Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 531062, 7 pages
http://dx.doi.org/10.1155/2013/531062

Research Article

Generalized Linear Spatial Models to Predict Slate Exploitability

Angeles Saavedra,[1,2] Javier Taboada,[3] María Araújo,[3] and Eduardo Giráldez[3]

[1] *Department of Statistics, University of Vigo, 36310 Vigo, Spain*
[2] *E.T.S.I. MINAS, Universidad de Vigo, Campus Lagoas-Marcosende, Rúa Maxwell, 36310 Vigo, Spain*
[3] *Department of Natural Resources, University of Vigo, 36310 Vigo, Spain*

Correspondence should be addressed to Angeles Saavedra; saavedra@uvigo.es

Received 11 November 2012; Revised 3 May 2013; Accepted 26 May 2013

Academic Editor: Zhiping Qiu

The aim of this research was to determine the variables that characterize slate exploitability and to model spatial distribution. A generalized linear spatial model (GLSMs) was fitted in order to explore relationship between exploitability and different explanatory variables that characterize slate quality. Modelling the influence of these variables and analysing the spatial distribution of the model residuals yielded a GLSM that allows slate exploitability to be predicted more effectively than when using generalized linear models (GLM), which do not take spatial dependence into account. Studying the residuals and comparing the prediction capacities of the two models lead us to conclude that the GLSM is more appropriate when the response variable presents spatial distribution.

1. Introduction

The exploitability of a slate deposit depends on many quality-determining factors that are spatially correlated. Knowledge and study of these factors are essential for the evaluation of deposits [1, 2]. Therefore, it can be fairly safely assumed that better evaluations of quality parameters and better predictions of slate exploitability could be obtained by using specific statistical models that take spatial correlation into account.

Traditionally, the main aim of geostatistical models has been to predict a spatially correlated response variable. Under this approach, estimating the parameters of the geostatistical model is not usually the main interest. However, estimating and inferring parameters enables a more precise identification of the factors influencing the geographical distribution of exploitable slate, thus allowing greater knowledge to be gained regarding the response variable of interest.

In our research, the model-based geostatistics methodology was adapted in the analysis of slate exploitability using a generalized linear spatial model (GLSM). With this type of model, the objective of inference can be focused on the parameters of the regression function, on the properties of the residuals, or on the distribution of the residuals conditionally on the response variable.

A brief description of the statistical models used in this study is given in Section 2. The data studied and the formulated model are described in Section 3. The statistical results and analyses are presented in Section 4, and some comments and the discussion can be found in Section 5.

2. Statistical Analysis Methods

2.1. Generalized Linear Spatial Models. Generalized linear models (GLMs) were introduced by [3] and studied in depth by [4] and later by several authors (see [5–9]).

In a GLM, a response variable $Y = (Y_1, Y_2, \dots, Y_n)$ is assumed so that the variables $Y_1, Y_2, \dots, Y_n$ are mutually independent and with its expected value related to a linear predictor $E[Y] = g^{-1}(d^T\beta)$, where $\beta \in \mathfrak{R}^p$ is a vector of unknown regression parameters, d are known explanatory variables, and g is a known function called link function.

An important extension of the GLM is the generalized linear mixed model (GLMM) [10], in which the response variables are considered independent of one another conditionally on the values for a set of latent variables. The generalized linear spatial model (GLSM) [11] is basically a GLMM, in as much as the latent variables derive from a spatial process. The term "model-based geostatistics" was first used by these authors to describe an approach to geostatistical problems based on formal statistical models and inference procedures. This leads to the following specification of the model.

Consider n different locations $\{x_1, \dots, x_n\} \subset I \subset \mathfrak{R}^2$ and assume that a realization $y = (y_1, \dots, y_n)^T$ of Y is observed, where $y = Y(x_i)$.

