# Comparison of Performances of Plug-in Spatial Classification Rules based on Bayesian and ML Estimators

Kestutis Ducinskas, Egle Zikariene and Lina Dreiziene
*Department of Matematics and Statistics, Klaipeda University, Herkaus Manto str. 84, Klaipeda, Lithuania*

Keywords: Bayes Rule, Spatial Discriminant Function, Gaussian Random Field, Actual Risk, Training Labels Configuration.

Abstract: The problem of classifying a scalar Gaussian random field observation into one of two populations specified by a different parametric drifts and common covariance model is considered. The unknown drift and scale parameters are estimated using given a spatial training sample. This paper concerns classification procedures associated to a parametric plug-in Bayes Rule obtained by substituting the unknown parameters in the Bayes rule by their estimators. The Bayesian estimators are used for the particular prior distributions of the unknown parameters. A closed-form expression is derived for the actual risk associated to the aforementioned classification rule. An estimator of the expected risk based on the derived actual risk is used as a performance measure for the classifier incurred by the plug-in Bayes rule. A stationary Gaussian random field with an exponential covariance function sampled on a regular 2-dimensional lattice is used for the simulation experiment. A critical performance comparison between the plug-in Bayes Rule defined above and a one based on ML estimators is performed.

## 1 INTRODUCTION

It is known that for completely specified populations and loss function, Bayes Rule (BR) is an optimal classification procedure in the sense of minimum risk (expected losses) (Anderson, 2003). When this is not the case, the missing information is usually provided by a training sample. For parametrically specified populations, the training sample is used to obtain the estimators of statistical parameters and plugging them into the BR. The resulting classification rule is usually called a plug-in Bayes rule (PBR). Actual risk (ACR) or conditional risk is usually used as a performance measure for the PBR. Performance comparison of the PBR based on the different types of estimators can easily be done by the closed-form expression of the ACR.

Many authors have investigated the performance of the PBR when the parameters are estimated from training samples consisting of dependent observations by using the frequentist approach for the estimation (Kharin, 1996; Saltyte-Benth and Ducinskas, 2005; Batsidis and Zografos, 2011). A closed-form expression for the ACR in supervised classification of Gaussian random field observations is derived by Ducinskas (2009). Only the ML estimators of the drift parameters and the scale parameter of covariance function are considered.

In the present paper we use Bayesian estimators instead of the ML estimators for the classification problem described above. Proposed methodology is useful for classification of images corrupted by the Gaussian spatial correlated noise (Ducinskas, Stabingiene and Stabingis, 2011).

The closed-form expression for the ACR associated with the PBR is derived. The estimator of expected risk is based on the derived actual risk and is used as the performance of the PBR which is measured by the average of the ACR usually called an empirical estimator of expected risk.

This estimator of expected risk is a case of the stationary Gaussian random field on 2-dimentional regular lattice with an exponential covariance function. The dependence of the ACR values on the statistical hyperparameters is investigated. Numerical comparison for the case of ML estimators is implemented.

Ducinskas K., Zikariene E. and Dreiziene L..
Comparison of Performances of Plug-in Spatial Classification Rules based on Bayesian and ML Estimators.
DOI: 10.5220/0004760701610166
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 161-166
ISBN: 978-989-758-018-5

## 2 THE MAIN CONCEPTS AND DEFINITIONS

This paper concerns classifying a Gaussian random field (GRF) $\{Z(s): s \in D \subset R^p\}$ observations into one of two populations denoted by $\Omega_1, \Omega_2$.

The model of an observation $Z(s)$ in a population $\Omega_j$ is

$$Z(s) = \mu_j(s) + \varepsilon(s) \quad (1)$$

where $\mu_j(s)$, is a drift (a deterministic function of locations) and $\varepsilon(s)$ is the stochastic error term, j=1,2.

Suppose the drift $\mu_j(s)$ can be represented as $\mu_j(s) = x'\beta_j$, where x is a $q \times 1$ vector of non-random regressors and $\beta_j$ is a $q \times 1$ vector of the parameters, $j = 1, 2$.

