Estimating Binary Spatial Autoregressive Models for Rare Events*

Raffaella Calabrese (University of Essex)
Johan A. Elkink (University College Dublin)

December 28, 2014

Abstract

This paper proposes a new statistical estimator, to be applied to the prediction of state failures. State failures are typically conceptualised in a binary fashion—a state fails or it does not—and are rare events. Furthermore, state failures are not geographically independent events. The failure of one state can be expected to have an impact on the stability and peace in neighboring states, increasing the probability of state failures in geographical contiguous regions. Currently there is mixed evidence of such diffusion of state failure taking place Iqbal and Starr (2008). This paper proposes a new estimator to be used for estimates of the spatial interdependence among state failures and focuses on the ability to predict state failures in an international context.

The most used spatial regression models for binary dependent variable consider a symmetric link function. When the dependent variable represents a rare event, a symmetric link function is not coherent. Following Calabrese and Osmetti (2013), we suggest the quantile function of the Generalized Extreme Value (GEV) distribution as link function in a spatial generalised linear model and we call this model the Spatial GEV (SGEV) regression model. To estimate the parameters of such model, a modified version of the Gibbs sampling method of LeSage (2000) and Wang and Dey (2010) is proposed. We analyse the performance of our model by Monte Carlo simulations and evaluating the prediction quality in empirical data on state failure.

*Paper to be presented at the 2015 Asian Political Methodology Conference, January 9–10.

1 Introduction

State failures are typically conceptualised in a binary fashion—a state fails or it does not—and are rare events. Furthermore, state failures are not geographically independent events. The failure of one state can be expected to have an impact on the stability and peace in neighboring states, increasing the probability of state failures in geographical contiguous regions. Currently there is mixed evidence of such diffusion of state failure taking place Iqbal and Starr (2008). This paper proposes an estimator the the prediction of state failures in an international context, taking the rare nature of the dependent variable and spatial interdependence into account.

Both the rare nature of the dependent variable and the spatial autoregressive nature generate special challenges for the statistical estimation of the model. The most used spatial regression models for binary dependent variable consider a symmetric link function (logit or probit functions). When the dependent variable represents a rare event, a symmetric link function is not coherent. King and Zeng (2001*a*,*b*) discuss the estimation of binary dependent variable models for rare events, applying a selection and subsequent correction strategy to improve the prediction quality of such models and apply this model to the study of state failures. We suggest the quantile function of the Generalized Extreme Value (GEV) distribution as link function (Calabrese and Osmetti, 2013) in a spatial generalized linear model and we call this model the Spatial GEV (SGEV) regression model. To estimate the parameters of such model, a modified version of the Gibbs sampling method of LeSage (2000) and Wang and Dey (2010) is proposed. LeSage (2000) proposes a Gibbs sampler for estimating binary dependent variable models with spatial interdependence, which comparative analysis shows is one of the best estimators available in this context (Calabrese and Elkink, 2014). Wang and Dey (2010) propose a Gibbs sampler for rare events using the GEV distribution. We merge the two efforts to develop a new estimator for rare events with spatial autocorrelation.

While this paper focuses on the application of state failures, building on the work by King and Zeng (2001*a*,*b*), the estimator can be applied in a wide range of different contexts. Since the focus is on the quality of prediction, the estimator is of particular relevance to areas such as credit risk, risk of defaults, but also other political science areas such as the prediction of international conflict or economic crisis. The interdependence here can be spatial, specified in an exogenously given spatial weights matrix, or any other kind of network structure, such as for example credit lines (Calabrese, Elkink and Guidici, 2014).

The proposed estimator will be evaluated in a series of Monte Carlo simulations, comparing the performance to a number of existing estimators for similar data structures, where we evaluate both the quality of the estimation of the model coefficients and the accuracy of the prediction. In so doing, this research makes a significant contribution to the literatures on spatial econometrics, rare events modeling, and the study of state failure and international factors in domestic violent conflict.

Section 2 provides a brief overview of binary regression models with spatially interdependent data. Section 3 will outline the estimation complications arising from the nature of rare events data and proposes the use of an asymmetric link function in the binary regression model. Section 4 proposes our Spatial Generalized Extreme Value model for the estimation of rare events data with spatial or network interdependence. Section 5 provides a Monte Carlo analysis to evaluate the statistical performance of the proposed estimator. An initial empirical application is presented in Section 6 and Section 7 concludes.

