MOAD'2024
The Sixth International Colloquium on
Methods and Tools for Decision Support
October 20th–22nd 2024, Tizi-Ouzou, Algeria

Impact of missing data on the prediction of stationary random fields

Abdelghani Hamaz
Laboratoire de Mathématiques Pures et Appliquées. Mouloud Mammeri University of Tizi-Ouzou. Algeria
Farida Achemine
Laboratoire de Mathématiques Pures et Appliquées. Mouloud Mammeri University of Tizi-Ouzou. Algeria

Mohamed Hemou Ibazizen
Laboratoire de Mathématiques et Applications. University of Poitiers. France.

Abstract

The purpose of this paper is to treat the prediction problems where a number of observations are missing to the quarter-plane past of a stationary random field. Our aim is to quantify the influence of missing values on the prediction by giving the simple bounds for the prediction error variance. These bounds allow to characterize the random fields for which the missing observations do not affect the prediction. Simulation experiments and an application to real data are presented.

Keywords and phrases Stationary random fields; moving average representation; autoregressive representation; prediction; interpolation.

Received 16th March 2024 **Accepted** To be completed by MOAD editorial office

1 Introduction

The problem of linear prediction of stationary processes requires knowledge of the observed past and their covariance function. When the data is coming from physical and natural sciences it is common to have irregularities, missing or outlying observations. The problem of spatial prediction based on incomplete past was considered by Kohli and Pourahmadi [4] to provide estimates for the missing values as well as the predictors. The original impetus for their work came from the interpolation results in [1] and prediction based on incomplete past in [2] for a second order stationary time series. The key idea in both these methods is the appropriate orthogonalization of the "past" and "future" of the time series, where past corresponds to the infinite past of the first missing value and future to all the values observed between missing values and the time point at which we need to predict.
In this paper, we investigate the problem of linear prediction of stationary random fields with non-symmetrical half-plane past. Our main contribution lies in finding an explicit formula of the mean square convergent autoregressive series representation for all (h_1, h_2)-step ahead linear predictors, $(h_1, h_2) \geq (0, 0)$. In order to calculate explicitly the prediction coefficients of our new expression

Prelimanries

Let H be a Hilbert space (e.g, $H = L^2(\Omega, F, P)$) the space of all random variables on (Ω, F, P) with finite second order moments and zero mean, endowed by the inner product $< X, Y >= E(XY)$ with norm $\|X\| = \sqrt{EX^2}$. Let $\{X(s, t), (s, t) \in \mathbb{Z}^2\}$ on (Ω, F, P) a random field. We assume that $E(X(s, t)) = 0$, $(s, t) \in \mathbb{Z}^2$, $\{X(s, t), (s, t) \in \mathbb{Z}^2\}$ is said to be a second order random field if $E|X(s, t)|^2 < \infty$, $(s, t) \in \mathbb{Z}^2$. That is, $X(s, t) \in H$, moreover, if for all integers s_1, s_2, t_1 and t_2, the covariance of the $X(s_1, t_1)$ and $X(s_2, t_2)$ depends on the lags $(s_1 - s_2, t_1 - t_2)$, namely,

$$cov(X(s_1, t_1), X(s_2, t_2)) = \gamma(s_1 - s_2, t_1 - t_2)$$

then

- $\{X(s, t), (s, t) \in \mathbb{Z}^2\}$ is called a second order stationary random field.
- $\gamma(., .)$ is a positive-definite function on the group of lattice points $\mathbb{Z}^2$, by the Bochner's theorem.

Unlike the 1-D case, there is no natural order definition in the 2-D domain. However, the totally ordered NSHP support is a favorable type of support in the sense that it yields a natural extension to the 1-D results.

▶ **Definition 1.** *We call a nonsymmetrical half-plan past (NSHP) any subset S of $\mathbb{Z}^2$ satisfying*

1. *S stable under addition*
2. *$S \cup -S = \mathbb{Z}^2$*
3. *$S \cap -S = \{(0, 0)\}$.*

For a purely nondeterministic stationary (PND) random field $X(s, t)$ (the case when $V(s, t) = 0$, i.e $X(s, t) \in \overline{sp}\{\varepsilon(s, t), (s,t) \in S\}$). $X(s, t)$ have a mean square convergent infinite moving average $MA(\infty)$ representation [4]

$$X(s, t) = \sum_{k=0}^{+\infty} \sum_{l=0}^{+\infty} b_{kl}\varepsilon(s - k, t - l), \tag{1}$$

the sequence $\{b_{k,\,l},\ (k,\ l) \in \mathbb{Z}^2\}$ is called the MA(∞) parameters.

