Chemosphere 177 (2017) 326–333

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Evaluation of a multiple linear regression model and SARIMA model in forecasting ^{7}Be air concentrations

María del Carmen Bas [a, b, *], Josefina Ortiz [a], Luisa Ballesteros [a], Sebastián Martorell [a]

[a] *Laboratorio de Radiactividad Ambiental, Universitat Politècnica de València, Spain*
[b] *Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, Spain*

HIGHLIGHTS

- A SARIMA and a MLR models were proposed to forecast ^{7}Be activity.
- The SARIMA model provides good forecasts of ^{7}Be air concentrations in the out-of-sample year.
- The MLR model provides information on the significant meteorological variables that affect ^{7}Be concentrations.

ARTICLE INFO

Article history:
Received 9 December 2016
Received in revised form
6 March 2017
Accepted 8 March 2017
Available online 15 March 2017

Handling Editor: R. Ebinghaus

Keywords:
^{7}Be
Time series
Multiple linear regression
Forecasting

1. Introduction

^{7}Be is widely used as an atmospheric radiotracer due to its relatively short life ($T_{1/2} = 53.3$ days) and ease of measurement by γ-spectrometry, which provides important information on atmospheric air mass motions. A better understanding of its distribution would facilitate refinement and validation of global atmospheric circulation models (Dueñas et al., 2015). ^{7}Be forecasting can thus be adopted as a target value in analyzing fluctuations or deviations that could imply important atmospheric changes.

It is generally accepted that the ^{7}Be production rate depends on a number of atmospheric factors. Several studies have pointed out that the intensity of galactic cosmic rays in the Earth's orbit is affected by solar activity and the geomagnetic field, which is under constant cosmic ray bombardment from space (O'Brien, 1979; Vogt et al., 1990; Hötzl et al., 1991; Ioannidou and Papastefanou, 1994). In particular, an increase in solar activity and the geomagnetic field reduce the galactic cosmic ray flux, which is followed by reduced ^{7}Be production.

In addition to the above-mentioned sources of variability, ^{7}Be concentrations in the lower layers of the atmosphere present temporal variations caused by solar radiation and meteorological parameters that can affect regional weather patterns (temperature, relative humidity, precipitations, wind speed and wind direction) (Feely et al., 1989; Baeza et al., 1996).

Many research studies have analyzed the relation between ^{7}Be air concentrations and the meteorological and atmospheric variables using a simple correlation analysis (e.g. Dueñas et al., 1999; Ioannidou and Papastefanou, 2006; Piñero-García and Ferro-García, 2013; Ceballos et al., 2016; Neroda et al., 2016). Furthermore, some of these studies have applied Multiple Linear Regression (MLR) analysis to develop an explanatory and predictive model for ^{7}Be air concentrations using the atmospheric and meteorological variables as predictors (Table 1).

Each study uses several predictors to explain ^{7}Be air concentration in different time periods at different locations. The explicative power of the model, measured by the R square coefficient, is, in general, less than 50%. The studies that get the highest R^2, use a historical data range of less than five years, which may not be enough information to forecast the ^{7}Be air concentration for the following year. In addition to explanatory power, it is very important to compute accuracy measurements with data that have not been used to develop the model. This procedure is not applied in the above MLR models and is important in measuring the validity and forecasting power of the model, which is one of the aims of the present study.

* Corresponding author. Laboratorio de Radiactividad Ambiental, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain.
E-mail address: maibacer@eio.upv.es (M.C. Bas).

http://dx.doi.org/10.1016/j.chemosphere.2017.03.029
0045-6535/

Table 1
^{7}Be predictive models for different time periods at different locations.

