SPATIAL MODEL SELECTION BASED ON HYBRID PERFORMANCE MEASURE OF LINEAR CLASSIFIER

K. Dučinskas[1,2], L. Dreiziene[1,3]
[1]*Klaipeda University*
[2]*Vilnius University*
[3]*Lithuanian Maritime Academy*
Klaipeda, Vilnius, LITHUANIA
e-mail: k.ducinskas@gmail.com, l.dreiziene@gmail.com

Abstract

Assuming that spatial data is generated by Gaussian random field (GRF), the problem of classifying its observation into one of two populations is considered. Populations are specified by the common regressors but different regression parameters. Authors concern with classification procedures associated with Bayes discriminant function (BDF) and its sample version (SDF). The average of plug-in and apparent correct classification rates is considered as performance measure of classifier based on SDF. Various types of spatial data models for invasive species (*zebra mussels)* distributed in the Curonian Lagoon are considered and ranked by the defined criterion. Advanced models are proposed to the mapping of presence and absence of zebra mussels in the Curonian Lagoon.
Keywords: spatial model, data science, linear classifier, Gaussian random field

1 Introduction

Classification of spatial data has been mentioned in the ecological literature, but lacks full mathematical treatment and easily available algorithms and software. This paper fills this gap by defining the method of statistical classification based on BDF by providing novel formulas and algorithms, which allows to evaluate the influence of spatial information to the performance of proposed classifier. Performance of the classifier based on SDF in the complete parametric uncertainty case is implemented by Ducinskas and Dreiziene (2011). Numerical comparison of the performances for different spatial classification rules is performed by Berrett and Calder (2016). In the present paper we focus on linear classification problem of GRF observation for the so-called geostatistical model (GS) with continuous spatial index and directly specified parametric covariance functions. It should be noted that classification of spatial lattice data modeled by conditionally autoregressive models is recently explored by Ducinskas and Dreiziene (2018). The average of the plug-in and apparent correct classification rates (AVER) is considered as an hybrid estimator for the classifiers based on SDF. These are used in comparison and selection of the spatial linear models for spatial ecological data. Spatial distribution and spread of invasive species (zebra mussels) in lagoons and bays are interested a lot of ecologists (see, e.g. Zaiko, Daunys 2015). In the present paper three spatial linear models for zebra mussels distributed in the Curonian Lagoon are considered and compared by proposed performance measure.

2 The Main Concepts and Definitions

In this paper we focus on classification of a single scalar GRF $\{Z(s) : s \in D \subset R^2\}$ observation, when training sample is given. The model of observation $Z(s)$ in population Ω_l is $Z(s) = x'(s)\beta_l + \varepsilon(s)$, where $x(s)$ is a $q \times 1$ vector of non-random regressors and β_l is a $q \times 1$ vector of parameters, $l = 1, 2$, and $\beta_1 \neq \beta_2$. The error term $\varepsilon(s)$ is generated by zero-mean GRF $\{\varepsilon(s) : s \in D\}$ with covariance function $\sigma(s,t) = cov(\varepsilon(s), \varepsilon(t))$, for $s, t \in D$.

Suppose that $\{s_i \in D, i = 0, 1, \ldots, n\}$ is the set of spatial sites where the observations of GRF are taken. Indexing spatial sites by integers i.e. $s_i = i, i = 0, 1, \ldots, n$, denote the set of training sites by $S_n = S^{(1)} \cup S^{(2)}$, where $S^{(1)} = \{1, 2, \ldots, n_1\}$ and $S^{(2)} = \{n_1 + 1, \ldots, n_1 + n_2\}$, $n = n_1 + n_2$, are the subsets of S_n that contains n_l observations of $Z(s)$ from $\Omega_l, l = 1, 2$. The location of the observation to be classified is indexed by $\{0\}$.

In what follows we use the notations $Z(i) = Z_i$, $\varepsilon(i) = \varepsilon_i$, $x(i) = x_i$, $\sigma_{ij} = cov(Z_i, Z_j)$, $i, j = 0, 1, \ldots, n$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, $Z = (Z_1, \ldots, Z_n)'$. Define n-vector c_0 and $n \times n$ matrix Σ by $c_0 = (\sigma_{01}, \sigma_{02}, \ldots, \sigma_{0n})'$ and $\Sigma = (\sigma_{ij}, i, j = 1, \ldots, n)$.

