 5.4.0 20160609, 64-bit) and did the analysis with R functions and SQL scripts.

## 3.3 Workflow

Workflow orchestration and statistical analysis was implemented in R:

- R language for statistical computing www.r-project.org
- and various R packages installed from the CRAN public repository

We have set up a PostGIS database incorporating the required data, and along with the R scripts and readme in the open source software repository https://github.com/cardat/GIS_regression_mapping_scripted_workflow. In the 'code' folder can be found the codes to run the prediction analysis.

The "main.R" script in the project directory is set up so that the user can declare specific inputs required for the specific update, and then run the component steps in sequence.

We also include some helper functions written to enable user authentication on the database and database management.

### 3.4 Data sources

The data sources used for the code in this study are outlined in Table 1.

Table 1: Data sources

| Variable | Source |
|---|---|
| OMI tropospheric NO2 column density (molecules × 1015 / cm2) | Aura OMI level-3 NO2 product via NASA Giovani interface. http://gdata1.sci.gsfc.nasa.gov/daac-bin/G3/gui.cgi?instance_id=omi Acker and Leptoukh (2007). |
| impervious surfaces | NOAA constructed impervious surface area product 2000-2001. http://ngdc.noaa.gov/eog/dmsp/download_global_isa.html. Elvidge et al. (2007) |
| major roads (km)b | PSMA Australia Transport and Topography product** http://www.psma.com.au/?product=transport-topography PSMA (2013) |
| non-vehicle point source NOX emissions (kg/yr)e,f | Australia National Pollutant Inventory 2008/9 http://www.npi.gov.au/ |
| Land use (open space, Industrial) | Australian Bureau of Statistics Mesh Blocks http://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Geography |

# 4 Results

The resulting prediction datasets can be exported to a variety of GIS formats including ESRI Shapefiles and GeoTIFF raster files Figure 1.

### 4.1 Current case study application

The scripted workflow has been developed through the Australian National Health and Medical Research Council (NHMRC) centre of research excellence Centre for Air pollution, energy and health Research (CAR). CAR is a collaboration between more than 30 researchers from around Australia and internationally in diverse but related disciplines who are at the forefront of their scientific fields. The Centre is based in eight of Australia's most prestigious research institutions. The Centre provides an example of a current application for this scripted air pollution prediction models.

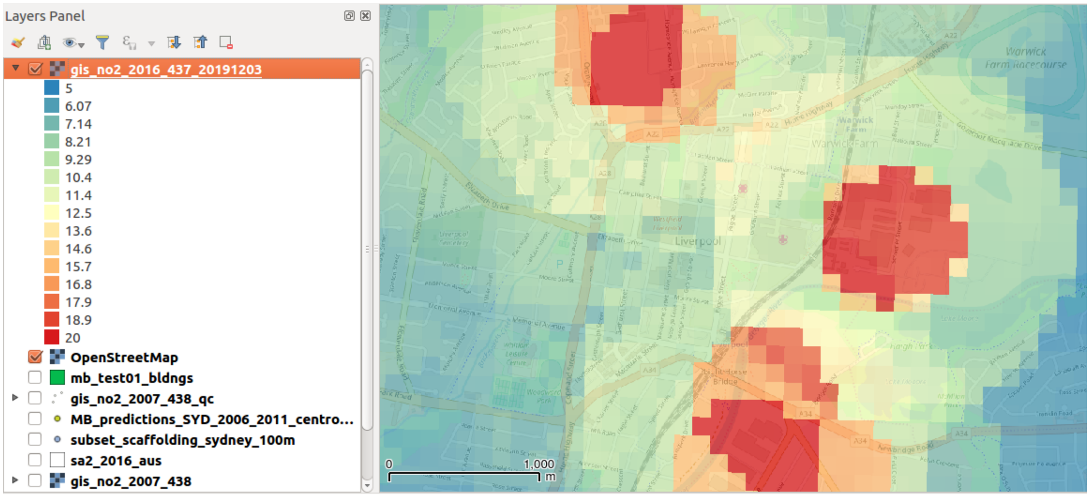


Figure 1: Western Sydney Liverpool case study 2016 NO2 predicted at a grid of evenly spaced points 100m by 100m and stored as a GeoTIFF file

## 5 Discussion

Our scripted workflow tools are aimed at GIS specialists and uses the PostGIS and R software environments which are technically sophisticated computing tools. These can be complicated to install and configure so we also provide an online cloud-based virtual desktop environment with these tools pre-configured so that users can develop the updates without having to install the tools on their local computers (https://cardat.github.io/). All the software is published as free and open source software on Github (https://github.com/ivanhanigan/GIS_regression_mapping_scripted_workflow). Some of the data inputs require additional data provision agreements to be made between data owners and data users.