Let $S = \{S(x) : x \in I\}$, $I \subset \mathfrak{R}^2$ be a Gaussian process with mean function $E[S(x)] = d(x)^T \beta$ and covariance $\operatorname{cov}(S(x), S(x')) = \sigma^2 \rho(x, x'; \varphi) + \tau^2 1\{x = x'\}$, where $\beta \in \mathfrak{R}^p$ is, as in the GLM case, a vector of unknown regression parameters, $d(x)$ are known explanatory variables now with spatial dependence, $\rho(x, x'; \varphi)$ is a correlation function in $\mathfrak{R}^2$, φ is a scale parameter that controls the speed at which spatial correlation approaches 0 as the distance between locations grows, and, finally, $\tau^2 \geq 0$ is known as the nugget effect, in accordance with the usual geostatistical terminology. The nugget effect can be interpreted as a measurement error or a microscale variation or a combination of both.

Conditionally on S, the process $\{Y(x), x \in I\}$ consists of random mutually independent variables and, for each location $x \in I$, the error distribution or $[Y(x) \mid S]$ has a density that depends only on the conditional mean $E[Y(x_i) \mid S(x_i)]$. A known link function g relates the conditional mean and $S(x)$ so that $E[Y(x_i) \mid S(x_i)] = g^{-1}(S(x_i))$.

When the regression parameters β are of interest, it is important to remember that their interpretation is more conditional than marginal. In particular, $E[Y(x_i) \mid S(x_i)]$ and $E[Y(x_i)]$ differ in terms of the structural dependence of the explanatory variables $d(x_i)$; thus, the interpretation of β calls for caution. Only in the case where $Y(x_i) \mid S(x_i)$ is Gaussian and the link function is identity parameter, comparison is direct. The need to distinguish between conditional and marginal regression parameters, which is not possible in Gaussian linear models, is well known in the context of GLMs for longitudinal data (see, e.g., [12]).

To estimate the parameters for the GLSM and due to the fact that the stationary Gaussian process $S(x)$ is not observable, it is not possible to obtain a closed-form likelihood function except as a high-dimension integral. Reference [11] suggests using algorithms based on Markov chain Monte Carlo (MCMC) to calculate GLSM parameters in a Bayesian framework. This is the approach used in our analysis, implemented using geoR and geoRglm packages (free open-source programs for use with R statistical software [13]).

2.2. ROC Curves. When the marginal distribution (in the GLM) or conditional distribution (in the GLSM) of the response variable Y follows a binomial distribution, the models can be called binary classification systems. The exactitude of a diagnostic test for a binary classification system can be summarized as a receiver operating characteristic (ROC) curve, which is a graphic representation of true positive versus false positive rates when the discrimination threshold is varied. Within the framework of binary GLMs, it is normal to estimate the ROC curves of models in which one or more explanatory variables have been excluded so as to evaluate the effects of these variables. Analysing ROC curves provides tools for comparing and selecting the best models. More precisely, the area under the curve (AUC) of the ROC curve is usually calculated in order to compare the different binary models and thereby select the explanatory variables to be included in the model. Reference [14] described a bootstrap-based method for testing the significant effect of dependent variables on the ROC curve.

We used the AUC and residual semivariograms to demonstrate the goodness-of-fit of the binary GLSM compared with the binary GLM when working with spatially correlated data.

3. Data Description and Model Formulation

3.1. The Studied Area and the Geographic Database. The data used to build the proposed model was collected from borehole samples taken from slate deposits in Baja Cabrera Leonesa (northwest Spain), an area with a long tradition of extracting, processing, and exporting roofing slate.

When surveying a slate deposit, in-depth studies of the rock are performed by taking continuous borehole samples, which enable geologists to study the living rock and analyse the possibility of using it as ornamental slate, see [15]; these samples also reveal the degree of fracturation inside the rock mass.

The specific borehole logging process was based on manual and visual inspection of the borehole by an expert who, after evaluating the aesthetic and functional defects and properties of the slate, differentiated between seams of commercial and unusable slate. The survey was performed by taking a control sample every 25 centimetres; rock quality designation (RQD), however, was defined by homogeneously fractured sections.