2 Main result:Autoregressive representation of the Multi-step ahead linear predictor

The multi-step ahead prediction problem of stationary random fields has been studied by .... when the third quadrant is used as the past. We extends their pioneer work to random fields with nonsymmetrical half-plane past. Let $\{X(s,\ t); (s,\ t) \in \mathbb{Z}^2\}$ is a PND stationary random field. The procedure for solving the $(h_1,\ h_2)-$step ahead linear prediction problem with respect to the total order and nonsymmerical half-plane (NSHP) support defined by (8) involves the construction of predictor of future values as a linear combination of $\{X(k,\ l),\ (k,\ l) \in S\}$ which are close to $X(s + h_1,\ t + h_2),\ (h_1,\ h_2) \geq (0,\ 0)$ in the sense of mean squared error. The representation (1) is inverted to give

$$\varepsilon(s,\ t) = \sum_{k=0}^{+\infty} \sum_{l=0}^{+\infty} a_{k,\,l} X(s - k,\ t - l). \tag{2}$$

The representation (2) converges in mean square [3] . The collection of all finite linear combinations of elements in the space and its closure are also included in the space. At first we fix our attention on the problem of finding convergent representation for the one-step ahead linear predictor $P_{H^s}X(s,\ t)$, i.e the minimum norm linear causal and continuous support predictor of $X(s, t)$. We show that when (8) converges, such a representation exists.

▶ **Theorem 2.** *Let $\{X(T);\ T \in \mathbb{Z}^2\}$ be a PND stationary random field. The one step ahead linear predictor $P_{H^s}X(s,\ t)$ of $X(s, t)$ possesses a convergent serie representation given by*

$$P_{H^s}X(s,\ t) = -\sum_{\substack{k=0 \\ (k,\,l)\neq(0,\,0)}}^{+\infty} \sum_{l=0}^{+\infty} a_{k,\,l} X(s - k,\ t - l), \tag{3}$$

if and only if $\varepsilon(s, t)$ has the convergent representation series.

Proof. We have (6) implies that

$$E(\varepsilon(s,\ t)X(s,\ t)) = E(\varepsilon(s,\ t))^2.$$

From (1) we deduce that

$$E(\varepsilon(s,\ t))^2 = a_{00}E(\varepsilon(s,\ t)X(s,\ t)),$$

and necessarily $a_{00} = 1$. Thus, (2) may be rewritten as

$$X(s,\ t) = \varepsilon(s,\ t) - \sum_{\substack{k=0 \\ (k,\,l)\neq(0,\,0)}}^{+\infty} \sum_{l=0}^{+\infty} a_{k,\,l} X(s - k,\ t - l).$$

◀

Now, we are interested in Predicting future values other than $X(s,\,t)$ which is greatly important in the theory and applications of stationary random fields. The next lemma is useful for computing the predictor.

▶ **Lemma 3.** *Let* $\{X(T);\, T \in \mathbb{Z}^2\}$ *be a PND stationary random field, the MA and the AR parameters are* $\{b_{k,\,l},\, (k,\,l) \in \mathbb{Z}^2\}$ *and* $\{a_{k,\,l},\, (k,\,l) \in \mathbb{Z}^2\}$*, respectively, then the following equation is satisfied for all* $(k,\,l) \geq (0,\,1)$

$$\sum_{i=0}^{k}\sum_{j=0}^{l} a_{ij} b_{k-i,\,l-j} = 0. \tag{4}$$

Proof. By substituting (2) we obtain for all $(k,\,l) \geq (0,\,1)$

$$\begin{aligned} 0 = E(\varepsilon(s-k,\,t-l)\varepsilon(s,\,t)) &= E(\varepsilon(s-k,\,t-l)\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty} a_{ij} X(s-i,\,t-j)) \\ &= E(\varepsilon(s-k,\,t-l)\sum_{i=0}^{k}\sum_{j=0}^{l} a_{ij} X(s-i,\,t-j)) \\ &= \sum_{i=0}^{k}\sum_{j=0}^{l} a_{ij} E(\varepsilon(s-k,\,t-l) X(s-i,\,t-j)) \\ &= \sum_{i=0}^{k}\sum_{j=0}^{l} a_{ij} b_{k-i,\,l-j} E(\varepsilon(s-k,\,t-l))^2. \end{aligned}$$