2 Spatial binary regression models

A widely used representation of a regression model for a binary response Y is the latent response model (Verbeek, 2008). A continuous variable Y^* is the dependent variable with the observation mechanism

$$Y_i = \begin{cases} 1, & Y^* > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

A linear model is specified for this latent response, so the model specification is

$$\mathbf{Y}^* = \rho \mathbf{W}\mathbf{Y}^* + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where the error term $\boldsymbol{\varepsilon}$ can follow a multivariate normal distribution in a probit model or a multivariate logistic distribution in a logit model. $\mathbf{W}$ is a spatial lag weights matrix and ρ the associated scalar parameter. This corresponds to the lattice perspective on spatial data (Anselin, 2002, 255),[1] which can be directly applied to any other (social) network data—any application where the dependent variable represents a binary characteristic of the nodes and the edges (i.e. $\mathbf{W}$) are exogenously given.

From the model (2.2), the Binary Spatial AutoRegressive model (BSAR) is obtained

$$\mathbf{Y}^* = (I - \rho \mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2.3}$$

with $\mathbf{A} = I - \rho \mathbf{W}$ and $\mathbf{e} = \mathbf{A}^{-1}\boldsymbol{\varepsilon}$ (see also McMillen, 1992, 1995; Fleming, 2004).

The variance of the error term follows as

$$var(\mathbf{e}) = var\left[\mathbf{A}^{-1}\boldsymbol{\varepsilon}\right] = \sigma^2_{\boldsymbol{\varepsilon}}\left[\mathbf{A}'\mathbf{A}\right]^{-1}. \tag{2.4}$$

We further specify

$$\mathbf{D} = diag(\boldsymbol{\sigma}_{\mathbf{e}}) \tag{2.5}$$

to be the diagonal matrix with diagonal elements $\boldsymbol{\sigma}_{\mathbf{e}}$ that represent the root square of the diagonal elements in matrix (2.4) and

$$\mathbf{q} = \mathbf{D}^{-1}\mathbf{A}^{-1}\mathbf{X}\boldsymbol{\beta}. \tag{2.6}$$

The inherent heteroskedasticity present in matrix (2.4) renders standard binary regression models inconsistent and inefficient (McMillen, 1992), a problem which has been addressed in the literature by the development of various different estimators for BSAR models. We can identify five main estimators of the spatial autocorellation parameter ρ in this context (see also Fleming, 2004). In the first method, McMillen (1992, 1995) uses an EM algorithm. The latent variable Y^* is replaced with its expected value and the Maximum Likelihood (ML) method is applied. Analogously to McMillen (1992), LeSage (2000) also replaces the latent variable Y^* with its expected value, but a Gibbs sampling approach is applied for the parameter estimation. In the third method, since the likelihood function is a multivariate normal distribution, Beron and Vijverberg (2004) suggest to apply the recursive importance sampling (RIS) to the ML estimation. Pinkse and Slade (1998) apply a Generalized Method of Moments (GMM). Finally, Klier and McMillen (2008) suggest an approximation of the method proposed by Pinkse and Slade (1998), whereby an extrapolation is applied based on the estimate of β when $\rho = 0$.

[1] The altenative, geostatistics, perspective concerns spatial data where space is seen as continuous and observations measured at specific coordinates (Bivand, 1998; Anselin, 2002, 255).

Calabrese and Elkink (2014) analyse the properties of the above estimators by Monte Carlo simulations and an empirical application. This study shows that the Gibbs sampler performs best for low values of the spatial autocorrelation parameter and the RIS estimator for high values of ρ. The computationally much more efficient linearized GMM estimator of Klier and McMillen (2008) performs well under low autocorrelation and large sample size conditions. Because of these properties, we propose a modified version of the Gibbs sampling approach to binary spatial autoregression models, modified for rare events data using an asymmetric link function.

3 Rare events and symmetric link functions

Let Y be a Bernoulli random variable with parameter π and $\mathbf{x}$ a covariate vector. The most used regression models for a binary dependent variable are the Generalized Linear Models (GLMs). In GLMs the link function $g(\cdot)$ is a monotonic function such that

$$g(\pi) = \beta'\mathbf{x}.$$

For simplicity, the most used link functions are symmetric functions. For example, in the logistic regression model the probability $P\{Y = 1\} = \pi$ is the logistic cumulative distribution function (cdf)

$$\pi(\mathbf{x}) = \frac{\exp(\beta'\mathbf{x})}{1+\exp(\beta'\mathbf{x})}. \tag{3.7}$$

We apply the inverse function to equation (3.7) to obtain the logit link function

$$g(\mathbf{x}) = logit(\pi(\mathbf{x})) = ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta'\mathbf{x},$$

which is a symmetric function.

