Nonlinear Analysis: Modelling and Control, Vol. 26, No. 2, 363–374
https://doi.org/10.15388/namc.2021.26.22359

Vilnius University Press

Classification of Gaussian spatio-temporal data with stationary separable covariances

Marta Karaliutė[a], **Kęstutis Dučinskas**[a,b]

[a]Institute of Data Science and Digital Technologies, Vilnius University
Akademijos str. 4, LT-08412 Vilnius, Lithuania
marta.karaliute@mif.vu.lt

[b]Faculty of Marine Technologies and Natural Sciences, Klaipėda University
Herkaus Manto str. 84, LT-92294 Klaipėda, Lithuania

Received: March 23, 2020 / **Revised:** November 23, 2020 / **Published online:** March 1, 2021

Abstract. The novel approach to classification of spatio-temporal data based on Bayes discriminant functions is developed. We focus on the problem of supervised classifying of the spatio-temporal Gaussian random field (GRF) observation into one of two classes specified by different drift parameters, separable nonlinear covariance functions and nonstationary label field. The performance of proposed classification rule is validated by the values of local Bayes and empirical error rates realized by leave one out procedure. A simulation study for spatial covariance functions belonging to powered-exponential family and temporal covariance functions of AR(1) models is carried out. The influence of the values of spatial and temporal covariance parameters to error rates for several label field models are studied. The results showed that the proposed classification methodology can be applied successfully in practice with small error rates and can be a useful tool for discriminant analysis of spatio-temporal data.

Keywords: separable covariance function, Bayes discriminant function, powered-exponential family.

1 Introduction

Spatial supervised classification is a problem of labeling observations based on feature information and information about spatial adjacency relationships with training sample. Switzer [25] was the first to treat classification of spatial data. Atkinson and Lewis [1] reviewed geostatistical techniques for classification of remotely sensed images. De Oliveira [20] proposed spatial classification techniques based clipping of Gaussian random fields. Spatial contextual classification problems arising in geospatial domain is considered by Shekhar et al. [22]. It is usually assumed that feature observations conditional on labels are independent (conditional independence) and normally distributed and the labels follow the random field (RF) model. This approach is widely used in image classification Nishii and Eguchi [19]. Dučinskas [9, 10] proposed and explored Bayes classification rules for spatial Gaussian data by avoiding the assumption of conditional independence. Comprehensive overview of methods for statistical classification and discrimination of Gaussian spatial data is provided by Berrett and Calder [3]. The novel approach to classification of Gaussian Markov random fields observation is developed by Dučinskas and Dreižienė [11]. Critical comparison of spatial linear mixed models for ecological data based on the correct classification rates is performed by Dreižienė and Dučinskas [8].

Some authors have investigated the performance of the Bayes classification rules (BCR) when training samples consist of temporally dependent observations (see, e.g., [16, 18]).

Spatio-temporal data are often collected at monitored discrete time lags in locations belonging to continuous area. Such type of data sets is usually viewed as a spatial time series (see, e.g., [7]).

Valid and practical covariance structures are needed to model these types of data sets in various disciplines such as environmental science, climatology and agriculture. Usually, in environmental and agricultural research, the data are recorded at regular time intervals (time lags) and at irregular stations (locations) in compact area (see, e.g., [14]). Recently, deep learning methods via convolutional neural networks have been intensively explored and used in image analysis and spatial data mining (see, e.g., [2, 27–31]).

However, statistical discriminant analysis of spatio-temporal data has been rarely considered previously (see, e.g., [15]). Šaltytė-Benth and Dučinskas [26] considered classification of spatio-temporal data modeled by GRF in particular case when observation of feature at focal location is uncorrelated with the training sample that consists of interdependent feature variables.

In the present paper, avoiding this restriction, we focus on the classification of data modeled by random fields with separable spatio-temporal covariance structures specified by geostatistical spatial margins and discrete temporal margins (see, e.g., [6]). Separability of covariances was assumed for the sake of reduction of complexity due to interdependencies between features.

The main distinctive feature of proposed approach is the allowing label field to be nonstationary in time for each location, i.e., class label at each location can vary in time. That essentially widens the application area of presented investigations.

