COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 1, Number 2, 2013, 69-78

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966

STATISTICAL CLASSIFICATION OF GAUSSIAN SPATIAL DATA GENERATED BY CONDITIONAL AUTOREGRESSIVE MODEL

Kęstutis Dučinskas, Ingrida Borisenko, Indrė Šimkienė
Department of Statistics, Klaipėda University, Herkaus Manto str. 84,
LT-92294 Klaipėda, Lithuania
kestutis.ducinskas@ku.lt, ingrida.borisenko@ku.lt, indre.simkiene@gmail.com

Abstract: Given training sample, the problem of classifying Gaussian spatial data into one of two populations specified by conditional autoregressive model (CAR) with different mean functions is considered. This paper concerns classification procedures associated with Bayes Discriminant Function (BDF) under deterministic spatial sampling design. In the case of complete parametric certainty, the overall misclassification probability associated with BDF is derived. In this paper we develop further our methods of spatial classification to apply to the CAR case. Spatial weights based on inverse of Euclidean distance and the second and third order neighbourhood schemes on regular 2-dimensional lattice are used for illustrative examples. The effect of the spatial sampling design, Mahalanobis distances and prior probabilities on the performance of proposed classification procedure is numerically evaluated.
Keywords: Bayes Discriminant Function, Covariance function, Gaussian random field, misclassification probability, Training labels configuration.

Introduction

It is well known that in case of completely specified populations and known loss function, an optimal classification rule in the sense of minimum overall misclassification probability (OMP) is based on BDF (Anderson, 2003). Many authors (see e. g. Lawoko and Mclachlan, 1985; Kharin, 1996) have investigated the performance of the BDF in classification of dependent observations (stationary time series, equicorrelaton, Markov dependence, autoregressive models). Switzer (1980) was the first to treat classification of spatial data. Spatial discrimination for feature observations having elliptically contoured distributions is implemented in Batsidis and Zografos (2011). Saltyte and Ducinskas (2002) derived the formulas for error rates when classifying the observation of Gaussian random field into one of two classes with different regression mean models and common covariance function. However, in these publications, the observations to be classified are assumed to be independent from training samples. This is unrealistic assumption particularly when the locations of observations to be classified are close to ones of training sample.

The first extensions to the case when spatial correlations between Gaussian observations to be classified and observations in training sample are not assumed equal zero is done in Ducinskas (2009). Here only the spatial covariance functions belonging to Mattern class are considered. In this paper we develop further the latter investigation to the case of spatial Gaussian data specified by widely used CAR model, pioneered by Besag (1974).

We have derived an explicit expression of the OMP (Lema) and proposed to use it as performance measure of classification procedure. By using the derived OMP, the performance of the BDF is numerically analyzed in the case of stationary Gaussian random field on 2-dimentional regular lattice. The dependence of the values of obtained OMP on the Mahalanobis distance for different spatial sampling designs and prior distributions for class membership is investigated.

By applying the proposed criterion, the numerical comparison of some training labels configurations (TLC) is carried. That gives us the strong arguments for suggestion to include derived formulas of error rates in the geospatial data mining (Shekhar et al, 2002). The proposed BDF could also be considered as the extension of widely used Bayesian methods to the restoration of image corrupted by spatial Gaussian noise (Cressie, ch. 7.4, 1993).

1. The main concepts and definitions

The main objective of this paper is to classify the observations of Gaussian random field (GRF)

$$\{Z(s): s \in D \subset R^p\}.$$

The model of observation $Z(s)$ in population Ω_j is

$$Z(s) = \mu_j(s) + \varepsilon(s), \tag{1}$$

where $\mu_j(s)$ is a deterministic mean function , $j = 1, 2$ and error term $\varepsilon(s)$ is generated by zero-mean stationary Gaussian random field $\{\varepsilon(s): s \in D\}$ with covariance function defined by model for all $s, u \in D$:

$$\operatorname{cov}\{\varepsilon(s), \varepsilon(u)\} = C(s-u;\theta), \tag{2}$$

where $\theta \in \Theta$ is a $p \times 1$ parameter vector, Θ being an open subset of $\boldsymbol{R}^k$.

