Using multiple imputation to treat missingness in air pollution time series

Michela Baccini
Dipartimento di Statistica, Informatica e Applicazioni, Università di Firenze
Unità di Biostatistica, ISPO, Firenze

joint work with

Simone Giannini, Andrea Ranzi, Paolo Lauriola
ARPA Emilia Romagna

Background

- The presence of missing values is common in air pollution time series.

- Complete case analysis:
  - is biased unless missingness is competely random (MCAR).
  - breaks the time structure of the data.

- Missingness in this context is usually faced by single imputation:

  Given H available monitors, if a specific measurement *p* from monitor *h* is missing in a certain day *i* (X_{phi} is missing)

$$\hat{X}_{phi} = f\left(X_{p1i},\ldots,X_{p(\mathrm{h}-1)i},X_{p(\mathrm{h}+1)i}..,X_{pHi}\right) \qquad \textit{assuming that } X_{pki} \textit{ is not missing } \forall k \neq h$$

- In general, this procedure can bring to biased results if the imputation model is not correct, in particular if other relevant covariates should be considered.
- Single imputation does not account for uncertainty due to missingness, producing uncorrect standard errors of the parameters estimates.

Background

- Bias →Missing At Random (MAR) assumption:

$$P\left(Y\,missing \middle| X, Y\right) = P\left(Y\,missing \middle| X\right)$$

MAR cannot be tested on the data.

MAR is more plausible if we include in the imputation model a large number of covariates.

- Uncertainty due to missingness → Multiple Imputation (MI) (Rubin, 1987)

M imputations instead of 1 to account for uncertainty.

There are few examples of MI on air pollution time series (e.g. Lertxundi et al. 2015).

Multiple imputation

N units (days)*; K* variables $Y_1,..., Y_K$ where Y_{ki} value of *k*th variable in the *i*th unit

- We specify:

  1) a model for the joint distribution of the variables: p($Y_1, ..., Y_K \mid \vartheta$), where ϑ *is a* vector of model parameters.

  2) a prior distribution p(ϑ) for ϑ .

- Under the MAR assumption, the following Gibbs sampler is defined:

  1. Start at preliminary values for the missing data of the variables Y_{-k}, where Y_{-k} is defined to be all variables except for Y_k

  2. For *k =1, ..., K*: generate a random value, ϑ^*, from the posterior distribution p($\vartheta \mid Y_k, Y_{-k}$)

  3. Simulate the missing values of Y_k *from the* posterior predictive distribution p($Y_k \mid Y_{-k}, \vartheta^*$)

  Repeat steps 2 and 3 up to convergence.

- **At convergence, we consider M simulated values for each missing value, obtaining M completed data sets.**

Different MI procedures

- In the literature, different MI procedures have been proposed, which mainly differ for the specification of the imputation model.

- Specifying the joint model $p(Y_1, ..., Y_K \mid \vartheta)$ can be a complex task; a simpler approach consists in defining an univariate conditional distribution $p(Y_k \mid Y_{-k}, \vartheta)$, in the form of a regression model, for each variable with missing values (Multiple Imputation by Chained Equation, MICE).

  - Problem: incompatibility between conditional distributions and the joint distribution.

- A recent alternative approach which reduces incompatibility: Multiple Imputation by Ordered Monotone Blocks (Li et al. 2014).

Combining the results

Following Rubin (1987), let be *Q* the unknown parameter of interest and $\hat{Q}_i$ the estimate of the parameter of interest in the *i*th imputed dataset (*i*=1,…*M*)

The final estimate of *Q* is given by:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q}_i$$

The estimated variance is the sum of two components:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \widehat{U}_i \qquad B = \frac{1}{m-1}\sum_{i=1}^{m} (\widehat{Q}_i - \bar{Q})^2 \qquad T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

where $\widehat{U}_i$ is the estimated variance in the *i*th imputed data set.

Missingness in air pollution time series

- Reasons of missingness can be various:
  - Missing values due to technical problems (e.g. measuring instrument failures);
  - Missing values due to errors in reporting measurements;
  - Missing values by design (moving monitors or specific analyses planned to be performed only in a subset of days).
- Missin Completely At Random (MCAR) may hold in the two last situations, making complete case analysis unbiased.
- However, if the planned analyses require that daily time series structure is preserved (e.g. seasonal adjustement through non parametric functions is needed or lagged exposures/covariates are to be included in the model), complete case analysis is not appropriate and MI procedure should be anyway applied.
- Missingness could be due to a combination of different mechanisms. In this case, violation of MCAR cannot be excluded.

Motivating examples

1) March 2002 - October 2013

   1 monitor in San Pietro Capofiume (BO) provided daily levels of airborne particle.

   Different fractions of particle were considered.

   Missing values mainly due to failures of the instrument used to determine fractions.

2) November 2011 - March 2015 Supersito Project

   4 monitors: Bologna (main site), Parma, Rimini, S. Pietro Capofiume (rural site)

   Particle speciation and dimension were considered.

   Missing values by design (laboratory analysis on one particles sample every 3 days)

In both cases a large number of additional meteorological and air quality variables were available.

Points to be addressed

- **Use of reliable methods to select the predictors to be included in the imputation models** (penalized regression approach: Tibshirani 1996; Zou and Hastie 2005)

- **Inclusion of lagged variables in models specifications**

- **Evaluation of MI quality/performance**

  Several methods have been proposed (), but they are not frequently used in practice.

  We plan to perform a simulation study (Baccini et al. 2010):

  - Step1: estimate the probability of each specific pattern of missingness, given covariates;
  - Step2: generate missing values on the subset of complete data, according to these estimated probabilities;
  - Step3: perform MI on the simulated data set.
  - Step4: Calculate reference quantities from the M imputed data sets and combine the M results.
  - Step 5: Repeat Step 2-4 many times.
  - Step 6: compare results from the simulated data sets and from the subset of complete data.

References

- Baccini M, Cook S, Frangakis CE, et al. (2010). Multiple Imputation in the Anthrax Vaccine Research Program. Chance, 23:16:23.
- Lertxundi A, Baccini M, Lertxundi N, et al. (2015). Exposure to fine particle matter, nitrogen dioxide and benzene duirng pregnancy and cognitive and psychomotor developments in children at 15 months of age. *Environ Int*, 80:33-40. doi: 10.1016/j.envint.2015.03.007.
- Li F, Baccini M, Mealli F, et al. (2014). Multiple imputation by ordered monotone blocks with application to the Anthrax Vaccine Adsorbed Trial. *Journal of Computational and Graphical Statistics,* 23:877-892.
- Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.
- Zou and Hastie (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B,67:301-320.

Gli autori dichiarano di non avere conflitto di interessi