TODO we should describe how an extension case study would gather new data and insert into the database.

## 6 Conclusion

TODO

## 7 Acknowledgments

The work by LK, IH and CG was funded by the Clean Air and Urban Landscapes (CAUL) hub of the Australian Government's National and Environmental Science Programme (NESP). TODO Acknowledge support from Prof Jane Heyworth and Prof Geoff Morgan. Financial support was also provided by the Centre for Air pollution, energy and health Research (CAR, http://www.car-cre.org.au) an Australian National Health and Medical Research Council (NHMRC) Centre of Research Excellence; and the Air Pollution, Traffic Exposures and Mortality and Morbidity in Older Australians (APTEMA) Study.

This research is undertaken with the assistance of resources from the Collaborative Environment for Scholarly Analysis and Synthesis (CoESRA; https :// coesr a.tern.org.au).

## References

Briggs, D.J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K. & Van Der Veen, A. (1997). Mapping urban air pollution using gis: A regression-based approach. *International Journal of Geographical Information Science*, 11(7),: 699–718.

Cowie, C.T., Garden, F., Jegasothy, E., Knibbs, L.D., Hanigan, I., Morley, D., Hansell, A., Hoek, G. & Marks, G.B. (2019). Comparison of model estimates from an intra-city land use regression model with a national satellite-LUR and a regional Bayesian Maximum Entropy model, in estimating NO2 for a birth cohort in Sydney, Australia. *Environmental Research*, 174: 24–34.

Dirgawati, M., Hinwood, A., Nedkoff, L., Hankey, G.J., Yeap, B.B., Flicker, L., Nieuwenhuijsen, M., Brunekreef, B. & Heyworth, J. (2019). Long-term Exposure to Low Air Pollutant Concentrations and the Relationship with All-Cause Mortality and Stroke in Older Men. *Epidemiology (Cambridge, Mass)*, 30(July),: S82–S89.

Hoogh, K. de, Gulliver, J., Donkelaar, A. van, Martin, R.V., Marshall, J.D., Bechle, M.J., Cesaroni, G., Pradas, M.C., Dedele, A., Eeftens, M., Forsberg, B., Galassi, C., Heinrich, J., Hoffmann, B., Jacquemin, B., Katsouyanni, K., Korek, M., Künzli, N., Lindley, S.J., Lepeule, J., Meleux, F., Nazelle, A. de, Nieuwenhuijsen, M., Nystad, W., Raaschou-Nielsen, O., Peters, A., Peuch, V.H., Rouil, L., Udvardy, O., Slama, R., Stempfelet, M., Stephanou, E.G., Tsai, M.Y., Yli-Tuomi, T., Weinmayr, G., Brunekreef, B., Vienneau, D. & Hoek, G. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research*, 151(2),: 1–10.

Knibbs, L.D., Hewson, M.G., Bechle, M.J., Marshall, J.D. & Barnett, A.G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental Research* (Epub ahead of print 2014). doi: 10.1016/j.envres.2014.09.011

Knibbs, L.D., Van Donkelaar, A., Martin, R.V., Bechle, M.J., Brauer, M., Cohen, D.D., Cowie, C.T., Dirgawati, M., Guo, Y., Hanigan, I.C., Johnston, F.H., Marks, G.B., Marshall, J.D., Pereira, G., Jalaludin, B., Heyworth, J.S., Morgan, G.G. & Barnett, A.G. (2018). Satellite-Based Land-Use Regression for Continental-Scale Long-Term Ambient PM 2.5 Exposure Assessment in Australia. *Environmental Science and Technology*, 52(21),: 12445–12455.

Pereira, G., Lee, H.J., Bell, M., Regan, A., Malacova, E., Mullins, B. & Knibbs, L.D. (2017). Development of a model for particulate matter pollution in Australia with implications for other satellite-based models. *Environmental Research*, 159(July),: 9–15.