For the performance of classifiers, the values of derived in local Bayes error rates and empirical error rates are used. Empirical error rates are validated by modified leave-one-out method when all but one observation is used to when complete the classification rule, and this rule is then used to classify the omitted observation (see, e.g., [12]). For numerical illustrations, the two powered-exponential isotropic models for spatial covariance are considered. Temporal covariance is obtained by the Yule–Walker equations for AR(1) models. Performance of proposed classification rule is compared for different parameters of pure spatial and temporal covariances and prior class probabilities models.

This paper is organized as follows: proposed spatio-temporal data models and conditional distributions are delivered in the next section; in Section 3, conditional Bayes classification rules and its error rate is presented; in Section 4, the numerical illustrations and simulations for various separable stationary spatio-temporal covariance and prior probabilities models are displayed, and finally, the conclusions are in the last section.

2 Spatio-temporal data models and conditional distributions

The main objective of this paper is to classify observations of GRF $\{Z(s;t)\colon s \in D \subset \mathbb{R}^2, t \in D_T = [0, \infty]\}$, where s and t define spatial and temporal coordinates, respectively. Let $\{Y(s;t)\colon s \in D \subset \mathbb{R}^2, t \in D_T\}$ be a random field that represents class label and takes only the value 0 or 1 (see, e.g., [23]).

In this study, we assume that for $l = 0, 1$, the model of observation $Z(s;t)$ *conditional on* $Y(s;t) = l$ is $Z(s;t) = \mu_l(s;t) + \varepsilon(s;t)$, where $\mu_l(s;t)$ – deterministic spatio-temporal trend. The error term is assumed to be generated by the univariate zero-mean GRF $\{\varepsilon(s;t)\colon s \in D \subset \mathbb{R}^2, t \in T\}$ with covariance function defined by model $\operatorname{cov}(\varepsilon(s;t), \varepsilon(u;r)) = C(s,u;t,r)$ for all $s, u \in D$ and $t, r \in T$.

In present paper, we restrict our attention to the separable spatio-temporal covariance model $C(s,u;t,r) = C_S(s,u)C_T(t,r)$, *where* $C_S(s,u)$ *denotes pure spatial covariance between observations in locations* s *and* u, *and* $C_T(t,r)$ *denotes pure temporal covariance between observations at time points* t *and* r. Under this assumption, the spatio-temporal covariance structure factors into a purely spatial and a purely temporal component, which allows for computationally efficient estimation and inference. Consequently, separable covariance models have been popular even in situations in which they are not physically justifiable. Many statistical tests for separability have been proposed recently and are based on parametric models (see, e.g., [5, 13]) or spectral methods [21].

Let $S_n = \{s_i \in D,\ i = 1, \dots, n\}$ be a set of locations, where observations are taken at time $t \in D_p = \{1, 2, \dots, p, p+1\}$. At every moment of time $t \in D_p$, the set S_n is split into two classes, $S_t^{(0)}$ and $S_t^{(1)}$ (i.e., $S_n = S_t^{(0)} \cup S_t^{(1)}$): $S_t^{(l)} = \{s \in S_n\colon Y(s,t) = l\}$, $l = 0, 1$.

Denote n_{lt} the number of locations (of n) at time t that belong to class l; thus n_{lt} is the number of points in the set $S_t^{(l)}$, and $n = n_{0t} + n_{1t}$ for every $t \in D_p$. Hence a set of class labels at any time moment can differ in composition.

Joint training sample Z is stratified training sample specified by $n \times p$ matrix $Z = (Z_1, \dots, Z_p)$, where $Z_t = (Z(s_1,t), \dots, Z(s_n,t))'$. This structure of data presentation is motivated by a model that assumes multivariate (in space) time series. Denote by $z_t = (z_t^1, \dots, z_t^n)$ and $y_t = (y_t^1, \dots, y_t^n)$ the realized value of Z_t and $Y_t = (Y(s_1,t), \dots, Y(s_n,t))'$, respectively.

In what follows, with an insignificant loss of generality, we focuse on the linear independent of time drift $\mu_l(s;t) = \beta_l' x(s)$, *where* $x(s) = (x_1(s), \dots, x_q(s))'$ *is the vector of a spatial covariates, and* β_l *is a q-dimensional vector of parameters,* $l = 0, 1$.

Denote by X the $n \times 2qp$ matrix $X = (X_{(1)}, X_{(2)}, \dots, X_{(p)})$, where

$$X_{(t)} = \begin{pmatrix} x_1'(1 - y_t^1) & x_1' y_t^1 \\ x_2'(1 - y_t^2) & x_2' y_t^2) \\ \vdots & \vdots \\ x_n'(1 - y_t^n) & x_n' y_t^n \end{pmatrix},$$

and $x_i = x(s_i)$, $i = 1, \dots, n$.

