International Journal of Population Data Science

Journal Website: www.ijpds.org

Cognitive development Respiratory Tract Illness and Effects of eXposure (CORTEX) project: Data processing challenges in combining high spatial resolution pollution level data with individual level health and education data

Lyons, J[1*], Mizen, A[1], Rodgers, S[1], Berridge, D[1], Akbari, A[1], Wilkinson, P[2], Milojevic, A[2], Doherty, R[3], Dearden, L[4], Lake, I[5], Carruthers, D[6], Strickland, S[6], Mavrogianni, A[7], and Davies, G[1]

[1]Swansea Unviersity
[2]The London School of Health and Tropical Medicine
[3]The University of Edinburgh
[4]Institute of Fiscal Studies
[5]University of East Anglia
[6]CERC
[7]University College London

Background and Objectives

There is a lack of evidence of the adverse effects of air pollution and pollen on cognition for people with air quality-related health conditions. The CORTEX project combined routinely collected health and education data, high spatial resolution air pollution modelling, and daily pollen measurements for 18,241 pupils living in Cardiff, UK, between 2009 and 2015, to investigate the acute effects of air quality and respiratory conditions on education attainment.

Datasets

Air pollutants PM2.5, PM10, NO2, and ozone levels were modelled for 157,361 home and school locations, anonymised into the Secure Anonymised Information Linkage (SAIL) Databank, and summarised into minimum, average and maximum readings for 4 daily time periods reflecting pupil home/school exposure. Adding a unique Residential Anonymised Linking Field (RALF) allowed linkage of pollution estimates to individual level data. Annual pollution datasets contained 369 columns and 472,083-rows, with one column per location, pollutant, daily time-period and day of year. Dataset transformation produced a 5 column, 3,446,205,900-row matrix per year.

Methods and Conclusions

An algorithm using Structured Query Language (SQL) to manage data held within a relational database management system, was designed to reduce dimensionality from 24 billion to 18,241 rows of data. The algorithm calculated average means for each pollutant (PM2.5, PM10, NO2, and ozone levels) over the revision and examination periods, and summarised data into one row per pupil. The algorithm adjusted for weekends, school, and bank holidays, it calculated daily pollutant exposure for each pupil, and successfully linked 95% of pupil pollution exposures to their health and education data.

*Corresponding Author:
Email Address: j.lyons@swansea.ac.uk (L Lyons)

http://dx.doi.org/10.23889/ijpds.v3i2.534