Spatio-Temporal Multivariate Imputation of Missing Air Quality Data

Dr. Şebnem Er

University of Cape Town
Statistical Sciences Department
Sebnem.Er@uct.ac.za

October 6, 2018

www.igs2018.mu.edu.tr

Introduction

- Why air pollution (quality) data?

- Why air pollution (quality) data?
  - Well-recognized adverse effect on health of individuals
  - Several short and long-term effects such as lung and cardiovascular problems amongst people (if high levels).
  - Consequently an adverse effect on the economy from a public health point of view
  - As a result, air pollution has become an increasingly concerning global problem

- Why air pollution (quality) data?
  - Well-recognized adverse effect on health of individuals
  - Several short and long-term effects such as lung and cardiovascular problems amongst people (if high levels).
  - Consequently an adverse effect on the economy from a public health point of view
  - As a result, air pollution has become an increasingly concerning global problem
- Very little has been done on the effects of outdoor air pollution in SA. Mainly due to the fact that malfunctions and communication errors cause usually large amounts of missing data

- Why air pollution (quality) data?
  - Well-recognized adverse effect on health of individuals
  - Several short and long-term effects such as lung and cardiovascular problems amongst people (if high levels).
  - Consequently an adverse effect on the economy from a public health point of view
  - As a result, air pollution has become an increasingly concerning global problem
- Very little has been done on the effects of outdoor air pollution in SA. Mainly due to the fact that malfunctions and communication errors cause usually large amounts of missing data
- The aim of the research is to determine which of the imputation methods will be most appropriate to use for the data (air quality) to have minimal error when the data is modelled.

Outline

- Section 2 describes the data
- Section 3 describes the imputation methods used in time series and space-time series.
- Section 4 presents the imputation results and discussion.

Software used: R

- imputeTS
- SpatioTemporal
- raster
- mapview
- maps

The Data

Variable: Nitrogen Oxides

- NO_x is a generic term for the nitrogen oxides that are most relevant for air pollution namely

Variable: Nitrogen Oxides

- NO_x is a generic term for the nitrogen oxides that are most relevant for air pollution namely
  - NO: nitric oxide
  - NO_2: nitrogen dioxide

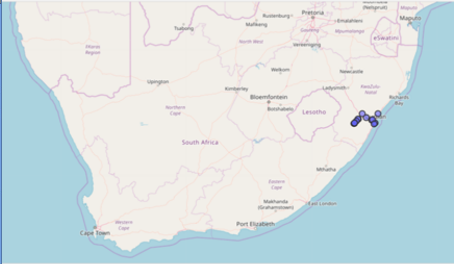
Variable: Nitrogen Oxides

- NO_x is a generic term for the nitrogen oxides that are most relevant for air pollution namely
  - NO: nitric oxide
  - NO_2: nitrogen dioxide

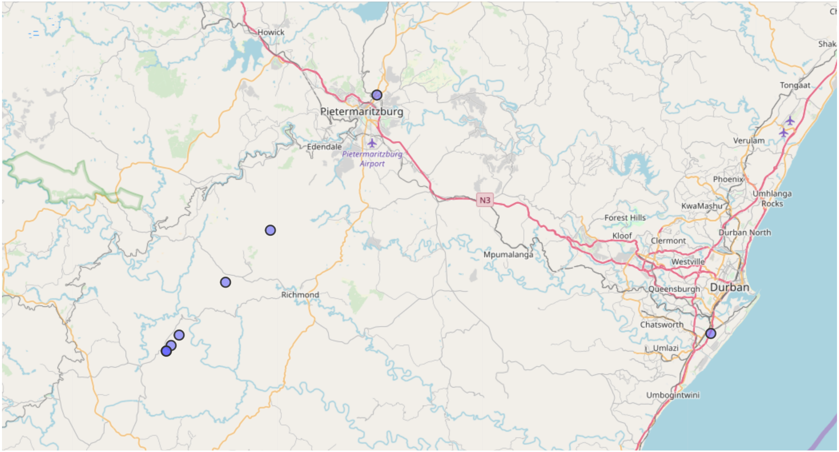
These gases contribute to the formation of smog and acid rain, as well as affecting tropospheric ozone in the formation of fine particles (PM - particulate matter) and ground level ozone, both of which are associated with adverse health effects.