Then the matrix model for Z conditional on $\{Y_t = y_t,\ t = 1, \dots, p\}$ is $Z = XB + E$, where $B = I_p \otimes \beta$ with $\beta = (\beta_0', \beta_1')'$, and $n \times p$ matrix of Gaussian errors $E = (\varepsilon(s_i; t)\colon\ i = 1, \dots, n,\ t = 1, \dots, p)$. Here I_p is $p \times p$ identity matrix, and β is $2q \times 1$ vector of parameters.

Denote pure spatial covariance $n \times n$ matrix by $C_S = (c_S^{ij} = C_S(s_i, s_j),\ i, j = 1, \dots, n)$. In practice, it usually belongs to Matern class [17, Sect. 3.2] or powered-exponential class of covariance functions (see, e.g., [24, p. 31], [6, Sect. 4.1.1]).

Then the model of training sample $M = \operatorname{vec}(Z)$ conditional on $Y_t = y_t$, $t = 1, \dots, p$, is

$$M = \operatorname{vec}(XB) + \operatorname{vec}(E), \tag{1}$$

where $\operatorname{vec}(E)$ is the $np \times 1$ vector of random errors that has normal distribution, i.e., $\operatorname{vec}(E) \sim N_{np}(0, \Sigma)$ with $\Sigma = \operatorname{var}(\operatorname{vec}(E)) = C_T \otimes C_S$, and C_T is a $p \times p$ matrix of pure temporal covariances, $C_T = (c_T^{tr} = C_T(t, r),\ t, r = 1, \dots, p)$.

In present paper, we concern with the problem of classification of the observations $Z(s_i,\, p+1)$, $i = 1, \dots, n$, into one of two classes with given joint training sample M or, in other words, based on training sample information we want to predict label at an unobserved location $t = p + 1$.

Set $c_T^{p+1,r} = C_T(p+1, r)$, $r = 1, \dots, p$, $c_T^{p+1} = (c_T^{p+1,1}, \dots, c_T^{p+1,p})'$ and e_i' – the ith row of identity matrix I_n.

Under spatio-temporal data model specification, we can conclude that in $l = 0, 1$, the conditional distribution of $Z(s_i,\, p+1)$ given $M = m$ and $Y(s_i,\, p+1) = l$, is Gaussian, i.e.,

$$\bigl(Z(s_i,\, p+1) \bigm| M = m;\ Y(s_i,\, p+1) = l\bigr) \sim N\bigl(\mu_{li(m)}^{p+1}, \Sigma_{p+1,\, i(m)}\bigr), \tag{2}$$

where

$$\mu_{li(m)}^{p+1} = \beta_l' x_i + \bigl(\bigl(c_T^{p+1}\bigr)' C_T^{-1} \otimes e_i'\bigr) \operatorname{vec}(E),$$
$$\Sigma_{p+1,\, i(m)} = \operatorname{var}\bigl(Z(s_i,\, p+1)\bigr) - c_s^{ii} \bigl(c_T^{p+1}\bigr)' C_T^{-1} c_T^{p+1} = c_S^{ii} \rho_{p+1}$$

with $\rho_{p+1} = c_T^{p+1,p+1} - (c_T^{p+1})' C_T^{-1} c_T^{p+1}$.

In this study, we assume that the conditional distribution of label $Y(s_i,\, p+1)$, $i = 1, \dots, n$, given joint training sample M depends only on class labels values, i.e., conditional distribution of $(Y(s_i,\, p+1) = l \mid M = m)$ is identical to conditional distribution of $(Y(s_i,\, p+1) = l \mid \{Y_t = y_t,\ t = 1, \dots, p\})$.

This assumption is quite frequently used by image classification researches (see, e.g., [19]). Set $\mathbf{P}(Y(s_i,\, p+1) = l \mid M = m) = \pi_l(s_i,\, p+1)$, $l = 0, 1$, and shortly call them *prior class probabilities*.