The error term is generated by a zero mean stationary GRF $\{\varepsilon(s): s \in D\}$ with the covariance function defined by the model for all $s, u \in D$

$$\operatorname{cov}\{\varepsilon(s), \varepsilon(u)\} = \sigma^2 C(s-u), \quad (2)$$

where $\sigma^2$ is the variance or the scale parameter and $C(\cdot)$ is the spatial correlation function.

In the case when the covariance function parameters are known, the model (1), (2) is called a universal kriging model (Cressie, 1993).

For the given training sample, consider the classification problem of the $Z_0 = Z(s_0)$ into one of two populations when $x'(s_0)\beta_1 \neq x'(s_0)\beta_2, s_0 \in D$. Denote by $S_n = \{s_i \in D; i = 1, ..., n\}$ the set of locations where the training sample $T' = (Z(s_1), ..., Z(s_n))$ is taken. It specifies the spatial sampling design or the spatial framework for the training sample (Shekhar et al., 2002).

We shall assume the deterministic spatial sampling design and all analyses are carried out conditional on $S_n$.

Assume that each training sample realization $T=t$ and $S_n$ are arranged in the following way. The first $n_1$ components are the observations of $Z(s)$ from $\Omega_1$ and the remaining $n_2 = n - n_1$ components are the observations of $Z(s)$ from $\Omega_2$. So $S_n$ is partitioned into a union of two disjoint subsets, i. e. $S_n = S^{(1)} \cup S^{(2)}$, where $S^{(j)}$ is a subset of $S_n$ that contains $n_j$ locations of the feature observations from $\Omega_j$, $j = 1, 2$. So each partition $\xi(S_n) = \{S^{(1)}, S^{(2)}\}$ with marked labels determines the training labels configuration (TLC).

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

The $n \times 2q$ design matrix of the training sample $T$ denoted by $X$ is specified by $X = X_1 \oplus X_2$, where the symbol $\oplus$ denotes a direct sum of the matrices and $X_j$ is a $n_j \times q$ matrix of the regressors for the observations from $\Omega_j$, $j = 1, 2$.

As it follows, we assume that $S_n$ and TLC $\xi(S_n)$ are fixed. This is the case, when spatial classified training data are collected at the fixed locations (stations).

So the model of the training sample is

$$T = X\beta + E, \quad (3)$$

where $\beta = (\beta_1', \beta_2')'$ is a $2q \times 1$ vector of the regression parameters and $E$ is the $n$ - vector of the random errors that has a multivariate Gaussian distribution $N_n(0, \sigma^2 C_n)$.

Here $C_n$ denotes a spatial correlation matrix among the observations forming the training sample $T$.

Denote by $c_0$ the vector of correlations among $Z_0$ and the components of $T$. Let $t$ denote the realization of $T$.

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

$$\mu_{lt}^0 = E(Z_0 | T = t; \Omega_j) = x_0'\beta_j + \alpha_0'(t - X\beta) \quad (4)$$

and the variance

$$\sigma_t^2 = \operatorname{var}(Z_0 | T = t; \Omega_j) = \sigma^2 \rho_0, \quad (5)$$

where $x_0' = x'(s_0)$, $\alpha_0' = c_0' C_n^{-1}$, $\rho_0 = (1 - c_0' C_n^{-1} c_0)$.

Under the assumption of complete parametric certainty of populations and for known finite non-negative losses $\{L(i, j),\ i, j = 1, 2\}$, BR minimizing the risk of classification is associated with the spatial

discriminant function (SDF) formed by a log ratio of conditional likelihoods (McLachlan, 2004).

$$W_t(Z_0,\Psi)=\left(Z_0-\frac{1}{2}\left(\mu_{1t}^0+\mu_{2t}^0\right)\right)\times \times\left(\mu_{1t}^0-\mu_{2t}^0\right)\Big/\sigma_t^2+\gamma, \quad (6)$$

where $\gamma=\ln\left(\pi_1^*/\pi_2^*\right)$ and $\Psi=\left(\beta',\sigma^2\right)'$.