A total of 313 equally spaced in-depth observations were obtained, resulting from prior evaluation of various parameters affecting the ornamental quality of the slate and from direct binary values (0 or 1) assigned by the expert to indicate exploitation potential. The 9 specific variables that affected the results of borehole logging were as follows.

(i) RQD: borehole core samples recovered in pieces greater than 10 cm long as a percentage of the total borehole length. This is an indicator of the degree of rock mass fracturing.

(ii) Veins: presence of microfractures filled with quartz that determine the breakage resistance of a commercial slab.

(iii) Crenulations: effect of crenulation cleavage on the main schistosity planes. This increases the roughness of the foliation surfaces of the slate and reduces fissility.

(iv) Kink bands: Presence of microfolding caused by late Variscan deformations.

(v) Sandy laminations: presence of sedimentary sand layers which cut the schistosity planes and have a negative effect on fissility.

Table 1: Correlation matrix of the $p = 9$ explanatory variables that affected the results of borehole logging.

(a)

	RQD	Veins	Crenulation	Kink bands	Sandy laminations
RQD	1				
Veins	0.1624	1			
Crenulation	0.3038	0.1631	1		
Kink bands	0.1040	0.0408	0.1962	1	
Sandy laminations	0.0177	0.0187	0.0200	0.0736	1
Microfractures	0.6696	0.0035	0.1284	−0.0590	−0.0138
Pyrite	−0.1953	−0.0403	−0.1221	−0.0891	−0.1202
Oxidation	0.2053	0.0837	0.1722	0.0821	−0.0645
Rough cleavage	−0.0908	−0.0593	0.2096	−0.0058	−0.0454

(b)

	Microfractures	Pyrite	Oxidation	Rough cleavage
Microfractures	1			
Pyrite	−0.1587	1		
Oxidation	0.1162	−0.1152	1	
Rough cleavage	−0.0989	0.0065	−0.0366	1

(vi) Microfractures: presence of barely visible fractures which determine the breakage resistance of slabs measuring 3–5 mm thick.

(vii) Pyrite: presence of iron sulphides.

(viii) Oxidation: degree of oxidation of iron sulphides in the slate.

(ix) Rough cleavage: slate with poor fissility due to textural heterogeneity.

In-depth knowledge of the variability and distribution of exploitable slate and possible correlation between properties are conducive to the use of GLSM to spatially model the geographic database.

Table 1 shows the correlation matrix of the explanatory variables given above in order to know the degree of dependence between them.

3.2. Model Formulation. The response variable, $Y(x)$, takes the values 0 or 1 to indicate disposable or exploitable slate respectively, in a particular location x. It is assumed in what follows that slate exploitability is a spatial phenomenon that can be modelled using a GLSM. In other words, conditional to the Gaussian process $S(x)$, the data $Y(x_i)$, $i = 1, \dots, n$ follow the classic GLM. The role of $S(x)$ is, therefore, to explain the residual spatial variation after considering all the known explanatory variables. It is also reasonable to assume that the conditional distribution of exploitability can be modelled as a binomial distribution, which is why a binomial error distribution was considered in our study.

Binomial error distribution was used by Diggle et al. [6] and Zhang [8]. A class of transformations that can be used as link functions for this distribution was described by [16]. We assume the stationary Gaussian process $S(\cdot)$ to be the basis for a model of spatial variation in the probability, $P(x)$, that the slate in x is exploitable, but with a logit transformation to map the domain of $S(\cdot)$ onto the unit interval. Thus,

$$g\left[P(x)\right] = \log\left\{\frac{P(x)}{\left[1 - P(x)\right]}\right\} = \mu + S(x). \tag{1}$$

The regression function $E[Y(x_i) \mid S(x_i)]$ varies spatially only through $S(x)$ in the locations x_i.