For given training sample, consider the problem of classification of the $Z_0 = Z(s_0)$ into one of two populations when

$$\mu_1(s_0) \neq \mu_2(s_0), s_0 \in D.$$

Denote by $S_n = \{s_i \in D;\ i = 1,..., n\}$ the set of locations where training sample $T' = (Z(s_1),..., Z(s_n))$ is taken, and call it the set of training locations (STL). It specifies the spatial sampling design or spatial framework for training sample. We shall assume the deterministic spatial sampling design and all analyses are carried out conditional on S_n.

Assume that S_n is partitioned into union of two disjoint subsets, i. e. $S_n = S^{(1)} \cup S^{(2)}$, where $S^{(j)}$ is the subset of S_n that contains n_j locations of feature observations from Ω_j, $j = 1, 2$. So each partition $\xi(S_n) = \{S^{(1)}, S^{(2)}\}$ with marked labels determines TLC.

For TLC $\xi(S_n)$, define the variable $d = |D^{(1)} - D^{(2)}|$, where $D^{(j)}$ is the sum of distances between the location s_0 and locations in $S^{(j)}$, $j = 1, 2$.

As it follows, we assume that STL S_n and TLC ξ are fixed. This is the case, when spatial classified training data are collected at fixed locations.

For notational convenience, the argument θ in all its functions is now dropped.
So the model of training sample is

$$T = M + E,$$

where M is the vector of the training sample mean and E is the $n \times 1$ – vector of random errors that follows multivariate Gaussian distribution $N_n(0,V)$.

Denote by c_0 the vector of covariances between Z_0 and T.

Set $S_n^0 = S_n \cup s_0$ and $T_0' = (Z(s_0), Z(s_1),..., Z(s_n))$.

Then the variance-covariance matrix of vector T_0 is

$$V^+ = \mathrm{var}(T_0) = \begin{pmatrix} C(0) & c_0' \\ c_0 & V \end{pmatrix}. \tag{3}$$

Let *t* denote the realization of T .

Since Z_0 follows model specified in (1), (2), the conditional distribution of Z_0 given $T = t,\ \Omega_j$ is Gaussian with mean

$$\mu_{jt}^0 = E(Z_0 | T = t; \Omega_j) = \mu_j(s_0) + \alpha_0'(t - M),\ \ j = 1,2 \tag{4}$$

and variance

$$\sigma_0^2 = \mathrm{var}(Z_0 | T = t; \Omega_j) = C(0) - c_0' V^{-1} c_0, \tag{5}$$

where $\alpha_0' = c_0' V^{-1}$.

Under the assumption of complete parametric certainty of populations the BDF minimizing the OMP is formed by log ratio of conditional likelihoods.

Then BDF is specified by (McLachlan, 2004)

$$L_t(Z_0) = \left(Z_0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0\right)\right)\left(\mu_{1t}^0 - \mu_{2t}^0\right) / \sigma_0^2 + \gamma, \tag{6}$$

where $\gamma = \ln(\pi_1 / \pi_2)$.

Here $\pi_1, \pi_2\ (\pi_1 + \pi_2 = 1)$ are prior probabilities of the populations Ω_1 and Ω_2 for observation at location s_0 .They specified the prior distribution for class membership for observation at location s_0 .

So BDF allocates the observation in the following way:

Classify observation Z_0 given $T = t$ to population Ω_1, if $L_t(Z_0) \geq 0$, and to population Ω_2, otherwise.

Definition. The OMP for the BDF $L_t(Z_0)$ specified in (6) is defined as

$$PB = \sum_{i=1}^{2} \sum_{j=1, j \neq i}^{2} \pi_i P_{ij},$$

where, for $i, j = 1, 2$,

$$P_{ij} = P_{it}\left((-1)^j L_t(Z_0) < 0\right).$$

Here, for $i=1,2$, the probability measure P_{it} is based on conditional distribution of Z_0 given $T=t,\Omega_i$ specified in (4), (5). As it follows, P will be called Bayes OMP (BOMP).

Note that the squared Mahalanobis distance between marginal distributions of Z_0 and the squared Mahalanobis distance between conditional distributions of Z_0 given $T=t$ are specified by $\Delta^2=\left(\mu_1^0-\mu_2^0\right)^2/C(0)$ and $\Delta_0^2=\left(\mu_{1t}^0-\mu_{2t}^0\right)^2/\sigma_0^2$, respectively.