3 Conditional Bayes discriminant functions and its error rate

Under the assumption that the classes are completely specified, the conditional Bayes discriminant function (CBDF) minimizing the probability of misclassification is formed by the log-ratio of conditional likelihood of distribution specified in (1)–(2), that is

$$W\big(Z(s_i,\, p+1)\big) = \left(Z(s_i,\, p+1) - \frac{\mu^{p+1}_{1i(m)} + \mu^{p+1}_{0i(m)}}{2}\right) \times \Sigma^{-1}_{p+1,\, i(m)}\big(\mu^{p+1}_{1i(m)} - \mu^{p+1}_{0i(m)}\big) + \gamma_i(p+1), \tag{3}$$

where $\gamma_i(p+1) = \ln(\pi_1(s_i,\, p+1)/\pi_2(s_i,\, p+1))$.

It is easy to deduce that discriminant function $W(Z(s_i,\, p+1))$ is optimal under the criterion of the minimum of misclassification probability (see [18]).

Call the probability of misclassification for $W(Z(s_i,\, p+1))$ as local Bayes error rate and denote it by P_i. Also, denote squared *Mahalanobis distance between conditional distributions by*

$$\Delta^2_{p+1,\, i(m)} = \big(\mu^{p+1}_{1i(m)} - \mu^{p+1}_{0i(m)}\big)' \Sigma^{-1}_{p+1,\, i(m)}\big(\mu^{p+1}_{1i(m)} - \mu^{p+1}_{0i(m)}\big).$$

Lemma 1. *The local Bayes error rate is*

$$P_i = \pi_1(s_i,\, p+1)\Phi\left(-\frac{\Delta_{p+1,i(m)}}{2} - \frac{\gamma_i(p+1)}{\Delta_{p+1,\, i(m)}}\right) + \pi_0(s_i,\, p+1)\Phi\left(-\frac{\Delta_{p+1,\, i(m)}}{2} + \frac{\gamma_i(p+1)}{\Delta_{p+1,\, i(m)}}\right),$$

where $\Phi(x)$ is the standard normal cumulative distribution function.

Proof. It is easy to derive that conditional distribution of $W(Z(s_i,\, p+1))$ given $M = m$, $Y(s_i,\, p+1) = l$ is univariate Gaussian distribution with mean

$$\mathbf{E}\big(W\big(Z(s_i,\, p+1)\big) \mid M = m,\ Y(s_i,\, p+1) = l\big) = (-1)^{l+1}\frac{\Delta^2_{p+1,\, i(m)}}{2} + \gamma_i(p+1)$$

and variance

$$\operatorname{var}\big(W\big(Z(s_i,\, p+1)\big) \mid M = m,\ Y(s_i,\, p+1) = l\big) = \Delta^2_{p+1,\, i(m)}, \quad l = 1, 2.$$

Using properties of the multivariate Gaussian distribution, we complete the proof. □

Error estimation is critical to classification because the validity of the resulting classifier model, composed of the classifier and its error estimate, is based on the accuracy of the error estimation procedure. Given a set of sample data, the data can be split between training and test data with a classifier being designed on the training data and its error being validated on the test data. In this paper, our focus is on using p temporal observations for training and the observations at $p + 1$th time moment is using for testing.

Performance of the classification rule based on $W(Z(s_i,\, p+1))$ *could be evaluated by several methods* (e.g., [12]). In the present study, we prefer the leave-one-out estimator or procedure when all but one (test observation) observation is used to complete the classification rule, and this rule (based on CBDF) is then used to classify the omitted observation. This procedure consists of simulating a sample of v independent values of $Z(s_i,\, p+1)$, denoted by $\{Z^j(s_i,\, p+1),\ j = 1,\dots,v\}$, drawn from conditional distribution specified in (2) with prescribed labels $Y(s_i,\, p+1)$.

For $i = 1,\dots,n$, define the empirical error rate by

$$LO_i = \frac{1}{v}\left(\sum_{j=1}^{v} \bigl|Y(s_i,\, p+1) - \widehat{Y}^j(s_i,\, p+1)\bigr|\right),$$

where $\widehat{Y}^j(s_i,\, p+1) = H(W(Z^j(s_i,\, p+1)))$, and $H(\cdot)$ is the Heaviside step function.

4 Numerical illustrations and simulations

For numerical illustrations of obtained results, we considered the Gaussian spatio-temporal model with pure spatial covariances belonging to the family of powered-exponential isotropic models and with pure temporal covariance of AR(1) model. It is known that for this model, $c_T^{1,1} = c_T^{t,t}$ for $t = 2,\dots,p+1$, parameter α quantifies temporal dependency by equation $c_T^{1,1} = \sigma_T^2/(1-\alpha^2)$, where σ_T^2 is the white noise variance.