Location	Period	Significant variables used in MLR	R^2	Source
Málaga, Spain	1992–1995	- Maximum Temperature - Rainfall - Relative Humidity - Hours of sunshine	27%	Dueñas et al. (1999)
Thessaloniki, Greece	1987–2001	- Temperature - Relative Humidity - Sunspot Number	38.5%	Ioannidou and Papastefanou (2006)
Granada, Spain	1993–2001	- Temperature - Rainfall - Sunspot Number	71%	Azahra et al. (2004)
Málaga, Spain	1997–2007	- Solar energetic proton - Aerosol optical depth	34%	Dueñas et al. (2015)
Granada, Spain	1996–2010	- Temperature - Relative Humidity - Sunspot Number	52%	Piñero-García and Ferro-García (2013)
Granada, Spain	2005–2009	- Temperature - Relative Humidity - Rainfall	72.16%	Piñero-García et al. (2012)
Plymouth, UK	2009–2010	- Rainfall	94%	Taylor et al. (2016)
Granada, Spain	2011–2014	- Solar Irradiance - Total suspended particles	66.9%	Essaid et al. (2015)
Vladivostok, Russia	2013–2014	- Altitude - Precipitation - Temperature - Aerosol concentration - Trajectories in the pacific (North-East)	55%	Neroda et al. (2016)

Several authors recommend the use of time series modelling techniques instead of multiple linear regression when monitoring correlated process data (Alwan and Roberts, 1988; Harris and Ross, 1991; Wardell et al., 1994). Classical regression is often insufficient for explaining all the interesting dynamics of a time series. For instance, the estimated autocorrelation function (ACF) of the residuals of the regression model could reveal additional structure in the data that the regression did not capture. Instead, the introduction of Box-Jenkins models could deal with the limitations of classical regression in time series (Shumway and Stoffer, 2006).

A recent study applied a decomposition of the ^{7}Be time series into a trend-cycle, a seasonal and an irregular component in order to separate the inter- and intra-annual patterns of ^{7}Be variability (Bas et al., 2016). The results of this study showed the suitability of applying time series analysis to correlated data in order to separate the different sources of variability of ^{7}Be concentrations and to develop a forecasting model.

The aim of this study is to propose two models to explain and forecast ^{7}Be air concentrations: i) a Seasonal Autoregressive Integrated Moving Average (SARIMA) model and ii) a Multiple Linear Regression (MLR) model using meteorological and atmospheric variables. Both the time series and multiple linear regression models are evaluated by comparison with real ^{7}Be air concentrations for the city of Valencia in 2007–2014 and with out-of-sample tests for the 12 months of the year 2015, using the Root Mean Square Error (RMSE) and the Adapted Mean Absolute Percentage Error (AMAPE) as forecasting accuracy measures. Finally, the results of the accuracy measurements of both models are compared.

2. Material and methods

2.1. Study area and sampling

Airborne particulate samples were collected weekly on the campus of the Universitat Politècnica de Valencia from January 2007 to December 2015. Valencia is situated on the east coast of Spain (15 m above sea level) in the western Mediterranean Basin (39°28′50″ N, 0°21′59″ W) and has a relatively dry subtropical Mediterranean climate with very mild winters and long hot summers. The sampling point was located approximately 2 km away from the coastline.

Aerosol samples were collected using Eberlyne G21DX and Saic AVS28A air samplers placed approximately 1 m above ground level. The aerosol particles were retained on a cellulose filter of 4.2×10^{-2} m effective diameter and 0.8 µm pore size. The filters were changed weekly and the average volume ranged from 300 to 400 m^3 per week. Each filter was put inside a plastic box and kept in a desiccator until it was measured.

2.2. ^{7}Be activity measurements

A monthly composite sample containing 4–5 filters was measured by γ-spectrometry to determine specific ^{7}Be activities using an HPGe detector (ORTEC Industries, USA) n-type with relative efficiency of 18% for 60Co gamma-ray. A certificated standard containing radionuclides with energies ranging from 59 to 1836.1 keV was used for preparing the calibrated filters, which were placed inside their plastic boxes on the top of the detector. The counting time was 60000s and the γ-line 477.7 KeV was used to calculate the activity. ORTEC Gamma-Vision software was used for acquisition and analysis. Concentration activities were corrected for the radioactive decay to the mid-collection period. The mean measured uncertainties (K = 2) were around 10%.

2.3. Statistical analysis

2.3.1. SARIMA model

The SARIMA model building process is designed to take advantage of the association in the sequentially lagged relationships that usually exists in data collected periodically. A time series $\{z_t,\ t = 1, \ldots, N\}$ is generated by a SARIMA$(p, d, q)(P, D, Q)_s$ model if: