Imputation of missing values in air quality data sets (⋆)

Stima di dati mancanti in dataset sulla qualità dell'aria

Antonella Plaia and Anna Lisa Bondì
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo
e-mail: plaia@unipa.it

Riassunto: Uno dei principali problemi che ci si presentano quando si studiano dati longitudinali riguarda la presenza di dati mancanti. Negli studi in ambito ambientale ciò può essere dovuto a errori di misura o a mancata acquisizione del dato. Due sono le possibili strade da percorrere: utilizzare una metodologia statistica che tenga espressamente conto della presenza di dati mancanti, ovvero rendere il dataset "completo" stimando i dati mancanti. Il presente lavoro intende seguire questo secondo approccio, proponendo un metodo per l'imputazione di dati mancanti da utilizzare in un contesto multilivello. Il comportamento del metodo verrà confrontato con quello di metodi presenti in letteratura in presenza di diversi schemi (simulati) di dati mancanti.

Keywords: missing data imputation, multilevel data.

1. Introduction

Missing data is a very frequent problem in environmental research, usually due to faults in data acquisition. Referring to the standard classification of missingness (Little and Rubin, 1987), we can cope with data: (1) missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Usually, the missing data mechanism of air quality data is MAR, as, being due to the monitoring site down, the probability that a value is missing does not depend on the missing value (as it would be, for example, if the monitoring site could not measure values below a given threshold).

Occurring missing values, a possibility is to use statistical methods that take the missing data into account during the analysis, otherwise we can discard units whose information is incomplete, considering a *listwise deletion* (that is the complete discard of the unit), or a *pairwise deletion* (where units are excluded from any calculations involving variables for which they have missing data). A different approach consists in inputing missing values, by single or multiple imputation. In single imputation each missing value is substituted by an inputed one, while, with multiple imputation, multiple simulated values are generated for each missing value in order to consider the uncertainty attached to missing data.

2. Imputing missing values in multilevel data

Air quality data sets consist of pollutant concentrations measured, for example on a time-scale of one per hour (or one per two hours), by monitoring stations distributed over a certain area. These longitudinal data show a multilevel structure (the term *multilevel*

(⋆) Work partially supported by a University of Palermo grant (ex 60%)

must be referred to the data structure and not to the methodology to analyze them, at least at the moment) with monitoring sites as second level units and single measure as first level units. Missing values are usually due to monitoring station being down, and this can cause a gap in the data a single value to many consecutive values long. Many single imputation methods have been proposed to estimate missing data, going from simple mean substitution to regression based imputation and neural network (Junninen et al 2004, Xia et alt. 1999, Twisk and de Vente 2002). While, in order to impute the missing value of a meteorological variable in a monitoring site, a multiple regression model which assumes, as covariates, other meteorological variables measured in the same site, can be appropriate, with pollution data we suppose to be more efficient an imputation based on the same pollution variable, measured in the same as well as in other monitoring sites. In the present paper we propose a new imputation method that considers the peculiarity of pollution data and therefore site-specific effects, for example week, week-day or day-hour site specific effects. In order to illustrate it and compare its performance with other methods proposed in literature, we will refer to a data set, illustrated in Table 1[(1)], consisting of PM_{10} concentration levels measured every two hours in eight monitoring stations in Palermo in 2003. The actual percentage of missing data goes from 3.5% of Station 3 to 13% of Station 2. About 93% of the sequences of missing values has a length of 3 or smaller, 5% has a length between 3 and 12 , and only 2% are longer that 12 (that is longer than one day): in particular we found only three sequences with a length between 57 and 67 in a single station.

Table 1: *Data set structure.*

W	W-D	H	St1	St2	St3	St4	St5	St6	St7	St8
3	1	2	23.32	12.47	108.02	30.46	NA	4.29	45.05	50.91
3	1	4	34.68	27.09	28.17	9.34	20.10	26.65	NA	26.36
3	1	6	26.86	11.10	34.70	24.17	29.76	21.94	NA	44.46
3	1	8	19.81	24.62	31.51	27.23	34.96	26.20	NA	45.89
3	1	10	24.38	14.84	23.91	18.50	15.21	20.10	20.94	NA
3	1	12	22.48	16.46	30.72	38.54	35.87	16.76	15.20	NA
...	...	...	...	...	...	...	...	...	...	...