We adopted a Bayesian framework for inference and prediction of the parameters, using algorithms based on MCMC.

The parameters of this binomial GLSM are $\theta = (\sigma^2, \varphi)$ and $\beta = (\beta_0, \dots, \beta_p)$, where β_0 is the independent term and $\beta_1, \dots, \beta_p$ are the regression coefficients corresponding to each known dependent variable.

4. Statistical Analysis

We initially included all the variables that characterize slate exploitability, namely, RQD, veins, crenulations, kink bands, sandy laminations, microfractures, pyrite, oxidation, and poor fissility. Taking this data and, considering slate exploitability as the response variable, we fitted a binary GLM, called GLM1. A ROC curve was estimated for this complete binary model and the AUC was 0.99.

Next, binary GLMs were fitted to different groups of dependent variables in an attempt to find the minimum number of variables that would provide a high AUC value, that is, close to 0.99. The model, called GLM2, fitted with the RQD, crenulation, kink band, and microfracture variables obtained an AUC of 0.92. Figure 1 shows the ROC curves and the corresponding AUC values for both GLM1 and GLM2. The study was continued with the four variables included in GLM2, given that the reduction in the number of variables did not overly affect the accuracy of the model. A binary nonspatial GLM was fitted using Bayesian methods and the

Table 2: Estimated coefficients and 2.5%, 25%, 75%, and 97.5% quantiles for the GLM.

	Coefficient	2.5% quantile	25% quantile	75% quantile	97.5% quantile
β_0 (intercept)	−35.6747	−77.6804	−49.0098	−26.0904	−18.6385
β_1 (RQD)	1.2826	−1.1476	0.4192	2.0516	3.9633
β_2 (crenulation)	0.5875	0.3288	0.5089	0.6761	0.8352
β_3 (kink band)	0.5668	0.3965	0.5019	0.6395	0.7857
β_4 (microfracture)	2.5815	0.9221	1.6093	3.9181	6.7967

Table 3: Estimated coefficients and 2.5%, 25%, 75%, and 97.5% quantiles for the GLSM.

	Coefficient	2.5% quantile	25% quantile	75% quantile	97.5% quantile
β_0 (intercept)	−27.7021	−32.0664	−29.6753	−24.7446	−17.7297
β_1 (RQD)	0.2521	−0.0385	0.1291	0.3726	0.5388
β_2 (crenulation)	0.7334	0.3762	0.6121	0.8283	1.0691
β_3 (kink band)	0.9189	0.5881	0.7802	1.0274	1.2261
β_4 (microfracture)	1.2911	0.8834	1.0125	1.4480	1.6261

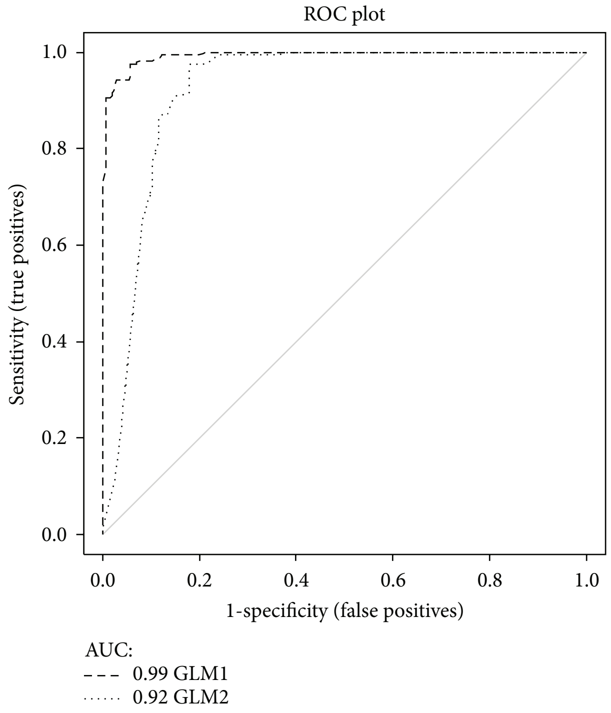

Figure 1: ROC curves and the corresponding AUC values for the two binary GLMs included in the study.