Here $\pi_j^*=\pi_j\left(L\left(j,3-j\right)-L\left(j,j\right)\right)$, $j=1,2$, where $\pi_1,\pi_2$ $\left(\pi_1+\pi_2=1\right)$ are prior probabilities of the populations $\Omega_1$ and $\Omega_2$, respectively.

Note, that in the present paper we implemented the following values of the prior probabilities $\hat{\pi}_j=n_j/n,\ j=1,2.$

So the classification procedure based on the SDF allocates the observation in the following way:

Classify the observation $Z_0$ given $T=t$ to the population $\Omega_1$ if $W_t\left(Z_0,\Psi\right)\geq 0$, and to the population $\Omega_2$, otherwise.

**Definition 1.** *The risk for the classification rule based on the SDF $W_t\left(Z_0;\Psi\right)$ is defined as*

$$R_0=\sum_{i=1}^{2}\sum_{j=1}^{2}\pi_i L(i,j)P_{ij}, \quad (7)$$

*where, for $i,j=1,2$, $P_{ij}=P_{it}\left(\left(-1\right)^j W_t\left(Z_0;\Psi\right)<0\right)$.*

Here, for $i=1,2$, the probability measure $P_{it}$ is based on the conditional distribution of $Z_0$ given $T=t$, $\Omega_i$ specified in (4), (5).

Note that under the condition (3), the squared Mahalanobis distance between $\Omega_1$ and $\Omega_2$ in the location $s_0$ based on marginal distributions of $Z_0$ and the same squared Mahalanobis distance given $T=t$ are specified by $\Delta^2=\left(\mu_1^0-\mu_2^0\right)^2/\sigma^2)$ and $\Delta_0^2=\left(\mu_{1t}^0-\mu_{2t}^0\right)^2/\sigma_t^2$, respectively.

From (4), (5) it follows that $\Delta_0^2=\Delta^2/\rho_0$. In the population $\Omega_j$, the conditional distribution of $W_t\left(Z_0;\Psi\right)$ given $T=t$ is the normal distribution with the mean $E_j\left(W_t\left(Z_0;\Psi\right)\right)=(-1)^{j+1}\Delta_0^2/2+\gamma$ and the variance $Var_j\left(W_t\left(Z_0;\Psi\right)\right)=\Delta_0^2, j=1,2.$

By using the properties of normal distributions we obtain

$$R_0=\sum_{j=1}^{2}\left(\pi_j^*\Phi\left(-\Delta_0/2+\left(-1\right)^j\gamma/\Delta_0+\pi_j L(j,j)\right)\right),$$

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

In practical applications not all statistical parameters of populations are known. In such cases the estimators of the unknown parameters can be found from the training sample. When the estimators of the unknown parameters are plugged into the SDF, the plug-in SDF (PSDF) is obtained. In this paper we assume that the true values of the parameters $\beta$ and $\sigma^2$ are unknown (complete parametric uncertainty).

Let $\hat{\beta},\hat{\sigma}^2$ be the estimators of the corresponding parameters from the training sample. Put $\hat{\Psi}=\left(\hat{\beta}',\hat{\sigma}^2\right)'$. After replacing the parameters with their estimates in (6) the PSDF gets the following form

$$W_t\left(Z_0;\hat{\Psi}\right)=\left(Z_0-\alpha_0'\left(t-X\hat{\beta}\right)-\frac{1}{2}x_0'H\hat{\beta}\right)\times \times\left(x_0'G\hat{\beta}\right)\Big/\left(\left(\hat{\sigma}^2\rho_0\right)+\gamma\right) \quad (8)$$

with $H=\left(I_q,I_q\right)$ and $G=\left(I_q,-I_q\right)$, where $I_q$ denotes the identity matrix of the order $q$.

**Definition 2.** *The expectation of the ACR with respect to the distribution of T is called the expected risk (ER).*

Recall that the actual risk incurred by the PSDF is obtained by replacing $\Psi$ by the ML estimator $\hat{\Psi}$ in (6) (Ducinskas and Dreiziene, 2011).