From (4), (5) it is easy to derive that

$$\Delta_0^2=\Delta^2C(0)/\sigma_0^2. \tag{7}$$

Thus Δ_0 does not depend on realizations of T.

In population Ω_j, the conditional distribution of $L_t(Z_0)$ given $T=t$ is normal distribution with mean

$$E_j\left(L_t(Z_0)\right)=(-1)^{j+1}\Delta_0^2/2+\gamma$$

and variance

$$\mathrm{var}_j\left(L_t(Z_0)\right)=\Delta_0^2,\ j=1,2.$$

By using the properties of normal distribution we obtain

$$PB=\sum_{j=1}^{2}\left(\pi_j\Phi\left(-\Delta_0/2+(-1)^j\,\gamma/\Delta_0\right)\right) \tag{8}$$

where $\Phi(\cdot)$ is the standard normal distribution function. So it is obvious from (8) that OMP does not depend on the realization of T.

The OMP is one of the natural performance measures to the BDF similar as the mean squared prediction error (MSPE) is the performance measure to the kriging predictor (see Diggle et al, 2002). MSPE are usually used for spatial sampling design criterion for prediction (see Zhu and Stein, 2006). These facts strengthen the motivation for the deriving an explicit expression of the OMP for spatial classification procedures.

2. Classification based on BDF for CAR model

For data collected over geographic regions such as counties, census tracts, zip codes, and so on, the most commonly used are CAR specifications (see Haining, 1990). CAR distributions are sometimes used as the distribution of random effects in the mean structure in hierarchical models (see Anselin, 1988).

Denote by w_{ij} a spatial weight specifying the interconnection between locations s_i and s_j ($w_{ii}=0$, and $w_{ij}\neq 0$ if $i\approx j$, and 0 otherwise) for $i,j=0,...,n.$ Here $i\approx j$ denotes that location s_j is a neighbour (typically defined in terms of spatial adjacency) of location s_i. Spatial weights can also be based on economic distance (Case, Rosen, Hines, 1993) or on trade-based interaction measures (Aten, 1996).

In the present paper we consider the case of a single parameter CAR model with full conditional distribution of Z_0 given $T_i\Omega_j$ with moments specified as

$$E(Z_0 \mid T_i \Omega_j) = \mu_j(s_0) + \tau \sum_{i \approx 0} w_{0i}(Z(s_i) - \mu(s_i))$$

and $\mathrm{var}(Z_0 \mid T_i \Omega_j) = \sigma_0^2$, where τ is a smoothing parameter and is often interpreted as measuring spatial association. Several authors prefer $0 < \tau \leq 1$ to avoid singularity of matrices. Then covariance matrix of vector T_0 specified in (3) having the following form

$$V^+ = \sigma_0^2 (I - \tau W^+)^{-1} \tag{9}$$

where W^+ is a $n+1$ by $n+1$ spatial weights matrix for set of locations S_n^0 . Denote by w_0 the *n* vector of spatial weights between s_0 and S_n, i.e.

$$w_0' = (w_{01}, w_{02}, \ldots, w_{0n}).$$

Set $B = I - \tau W$, where W is n by *n* spatial weights matrix for the set of locations S_n. Make the following assumptions:

(A1) The set of locations S_n^0 forms a clique of size $n+1$, i.e spatial weights between all locations are not zero.

(A2) Spatial weights for S_n and S_n^0 are based on the Euclidean distance between different locations.

Lema. Suppose that observation Z_0 to be classified by BDF and let covariance matrix of T_0 to be specified in (9).Then under the assumptions (A1) and (A2), OMP takes the form

$$PB = \sum_{j=1}^{2} \left(\pi_j \Phi\left(-\Delta k/2 + (-1)^j \gamma/(\Delta\, k) \right) \right), \tag{10}$$

where

$$k = 1/\sqrt{1 - \tau^2 w_0' B^{-1} w_0}. \tag{11}$$

Proof. Under assumptions (A1) and (A2), we can easily derive that

$$W^+ = \begin{pmatrix} 0 & w_0' \\ w_0 & W \end{pmatrix}. \tag{12}$$

By using some matrix inversion properties in (9), (12), we can conclude that conditional distribution of Z_0 given $T = t$ in Ω_j is Gaussian with mean

$$\mu_{jt}^0 = \mu_j(s_0) + \tau w_0'(t - M), \quad j = 1, 2,$$

and variance

$$\sigma_0^2 = C(0)(1 - \tau^2 w_0' B^{-1} w).$$

After inserting these expression into (7) and using (8) we obtain (10), (11) and complete the proof of Lema.