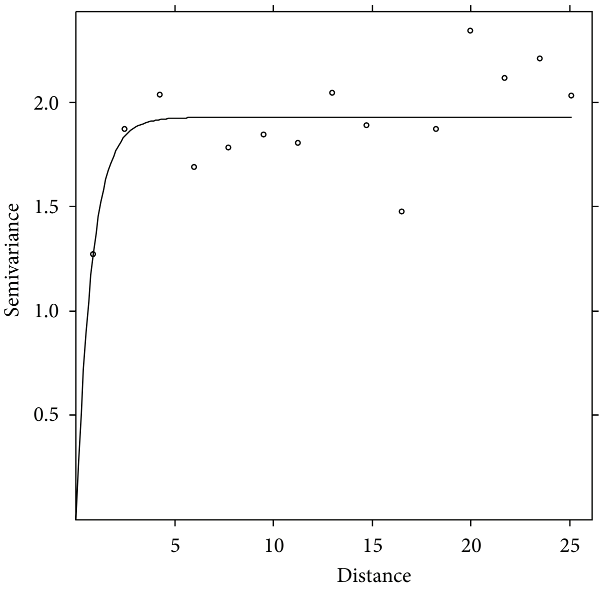

Figure 2: Experimental semivariogram for the GLM2 residuals (points), together with the fitted theoretical model fitted according to an exponential model (continuous line).

MCMClogit function from the MCMCpack (R language). The vector β of the parameters fitted for the GLM2 model was $\beta = (-35.6747, 1.2826, 0.5875, 0.5668, 2.5815)$, where the first value is the independent term and the remaining values are the corresponding coefficients of the explanatory variables in the same order as they were mentioned above.

Table 2 shows the estimated coefficients and the 2.5%, 25%, 75%, and 97.5% quantiles for the fitted model. It can be observed that, for a significance level of $\alpha = 0.05$, the only nonsignificant variable is the RQD.

The study of the residuals of the GLM2 model fitted with four explanatory variables detected a spatial dependence that could be modelled using an exponential theoretical semivariogram with range 0.81 and sill 1.61. Figure 2 shows the experimental semivariogram for the GLM2 residuals, together with the corresponding fitted theoretical model.

Given the presence of spatial dependence in the residuals, it then made sense to fit a GLSM, maintaining four explanatory variables and assuming an exponential model for the process $S(x)$. The correlation function considered was of the type $\operatorname{cov}(S(x), S(x')) = \sigma^2 \rho(x, x'; \varphi) + \tau^2 1\{x = x'\}$, with $\rho(u; \varphi) = \exp(-u/\varphi)$. The fit was made using Bayesian inference implemented via the MCMC algorithms. The first 10,000 sample observations from the simulation were ignored as burn-in, at which point it was considered that convergence time had been achieved. The subsequent samples were used to obtain the subsequent distribution of the parameters of

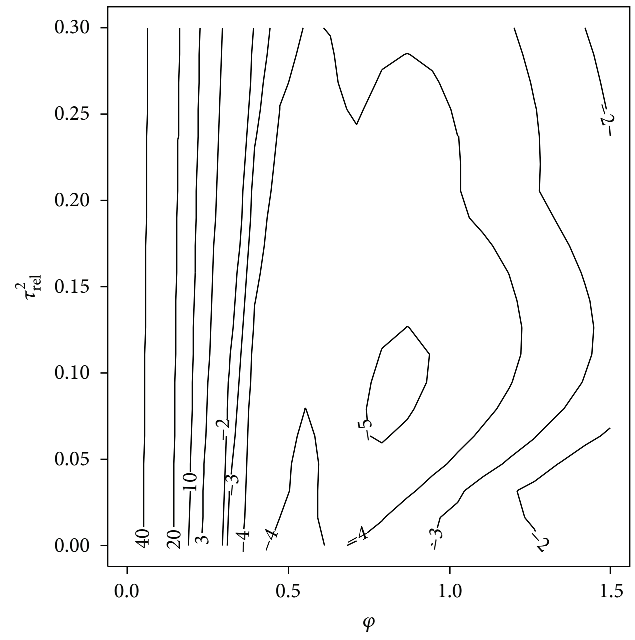

Figure 3: Two-dimensional likelihood profile for (φ, τ^2) for the GLSM.