**Lemma 1.** *The actual risk for $W_t\left(Z_0;\hat{\Psi}\right)$ specified in (8) is*

$$R_A\left(\hat{\Psi}\right)=\sum_{j=1}^{2}\left(\pi_j^*\Phi\left(\hat{Q}_j\right)+\pi_j L(j,j)\right), \quad (9)$$

*where*

$$\hat{Q}_j=\left(-1\right)^j\left((a_j-\hat{b})\,\mathrm{sgn}(x_0'G\hat{\beta})/\sigma_0+\hat{\sigma}_0^2\gamma/\left(\sigma_0\left|x_0'G\hat{\beta}\right|\right)\right)$$

*and* $a_j=x_0'\beta_j+\alpha_0'(t-X\beta)$,

$\hat{b}=\alpha_0'(t-X\hat{\beta}+x_0'H\hat{\beta}/2)$ *for* $j=1,2$.

The proof of the lemma is presented in the appendix.

The ACR is useful in providing a guide to the performance of the plug-in classification rule when it actually formed from the training sample. The ER is the performance measure to the PBDF similar as the mean squared prediction error (MSPE) is the performance measure to the plug-in kriging predictor (Diggle, Ribeiro and Christensen, 2002). The estimators of the MSPE are used for the spatial sampling design criterion for the prediction

(Zimmerman, 2006; Zhu and Zhang, 2006). These facts strengthen the motivation for the deriving of the estimators of the ER associated with the PSDF.

In this paper we propose an empirical estimator of the ER incurred by the rule based on the proposed PSDF. The following steps are performed to construct this estimator:

1. Simulate $M$ training sample $T$ realizations according to the model specified in (3).
2. For each simulated realization of $T = t_l$, $l = \overline{1, M}$ compute the appropriate estimates $\hat{\beta}_{(l)}$ and $\hat{\sigma}^2_{(l)}$, respectively. Set $\hat{\Psi}'_{(l)} = \left(\hat{\beta}'_{(l)}, \hat{\sigma}^2_{(l)}\right)$
3. By using (9) compute the empirical estimator of the ER

$$\hat{R}^{(\kappa)} = \sum_{l=1}^{M} R_A\left(\hat{\Psi}_{(l)}\right) \Big/ M, \qquad (10)$$

where $\kappa$ - denotes the abbreviation of the estimator type, i.e. takes the values BA or ML.

## 3 THE ESTIMATORS OF PARAMETERS

It is known, that the ML estimators of $\beta$ and $\sigma^2$ based on T are $\hat{\beta}_{ML} = \left(X'R^{-1}X\right)^{-1} X'R^{-1}T$, $\hat{\sigma}^2_{ML} = \left(T - X\hat{\beta}_{ML}\right)' R^{-1}\left(T - X\hat{\beta}_{ML}\right)\Big/ n$.

Using the properties of the multivariate Gaussian distribution is easy to prove that $\hat{\beta}_{ML} \sim N_{2q}\left(\beta, \Sigma_\beta\right)$, $\Sigma_\beta = \sigma^2\left(X'R^{-1}X\right)^{-1}$ $\hat{\sigma}^2_{ML} \sim \sigma^2 \chi^2_{n-2q} \Big/ \left(n-2q\right)$.

The ML estimator of $\beta$ and the bias adjusted ML estimator of $\sigma^2$ are used in the PLDF, i.e. $\hat{\beta} = \hat{\beta}_{ML}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{ML} n \Big/ \left(n-2q\right)$ (Ducinskas, 2009).

In the Bayesian approach the likelihood is given by $T|\beta, \sigma^2 \sim N_n\left(\beta, \sigma^2 R\right)$. The conjugate prior are chosen for the parameters so $p\left(\beta, \sigma^2\right) = p\left(\beta \middle| \sigma^2\right) p\left(\sigma^2\right)$, where $p\left(\beta \middle| \sigma^2\right) \sim N_{2q}\left(\beta^{(0)}, \sigma^2 \Sigma^{(0)}\right)$ is the Gaussian prior for $\beta$ conditional on $\sigma^2$, $p\left(\sigma^2\right) \sim IG\left(u^{(0)}, v^{(0)}\right)$ – the prior density for $\sigma^2$, where $u^{(0)}, v^{(0)} > 0$.
So the conjugate prior is the Normal–Inverse Gamma (NIG) and denotes as $NIG\left(\beta^{(0)}, \Sigma^{(0)}, u^{(0)}, v^{(0)}\right)$. Combining the prior with the likelihood gives a joint Normal–Inverse Gamma posterior (Diggle, Ribeiro and Christensen, 2002):