Then $c_T^{p+1} = (\sigma_T^2/(1-\alpha^2))(\alpha^p,\ \alpha^{p-1}\ \cdots\ \alpha)$, and the inverse of temporal covariance matrix C_T is obtained by the Yule–Walker equations (see [4]).

Temporal covariance matrix C_T^{-1} is obtained by the Yule–Walker equations for AR(1) model, i.e.,

$$C_T^{-1} = \frac{1}{\sigma_T^2}\begin{pmatrix} 1 & -\alpha & \cdots & 0 & 0 \\ -\alpha & 1+\alpha^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1+\alpha^2 & -\alpha \\ 0 & 0 & \cdots & -\alpha & 1 \end{pmatrix}.$$

It is easy to derive that $(c_T^{p+1})' C_T^{-1} = \alpha e_p'$ and $\rho_{p+1} = \sigma_T^2$, where e_p' denotes the pth row of identity matrix I_p.

Hence $\mu_{li(m)}^{p+1} = \beta_l' x_i + \alpha(e_p' \otimes e_i')\operatorname{vec}(E)$ and $\Sigma_{p+1,i(m)} = c_S^{ii}\sigma_T^2$. Here α is AR(1) model parameter that quantifies temporal dependency, and σ_T^2 is the white noise variance for this model.

In the study, two isotropic nugetless spatial covariance structures belonging to the powered-exponential family are considered. Assuming that $C_s = \sigma_s^2 R$, where $R = (r_{ij})$ is spatial correlation matrix, we concern on the following two particular cases:

(i) exponential case with $r_{ij} = r(|s_i - s_j|) = \mathrm{e}^{-|s_i - s_j|/\varphi}$;

(ii) squared-exponential case $r_{ij} = r(|s_i - s_j|) = \mathrm{e}^{-(|s_i - s_j|/\varphi)^2}$.

Here φ is the so called range parameter that represents the spatial dependence.

This choice of is based on the smoothness level of sample paths. Sample paths of a GRF with the exponential covariance function are not smooth when the squared exponential covariance model has smooth sample paths.

Two methods for prior class probabilities is proposed.
First one is based on temporal weighted moving average (TWMA) method

$$\pi_{1t}(s_i,\, p+1) = \frac{\sum_{t=1}^{p} y_i^t t}{(1+p)p/2}.$$

Second one adds spatial correlations for weighting

$$\pi_{1ts}(s_i,\, p+1) = \frac{\sum_{t=1}^{p} y_i^t t + \sum_{t=1}^{p} y_{i_0}^t t r_{ii_0}}{(1+p)p/2 + r_{ii_0}(1+p)p/2},$$

where i_0 denotes the index of the nearest neighbor to s_i. Denote this method by (STWMA).

We have compared these four particular cases by calculating the P_i and LO_i for $i = 1, \dots, n$, and we have presented them in tables.

Numerical illustrations are performed on 20 locations on two dimensional area that are depicted in Fig. 1. Class labels for 20 locations and 4 time points in training sample is presented in Table 1.

Local Bayes error rates P_i and their averages $AP = \sum_{i=1}^{20} P_i/20$ for two cases of spatial covariances and two models for prior probabilities are presented in Table 2.

As it might be seen from Table 2, for $\alpha = 0.1, 0.3$, classifiers with STWMA priors in majority locations have an advantage against cases with TWMA priors for both spatial covariance models. For large α values, significant difference between these two is not observed.

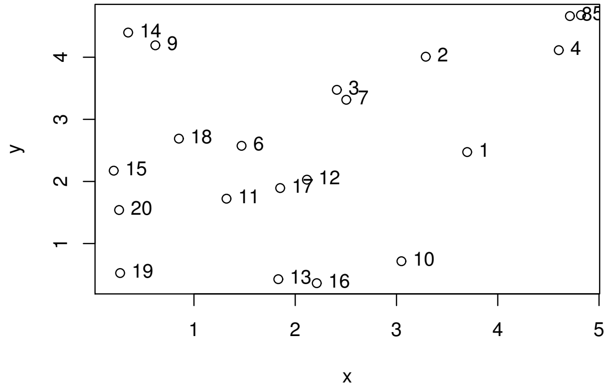

Figure 1. Spatial sampling set S_{20}.