interest. The chain was sampled for each 100 of the 50,000 iterations to obtain samples containing 500 values (for further details, see [17]).

Using this procedure, the parameters were estimated as $\varphi = 0.82$, $\sigma^2 = 1.46$, $\tau^2 = 0.09$, and $\beta = (-27.7021, 0.2521, 0.7334, 0.9189, 1.2911)$. Table 3 shows the estimated β coefficients and their 2.5%, 25%, 75%, and 97.5% quantiles. Once again we can see how, for a significance level of $\alpha = 0.05$, only the RQD variable was not significant.

It is important to remember that these parameters should be conditionally and not marginally interpreted and so should not be directly compared with the parameters estimated for the GLM. Direct comparison of a spatial and nonspatial GLM could lead to erroneous conclusions, as the estimation methods are fundamentally different. Nonetheless, there is a certain correlation in the conclusions to be drawn from these tables, with variables such as crenulation, kink band, and microfracture remaining significant; RQD, on the other hand, was not significant at 5% level.

Figure 3 shows the likelihood profile in two dimensions for parameters (φ, τ^2) of the model, illustrating the flatness of the likelihood surface obtained using the MCMC algorithms.

A study of the spatial dependence of the GLSM residuals indicated that the spatial component had been correctly modelled on this occasion. Figure 4 shows an experimental semivariogram of the GLSM residuals and a theoretical model fitted according to a nugget effect model. It is clear that the empirical semivariogram is essentially flat, which suggests a suitable fit to the spatial structure. A direct comparison between Figures 2 and 4 leads to the conclusion that the spatial dependence has been properly captured by the stationary Gaussian process S.

The ROC curve and AUC were calculated for the GLSM. The AUC of the binary spatial model was 0.99, which indicates a substantial improvement in the precision of the GLSM. This improvement is reflected in Figure 5, which depicts the ROC curves and AUC values for the GLM2 and GLSM binary models.

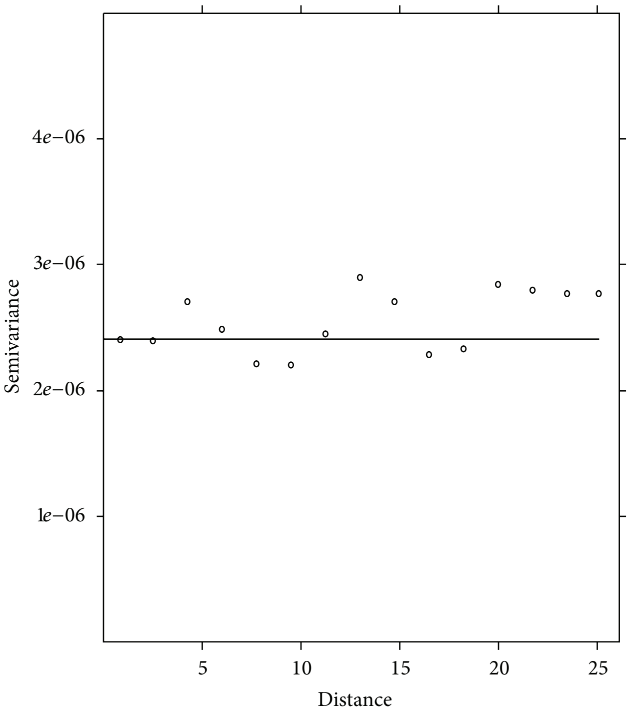

Figure 4: Experimental semivariogram for the GLSM residuals (points), together with the theoretical model fitted according to a nugget effect model (continuous line).