$$p\left(\beta, \sigma^2 \middle| T\right) = \frac{p\left(\beta, \sigma^2\right) p\left(T \middle| \beta, \sigma^2\right)}{p(T)} = NIG\left(\beta^{(1)}, \Sigma^{(1)}, u^{(1)}, v^{(1)}\right)$$

where

$$\beta^{(1)} = \left(X^T R^{-1} X + \left(\Sigma^{(0)}\right)^{-1}\right)^{-1} \left(X^T R^{-1} T + \left(\Sigma^{(0)}\right)^{-1} \beta^{(0)}\right)$$

, $\Sigma^{(1)} = \left(X^T R^{-1} X + \left(\Sigma^{(0)}\right)^{-1}\right)^{-1}$, $u^{(1)} = u^{(0)} + \frac{n}{2}$,

$$v^{(1)} = v^{(0)} + \frac{1}{2}\left(\left(\beta^{(0)}\right)' \left(\Sigma^{(0)}\right)^{-1} \beta^{(0)} + T'R^{-1}T - \left(\beta^{(1)}\right)' \left(\Sigma^{(1)}\right)^{-1} \beta^{(1)}\right).$$

The marginal posterior for $\beta$ on integrating out $\sigma^2$ is a multivariate Student distribution $p\left(\beta \middle| T\right) \sim t_p\left(\beta^{(1)}, \Sigma^*\right)$, where $p = 2u^{(1)}$, $\Sigma^* = \left(v^{(1)} \Big/ u^{(1)}\right)\Sigma^{(1)}$.

The marginal posterior for $\sigma^2$ is $p\left(\sigma^2 \middle| T\right) \sim IG\left(u^{(1)}, v^{(1)}\right)$, where $IG(\cdot,\cdot)$.

So the BA of $\beta$ and $\sigma^2$ are $\hat{\beta}_{BA} = \beta^{(1)}$, $\hat{\sigma}^2_{BA} = v^{(1)} \Big/ \left(u^{(1)} - 1\right)$, respectively.

## 4 EXAMPLE AND DISCUSSIONS

A numerical example is considered to investigate the influence of the parameter estimation methods to the proposed empirical estimator of the ER in the finite (even small) training sample case. With an insignificant loss of generality a case with $L(i,j) = 1 - \delta_{ij}$, $i, j = 1, 2$ is considered.

In this example, the observations are assumed to arise from the stationary Gaussian random field with the constant mean and the exponential correlation function specified by $c(h) = \exp\left\{-|h| \Big/ \phi\right\}$ is considered. Here $\phi$ denotes the range parameter.
Assume that $D$ is a regular 2-dimentional lattice with unit spacing. Consider the case of $s_0 = (1,1)$ and the fixed set of training locations $S_8$ containing 8 second-order neighbours of $s_0$. This example also illustrates the comparison of two different TLC.

Consider two TLC $\xi_1$, $\xi_2$ for $S_8$ specified by $\xi_1 = \left\{S^{(1)} = \{(0,2),(1,2),(2,2),(2,1)\}, S^{(2)} = \{(2,0),(1,0),(0,0),(0,1)\}\right\}$.

$\xi_2 = \{ S^{(1)} = \{(0,2),(1,2),(2,2),(2,1),(2,0)\}, S^{(2)} = \{(1,0),(0,0),(0,1)\}\}$.
These are illustrated in Figure 1.

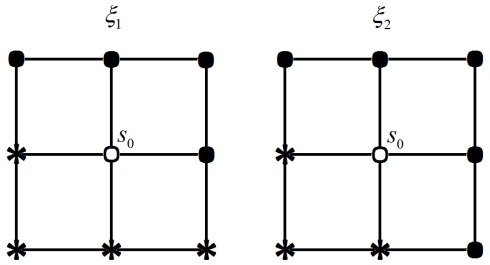


Figure 1: Two different TLC with $S^{(1)}$ and $S^{(2)}$ points marked as dots and asterisks.

