Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile

Quinteros, María Elisa
Lu, Siyao
Blazquez, Carola
Cárdenas-R, Juan Pablo
Ossa, Ximena
Delgado-Saborit, Juana María
Harrison, Roy M.
Ruiz-Rudolph, Pablo

© 2018 Elsevier Ltd

Missing data from air quality datasets is a common problem, but is much more severe in small cities or localities. This poses a great challenge for environmental epidemiology as high exposures to pollutants worldwide occur in these settings and gaps in datasets hinder health studies that could later inform local and international policies. Here, we propose the use of imputation methods as a tool to reconstruct air quality datasets and have applied this approach to an air quality dataset in Temuco, a mid-size city in Chile as a case-study. We attempted to reconstruct the database comparing five approaches: mean imputation, conditional mean imputation, K-Nearest Neighbor imputation, multiple imputation and Bayesian Principal Component Analysis imputation. As a base for the imputation methods, linear regression models were fitted for PM2.5 against other air quality and meteorological variables. Methods were challenged against validation sets where data was removed artif