Denoting the generic element of the dataset as x_{swdh}, where s refers to the monitoring site ($s = 1, 2, ...S$), w to the Week ($w = 1, 2, ..., 53$), d to the Week-Day ($d = 1, 2, ...7$) and h to the Hour ($h = 2, 4, ...24$), we will propose to use a Site-Dependent Effect Method (SDEM) that considers explicitly a week effect, a day effect and an hour effect (all site-dependent), that is to input a missing value as:

$$\hat{x}_{swdh} = \bar{x}_{\cdot wdh} + (\bar{x}_{sw\cdot\cdot} - \sum_{s=1}^{S} \frac{\bar{x}_{sw\cdot\cdot}}{S}) + (\bar{x}_{s\cdot d\cdot} - \sum_{s=1}^{S} \frac{\bar{x}_{s\cdot d\cdot}}{S}) + (\bar{x}_{s\cdot\cdot h} - \sum_{s=1}^{S} \frac{\bar{x}_{s\cdot\cdot h}}{S}) \quad (1)$$

The performance of this method will be compared to other methods in literature, and in particular to $\bar{x}_{s\cdot\cdot h}$ (Li et alt., 1999), and to *Last & Next* (Engels and Diehr, 2003).

(1) W = Week, W-D = Day of the week, H = Hour of the Day

3. Simulation of missing data and performance indicators

The imputation method performance, besides depending on the amount of missing data, is influenced by the characteristics of the missing data patterns. In other words, we expect the performance to change accordingly, most of all, to the length of the gaps. Referring to Table 1, we will consider 4 different missing data patterns that differ for the total percentage of missing data in the table, for the distribution of the gap length, and for the maximum number of missing values per row. Two different total amount of missing data have been considered: about 5% and about 15%. Two maximum number of missing values per row have been considered, 4 and 8. For each of the four missing data patterns 100 missing data indicator matrix M (that applied to the observed dataset creates "artificially" missing values) have been generated drawing *the gap length* from a mixture of two distributions, an Exponential of parameter $\lambda = 0.5$ (that produces the short gaps) and a Uniform with parameters (20, 70) and (40, 120) for the 5% and 15% missing data patterns respectively (to produce the long gaps), *the starting point* of each gap from a Uniform (1, 4380) and *the number of missing data per row* (in Table 1) from a Uniform (0,4) or (0,8).

The goodness of imputation has been evaluated both by the coefficient of correlation r between imputed (I_i) and observed (O_i) values, and by an index of agreement, d (Junninen et alt., 2004), as r may be unrelated to the sizes of the discrepancies between observed and imputed values:

$$d = 1 - \left[\frac{\sum_{i=1}^{N} (I_i - O_i)^2}{\sum_{i=1}^{N} (\left|I_i - \bar{O}\right| + \left|O_i - \bar{O}\right|)^2} \right] . \tag{2}$$

where N is the number of imputed values.

4. Results and discussion

Each of the fourth missing data patterns have been simulated one hundred times, and Table 2 shows $\bar{d}$ and σ_d for each monitoring site and each data pattern.

The results confirm the importance of considering site-specific effects. As Figure 1 shows, the Hour-Mean (H-M) does not perform well, both considering r and d, most of all because it does not consider the seasonality (week) and the weekend effect. The *Last & Next* performance is influenced by the gap length, as for the third and forth pattern of missing values (which presents gaps up to ten days long) both r and d are lower than in the other two patterns. SDEM appears not to be influenced by gap length and shows both a coefficient of correlation and an index of agreement always higher than *Last & Next* (except for Station 2 where d is lower). The different behavior of the method in Station 2 is probably due to the different location of this monitoring site: background urban area. The row-wise missing pattern (that is the number of sites with missing values at the same time) does not influence method performance, as shown both by Tab. 2 and by Fig. 1.

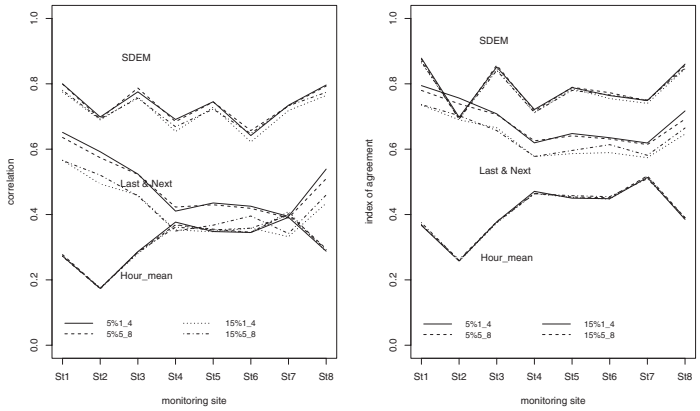
References

Engels J. M., Diehr P.. (2003) Imputation of missing longitudinal data: a comparison of methods, *Journal of Clinical Epidemiology*, 56, 968–976.