Table 1. Class labels for 20 locations (i) at 4 time moments (t).

t	i																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0

Table 2. Local and average Bayes error rates for $\Delta = |\mu^{p+1}_{1i(m)} - \mu^{p+1}_{0i(m)}| = 1$, $\varphi = 3$ and various α.

i	$\pi_{1t}(s_i, p+1)$				$\pi_{1ts}(s_i, p+1)$			
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
	exponential case							
1	0	0	0	0.23	0	0.01	0.1	0.3
2	0	0	0.02	0.02	0	0	0	0.19
3	0	0	0	0.35	0	0.02	0.16	0.35
4	0.02	0.02	0.02	0.32	0.02	0.02	0.15	0.32
5	0	0.23	0.23	0.23	0	0.11	0.28	0.28
6	0	0	0	0.32	0	0	0.14	0.34
7	0	0.11	0.35	0.35	0	0.02	0.18	0.35
8	0.02	0.02	0.32	0.32	0	0.11	0.28	0.28
9	0	0	0.32	0.32	0	0	0.12	0.35
10	0	0	0	0	0	0	0.04	0.04
11	0	0	0	0.02	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	0.35	0.35	0.35	0	0.19	0.32	0.32
14	0	0	0	0.32	0	0	0.1	0.35
15	0	0	0	0	0	0	0	0
16	0	0	0.23	0.23	0	0.15	0.31	0.31
17	0	0	0	0	0	0	0	0
18	0.02	0.02	0.35	0.35	0	0	0.2	0.35
19	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
AP	0.003	0.038	0.11	0.187	0.001	0.032	0.119	0.207
	squared-exponential case							
1	0	0	0	0.23	0	0.02	0.17	0.32
2	0	0	0.02	0.02	0	0	0	0.23
3	0	0	0	0.35	0	0.02	0.17	0.35
4	0.02	0.02	0.02	0.32	0.02	0.02	0.17	0.32
5	0	0.23	0.23	0.23	0	0.11	0.28	0.28
6	0	0	0	0.32	0	0	0.17	0.34
7	0	0.11	0.35	0.35	0	0.02	0.17	0.35
8	0.02	0.02	0.32	0.32	0	0.11	0.28	0.28
9	0	0	0.32	0.32	0	0	0.11	0.35
10	0	0	0	0	0	0	0.06	0.06
11	0	0	0	0.02	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	0.35	0.35	0.35	0	0.17	0.32	0.32
14	0	0	0	0.32	0	0	0.11	0.35
15	0	0	0	0	0	0	0	0
16	0	0	0.23	0.23	0	0.17	0.32	0.32
17	0	0	0	0	0	0	0	0
18	0.02	0.02	0.35	0.35	0	0	0.17	0.34
19	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0
AP	0.003	0.038	0.11	0.187	0.001	0.032	0.125	0.211

For $v = 30$ *independent replications, local empirical error rates* LO_i *and their averages* $ALO = \sum_{i=1}^{20} LO_i/20$ *for two cases of spatial covariances and two models for prior class probabilities are presented in Table* 3.

Table 3. Local and average empirical error rates for $\Delta = |\mu^{p+1}_{1i(m)} - \mu^{p+1}_{0i(m)}| = 1$, $\varphi = 3$ and various α.

i	$\pi_{1t}(s_i,\, p+1)$				$\pi_{1ts}(s_i,\, p+1)$			
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
	exponential case							
1	0	0	0.03	0	0	0	0.03	0
2	0	0	0.03	0.10	0	0	0.03	0
3	0	0	0.03	0	0	0	0	0
4	0	0	0.03	0	0	0	0.03	0
5	0	0	0	0	0	0	0	0
6	0	0	0.03	0	0	0	0.03	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0.07	0.17	0	0	0.07	0.10
11	0	0	0.03	0.33	0	0	0.03	0.33
12	0	0	0	0.23	0	0	0	0.23
13	0	0	0	0	0	0	0	0
14	0	0	0	0.13	0	0	0	0.03
15	0	0	0.13	0.37	0	0	0.13	0.37
16	0	0	0	0	0	0	0	0
17	0	0	0.10	0.23	0	0	0.10	0.23
18	0	0	0	0	0	0	0	0
19	0	0	0.07	0.27	0	0	0.07	0.27
20	0	0	0.13	0.20	0	0	0.13	0.20
ALO	0	0	0.034	0.102	0	0	0.033	0.088
	squared-exponential case							
1	0	0	0.03	0	0	0	0.03	0
2	0	0	0.03	0.10	0	0	0.03	0
3	0	0	0.03	0	0	0	0	0
4	0	0	0.03	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0.03	0	0	0	0.03	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0.07	0.17	0	0	0.07	0.10
11	0	0	0.03	0.33	0	0	0.03	0.33
12	0	0	0	0.23	0	0	0	0.23
13	0	0	0	0	0	0	0	0
14	0	0	0	0.13	0	0	0	0
15	0	0	0.13	0.37	0	0	0.13	0.37
16	0	0	0	0	0	0	0	0
17	0	0	0.10	0.23	0	0	0.10	0.23
18	0	0	0	0	0	0	0.03	0
19	0	0	0.07	0.27	0	0	0.07	0.27
20	0	0	0.13	0.20	0	0	0.13	0.20
ALO	0	0	0.034	0.102	0	0	0.033	0.087