3. Example and discussions

Numerical example is considered to investigate the influence of the statistical parameters of populations to the proposed BDF in the finite (even small) training sample case. With an insignificant loss of generality the cases with $n_1 = n_2$ are considered. We also suppose that assumptions (A1) and (A2) hold.

In this example, observations are assumed to arise from stationary Gaussian random field with constant mean. The spatial weights are specified by $w_{ij} = 1/d_{ij}$, where d_{ij} is the Euclidean distance between different locations.

Assume D is regular 2-dimentional lattice with unit spacing. We consider two spatial structure schemes:

NN(2) denotes second order neighbourhood scheme with $s_0 = (1,1)$,

NN(3) denotes third order neighbourhood scheme with $s_0 = (2,2)$.

So for NN(2) STL consists of 8 second-order neighbours of s_0 and is denoted by $S_8 = \{(0,0),(1,0),(2,0),(0,1),(2,1),(0,2),(1,2),(2,2)\}$ and for NN(3) STL consists of 12 third-order neighbours of s_0 and is denoted by $S_{12} = \{(2,0),(1,1),(2,1),(3,1),(0,2),(1,2),(3,2),(4,2),(1,3),(2,3),(3,3),(2,4)\}$.

Set $M1 = \{i : s_i \in s^{(1)}\}$. Two cases of prior probabilities are considered:

C1. $\pi_1 = (\sum_{i \in M1} 1)/n$

C2. $\pi_1 = (\sum_{i \in M1} 1/d_{0i})/(\sum_{i=1,\ldots,n} 1/d_{0i})$

Case C1 is based only on the number of neighbours, while case C2 incorporated spatial adjacency (distances) also. So OMP is denoted PBN for the case C1 and PBD for the case C2.

Consider two TLC ξ_1, ξ_2 for S_8 specified by

$$\xi_1 = \left\{S^{(1)} = \{(0,2),(1,2),(2,1),(1,0)\},\quad S^{(2)} = \{(0,0),(0,1),(2,2),(2,0)\}\right\},$$

$$\xi_2 = \left\{S^{(1)} = \{(1,2),(2,1),(0,1),(1,0)\},\quad S^{(2)} = \{(0,0),(0,2),(2,0),(2,2)\}\right\}.$$

They are presented in Figure 1.

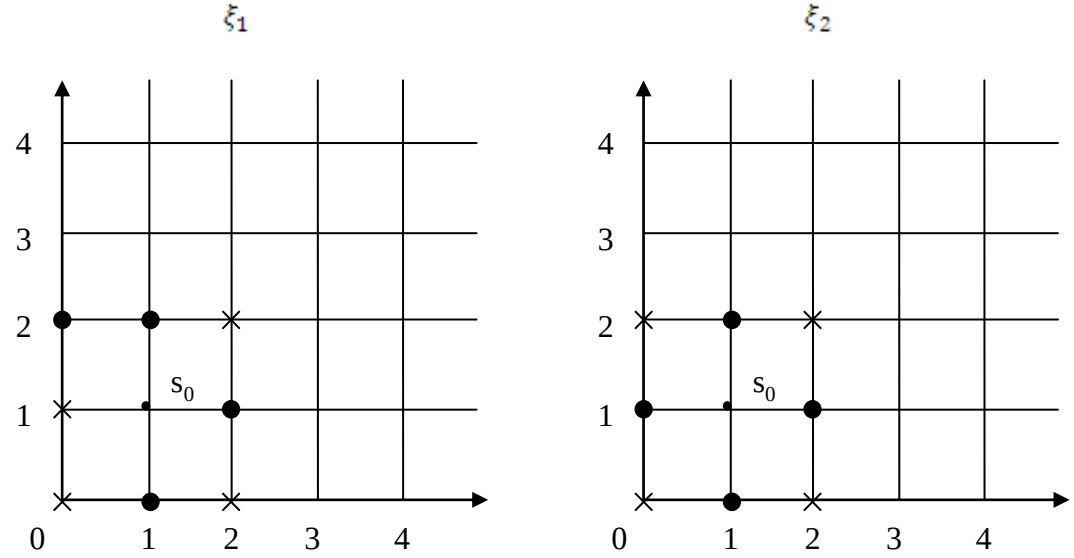

Figure 1. Two different TLC in second order neighbourhood scheme NN(2) with $S^{(1)}$ and $S^{(2)}$ points • and *, signed as respectively.

