COMPLETING THE MISSING DATA IN MARKOV SYMMETRIC STABLE TIME SERIES

S. N. Moiseev

Optimal algorithms for completing the missing data in Markov symmetric stable time series are proposed. We consider the cases where the data missing is caused by either a mechanism that is independent of the series values or a random censoring from below. The completing does not distort the probability properties of the series.

1. INTRODUCTION

Ionospheric propagation of ultrashort waves is significantly influenced by a sporadic $E$ layer ($E_s$ layer), for which an anomalously high electron density is typical. The appearance of the layer results in breaking the ultrashort-wave radio communication through the upper ionosphere and also in the formation of a sporadic channel of superlong-distance propagation of ultrashort radio waves through the lower ionosphere. When predicting the ionospheric propagation of radio waves, we allow for the influence of the $E_s$ layer using the analysis and prediction of its maximum electron density that is related in a unique fashion to the layer screening frequency. The screening frequency is measured round-the-clock by vertical-sounding stations. However, the measured results at these stations have numerous missing data due to technical breaks, limited range of the measuring instruments, and screening by the underlying regular layer $E$ of the sounding signal reflected from the $E_s$ layer. In summertime, the share of missing data is from 10 to 15%, while in wintertime it reaches 60%, which substantially impedes the data processing. To overcome this difficulty, we should perform a preliminary completing of the missing data to ensure further work with a complete data sample [1-3]. As a rule, the missing data in a sample are completed by the average values, which significantly distorts the probability properties of the data and can lead to erroneous statistical conclusions. In [4] we constructed a model of maximum electron density of the $E_s$ layer, which is a special case of the Markov symmetric stable process.

In this paper, we propose optimal algorithms for completing the missing data in the Markov symmetric stable time series, which do not distort their probability properties.

2. FORMULATION OF THE PROBLEM

The Markov symmetric stable time series $y_t$, where $t = 0, \pm 1, \pm 2, \ldots$, is described by the first-order autoregressive equation

$$y_t - \phi y_{t-1} = \alpha_t,$$

where $\phi \in (-1, 1)$ and $\alpha_t$ are independent, identically distributed symmetric random quantities with a zero median and the characteristic function

$$\theta(u) = \exp(-b|u|^\alpha),$$

where $b > 0$ and $\alpha \in (0, 2]$. The Gaussian Markov processes are obtained from Eq. (1) as a special case for $\alpha = 2$, while the linear Markov Cauchy processes are obtained for $\alpha = 1$. By analogy with the Gaussian
processes, the process $y_t$ is invariant with respect to linear transformations, so that an arbitrary linear functional of $y_t$ is a symmetric stable random quantity.

Let $n_{\text{obs}}$ readings $y_{\text{obs}}$ be observed and $n_{\text{mis}} = n - n_{\text{obs}}$ readings $y_{\text{mis}}$ be missing from $n$ readings $y = \|y_1, \ldots, y_n\|$ of the sample of the time series $y_t$. First, we assume that the data-missing generation process is independent of the values of the series $y_t$. We should complete the missing data by such values of $y_{\text{mis}}^*$ that the probability properties of the series $y_t$ have minimum distortions.

### 3. OPTIMAL COMPLETING

The breaking of the probability properties of the series caused by the missing-data completing is characterized by the loss functional

$$C(W, p) = |W(y) - p(y_{\text{mis}}|y_{\text{obs}})W(y_{\text{obs}})|,$$

where $W(y)$ is the $n$-dimensional probability density of the sample $y$, $W(y_{\text{obs}})$ is the $n_{\text{obs}}$-dimensional probability density of the observed readings $y_{\text{obs}}$ and $p(y_{\text{mis}}|y_{\text{obs}})$ is the conditional probability density of the values of $y_{\text{mis}}^*$, which are used to complete the misses for the fixed values of $y_{\text{obs}}$. Averaging the misses in Eq. (2), we obtain the expression for the average risk

$$R(p) = \int_S |W(y) - p(y_{\text{mis}}|y_{\text{obs}})W(y_{\text{obs}})|W(y) \, dy,$$

where $S = \{y : W(y) > 0\}$ is the support of the density $W(y)$. The specific feature of the average risk (3) is that it is a nonlinear functional with respect to the function $p(y_{\text{mis}}|y_{\text{obs}})$. In accordance with the theory of statistical decisions [5], an optimal rule for completing the missing data is determined by choosing such a function $p(y_{\text{mis}}|y_{\text{obs}})$ that minimizes the average risk (3). In the simplest case, when the probability model of the time series $y_t$ is fully known, the minimum of the functional (3) is reached at the point

$$p(y_{\text{mis}}|y_{\text{obs}}) = W(y_{\text{mis}}|y_{\text{obs}}).$$

From Eq. (4) it is obvious that the optimal completing rule is randomized. In other words, the missing data are completed by randomly chosen values from the distribution (4). Let us note that the discussed, rather general, approach to the problem of completing the missing data allows us to estimate the losses due to nonoptimal completing and develop quasi-optimal adaptive completing in the presence of an a priori uncertainty [5].

Assume that one of the missing readings occurs at time $t = k$. Then, using the expression

$$p(y_{\text{mis}}|y_{\text{obs}}) = p(y_k|y_{\text{obs}})p(y_{\text{mis}}-k|y_k, y_{\text{obs}}),$$

where $y_{\text{mis}}-k$ is the vector of the missed readings from the series $y_t$ except for the reading $y_k$, we easily show that the minimum of the average risk (3) is attained at the two points

$$p(y_k|y_{\text{obs}}) = W(y_k|y_{\text{obs}}) \quad \text{and} \quad p(y_{\text{mis}}-k|y_{\text{obs}}) = W(y_{\text{mis}}-k|y_{\text{obs}}).$$

Therefore, we draw two important conclusions. First, the joint completing the missing readings $y_{\text{mis}}$ is reduced to the following sequential process. We complete the missing data $y_k$ at time $t = k$ by the value $y_k^*$, which is randomly chosen (simulated) from the distribution $p(y_k|y_{\text{obs}})$. Then the reading $y_k^*$ is assumed to be known ($y_k^* \in \{y_{\text{obs}}\}$), and the subsequent missing reading from the set $\{y_{\text{mis}}-k\}$ is completed. Second, the data-completing sequence and the location of the current completed reading with respect to the readings $y_{\text{obs}}$ are arbitrary and have no influence on the completing quality.