South Africa and AQS

South Africa and AQS



AQS Stations Used for NOX imputation - 9 Stations



Data and Missing Value Patterns

Variable	Station	Data Missing
NO	City Hall	16.34%
NO	Ferndale	49.54%
NO	Ganges	10.2%
NO	Jacobs AQ	23.88%
NO	Southern Works 1	12.94%
NOX	Southern Works 2	12.94%
NO2	Southern Works 3	12.94%
NOX	Warwick Reservoir	13.33%
NOX	Wentworth Reservoir	11.59%

Data is observed between 2004-1-1 and 2011-01-31 on an hourly basis.

Data is observed between 2004-1-1 and 2011-01-31 on an hourly basis.

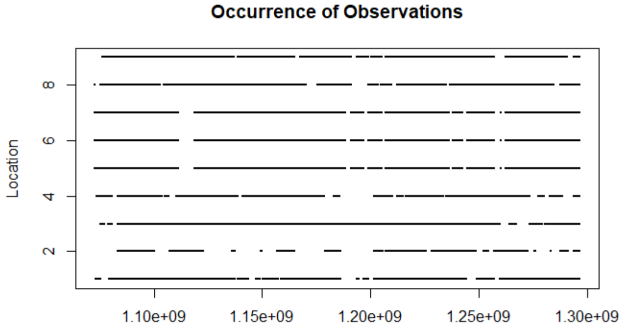
The data has been aggregated to daily averages.

Data is observed between 2004-1-1 and 2011-01-31 on an hourly basis.

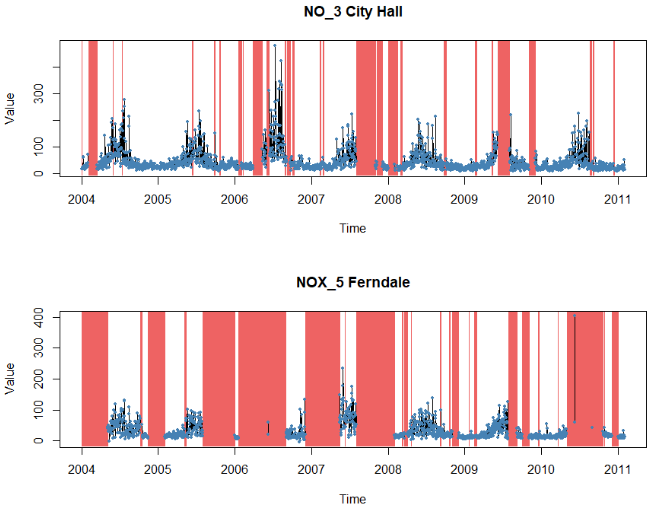
The data has been aggregated to daily averages.

Geographic covariates are utilized such as the distance to Durban.