Consider two TLC ξ_3, ξ_4 for S_{12}specified by

$\xi_3 = \{S^{(1)} = \{(1,3),(2,4),(2,3),(2,1),(3,2),(4,2)\}, \quad S^{(2)} = \{(0,2),(1,2),(2,0),(3,3),(3,1),(1,1)\}\}$,

$\xi_4 = \{S^{(1)} = \{(0,2),(1,2),(2,3),(2,1),(3,2),(4,2)\}, \quad S^{(2)} = \{(2,4),(1,3),(1,1),(2,0),(3,3),(3,1)\}\}$.

They are presented in Figure 2.

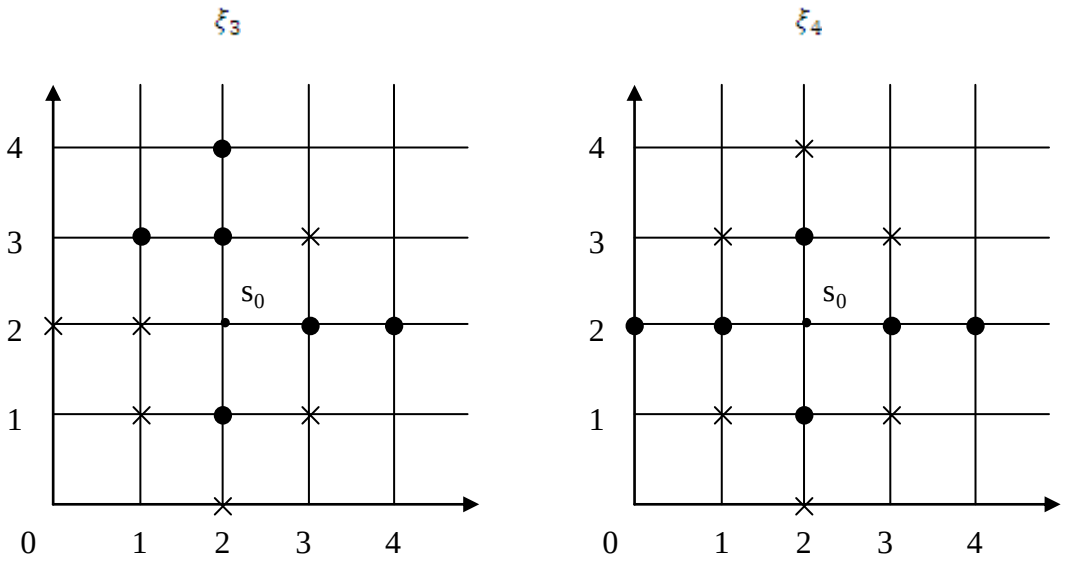

Figure 2. Two different TLC in third order neighbourhood scheme NN(3) with $S^{(1)}$ and $S^{(2)}$ points • and *, signed as respectively.

The comparison of two cases of prior distribution for each TLC is done by the values of index $\eta = PBD / PBN$. The results of calculations with $\tau = 0.3$ for ξ_1 and ξ_2 are shown in Figure 3 and for ξ_3 and ξ_4 in Figure 4.

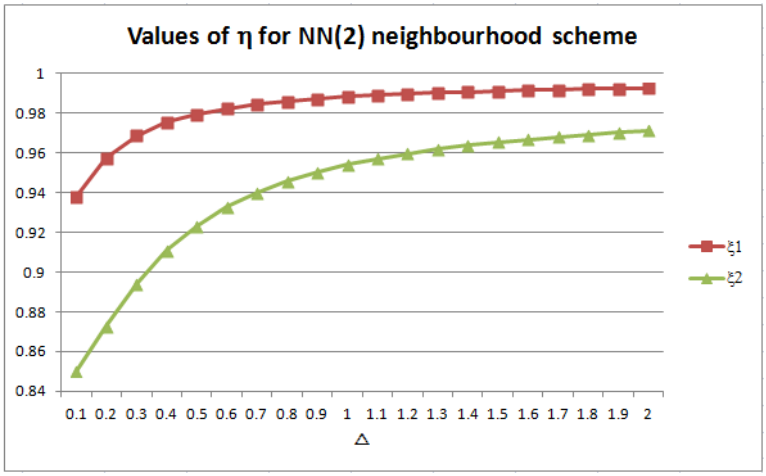

Figure 3. Values of η for NN(2) neighbourhood scheme.