Put $\beta' = (\beta_1', \beta_2')$, $\alpha_0 = \Sigma^{-1}c_0$, and denote by X the $n \times 2q$ design matrix of training sample Z. Then the training sample Z has multivariate Gaussian distribution $Z \sim N_n(X\beta, \Sigma(\theta))$.

The main objective of this paper is to classify the single observation of scalar GRF $\{Z(s) : s \in D \subset R^2\}$ at location s_0 given training sample Z. Let z denote the realization of Z. Then the conditional distribution of Z_0 given $Z = z$ in Ω_l is Gaussian with mean and variance

$$\mu_{lz}^0 = E(Z_0|Z = z; \Omega_l) = x_0'\beta_l + \alpha_0'(z - X\beta), \tag{1}$$

$$\sigma_{0z}^2(\theta) = \sigma_{00} - c_0'\Sigma^{-1}c_0. \tag{2}$$

For geostatistical data, spatial index s is assumed to vary continuously throughout the set D. Let $\Psi = (\beta', \theta')$ denote the combined vector of population parameters. Under the assumption of complete parametric certainty of populations and for known prior probabilities of the populations the BDF maximazing the probability of correct classification is formed by log ratio of conditional likelihoods of Z_0 at location s_0. Then BDF is specified by

$$W_z(Z_0, \Psi) = \Big(Z_0 - 1/2(\mu_{1z}^0 + \mu_{2z}^0)\Big)(\mu_{1z}^0 - \mu_{2z}^0)/\sigma_{0z}^2 + \gamma_0, \tag{3}$$

where $\gamma_0 = ln(\pi_1^0/\pi_2^0)$. π_1^0 and π_2^0 are prior probabilities, and $\pi_1^0 + \pi_2^0 = 1$. Suppose that for $l = 1, 2$, the probability measure P_{lz} based on conditional Gaussian distribution of Z_0 given $Z = z$, Ω_l i.e. $Z_0|Z = z, \Omega_l \sim N_l(\mu_{0z}^l, \sigma_{0z}^2)$.

Definition 1. The probability of correct classification for the BDF $W_z(Z_0, \Psi)$ is defined as $PC(\Psi) = \sum_{l=1}^2 \pi_l P_l$, where, for $l = 1, 2$, $P_l = P_{lz}((-1)^l W_z(Z_0, \Psi) < 0)$.

As it follows, $PC(\Psi)$ will be called Bayes probability of correct classification (BPCC).

Lemma 1. *Closed-form expression for BPCC is* $PC(\Psi) = \sum_{l=1}^{2} \pi_l^0 \Phi(\Delta_0/2 - (-1)^l \gamma_0/\Delta_0)$, *where* $\Phi(\cdot)$ *is the standard Gaussian distribution function and* Δ_0 *stands for conditional Mahalanobis distance between conditional distributions of* Z_0, *given* $Z = z$.

Proof of Lemma 1 follows from Definition 1 and properties of Gaussian distribution.

In practice it is rarely the case that regression parameters vector β and covariance parameter vector θ are known, and often we need to estimate these parameters from the training data. Here we use maximum likelihood (ML) method for estimation and corresponding estimators are denoted by $\hat{\beta}$, $\hat{\theta}$, and $\hat{\Psi} = (\hat{\beta}', \hat{\theta}')$.

Then using (1), (2) we get the estimators of conditional mean and conditional variance

$$\hat{\mu}_{lz}^0 = E(Z_0|Z = z; \Omega_l) = x_0'\hat{\beta}_l + \hat{\alpha'}_0(z - X\hat{\beta}), l = 1, 2, \hat{\sigma}_{0z}^2 = \sigma_{0z}^2(\hat{\theta}).$$

By replacing the parameters with their ML estimators in (3) we form the SDF $W_z(Z_0, \hat{\Psi})$.

