Seasonal Means Estimation and Missing Data in Real Data Time Series

Eljona Milo, Lorenc Ekonomi, Lorena Margo and Edlira Donefski

Department of Mathematics, Informatics and Physics
Fan S. Noli University, Korca, Albania

Abstract

In this paper we deal with missing data encountered in periodically correlated time series and we are interested in comparing the performance of different methods by impute these missing data. We study two periodically correlated time series with real data. We apply methods for missing the data in an appropriate percentage in one of the series, and the other time series has missing data. We use perARMA, Mice and Amelia packages from R programming language. Then we compare the results obtained in estimating the seasonal means before and after imputing the missing data. The results obtained show the same good performance in the case of using imputation methods for imputing the data compared with original time series.

Keywords Periodically correlated time series, missing data, imputation

Introduction

Missing data is a common occurrence in many areas of research and their existence sometimes create problems in the quality of conclusions. Data may be missing for a variety of reasons. Determining the appropriate studying method in the presence of incomplete observations is a major question for today researchers. The analysis of time series data is an important area of statistics. Because the data are records taken through time, the presence of missing observations in time series data are very common. Different time series may require different strategies for estimating these missing values. We emphasize especially the necessity of using these strategies as effectively as possible in order to obtain the best possible estimates.
In this paper we are focused in the performance of three methods in calculating seasonal means in the case of imputing the missing data using imputations methods and after replacing the missing data with seasonal means using perARMA package.

**Missing value mechanisms**

There are three important cases to distinguish for the responsible generating processes behind missing values (see Rubin (1987), Rubin and Little (2002)). Let $X = (x_{ij}), 1 \leq i \leq n, 1 \leq j \leq p$ denote the data, where $n$ is the number of observations and $p$ the number of observed variables (dimensions), and let $M = (M_{ij}), 1 \leq i \leq n, 1 \leq j \leq p$ be an indicator whether an observation is missing ($M_{ij} = 1$) or not ($M_{ij} = 0$). The missing data mechanism is characterized by the conditional distribution of $M$ given $X$, denoted by $f(M|X, \phi)$, where $\phi$ indicates unknown parameters. Then the missing values are Missing At Random (MAR) if it holds for the probability of missingness that $f(M|X, \phi) = f(M|X_{obs}, \phi)$ (*).

Here $X = (X_{obs}, X_{miss})$ denotes the complete data, and $X_{obs}$ and $X_{miss}$ are the observed and missing parts, respectively. Hence the distribution of missingness does not depend on the missing part $X_{miss}$.

If in addition the distribution of missingness does not depend on the observed part $X_{obs}$, the important special case of MAR called Missing Completely At Random (MCAR) is obtained, given by $f(M|X, \phi) = f(M|\phi)$.

If equation (*) is violated and the patterns of missingness are in some way related to the outcome variables, i.e., the probability of missingness depends on $X_{miss}$, the missing values are said to be Missing Not At Random (MNAR). This relates to the equation $f(M|X, \phi) = f(M|(X_{obs}, X_{miss}), \phi)$.

**Periodically correlated time series**

A special class of nonstationary time series has been defined by Gladyshev [3], called periodically correlated (PC) time series. These time series are nonstationary, but have periodic means and covariances. They have been known to be very useful in describing many time series.

The time series $\{X_t\}$ with finite first and second moments is said to be periodically correlated (PC) with period $T$, if there exists a positive integer $T$ such that the mean function $\mu_t = E(X_t)$, and the covariance function $\gamma(t, s) = Cov(X_t, X_s) = E((X_t - \mu_t)(X_s - \mu_s))$ satisfy $\mu_t = \mu_{t+T}$ and $\gamma(t, s) = \gamma(t + T, s + T)$ for all integers $t$ and $s$, and moreover $T$ is the smallest
number for which holds. The periodic mean is defined as
\[ \mu_s = E(X_{s+1} \mid T), \quad s = 1, 2, \ldots, T, \quad r \in \mathbb{Z} \]

and the autocovariance function defined
\[ \gamma_s(k) = \text{Cov}(X_{s+1}, X_{s+k}) = E[(X_{s+1} - \mu_s)(X_{s+k} - \mu_s)] \cdot \]

**Basic ideas behind perARMA, Mice and Amelia packages**

The package *perARMA* [1] provides procedures for periodic time series analysis. In this paper we used *permest* function, which assuming that the period T is known, procedure *permest* plots and returns the estimated periodic mean as a function of season. Missing data are permitted. If at time t there is a missing value, it is replaced with the periodic mean at (t mod T), provided the periodic mean exists (meaning there is at least one non-missing data for the season (t mod T)). Otherwise the periodic mean at (t mod T) will be set to "Missing" and in the output vectors xr and xd all the values whose times are congruent with (t mod T) will be set to "Missing". The series may contain missing values (the authors suggest using NaN) and the length of the series need not be an integer multiple of the period.

*Mice* (Multivariate Imputation via Chained Equations) [10] is one of the commonly used package by R users. Creating multiple imputations as compared to a single imputation takes care of uncertainty in missing values. The *Mice* package in R, helps us to impute missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing datapoint.