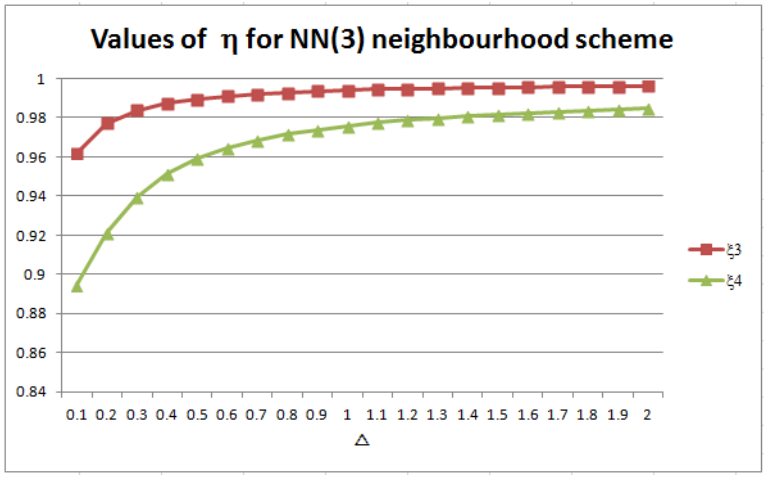

Figure 4. Values of η for NN(3) neighbourhood scheme.

By the definition variable d represents the asymmetry population labels distribution in training sample. It is easy to obtain that $d = 2(\sqrt{2}-1)$ for ξ_1 and ξ_3, and $d = 4(\sqrt{2}-1)$ for ξ_2 ξ_4. So ξ_1 and ξ_3 are less asymmetric TLC than ξ_2 and ξ_4.

Analyzing the contents of Figure 3 and Figure 4 we can conclude that prior distribution based on distances to neighbours outperforms the one based only on numbers of neighbours, because η values are smaller than one.

Figures also shows that for both neighbourhood schemes η increases with the increasing of Δ. Graphs also enable us to conclude that positive effect of the incorporation distances into prior distributions is stronger for more asymmetric TLC i.e. for ξ_2 and ξ_4, and for smaller Mahalanobis distances.

Conclusion

We have considered statistical classification of CAR observations based on BDF for two objectives: deriving of an explicit expression of the overall misclassification probability for proposed procedure, and numerical analysis of the influence of different prior distributions of class labels based on the values of OMP.

The first objective was reached in Lema by deriving an explicit expression of the OMP for the Gaussian case under slightly restrictive assumptions.

The examples considered for the realisation of the second objective shows the advantage of the prior distribution of the class labels with incorporated distance between locations against one based only on number of neighbours. The effect of the distance incorporation is evaluated for different spatial sampling designs. It was obtained that the effect is stronger for more asymmetric TLC.

Hence the results of numerical analysis give us strong arguments to expect that proposed derived formula of the OMP could be effectively used for performance evaluation of classification procedures and for the optimal designing of spatial training samples. The simulated annealing algorithm (see e.g., Lark, 2002) can be easily used in searching the optimal spatial sampling design for the considered spatial classification problem.

References

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

Anselin, L. (1988). Spatial Econometrics. Dordrecht: Kluwer Academic.

Aten, B. (1996). "Evidence of spatial autocorrelation in international prices", Review of Income and Wealth 42 (2): 149-63.

Batsidis, A., Zografos, K. (2011). "Errors of misclassification in discrimination of dimensional coherent elliptic random field observations". Statistica Neerlandica 65 (4): 446–461.

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems (with discussion)", Journal of the Royal Statistical Society. Series B (Methodological) 36 (2): 192–236.

