Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research 59 (2017) 2442–2457

Advances in Space Research
(a COSPAR publication)

www.elsevier.com/locate/asr

Data imputation analysis for Cosmic Rays time series

R.C. Fernandes [a,*], P.S. Lucio [b], J.H. Fernandez [b]

[a] Programa de Pós-Graduação em Ciências Climáticas, Universidade Federal do Rio Grande do Norte, Natal/RN 59078970, Brazil
[b] Departamento de Ciências Atmosféricas e Climáticas, Universidade Federal do Rio Grande do Norte, Natal/RN 59078970, Brazil

Received 30 July 2016; received in revised form 12 February 2017; accepted 13 February 2017
Available online 22 February 2017

Abstract

The occurrence of missing data concerning Galactic Cosmic Rays time series (GCR) is inevitable since loss of data is due to mechanical and human failure or technical problems and different periods of operation of GCR stations. The aim of this study was to perform multiple dataset imputation in order to depict the observational dataset. The study has used the monthly time series of GCR Climax (CLMX) and Roma (ROME) from 1960 to 2004 to simulate scenarios of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90% of missing data compared to observed ROME series, with 50 replicates. Then, the CLMX station as a proxy for allocation of these scenarios was used. Three different methods for monthly dataset imputation were selected: AMÉLIA II – runs the bootstrap Expectation Maximization algorithm, MICE – runs an algorithm via Multivariate Imputation by Chained Equations and MTSDI – an Expectation Maximization algorithm-based method for imputation of missing values in multivariate normal time series. The synthetic time series compared with the observed ROME series has also been evaluated using several skill measures as such as RMSE, NRMSE, Agreement Index, R, R^2, *F*-test and *t*-test. The results showed that for CLMX and ROME, the R^2 and R statistics were equal to 0.98 and 0.96, respectively. It was observed that increases in the number of gaps generate loss of quality of the time series. Data imputation was more efficient with MTSDI method, with negligible errors and best skill coefficients. The results suggest a limit of about 60% of missing data for imputation, for monthly averages, no more than this. It is noteworthy that CLMX, ROME and KIEL stations present no missing data in the target period. This methodology allowed reconstructing 43 time series.

Keywords: Bootstrap; Expectation maximization; Skill; Multivariate; Chained equations

1. Introduction

A major problem in the study of the Galactic Cosmic Rays (GCR) time series is the difficulty in finding a non-gapped long-term series. Data losses can be caused by mechanical or technical failure and human errors. Thus, several GCR studies are restricted to few stations distributed around the globe. This data missing problem is not a GCR time series privilege, but can also be observed in several other areas, like Meteorology, Biomedicine, Information Systems datasets, among others.

Over the past decades, the historical GCR time series has been reconstructed using sunspot numbers and cosmogenic ^{10}Be isotope levels found in both Earth Polar Caps (Usoskin et al., 2002, 2005; Mursula et al., 2003; McCracken, 2004; McCracken and Beer, 2007). However, it leads to some questions, such as: (a) Is it possible to create a synthetic series from another Neutron Monitor (NM) station? (b) What is the criterion for filling data gaps? and (c) Which GCR stations would be filled?

Therefore, the main aim of this study was to report the use of the multiple imputation method in order to analyze its efficiency on the reconstruction of observational GCR data.

* Corresponding author.
E-mail addresses: ronabson@hotmail.com (R.C. Fernandes), pslucio@ccet.ufrn.br (P.S. Lucio), jhenrix@gmail.com (J.H. Fernandez).

http://dx.doi.org/10.1016/j.asr.2017.02.022
0273-1177/

2. Material and methods

2.1. Dataset

GCR monthly databases from the Russian Academy of Sciences (http://www.wdcb.ru/stp/) and from the Geophysical World Data Center (GWDC) for Solar-Terrestrial Physics, for the 1960–2004 period (http://www.wdcb.ru/stp/data/cosmic.ray/Neutron_Monitors(monthly_values)/) were used. The spatial distribution of stations is shown in Fig. 1a. It was observed in Fig. 1b the Climax station (CLMX, Lat = 39.37°, Long = −106.1°, Alt = 3.400 m, Cut-Off Rigidity 2.99 GV, 17 NM64) and Rome (ROME, Lat. = 41.9°, Long. = 12.52°, Alt. 60 m, Cut-Off Rigidity 6.32 GV).