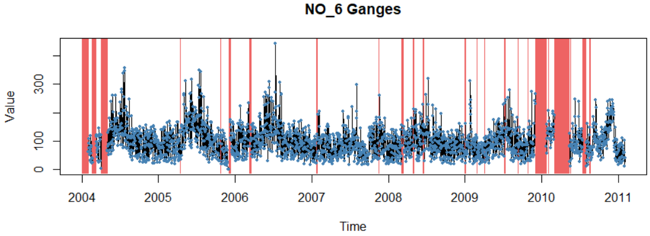
Occurrance of Observations

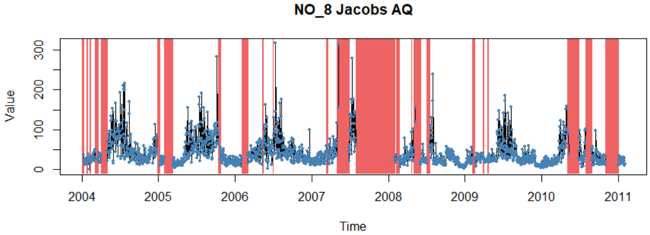


City Hall and Ferndale

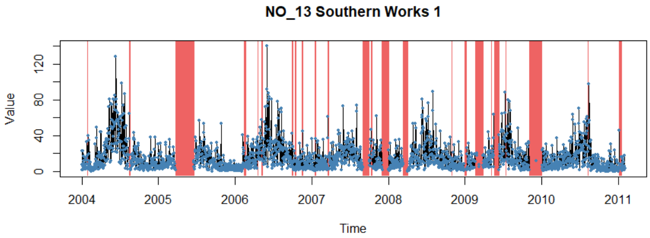


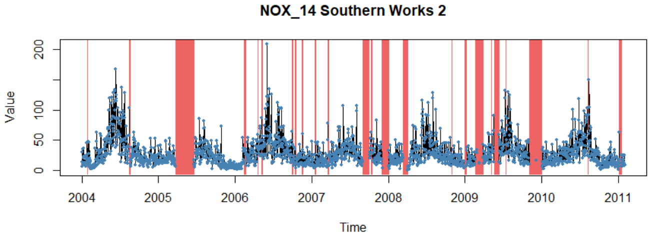
Ganges and Jacobs



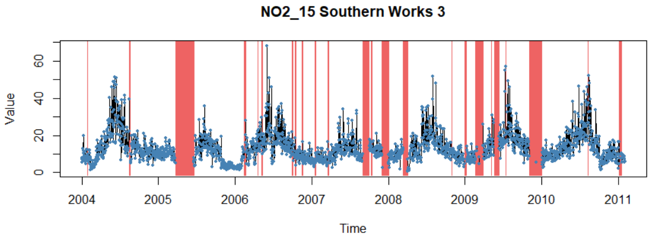


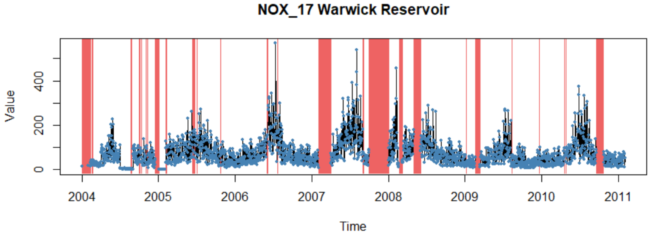
Southern Works 1 - 2



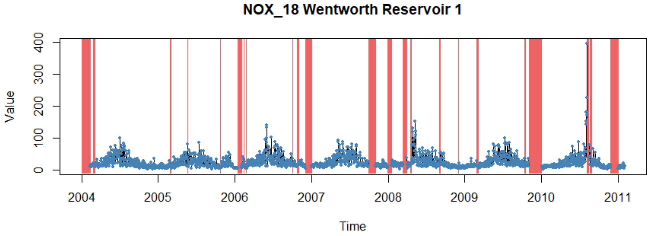


Southern Works 3 - Warwick Res





Wentworth Res



Imputation Techniques

Overview

- Imputation is a general and flexible method for handling missing-data problems. However, it has pitfalls. In the words of Dempster and Rubin (1983):

Overview

- Imputation is a general and flexible method for handling missing-data problems. However, it has pitfalls. In the words of Dempster and Rubin (1983):

  "*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor.*"

Overview

- Imputation is a general and flexible method for handling missing-data problems. However, it has pitfalls. In the words of Dempster and Rubin (1983):

  "*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor.*"

- Several methods have been proposed for air quality data.
  - Single imputation - one value per missing sample
  - Multiple imputation - multiple values per missing sample

Overview

- Imputation is a general and flexible method for handling missing-data problems. However, it has pitfalls. In the words of Dempster and Rubin (1983):

  "*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor.*"

- Several methods have been proposed for air quality data.
  - Single imputation - one value per missing sample
  - Multiple imputation - multiple values per missing sample
- Here the focus will be on single imputation.

Single imputation methods for time series

- Random value imputation
- Mean imputation
- Moving average imputation
- Last observation carried forward / Next observation carried backward
- Kalman smoothing imputation
- Interpolation by splines or linear imputation
- Seasonally adjusted linear imputation

For time series with a strong seasonality and trend, usually seasonally adjusted linear imputation results will yield the best results (least RSME).

Mean-Before-After Imputation

Let $y_1, y_2, \ldots, y_n, y_1^*, y_{n+1}, y_{n+2}, \ldots, y_{n_2}, y_2^*, \ldots, y_k^*, y_n$ be the data with y^* being the missing values.