Table 2: *Comparison of methods versus missing data patterns ($\bar{d} \pm \sigma_d$)*

	St1	St2	St3	St4	St5	St6	St7	St8
Missing data percentage of 5								
1 to 4 monitoring sites with simultaneously missing values								
H-M	0.37 ± 0.04	0.26 ± 0.03	0.38 ± 0.03	0.47 ± 0.07	0.45 ± 0.04	0.45 ± 0.07	0.51 ± 0.04	0.39 ± 0.03
L&N	0.79 ± 0.05	**0.76** ± 0.05	0.71 ± 0.05	0.62 ± 0.06	0.65 ± 0.06	0.64 ± 0.08	0.62 ± 0.05	0.72 ± 0.06
SDEM	**0.88** ± 0.02	0.70 ± 0.03	**0.85** ± 0.03	**0.72** ± 0.06	**0.79** ± 0.03	**0.76** ± 0.07	**0.75** ± 0.03	**0.86** ± 0.03
Up to 8 monitoring sites with simultaneously missing values								
H-M	0.37 ± 0.04	0.26 ± 0.03	0.38 ± 0.04	0.47 ± 0.08	0.45 ± 0.04	0.45 ± 0.06	0.52 ± 0.04	0.39 ± 0.04
L&N	0.78 ± 0.06	**0.74** ± 0.06	0.71 ± 0.06	0.63 ± 0.07	0.64 ± 0.06	0.63 ± 0.08	0.61 ± 0.06	0.69 ± 0.07
SDEM	**0.87** ± 0.03	0.69 ± 0.04	**0.86** ± 0.02	**0.72** ± 0.07	**0.79** ± 0.04	**0.77** ± 0.07	**0.75** ± 0.03	**0.86** ± 0.03
Missing data percentage of 15								
1 to 4 monitoring sites with simultaneously missing values								
H-M	0.37 ± 0.03	0.26 ± 0.02	0.37 ± 0.03	0.46 ± 0.06	0.45 ± 0.03	0.45 ± 0.05	0.52 ± 0.03	0.39 ± 0.03
L&N	0.73 ± 0.07	0.69 ± 0.07	0.67 ± 0.06	0.58 ± 0.06	0.59 ± 0.06	0.59 ± 0.07	0.57 ± 0.05	0.65 ± 0.06
SDEM	**0.86** ± 0.02	**0.70** ± 0.03	**0.84** ± 0.03	**0.71** ± 0.06	**0.79** ± 0.03	**0.75** ± 0.06	**0.74** ± 0.03	**0.85** ± 0.03
Up to 8 monitoring sites with simultaneously missing values								
H-M	0.38 ± 0.04	0.26 ± 0.03	0.38 ± 0.03	0.46 ± 0.07	0.46 ± 0.03	0.45 ± 0.05	0.51 ± 0.04	0.39 ± 0.03
L&N	0.74 ± 0.07	**0.70** ± 0.08	0.66 ± 0.08	0.58 ± 0.08	0.60 ± 0.07	0.61 ± 0.07	0.58 ± 0.07	0.66 ± 0.07
SDEM	**0.87** ± 0.02	0.69 ± 0.04	**0.84** ± 0.02	**0.71** ± 0.06	**0.78** ± 0.03	**0.77** ± 0.06	**0.75** ± 0.03	**0.85** ± 0.03

Figure 1: *Coefficient of correlation and index of agreement for the three methods and the four missing data patterns.*



Junninen H., Niska H., Tuppurainen K., Ruuskanen J, Kolehmainen M. (2004) Methods fo imputation of missing values in air quality data sets, *Atmospheric Environment*, 38, 2895–2907.

Li K. H., Le N. D.,Sun L., Zidek J. V. (1999) Spatial-temporal models for ambient hourly PM_{10} in Vancouver, *Environmetrics*, 10, 321–328.

Little R. J. A., Rubin D. B. (2002) *Statistical Analysis with Missing Data*, Wiley, New York.

Twisk J., de Vente W. (2002) Attrition in longitudinal studies: How to deal with missing data, *Journal of Clinical Epidemiology*, 55, 329–337.

Xia Y., Fabian P., Stohl A., Winterhalter M. (1999) Forest climatology: estimation of missing values for Bavaria, Germany, *Agricultural and Forest Meteorology*, 96, 131–144.