Case, A., Rosen, H. S., Rice, J. R. (1993). "Budget spillovers and fiscal policy interdependence: evidence from the States". Journal of public Economics 52: 285–307.

Cressie, N. A. C. (1993). Statistics for spatial data. New York: Wiley.

Diggle, P. J., Ribeiro P. J., Christensen O. F. (2002). "An introduction to model-based geostatistics", Lecture notes in statistics 173: 43–86.

Ducinskas, K. (2009). "Approximation of the expected error rate in classification of the Gaussian random field observations", Statistics and Probability Letters 79: 138–144.

Haining, R. (1990). Spatial Data analysis in the Social and Environmental sciences. Cambridge University Press.

Kharin, Yu. (1996). Robustness in Statistical Pattern Recognition. Dordrecht: Kluwer Academic Publishers.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood, Volume 105: 49–80.

Lawoko, C. R. O., McLachlan, G. L. (1985). "Discrimination with autocorrelated observations", Pattern Recognition 18 (2): 145–149.

McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. New York: Wiley.

Saltyte, J., Ducinskas, K. (2002). "Comparison of ML and OLS estimators in discriminant analysis of spatially correlated observations", Informatica 13 (2): 297–238.

Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., Chawla, S. (2002). "Spatial Contextual Classification and Prediction Models for Mining Geospatial Data", IEEE Transactions on Multimedia 4: 174–188.

Switzer, P. (1980). "Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery", Math. Geol. 12 (4): 367–376.

Zhu, Z., Stein, M. L. (2006). „Spatial Sampling Design for Prediction with Estimated Parameters", Journal of Agricultural, Biological, and Environmental Statistics 11 (1): 24–44.

K. Dučinskas. Education: 1983 – Vilnius University, Doctoral Thesis, Mathematics (01P); 1976 – Vilnius University, Graduate Studies of Mathematics (01P). Research Area: Spatial and Temporal Statistics, Methods of Applied Statistics.

I. Borisenko. Education: 2009 – Vilnius Gediminas Technical University, Doctoral thesis, Mathematics (01P). 2001 – Klaipeda University, Master degree in Mathematics (01P); 1999: Klaipeda University, Bachelor degree in Applied Mathematics (minor informatics) (01P, 09P). Research Area: Methods of Applied Statistics, Spatial Statistics, GIS, Distance Learning.

I. Šimkienė. Education: 2011 – Siauliai University, Master degree in Economics, 2008 – Klaipeda University, Bachelor degree in Mathematics. Research Area: Spatial Econometrics.

STATISTINIS GENERUOTŲ PAGAL SĄLYGINĮ AUTOREGRESINĮ MODELĮ ERDVINIŲ DUOMENŲ, PASISKIRSČIUSIŲ PAGAL GAUSO SKIRSTINĮ, KLASIFIKAVIMAS

Kęstutis Dučinskas, Ingrida Borisenko, Indrė Šimkienė

Santrauka

Darbe pasirinktoms mokymo imtims yra analizuojama erdvinių duomenų, pasiskirsčiusių pagal Gauso skirstinį, klasifikavimo į dvi populiacijas problema, teigiant, kad populiacijos apibrėžtos pagal sąlyginį auroregresinį modelį (CAR) su skirtingomis vidurkio funkcijomis. Straipsnyje sutelkiamas dėmesys ties klasifikavimo procedūra, susijusia su Bayeso diskriminantine funkcija (BDF) pagal deterministinę erdvinių imčių schemą. Šiuo atveju, kai visi parametrai yra žinomi, yra apibrėžta bendra klasifikavimo klaida susijusi su minėta BDF. Tai yra ankstesnių tyrimų tęsinys CAR atvejui. Erdviniai svoriai suteikiami naudojant atvirkštinio Euklidinio atstumo metodą, o erdvinės konfigūracijos pagrįstos antros ir trečios eilės kaimynų schemomis ant taisyklingos dvimatės gardelės. Taip pat straipsnyje yra skaitiškai įvertinami ir palyginami aprioriniai klasių skirstiniai, naudojant darbe išvestas bendros klasifikavimo klaidos tikimybės formules.

Pagrindiniai žodžiai: Bayes diskriminantinė funkcija, kovariacinė funkcija, Gauso atsitiktinis laukas, klaidingo suklasifikavimo tikimybė, mokymosi žymių konfigūravimas