◀

▶ **Corollary 4** (O. Arezki *at al*, [1])**.** *Let* $\{X(T);\, T \in \mathbb{Z}^2\}$ *be a PND stationary random field, the MA and the AR parameters are* $\{b_{k,\,l},\, (k,\,l) \in \mathbb{Z}^2\}$ *and* $\{a_{k,\,l},\, (k,\,l) \in \mathbb{Z}^2\}$*, respectively, then the following equation is satisfied for all* $(k,\,l) \geq (0,\,1)$

$$\sum_{i=0}^{k}\sum_{j=0}^{l} a_{ij} b_{k-i,\,l-j} = 0. \tag{5}$$

▶ **Theorem 5.** *Let* $\{X(T);\, T \in \mathbb{Z}^2\}$ *be a PND stationary random field. The one step ahead linear predictor* $P_{H^S} X(s,\,t)$ *of* $X(s,t)$ *possesses a convergent serie representation given by*

$$P_{H^S} X(s,\,t) = -\sum_{\substack{k=0 \\ (k,\,l)\neq(0,\,0)}}^{+\infty}\sum_{l=0}^{+\infty} a_{k,\,l} X(s-k,\,t-l), \tag{6}$$

if and only if $\varepsilon(s,t)$ *has the series representation* (2)*.*

Proof. By using Lemma 3, the equation (1) implies that

$$E(\varepsilon(s,\,t)X(s,\,t)) = E(\varepsilon(s,\,t))^2.$$

From (2) we deduce that

$$E(\varepsilon(s,\,t))^2 = a_{00} E(\varepsilon(s,\,t)X(s,\,t)),$$

and necessarily $a_{00} = 1$. Thus, (2) may be rewritten as

$$X(s,\,t) = \varepsilon(s,\,t) - \sum_{\substack{k=0 \\ (k,\,l)\neq(0,\,0)}}^{+\infty}\sum_{l=0}^{+\infty} a_{k,\,l} X(s-k,\,t-l).$$

Since $\varepsilon(s,\,t)$ is uncorrelated with $X(u,\,v)$, $(u,\,v) < (s,\,t)$, we deduce that the one-step predictor of $X(s,\,t)$ is given by

$$P_{H^S} X(s,\,t) = -\sum_{\substack{k=0 \\ (k,\,l)\neq(0,\,0)}}^{+\infty}\sum_{l=0}^{+\infty} a_{k,\,l} X(s-k,\,t-l). \tag{7}$$

▶ **Proposition 6.** *This is a proposition*

The existence of the convergent representation (6) for the one step predictorgiven in Proposition 6 is assured by the convergence of (2). Conversely, if the one-step predictor $P_{H^S_{s,\,t}}X(s,\,t)$ has the mean square representation (6), then the one-step prediction error satisfies

$$\varepsilon(s,\,t) = X(s,\,t) - P_{H^S}X(s,\,t) = \sum_{k=0}^{+\infty}\sum_{l=0}^{+\infty} a_{k,\,l}X(s-k,\,t-l),$$

with $a_{0,\,0} = 1$ and the sum convergent in mean square. Thus, we have shown that a necessary and sufficient condition for the existence of (6) as a mean square limit is the existence and the convergence of (2). ◀

▶ Remark 7. Theorem 5 implies that.

3 Example and Simulation study

Consider the stationary first order multiplicative spatial autoregressive model (MSAR(1)) defined by

$$X(s,t) = aX(s-1,t) + bX(s,t-1) - a.bX(s-1,t-1) + \epsilon(s,t) \tag{8}$$

where $\{\epsilon(s,t); (s,t) \in \mathbb{Z}^2\}$ are independent random variables with $\mathbf{E}\,(\epsilon(s,t)) = 0$, $\mathbf{Var}(\epsilon(s,t)) = \sigma^2$, $|a| < 1$ and $|b| < 1$.
By using the recursions given by (**??**) and the fact that $a_{10} = a$, $a_{10} = b$, $a_{11} = -a.b$ and $a_{ij} = 0$ if $(i,j) \notin \{(1,0),(0,1),(1,1)\}$, it can be shown that the MA representation of the MSAR(1) model is

$$b_{k,l} = \begin{cases} a^k b^l, & \text{if} \quad k \geq 0, l \geq 0 \\ 0 & \text{if} \quad k < 0 \text{ or } l < 0. \end{cases} \tag{9}$$