Table 4: Error rates for the binary models for three scenarios representing different levels of prediction difficulty.

	5% test	10% test	15% test
GLM	$6.6E-2$	$6.3E-2$	$5.6E-2$
GLSM	$4.1E-2$	$5.1E-2$	$4.6E-2$

The comparison between the two models was completed with a simulation study, designed to compare the reliability of the predictions in three scenarios of varying levels of difficulty. Randomly selected for the first scenario was 95% of the 313 initial observations, composing the training set that was used to fit the GLM and GLSM. The fitted models were then validated with the remaining 5% of the observations. This procedure was repeated 100 times and the number of errors in the slate exploitability prediction was recorded for each repetition. For the second scenario, we randomly selected 90% of the observations for the training set and the remaining 10% made up the test set. This simulation was also repeated 100 times. The procedure for the third scenario was similar, but this time 85% and 15% of observations made up the training and test sets, respectively.

Table 4 displays the error rates for the three scenarios described, with the error rate calculated as the ratio between the number of prediction errors and the total number of predictions in the test set.

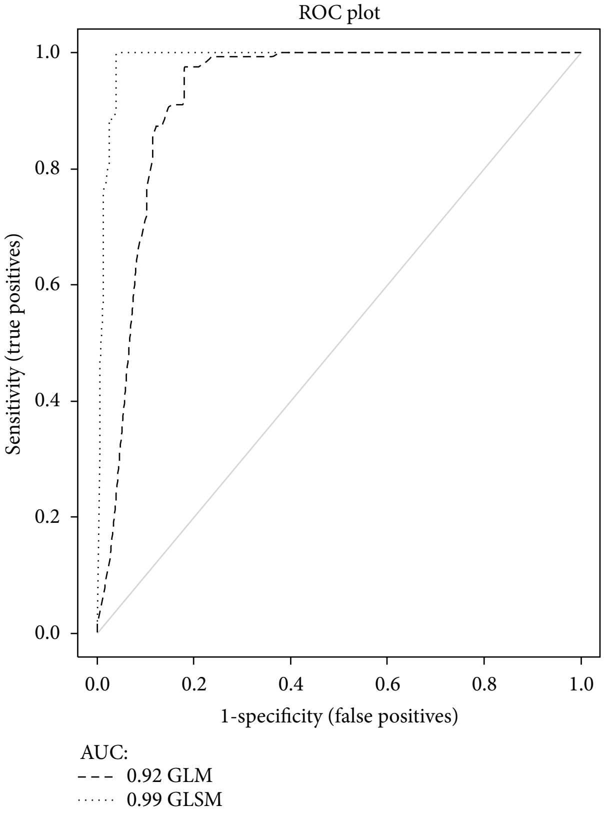

Figure 5: ROC curves and corresponding AUC values for the binary GLM2 and the GLSM.

In all the cases, it can be observed that the GLSM provided a better explanation not only of the effect of the variables determining slate quality, but also of the spatial behaviour of exploitable slate, thereby producing lower prediction error rates.

5. Conclusions

A general interpretation of the GLSM used in our analysis is that the spatial term S represents the accumulative effect of possible explanatory variables with an undetermined spatial structure, which have, therefore, not been observed.

In GLMs, the fact that the spatial correlation of the variables is not taken into account can significantly affect the quality of statistical results. Our study highlights the potential risk of using GLMs when the data is spatially structured.

The conclusion reached after comparing ROC curves and their corresponding AUCs is that GLSMs predict slate exploitability better than GLMs. Therefore, it would seem essential to include unexplained spatial variation when modelling spatially correlated variables.