By the definition, variable $d$ represents the asymmetry of the TLC with respect to the location $s_0$. It is easy to obtain that $d = 0$ and $n_1 = 4$ for $\xi_1$ and $d = 2\sqrt{2}$ with $n_1 = 5$ for $\xi_2$.
So we can conclude that $\xi_1$ is the symmetric TLC and $\xi_2$ is the asymmetric one. Denote by $\hat{R}^{(ML)}$ and $\hat{R}^{(BA)}$ the empirical estimators of the ER given in (10) with the implemented ML and BA parameter estimators, respectively.

The values of the empirical estimators of $\hat{R}^{(\kappa)}$ are presented for $\xi_1$ with $\pi_1 = 0,5$, and for $\xi_2$ with $\pi_1 = \frac{5}{8}$ in Table 1. Two cases of the simulated realizations of the training sample are selected here, i.e. $M = 10^2$ and $M = 10^4$. Analyzing the figures of the number of simulated realizations of the training sample in Table 1 we see that for all $\Delta = (0.5, 0.7, ..., 1.9)$ values $\hat{R}^{(ML)} \geq \hat{R}^{(BA)}$. So we can conclude that the BA case has an advantage against the ML case by the ER minimum criterion. The quantitative comparison of the two cases of the parameter estimators is also done by the values of the index $\eta = \hat{R}^{(ML)} / \hat{R}^{(BA)}$. The values of this index are shown in Figure 2 for $M = 10^4$ and $\xi_l$, $l = 1, 2$.

It is easy to see from Figure 2 that for both the TLC $\eta \geq 1$ and the values of this index increases when $\Delta$ increases. The same situation is for both TLC considered.

# 5 CONCLUSIONS

In this paper, the comparison of two approaches to parameter estimation is done based on the values of the ER incurred by the classification rule based on the PSDF.

The proposed optimality criterion is based on the derived formula of the actual risk.

The simulation experiment shows the advantage of Bayesian estimation approach against the frequentist (ML) approach. This advantage is greater for strongly separated populations (larger values of $\Delta$) than for the close populations. These conclusions are valid to the symmetric training labels configuration as well to the asymmetric one. Hence the results of this paper give us strong arguments to expect that Bayesian estimators of spatial population parameters could be effectively used in spatial Gaussian data classification incurred by plug-in Bayes rules.

Table 1: Values of $R_1^{(\kappa)}$ for the different estimators and the TLC $\xi_1$ and. $\xi_2$.

| | $\xi_1$ | | | | $\xi_1$ | | | |
|---|---|---|---|---|---|---|---|---|
| | ML | BA | ML | BA | ML | BA | ML | BA |
| $\Delta \setminus M$ | $10^2$ | | $10^4$ | | $10^2$ | | $10^4$ | |
| 0.5 | 0.352 | 0.335 | 0.365 | 0.349 | 0.359 | 0.340 | 0.328 | 0.315 |
| 0.7 | 0.275 | 0.262 | 0.277 | 0.256 | 0.310 | 0.280 | 0.257 | 0.239 |
| 0.9 | 0.195 | 0.177 | 0.201 | 0.180 | 0.196 | 0.182 | 0.189 | 0.171 |
| 1.1 | 0.143 | 0.123 | 0.140 | 0.122 | 0.143 | 0.119 | 0.136 | 0.118 |
| 1.3 | 0.094 | 0.083 | 0.098 | 0.083 | 0.097 | 0.085 | 0.097 | 0.081 |
| 1.5 | 0.067 | 0.053 | 0.067 | 0.055 | 0.061 | 0.051 | 0.066 | 0.053 |
| 1.7 | 0.049 | 0.035 | 0.047 | 0.036 | 0.044 | 0.034 | 0.047 | 0.035 |
| 1.9 | 0.034 | 0.024 | 0.033 | 0.024 | 0.033 | 0.024 | 0.032 | 0.023 |