In the same way, the best linear predictor of $X(0,0)$ based on future observations is

$$\hat{X}_{h_1,h_2}(0,0) = \frac{1}{a.b}\left(aX(0,1) + bX(1,0) - X(1,1)\right),$$

and then the suboptimal predictor is

$$\tilde{X}(0,0) = \alpha\hat{X}(0,0) + \beta\hat{X}_{h_1,h_2}(0,0). \tag{10}$$

The best linear interpolator of $X(0,0)$ by performing the extension of the prediction.
Their interpolation is a linear combination of 8 data points in the nearest neighborhood to the prediction point:

$$\tilde{X}(0,0) = \frac{1}{1+a^2+b^2+a^2b^2}\left\{\left(a - ab^2\right)\left(X(-1,0) + X(1,0) + \left(b + a^2b\right)\right.\right. \tag{11}$$

The values of a and b that satisfy (9) are obtained by numerical approximation using software R.

Listing 1 R code.

```
library(rootSolve)
fun<-function(x,y)  1/((1-x^2)*(1-y^2))-
((1-(1-((1-x^2)*(1-y^2)*(x^2+y^2-x^2*y^2)))^2)/
(1+(1-(1-x^2)*(1-y^2)*(x^2+y^2+x^2*y^2))^2))*
(2/((1-x^2)*(1-y^2)*(x^2+y^2-x^2*y^2)) - 1/(x^2+y^2+x^2*y^2))

for (x in seq(0.05, 0.95, by = 0.05)){
      myfun<- function(y) fun(x,y)
      Eq<- uniroot.all(myfun,c(-1,1))
      print(paste(x,Eq), sep=' ; ')
}
```

3.1 Simulation study

To demonstrate the validity of Theorem 5, we present here a simulation study carried out using the statistical software R 3.5.4. The steps involved in the computation of the estimates for the linear predictors are summarized in Algorithm 1.

Algorithm 1 Prediction Algorithm

for $a \in seq(0.05, 0.95, by = 0.05)$ **do**
  Set the values b using the program below.
  for $(n, m) \in \{(100, 150), (150, 200), (250, 250)\}$ **do**
    for $j \in \{1 : rep = 1000\}$ **do**
      - Generate a data as a $n \times m$ rectangular grid from a spatial models of the form (8) where $\{\epsilon(s,t)\}$ is Gaussian white noise process with mean 0 and variance $\sigma^2 = 1$.
      - Perform an indices change to locate $X(0,0)$ inside the $n \times m$ grid.
      - Calculate the value of the optimal linear predictor $\tilde{X}(0,0)$ by using (**??**) and calculate $\nu_j^{(1)} = \left(X(0,0) - \tilde{X}(0,0)\right)^2$.
      - Calculate the value of the suboptimal linear predictor $\tilde{X}(0,0)$ by using (10) where α and β are given by(11) and caluclate $\nu_j^{(2)} = \left(X(0,0) - \tilde{X}(0,0)\right)^2$.

    - Calculate the estimate of the prediction error variance (PEV) of $\tilde{X}(0,0)$ by $\nu_1 = \dfrac{1}{rep}\sum_{j=1}^{rep} \nu_j^{(1)}$ and the estimate of PVE of $\tilde{X}(0,0)$ by $\nu_2 = \dfrac{1}{rep}\sum_{j=1}^{rep} \nu_j^{(2)}$.

Not surprisingly, the predictions obtained based on $\{X(s,t); 1 \leq s \leq 144, 1 \leq t \leq 161\}$, after estimating the missing data, improve the quality of the prediction. The impact of these missing values depends on the horizons of the predictions. In fact, given the results in Table 3, it is clear that the impact of missing data increases as prediction steps h_1 and h_2 increase. Also, the examination of the results in Table 3 reveals a certain symmetry in the values of the impact considered. Indeed, the values of the impact for the horizons (h_1, h_2) and (h_2, h_1) are very close.