Set for $l = 1, 2$, $\hat{P}_{lz}((-1)^l W_z(Z_0, \hat{\Psi}) < 0)$. Then the actual correct classification rate for SDF $W_z(Z_0, \hat{\Psi})$ is $AR = \sum_{l=1}^{2} \pi_l^0 \hat{P}_l$.

Closed-form expression for AR is derived in Ducinskas and Dreiziene (2011).

Definition 2. Plug-in correct classification rates for the AR based on SDF is $PR = \sum_{l=1}^{2} \left(\pi_l^0 \Phi(\hat{\Delta}_0/2 - (-1)^l \gamma_0/\hat{\Delta}_0) \right)$.

Definition 3. Apparent correct classification rates are defined by $APR = \left(\sum_{i=1}^{n_l} H\left(W_z(Z_i, \hat{\Psi}) \right) + \sum_{i=n_l+1}^{n} H(-W_z(Z_i, \hat{\Psi})) \right)/n$, where $H(\cdot)$ is the Heaviside step function.

We propose $AVER = (PR + APR)/2$ consider as hybrid estimator of AR based on SDF for different linear models of spatial ecological data.

3 Model selection

In this section the application of the proposed estimators for model selection is considered. We use a real dataset of zebra mussels observed over the Curonian Lagoon, a large, shallow coastal waterbody connected to the Baltic Sea by the narrow Klaipeda Strait. Zebra mussels *(Dreissena polymorpha)* are one of the most widespread invasive freshwater animals in the world. Currently, zebra mussels are highly abundant in the Curonian Lagoon, occupying the littoral zone down to 3-4m depth and occurring on both hard substrates and soft bottoms (Zaiko, Daunys 2015). We have 39 spatial sites in Curonian Lagoon where salinity, depth and water renewal time were observed. We also have information about the absence and presence of zebra mussels at those sites. We treat water renewal time as dependent variable and the remaining two as explanatory variables. The main purpose is to select the most appropriate model to the mapping of presence and absence of zebra mussels in the Curonian Lagoon, that is, to build a model with the greatest correct classification probability.

Let M_T, M_R, M_M denote three candidate models with different mean structure, that is, a different mean component $X\beta$: M_T - 1st order trend surface model; M_R - regression model, that is represented as a function of two explanatory variables; M_M - mixed model which combines M_T and M_R. The design matrix for this model consists of intercept, coordinates of spatial sites and explanatory variables.

A different number of neighbours are used for the specifying the prior probabilities. Spatial correlation is modelled by isotropic exponential covariance function given by $\sigma(h) = \sigma^2 exp(-h/\eta) + \tau^2\delta(h)$, where h is a distance between spatial sites, η is a parameter of spatial correlation, τ^2 is a nugget effect, and $\delta(h) = 1$, if $h = 0$, and $\delta(h) = 0$, if $h \neq 0$.

The results show that the mixed model (M_M), including the set of closest neighbors for estimation of priors, gives the maximum of AVER (AVER=0.757). Salinity, depth and the coordinates of spatial sites are considered as covariates in the mean model.

4 Conclusions

This work describes a novel approach for spatial linear model selection, applicable to classified spatial data. This has several attractive features that make it compare favourably against other model selection approaches. First, it essentially incorporates the spatial information into data model and classification rule specification. Second, the approach provides an easily interpretable criterion of how strongly the data support each of the competing models. The best model has a mixed mean structure which includes coordinates of spatial sites and explanatory variables salinity and depth as covariates. The highest probability of correct classification could be approached using the set of nearest neighbours for estimating the prior probabilities.

References

[1] Berret C., Calder C.A. (2016). Bayesian spatial binary classification. *Spatial Statistics*. Vol. **16**, pp. 72-102.

[2] Ducinskas K., Dreiziene L. (2018). Risk of classification of the Gaussian Markov random field observations. *Journal of Classification*. Vol. **35**, pp. 422-436.

[3] Ducinskas K., Dreiziene L. (2011). Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. *Austrian Journal of Statistics*. Vol. **40(1&2)**, pp. 25-36.

[4] Zaiko A., Daunys D. (2015). Invasive ecosystem engineers and biotic indices: giving a wrong impression of water quality improvement. *Ecological Indicators*. Vol. **52**, pp. 292-299.