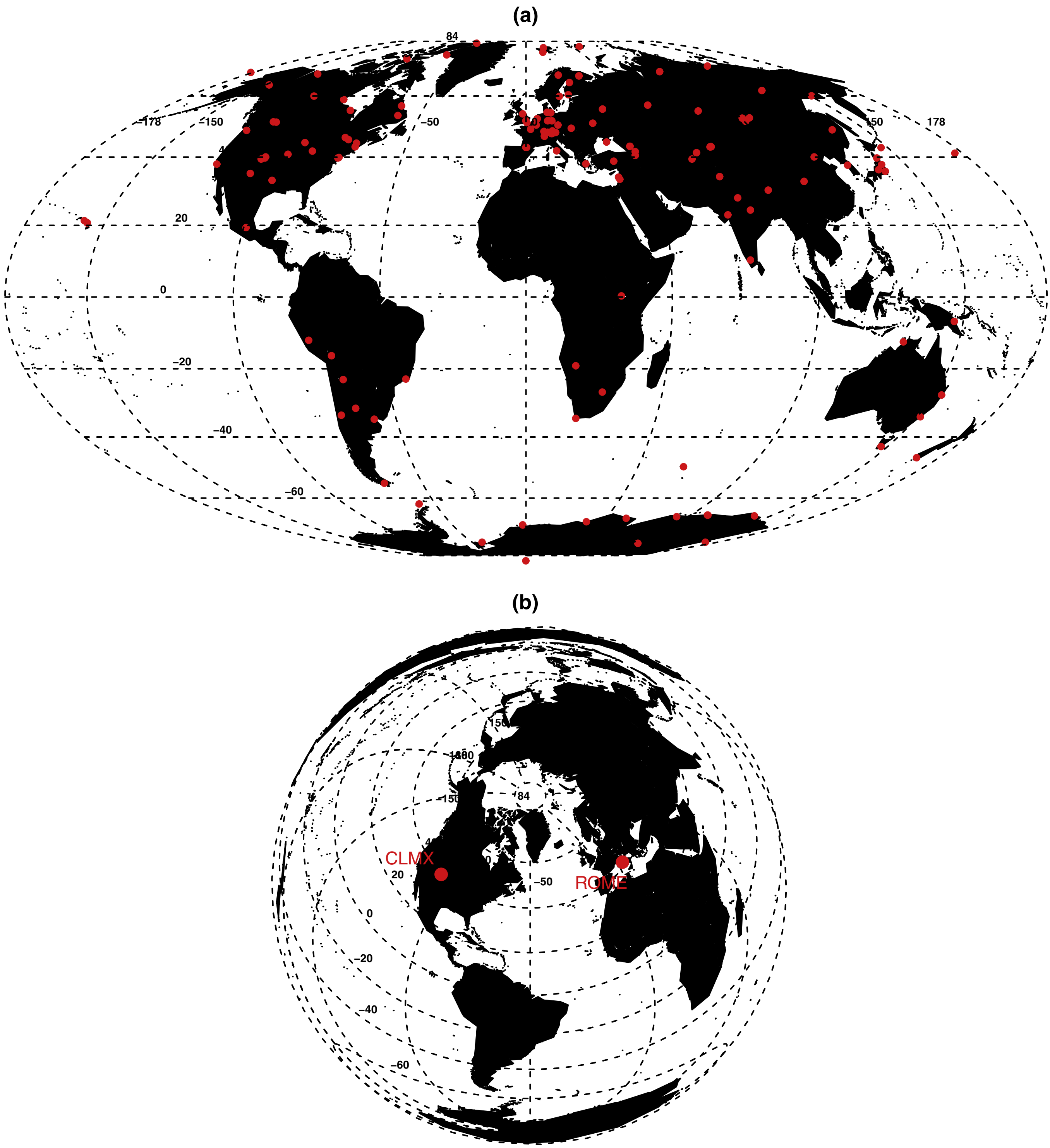

Fig. 1. (a) GCR NM stations spatial distribution in the globe, over the Mollweide projection and (b) Climax (CLMX) and Rome (ROME) GCR NM locations, according to the Azequalarea projection.