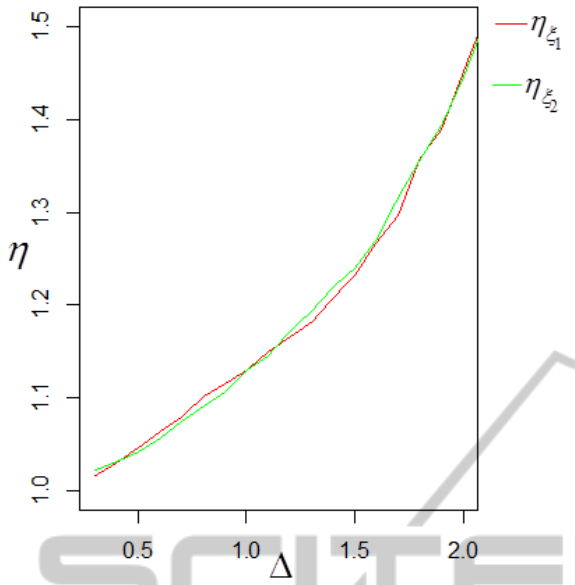


Figure 2: Values of $\eta$ for different TLC.



## REFERENCES


Anderson, T. W., 2003. *An Introduction to Multivariate Statistical Analysis*, Wiley. New York.

Batsidis, A. and Zografos, K., 2011. Errors of misclassification in discrimination of dimensional coherent elliptic random field observations. *Statistica Neerlandica*, 65, p. 446-461.

Cressie, N. A. C., 1993. *Statistics for spatial data*, Wiley. New York.

Diggle, P. J., Ribeiro, P. J. and Christensen, O. F., 2002. An introduction to model-based geostatistics. *Lecture notes in statistics*, 173, p. 43-86.

Ducinskas, K., 2009. Approximation of the expected error rate in classification of the Gaussian random field observations. *Statistics and Probability Letters*, 79, p. 138-144.

Ducinskas, K., Dreiziene, L., 2011. Supervised classification of the scalar Gaussian random field observations under a deterministic spatial sampling design. *Austrian Journal of Statistics.* 40, No. 1, 2, p. 25-36.

Ducinskas, K. Stabingiene, L., Stabingis, G., 2011. Image classification based on Bayes discriminant functions. *Procedia Environmental Sciences*, 7, p. 218-223.

Kharin, Y., 1996. *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers. Dordrecht.

McLachlan, G. J., 2004. *Discriminant analysis and statistical pattern recognition*, Wiley. New York.

Saltyte-Benth, J. and Ducinskas, K., 2005. Linear discriminant analysis of multivariate spatial-temporal regressions. *Scandinavian Journal of Statistics*, 32, p. 281 – 294.

Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W. and Chawla, S., 2002. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4, p. 174-188.

Zhu, Z. and Zhang, H., 2006. Spatial sampling design under infill asymptotic framework. *Environmetrics*, 17, p. 323-337.

Zimmerman, D. L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17, p. 635-652.


## APPENDIX

Proof of Lemma. Recall that the actual risk (ACR) for PSDF $W_t\left(Z_0;\hat{\Psi}\right)$ (Ducinskas and Dreiziene, 2011) is defined as $R\left(\hat{\Psi}\right)=\sum_{i=1}^{2}\sum_{j=1}^{2}\pi_i L(i,j)\hat{P}_{ij}$ where for $i,j=1,2$, $P_{ij}=P_{it}\left((-1)^j W_t\left(Z_0;\hat{\Psi}\right)<0\right)$.

In the population $\Omega_j$, it is easy to derive that the conditional distribution of $W_t\left(Z_0;\hat{\Psi}\right)$ given $T=t$ is normal distribution with the mean $E_j\left(W_t\left(Z_0;\hat{\Psi}\right)\right)=\left(a_j-\hat{b}\right)x_0'G\hat{\beta}/\hat{\sigma}_0^2+\gamma$ (14) and the variance $Var_j\left(W_t\left(Z_0;\hat{\Psi}\right)\right)=\left(x_0'G\hat{\beta}\right)^2\sigma_0^2/\hat{\sigma}_0^4$, $j=1,2$.

Then by using the properties of the normal distribution and we complete the proof of lemma 1.