The mean-before-after method replaces missing values with the mean of one date before and after the missing sample. Thus, here y_1^* will be replaced with

$$\bar{y}_1 = \frac{y_{n_1} + y_{n_1+1}}{2}$$

Mean-Before Imputation

The mean-before method replaces missing values with the mean of all data available before the missing sample. Thus, here y_1^* will be replaced with

$$\bar{y}_1 = \frac{y_{n_1} + y_{n_1+1}}{2}$$

y_2^* will be replaced with

$$\bar{y}_2 = \frac{1}{n_2 - n_1 - 1} \sum_{i=n_1+1}^{n_2} y_i$$

Linear Interpolation

In linear interpolation two data points are connected with a straight line and all missing values are imputed with data points along the line.

$$y = y_1 + k(x + x_1)$$

where $k = (y_2 - y_1)/(x_2 - x_1)$ and $x_1 < x < x_2$ and $y_1 < y < y_2$.

Univariate Nearest Neighbour Imputation

With nearest neighbour imputation, the endpoints of the missing samples are used as estimates for all the missing values. The equation is as follows with (y_1, x_1) the coordinates of the starting data points of the missingness and (y_2, x_2) the coordinates of the end data points of the missingness.

$$y = \begin{cases} y_1 & \textit{if} \quad x \leq x_1 + \frac{(x_2 - x_1)}{2} \\ y_2 & \textit{if} \quad x > x_1 + \frac{(x_2 - x_1)}{2} \end{cases}$$

Machine Learning Algorithms

- Self organising maps (clustering)
- K-Nearest Neighbours (clustering)
- Decision Trees (classification)
- Bayesion Networks

Inverse Distance Weighting Method

The inverse distance weighting method imputes missing data for one station using the weighted average of the values measured in the neighbours. The weights are the inverse distance matrix so that the values measured in the nearest stations will have a greater influence on the station of interest than those measured further away.

Spatio Temporal Imputation Methods

Quantity to be modelled, NO_x values at location s and time t, is composed of two parts:

$$Z(s,t) = \mu(s,t) + \epsilon(s,t)$$

mean field and the random space-time residual field.

Mean field

The mean field is modelled as follows:

$$\mu(s,t) = \sum_{l=1}^{L} \gamma_l \mathcal{M}_l(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t)$$

where $\mathcal{M}_l(s,t)$ are the spatio-temporal covariates, γ_l are the coefficients for the spatio-temporal covariates;

Mean field

The mean field is modelled as follows:

$$\mu(s,t) = \sum_{l=1}^{L} \gamma_l \mathcal{M}_l(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t)$$

where $\mathcal{M}_l(s,t)$ are the spatio-temporal covariates, γ_l are the coefficients for the spatio-temporal covariates;

m is the number of temporal functions, $\{f_i(t)\}_{i=1}^{m}$ set of smooth temporal functions, $\beta_i(s)$ spatially varying coefficients for the temporal functions.

Smooth Temporal Functions

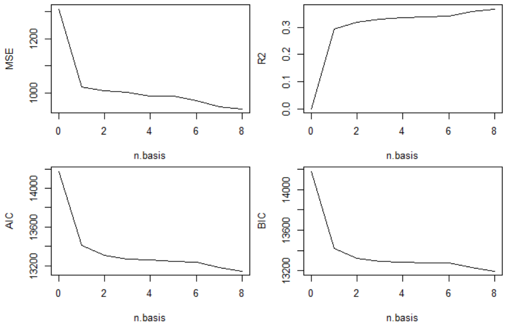
- The objective of these functions is to capture the temporal variability in the data, indicating that the residual space-time field, $v(s, t)$ are independent in time (with stationary, parametric spatial covariance).

Smooth Temporal Functions

- The objective of these functions is to capture the temporal variability in the data, indicating that the residual space-time field, $v(s, t)$ are independent in time (with stationary, parametric spatial covariance).
- Many air quality parameters display a dominant "seasonal" trend structure.
- How many? In order to determine the number of temporal functions that capture the temporal variability in the data, cross validation is used.
- We choose the number of temporal functions that minimises the MSE and maximises the R^2. For details see (Fuentes et all. 2006).