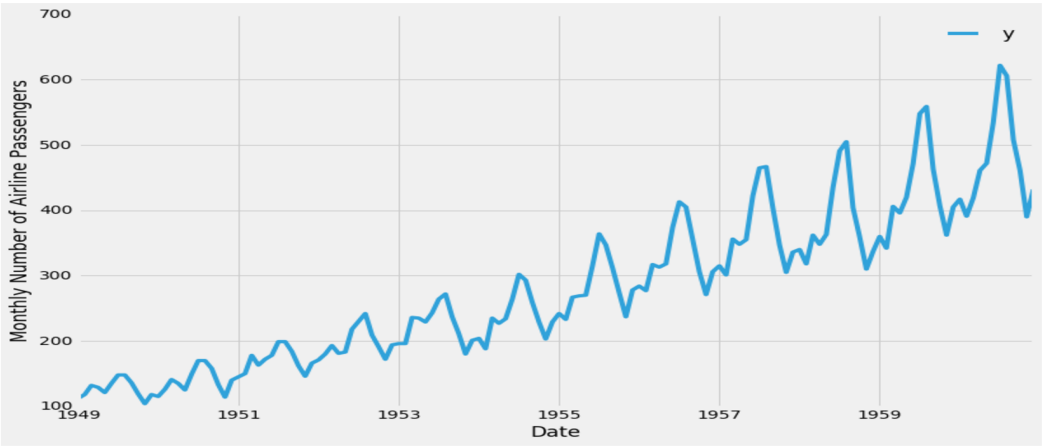

Figure 1 Error prediction

		(n,m)=(100,150)		(n,m)=(150,200)		(n,m)=(250,250)	
a	b	ν_1	ν_2	ν_1	ν_2	ν_1	ν_2
0.05	0.557	0.652	0.650	0.647	0.644	0.641	0.639
0.10	0.569	0.691	0.687	0.638	0.640	0.631	0.627
0.15	0.582	0.697	0.692	0.621	0.622	0.615	0.611
0.20	0.594	0.701	0.698	0.620	0.616	0.608	0.602
0.25	0.603	0.709	0.707	0.621	0.619	0.609	0.613
0.30	0.608	0.661	0.669	0.592	0.582	0.581	0.576
0.35	0.607	0.603	0.609	0.607	0.598	0.594	0.591
0.40	0.601	0.599	0.603	0.589	0.596	0.576	0.573
0.45	0.587	0.579	0.588	0.580	0.573	0.567	0.561
0.50	0.562	00567	0.571	0.559	0.562	0.546	0.549
0.55	0.516	0.553	0.560	0.601	0.609	0.556	0.559
0.60	0.406	0.542	0.549	0.536	0.540	0.531	0.532

Table 1 Prediction error variance for several values of a and b.

4 Conclusion

he purpose of this paper is to treat the prediction problems where a number of observations are missing to the quarter-plane past of a stationary random field. Our aim is to quantify the influence of missing values on the prediction by giving the simple bounds for the prediction error variance. These bounds allow to characterize the random fields for which the missing observations do not affect the prediction. Simulation experiments and an application to real data are presented.

References

1 O. Arezki and A. Hamaz. On linear prediction for stationary random fields with nonsymmetrical half-plane past. *Communications in Statistics - Theory and Methods*, 51(15):5298–5309, 2022. doi:10.1080/03610926.2020.1837880.

2 Lianfu Chen, Mohsen Pourahmadi, and Mehdi Maadooliat. Regularized multivariate regression models with skew-t error distributions. *Journal of Statistical Planning and Inference*, 149:125–139, 2014. doi:https://doi.org/10.1016/j.jspi.2014.02.001.

3 Abdelghani Hamaz, Ouerdia Arezki, and Farida Achemine. Impact of missing data on the prediction of random fields. *Journal of Applied Statistics*, 47(1):132–149, 2020. doi:https://doi.org/10.1080/02664763.2019.1633286.

4 P. Kohli and M. Pourahmadi. Some prediction problems for stationary random fields with quarter-plane past. *Journal of Multivariate Analysis*, 127:112–125, 2014. doi:https://doi.org/10.1016/j.jmva.2014.02.009.

A Appendix : Styles of lists, enumerations, and descriptions

List of different predefined enumeration styles:

- `\begin{itemize}...\end{itemize}`
- ...
- ...

1. `\begin{enumerate}...\end{enumerate}`
2. ...
3. ...

(a) `\begin{alphaenumerate}...\end{alphaenumerate}`
(b) ...
(c) ...

(i) `\begin{romanenumerate}...\end{romanenumerate}`
(ii) ...
(iii) ...