As it might be seen from Table 3, for all values of α, classifiers with STWMA and TWMA in majority locations have the similar empirical error rates for both spatial covariance models.

The last raw of Tables 2 and 3 (i.e., AP and ALO) allow us to compare averages of Bayes and empirical error rates for various combinations of spatial covariance and prior class probability models and to make optimal decisions in construction for the classifiers of spatio-temporal Gaussian data.

5 Conclusions

In this paper, we propose approach to classification of spatio-temporal data in the framework of Bayes discriminant for separable spatio-temporal covariance case stations. Several simulation studies were conducted to estimate and compare empirically the classifiers for various separable stationary spatio-temporal covariance and prior class probabilities models. Numerical analysis showed that:

(i) Bayes and empirical error rates increases when temporal correlation increases;
(ii) Incorporation spatial correlation in class prior probabilities improves the performances of classifiers;
(iii) Classifiers with spatial squared-exponential covariance have an advantage against classifiers with exponential covariance.

The results of performed calculations in all examples give us the strong argument to encourage the users do not ignore the spatial, temporal dependency and locational information from training sample in classification of spatio-temporal data and to apply the proposed approach in deep learning for spatio-temporal data mining.

References

1. P.M. Atkinson, P. Lewis, Geostatistical classification for remote sensing: An introduction, *Comput. Geosci.*, **26**(4):361–371, 2000, https://doi.org/10.1016/S0098-3004(99)00117-X.

2. G. Atluri, A. Karpatne, V. Kumar, Spatio-temporal data mining: A survey of problems and methods, *ACM Comput. Surv.*, **51**(4):83, 2018, https://doi.org/10.1145/3161602.

3. C. Berrett, C.A. Calder, Bayesian spatial binary classification, *Spatial Stat.*, **16**:72–102, 2016, https://doi.org/10.1016/j.spasta.2016.01.004.

4. P.J. Brockwell, R.A. Davis, *Time Series: Theory and Methods*, Springer, New York, 2009, https://doi.org/10.1007/978-1-4419-0320-4.

5. P.E. Brown, K.F. Karesen, G.O. Roberts, S. Tonellato, Blur-generated non-separable space-time models, *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, **62**(4):847–860, 2000, https://doi.org/10.1111/1467-9868.00269.

6. N. Cressie, C.K. Wikle, *Statistics for Spatio-Temporal Data*, Wiley, Hoboken, NJ, 2011.

7. S.S. Demel, J. Du, Spatio-temporal models for some data sets in continuous space and discrete time, *Stat. Sin.*, **25**:81–98, 2015, https://doi.org/10.5705/ss.2013.223w.