Based on the comparison of the semivariograms of the GLM and GLSM residuals, we would like to draw attention to the presence of spatial dependence in the GLM residuals, in contrast to what occurs when a GLSM is implemented. This indicates that spatial dependence has been captured correctly by the stationary Gaussian process S. We can, therefore, conclude that a GLSM is more suitable for modelling spatial dependence, which is overlooked by classic GLMs.

The simulation study demonstrates that, for varying levels of prediction difficulty, the GLSM had lower error rates than the GLM.

Although the parameters of the GLSM must be interpreted conditionally rather than marginally to S, the results of the statistical analysis denote the broader potential of the GLSM compared to the classic GLM in analysing spatial data. They also underline the potential risk of reaching erroneous conclusions when using nonspatial models to analyse spatially structured data.

Acknowledgments

This work was funded partly by the Project INCITE10REM304009PR of the Xunta de Galicia and by the Projects MTM2008-03129 and MTM2011-23204 of the Ministry of Science and Innovation.

References

[1] J. M. Matías, A. Vaamonde, J. Taboada, and W. González-Manteiga, "Support vector machines and gradient boosting for graphical estimation of a slate deposit," *Stochastic Environmental Research and Risk Assessment*, vol. 18, no. 5, pp. 309–323, 2004.

[2] F. G. Bastante, J. Taboada, L. Alejano, and E. Alonso, "Optimization tools and simulation methods for designing and evaluating a mining operation," *Stochastic Environmental Research and Risk Assessment*, vol. 22, no. 6, pp. 727–735, 2008.

[3] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society A*, vol. 135, pp. 370–384, 1972.

[4] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, London, UK, 1989.

[5] O. F. Christensen and R. Waagepetersen, "Bayesian prediction of spatial count data using generalized linear mixed models," *Biometrics*, vol. 58, no. 2, pp. 280–286, 2002.

[6] P. Diggle, R. Moyeed, B. Rowlingson, and M. Thomson, "Childhood malaria in the Gambia: a case-study in model-based geostatistics," *Journal of the Royal Statistical Society C*, vol. 51, no. 4, pp. 493–506, 2002.

[7] P. J. Diggle, P. J. Ribeiro,, and O. F. Christensen, "An introduction to model-based geostatistics," in *Spatial Statistics and Computational Methods*, J. Møller, Ed., pp. 43–86, Springer, New York, NY, USA, 2003.

[8] H. Zhang, "On estimation and prediction for spatial generalized linear mixed models," *Biometrics*, vol. 58, no. 1, pp. 129–136, 2002.

[9] H. Zhang, "Optimal interpolation and the appropriateness of cross-validating variogram in spatial generalized linear mixed models," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 698–713, 2003.

[10] N. Breslow and D. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, pp. 9–25, 1993.

[11] P. J. Diggle, J. A. Tawn, and R. A. Moyeed, "Model-based geostatistics," *Journal of the Royal Statistical Society C*, vol. 47, no. 3, pp. 299–350, 1998.

[12] P. J. Diggle, K. Y. Liang, and S. L. Zeger, *The Analysis of Longitudinal Data*, Clarendon, Oxford, UK, 1994.

[13] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, http://www.r-project.org/ .

[14] M. X. Rodríguez-Álvarez, J. Roca-Pardiñas, and C. Cadarso-Suárez, "ROC curve and covariates: extending induced methodology to the non-parametric framework," *Statistics and Computing*, vol. 21, no. 4, pp. 483–499, 2011.

[15] J. Taboada, A. Vaamonde, A. Saavedra, and A. Arguelles, "Quality index for ornamental slate deposits," *Engineering Geology*, vol. 50, no. 1-2, pp. 203–210, 1998.

[16] V. de Oliveira, B. Kedem, and D. A. Short, "Bayesian prediction of transformed Gaussian random fields," *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1422–1433, 1997.

[17] O. F. Christensen, "Monte Carlo maximum likelihood in model-based geostatistics," *Journal of Computational and Graphical Statistics*, vol. 13, no. 3, pp. 702–718, 2004.