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

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

[1] *The University of Queensland, School of Public Health*

[2] *Centre for Air pollution, energy and health Research (CAR)*

[3] *The University of Sydney, University Centre for Rural Health, School of Public Health*

[4] *The University of Canberra, Centre for Research and Action in Public Health*

[*] *Corresponding author: l.knibbs@uq.edu.au*

1 Summary

1.1 Introduction

In this study we aimed to develop a scripted workflow to update geographical information system (GIS) regression mapping models that use land use and other predictors. GIS regression mapping was first introduced by Briggs *et al.* (1997) and is a technique commonly used to create predictive spatial models (and spatiotemporal models) of pollutant levels. It is also known as land use regression (LUR), however, the label LUR does not reflect the fact that non-land use predictors are often used such as chemical transport models or remote sensing data (e.g. satellite aerosol optical depth). These models add more information than just land use data to the model as predictor variables. 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).

1.2 Purpose of the software

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.

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.

We developed a case study based on the methodology used by Knibbs *et al.* (2014). We also provide links to the protocols developed by the European Study of Cohorts for Air Pollution Effects (ESCAPE) (Eeftens *et al.* 2012; Beelen *et al.* 2013) as set out in the ESCAPE Exposure assessment manual (2010). Within air pollution research the ESCAPE methodology is used as the standard for developing GIS regression mapping models.

1.3 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.

1.4 Software implementation definitions

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

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

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

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. These are:

1. "01_test_connection_to_PostGIS.R"

2. “02_buffers.R”
3. “03_extract_OMI.R”
4. “04_overlay_IMPSA_with_1200M.R”
5. “05_majrds_intersect_with_500m_buffer.R”
6. “06_NPINOX_extract.R”
7. “07_industrial_or_open_in_buffer.R”

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

1. “function_get_pgpass.R”
2. “function_pgListTables.R”

1.5 Data inputs

TODO.

In the first case we should describe the data used for the Knibbs 2014 NO2 demo. Some of these are restricted access and we will need to discuss data provision agreements.

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

2 Acknowledgments

This project was funded by the Clean Air and Urban Landscapes (CAUL) hub of the Australian Government’s National and Environmental Science Programme (NESP). 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

Beelen, R., Hoek, G., Vienneau, D., Eeftens, M., Dimakopoulou, K., Pedeli, X., Tsai, M.Y., Kunzli, N., Schikowski, T., Marcon, A., Eriksen, K.T., Raaschou-Nielsen, O., Stephanou, E., Patelarou, E., Lanki, T., Yli-Tuomi, T., Declercq, C., Falq, G., Stempfelet, M., Birk, M., Cyrys, J., Klot, S. von, Nador, G., Varro, M.J., Dedele, A., Grazuleviciene, R., Molter, A., Lindley, S., Madsen, C., Cesaroni, G., Ranzi, A., Badaloni, C., Hoffmann, B., Nonnemacher, M., Kraemer, U., Kuhlbusch, T., Cirach, M., Nazelle, A. de, Nieuwenhuijsen, M., Bellander, T., Korek, M., Olsson, D., Stromgren, M., Dons, E., Jerrett, M., Fischer, P., Wang, M., Brunekreef, B. & Hoogh, K. de. (2013). Development of no2 and nox land use

regression models for estimating air pollution exposure in 36 study areas in europe - the escape project. *Atmospheric Environment*, 72: 10–23.

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.

Eeftens, M., Beelen, R., Hoogh, K. de, Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dedele, A., Dons, E., Nazelle, A. de, Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Grazuleviciene, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Molter, A., Nador, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.Y., Yli-Tuomi, T., Varro, M.J., Vienneau, D., Klot, S., Wolf, K., Brunekreef, B. & Hoek, G. (2012). Development of land use regression models for pm(2.5), pm(2.5) absorbance, pm(10) and pm(coarse) in 20 european study areas; results of the escape project. *Environ Sci Technol*, 46(20),: 11195–205.

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.

(2010). http://www.escapeproject.eu/manuals/ESCAPE_Exposure-manualv9.pdf