8. L. Dreižienė, K. Dučinskas, Comparison of spatial linear mixed models for ecological data based on the correct classification rates, *Spatial Stat.*, **35**, 2020, https://doi.org/10.1016/j.spasta.2019.100395.
9. K. Dučinskas, Approximation of the expected error rate in classification of the Gaussian random field observations, *Stat. Probab. Lett.*, **79**(2):138–144, 2009, https://doi.org/10.1016/j.spl.2008.07.042.
10. K. Dučinskas, Error rates in classification of multivariate Gaussian random field observation, *Lith. Math. J.*, **51**(4):477–485, 2011, https://doi.org/10.1007/s10986-011-9142-4.
11. K. Dučinskas, L. Dreižienė, Risks of classification of the Gaussian Markov random field observations, *J. Classif.*, **35**:422–436, 2018, https://doi.org/10.1007/s00357-018-9269-7.
12. I. Egbo, Evaluation of error rate estimators in discriminant analysis with multivariate binary variables, *Am. J. Theor. Appl. Stat.*, **5**(4):173–179, 2016, https://doi.org/10.11648/j.ajtas.20160504.12.
13. M. Fuentes, Testing for separability of spatial-temporal covariance functions, *J. Stat. Plann. Inference*, **136**:447–466, 2006, https://doi.org/10.1016/j.jspi.2004.07.004.
14. J. Haslett, A.E. Raftery, Space-time modelling with long memory dependence: Assessing Ireland's wind power resourse, *Appl. Stat.*, **38**(1):1–50, 1989, https://doi.org/10.2307/2347679.
15. B. Jeon, D. Landgrebe, Classification with spatiotemporal interpixel class dependency contexts, *IEEE Trans. Geosci. Remote Sens.*, **30**(4):663–672, 1992, https://doi.org/10.1109/36.158859.
16. C.R.O. Lawoko, G.L. McLachlan, Discrimination with autocorrelated observations, *Pattern Recognition*, **18**:145–149, 1985, https://doi.org/10.1016/0031-3203(85)90038-X.
17. B. Matern, *Spatial Variation,* 2nd ed., Springer, New York, 1986, https://doi.org/10.1007/978-1-4615-7892-5.
18. G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 2004, https://doi.org/10.1002/0471725293.
19. R. Nishii, S. Eguchi, Image classification based on Markov random field model with Jeffreys divergence, *J. Multivariate Anal.*, **97**(9):1997–2008, 2006, https://doi.org/10.1016/j.jmva.2006.01.009.
20. V. De Oliveira, Bayesian prediction of clipped Gaussian random fields, *Comput. Stat. Data Anal.*, **34**(3):299–314, 2000, https://doi.org/10.1016/S0167-9473(99)00103-6.
21. L. Scaccia, R.J. Martin, Testing axial symmetry and separability of lattice processes, *J. Stat. Plann. Inference*, **131**:19–39, 2005.
22. S. Shekhar, P.R. Schrater, R.R. Vatsavai, W. Wu, S. Chawla, Spatial contextual classification and prediction models for mining geospatial data, *IEEE Trans. Multimedia*, **4**(2):174–188, 2002, https://doi.org/10.1109/TMM.2002.1017732.

23. G. Stabingis, K. Dučinskas, L. Stabingienė, Comparison of spatial classification rules with different conditional distributions of class label, *Nonlinear Anal. Model. Control*, **19**(1): 109–117, 2014, https://doi.org/10.15388/NA.2014.1.7.

24. M.L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York, 1999, https://doi.org/10.1007/978-1-4612-1494-6.

25. P. Switzer, Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery, *Math. Geol.*, **12**(4):367–376, 1980, https://doi.org/10.1007/BF01029421.

26. J. Šaltytė Benth, K. Dučinskas, Linear discriminant analysis of multivariate spatial-temporal regressions, *Scand. J. Stat.*, **32**:281–294, 2005.

27. S. Wang, J. Cao, P. Yu, Deep learning for spatio-temporal data mining: A survey, *Deep Learning for Spatio-Temporal Data Mining: A Survey*, 2020, https://doi.org/10.1109/TKDE.2020.3025580.

28. Q. Zheng, X. Tian, N. Jiang, M. Yang, Layer-wise learning based stochastic gradient descent method for the optimization of deep convolutional neural network, *J. Intell. Fuzzy Syst.*, **37**(4): 5641–5654, 2019, https://doi.org/10.3233/JIFS-190861.

29. Q. Zheng, X. Tian, M. Yang, Y. Wu, H. Su, PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning, *Multidim. Syst. Signal Process.*, **31**(3): 793–827, 2020, https://doi.org/10.1007/s11045-019-00686-z.

30. Q. Zheng, M. Yang, X. Tian, N. Jiang, D. Wang, A full stage data augmentation method in deep convolutional neural network for natural image classification, *Discrete Dyn. Nat. Soc.*, **2020**: 4706576, 2020, https://doi.org/10.1155/2020/4706576.

31. Q. Zheng, M. Yang, J. Yang, Q. Zhang, X. Zhang, Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process, *IEEE Access*, **6**:15844–15869, 2018, https://doi.org/10.1109/ACCESS.2018.2810849.