Reconstructing missing data sequences in multivariate time series: an application to environmental data

Ricostruzione di sequenze di valori mancanti in serie storiche multivariate: un'applicazione a dati ambientali

Maria Lucia Parrella and Giuseppina Albano and Michele La Rocca and Cira Perna

Abstract Missing data arise in many statistical analyses, due to faults in data acquisition, and can have a significant effect on the conclusions that can be drawn from the data. In environmental data, for example, a common approach usually adopted by the Environmental Protection Agencies to handle missing values is by deleting those observations with incomplete information from the study, obtaining a massive underestimation of a lot of indexes usually used for evaluating air quality. In multivariate time series, moreover, it may happen that not only isolated values but also long sequences of some of the time series' components may miss. In such cases, it is quite impossible to reconstruct the missing sequences basing on the serial dependence structure alone. In this work, we propose an new procedure that aims to reconstruct the missing sequences by exploiting the spatial correlation and the serial correlation of the multivariate time series, simultaneously. The proposed procedure is based on a spatial-dynamic model and imputes the missing values in the time series basing on a linear combination of the *neighbor* contemporary observations and their lagged values. It is specifically oriented to spatio-temporal data, although it is general enough to be applied to generic stationary multivariate time-series. In this paper, the procedure has been applied to the pollution data, where the problem of missing sequences is of serious concern, with a remarkably satisfactory performance.

Abstract *I dati mancanti si presentano in molte analisi statistiche, a causa di errori nell'acquisizione dei dati e possono avere effetti significativi sulle conclusioni*

Maria Lucia Parrella
Dip. di Scienze Economiche e Statistiche, Università of Salerno, Italy,e-mail: mparrella@unisa.it

Giuseppina Albano
Dip. di Scienze Economiche e Statistiche, Università of Salerno, Italy, e-mail: pialbano@unisa.it

Michele La Rocca
Dip. di Scienze Economiche e Statistiche, Università of Salerno, Italy, e-mail: larocca@unisa.it

Cira Perna
Dip. di Scienze Economiche e Statistiche, Università of Salerno, Italy, e-mail: perna@unisa.it

che possono essere tratte dai dati. Nei dati ambientali, ad esempio, un approccio comunenemente adottato dalle agenzie per la protezione dell'ambiente per gestire i valori mancanti consiste nell'eliminare tali osservazioni dallo studio, ottenendo una massiccia sottovalutazione di molti indici solitamente usati per valutare la qualità dell'aria. Nelle serie temporali multivariate, inoltre, può accadere che non solo valori isolati, ma anche lunghe sequenze di alcune componenti delle serie temporali possano mancare. In questi casi, è assolutamente impossibile ricostruire le sequenze mancanti basandosi esclusivamente sulla struttura di dipendenza seriale. In questo lavoro, proponiamo una nuova procedura che mira a ricostruire le sequenze mancanti sfruttando contemporaneamente la correlazione spaziale e la correlazione seriale delle serie temporali multivariate. La procedura proposta si basa su un modello spaziale-dinamico e imputa i valori mancanti nelle serie temporali basandosi su una combinazione lineare delle osservazioni contemporanee di componenti vicine e dei loro valori ritardati.
In questo lavoro, la procedura è applicata ai dati sull'inquinamento, dove il problema delle sequenze mancanti diviene rilevante, con prestazioni notevolmente soddisfacenti.

Key words: Spatial correlation, missing values, PM_{10} data, time series.