Amelia II [4] performs multiple imputation, a general-purpose approach to data with missing values. Bootstrap-based EM algorithm is employed to impute missing values. The algorithm draws m (the number of imputation dataset) samples of size n (the size of original dataset) from original dataset. Point estimates of mean and variance (both are vectors) are performed in each sample by using EM method. Remember there are m sets of estimates. Then each set of estimates is used to impute the missing observations from original dataset. The result is m sets of imputed data that can be used for subsequent analyses. By assuming that time series data vary smoothly over time, observed values close in time to the missing value can greatly aid imputation of that value. The advantage of Amelia II is that it combines the comparative speed and ease-of-use of our algorithm with the power of multiple imputation, to let you focus on your substantive research questions rather than spending time developing complex application-specific models for nonresponse in each new data set.
Series of diabetic patients data

The first dataset used in this article was taken in collaboration with the Korca regional hospital. Pre and posture measurements in diabetic patients were taken over a three-year period 2015-2017. All the data were collected for 515 patients. These data have missing values. In the first measurements that are the glucose count, 8.5% of the data is missing and in the second measurements, glucose uptake after 10.6% of the data is missing. Based on the graph of the series we noted that the nature of this series is periodic. After a detailed analysis using perARMA package results that this time series is a periodically correlated time series with period 5. Using the perARMA package in the R environment as well as the "permest" function we will calculate the seasonal means and we will compare the results obtained with the seasonal means obtained after imputing the missing values with the help of the Amelia and Mice packages.

![Series of diabetic patients data with missing value](image)

Figure 1 Series of diabetic patients data with missing value

The results of seasonal means in the case of completing missing data with seasonal means and the values obtained after using packages Amelia and Mice for imputing the missing values, are presented in the following table:

Table 1: Seasonal means for diabetic patients data with three methods

<table>
<thead>
<tr>
<th>Seasonal mean</th>
<th>Seasonal mean after Amelia</th>
<th>Seasonal mean after Mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>216.7573</td>
<td>216.7573</td>
</tr>
<tr>
<td>2</td>
<td>187.9703</td>
<td>187.3548</td>
</tr>
<tr>
<td>3</td>
<td>164.4500</td>
<td>164.9881</td>
</tr>
<tr>
<td>4</td>
<td>146.6771</td>
<td>145.7026</td>
</tr>
<tr>
<td>5</td>
<td>126.0423</td>
<td>126.6670</td>
</tr>
</tbody>
</table>
We can notice a similar performance of the three methods used in calculating the seasonal means in the case of diabetes patients time series: SM-Seasonal Mean, SMAA- Seasonal Mean After Amelia, SMAM-Seasonal Mean After Mice (see figure below).

![Graph showing seasonal means of diabetic patients data using three methods.](image)

**Figure 2** Results for the seasonal means of diabetic patients data using three methods.

**Rainfall series**

We are examining a situation with real data from our country, which was obtained from the monthly meteorological records recorded on the rainfall in the Pogradec area for a period of 40 years.

This time series does not have any missing values, but in order to compare the proposed methods we will create 5% of missing data using the MCAR mechanism. In this series we know the seasonal means of the original series. Then we impute the missing values with “Amelia” and “Mice” packages in R program. And we will compare the seasonal means obtained after imputing the series with seasonal means of the original series that has not missing values.
Figure 3 A plot of time series and the Autocovariance function for Rainfall data

Table 2 Seasonal means for rainfall data using three methods

<table>
<thead>
<tr>
<th></th>
<th>Seasonal mean</th>
<th>Seasonal mean after Amelia</th>
<th>Seasonal mean after Mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>79.0925</td>
<td>79.75893</td>
<td>79.6800</td>
</tr>
<tr>
<td>2</td>
<td>85.1775</td>
<td>86.55016</td>
<td>85.2650</td>
</tr>
<tr>
<td>3</td>
<td>71.4425</td>
<td>71.81440</td>
<td>69.3500</td>
</tr>
<tr>
<td>4</td>
<td>58.4500</td>
<td>57.31504</td>
<td>60.5675</td>
</tr>
<tr>
<td>5</td>
<td>54.4275</td>
<td>54.42750</td>
<td>54.4275</td>
</tr>
<tr>
<td>6</td>
<td>34.8900</td>
<td>36.12676</td>
<td>37.5750</td>
</tr>
<tr>
<td>7</td>
<td>22.0250</td>
<td>22.96760</td>
<td>21.9700</td>
</tr>
<tr>
<td>8</td>
<td>27.1900</td>
<td>29.72380</td>
<td>36.3225</td>
</tr>
<tr>
<td>9</td>
<td>45.9400</td>
<td>45.55507</td>
<td>48.4775</td>
</tr>
<tr>
<td>10</td>
<td>82.9625</td>
<td>79.24671</td>
<td>76.7600</td>
</tr>
<tr>
<td>11</td>
<td>111.8825</td>
<td>109.67324</td>
<td>99.4800</td>
</tr>
<tr>
<td>12</td>
<td>96.8775</td>
<td>94.12641</td>
<td>91.7125</td>
</tr>
</tbody>
</table>

We can notice a similar performance of the three methods used in calculating the seasonal means in the case of rainfall time series (see figure below).
Conclusions

In this paper we studied two correlated periodically time series with real data. In the rainfall series that has no missing values and we know the seasonal means, we used the MCAR mechanism to create 5% of missing data. Then we imputed the data using multiple imputations functions from packages Amelia and Mice. From the results obtained, we notice that all these methods have a good performance. In the case of using Amelia, we can see that the seasonal means have a better approximation than the case when using Mice package. We used the same idea in the series with missing values and we noticed that the performance of the three methods was very good. Even in this case, the seasonal means of the series after imputing the missing data using Amelia and Mice packages are close to each other.

References


**Received: December 8, 2018; Published: January 11, 2019**