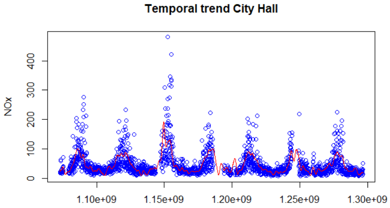
Results and Discussions

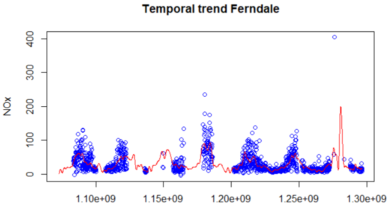
Number of Temporal Functions

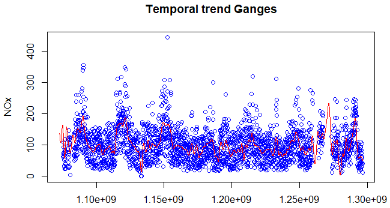


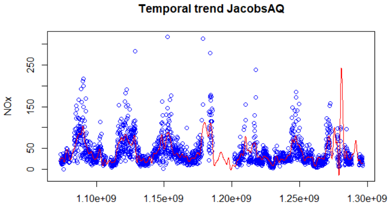
All four statistics flatten out after 4 basis functions, indicating that 4 basis functions is likely to provide the most efficient description of the temporal variability.

Data Driven Temporal Functions

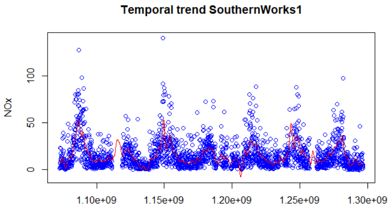


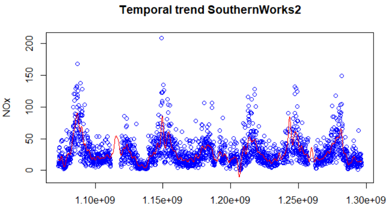


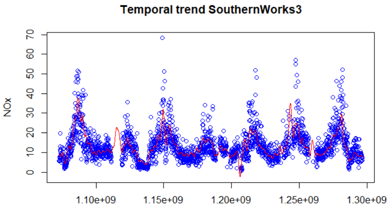


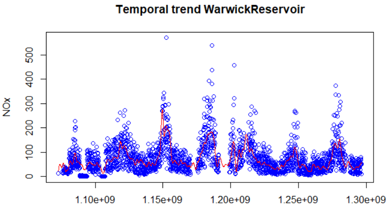


Data Driven Temporal Functions

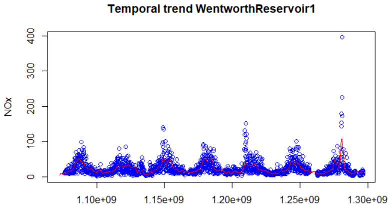








Data Driven Temporal Functions



- Transformation of the data is not considered, variance effects!
- To evaluate the methods and to compare their performance, a simulation study based on different missing data patterns will be conducted.
- Multiple imputation methods will be investigated.

- Teşekkürler....

References

References

https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf

Moritz, S. Beielstein T. imputeTS: Time Series Missing Value Imputation in R.

Little R., Rubin D. (2002) Statistical analysis with missing data, 2nd ed, Wiley, New York.

Norazian et all. (2008) Estimation of missing values in air pollution data using single imputation techniques. Science Asia, 34:341-345.

Liu, Y., Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. Data, 2,8. pp.1-15

Junninen et all. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 38, pp.2895-2907.

Plaia, A., Bondi, A.L. (2006). Imputation of missing values in air quality datasets

References

Junger, W.L., Leon, A.P. (2015). Imputation of missing data in time series for air pollutants. Atmospheric Environment, 102, pp.96-104.

Mwale, F.D. et all. (2012). Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi A self organizing map approach. Physics and Chemistry of the Earth. 50-52, pp.34-43.

Bergen, S., Lindstrom, J. (2018). Comprehensive tutorial for the spatio-temporal R-package.

Fuentes M, Guttorp P, Sampson PD (2006). Using transforms to analyze space-time processes. In B Finkenstadt, L Held, V Isham (eds.), Statistical Methods for Spatio-Temporal Systems, pp. 77-150. CRC-Chapman and Hall.