In the probit model the response curve $\pi(\mathbf{x})$ is the cdf of a standard normal random variable

$$\pi(\mathbf{x}) = \Phi(\beta'\mathbf{x}), \tag{3.8}$$

and the link function g of a probit model follows as

$$g(\mathbf{x}) = \Phi^{-1}(\beta'\mathbf{x}),$$

the quantile function of a standard normal random variable, which is also symmetric.

When the link function is symmetric, the response curve $\pi(\mathbf{x})$ approaches zero at the same rate it approaches one. When the dependent variable Y is a rare event, a symmetric link function is not appropriate. Since a Poisson distribution is usually used for counting a rare event and is positively skewed, it is more coherent to choose an asymmetric link function for binary regression models. Moreover, if the response curve $\pi(\mathbf{x})$ approaches zero and one at the same rate, this means that the ones and zeros of the dependent variable Y contain the same information. On the contrary, when Y is a rare event with a low frequency of ones in the sample, observed ones are more informative than observed zeros and should be weighted accordingly in the estimation of the model.

As a result of these characteristics, when the dependent variable Y is a rare event, a GLM with a symmetric link function potentially underestimates the probability π (Calabrese and

Osmetti, 2013; King and Zeng, 2001*b*). In order to overcome this drawback, a stratified sampling approach is commonly suggested (Manski and Lerman, 1977; McCullagh and Nelder, 1989; King and Zeng, 2001*a*,*b*).[2] Instead, Calabrese and Osmetti (2013) propose to focus the attention on the tail of the response curve for the values close to one. Since the GEV distribution function is used in the literature to represent the tail of a random variable, Calabrese and Osmetti (2013) propose the GEV cdf as a link function.

4 The Spatial Generalized Extreme Value (SGEV) regression model

To overcome the drawbacks of the spatial logit and probit models outlined in Section 3, we propose a new spatial regression model for binary responses of rare events. When the sample frequency of ones is very low, the observed ones are more informative than the observed zeros. The tail of the response curve for values close to one represent the features of ones. The GEV random variable is used in the literature (e.g. Kotz and Nadarajah, 2000; Falk, Haler and Reiss, 2010) to model the tail of a distribution. For this reason, we follow Calabrese and Osmetti (2013) in considering the GEV cumulative distribution function as response curve in a binary regression model and extend this to model spatial interdependence. Hence, we propose the Spatial Generalized Extreme Value (SGEV) regression model

$$\pi(\mathbf{x}) = \exp\left\{-\left[1+\tau\mathbf{q}\right]^{-1/\tau}\right\}, \tag{4.9}$$

where $\mathbf{q}$ is defined in equation (2.6).

Since a GEV response curve can be asymmetric, the underestimation of $P\{Y=1\}$ may be overcome. Another advantage of the GEV distribution is that it is very flexible with the tail shape parameter τ controlling the shape and size of the tails, with three different families of distributions subsumed under it. The Type II (Fréchet-type distribution) and the Type III (Weibull-type distribution) classes of the extreme value distribution correspond respectively to the case where $\tau > 0$ and $\tau < 0$, while the Type I class (Gumbel-type distribution) arises in the limit as $\tau \to 0$, Fréchet and Weibull distributions are related by a change of sign. For $\tau \to 0$ and $\rho = 0$, the SGEV regression model becomes the response curve of the log-log model, known in the literature (e.g Agresti, 2002). For $\tau > 0$, it becomes the Weibull response curve.

The main estimators proposed in the literature for spatial regression models are analysed in Section 2. Since the Gibbs sampling approach (LeSage, 2000) provides accurate estimates of the parameter ρ under a wide range of different Monte Carlo parameters (Calabrese and Elkink, 2014), we propose a modified version of this method to estimate the SGEV model (4.9). We make use of the Gibbs sampler for the non-spatial GEV model proposed by Wang and Dey (2010). While LeSage (2000) makes use of the Metropolis-Hastings algorithm for the estimation of the spatial parameter ρ, Wang and Dey (2010) use this algorithm for all model parameters. We follow Wang and Dey (2010)'s approach, which provides more accurate and computationally more efficient results.[3]

[2] See King and Zeng (2001*a*,*b*) for the appropriate subsequent correction on parameter estimates.

[3] This will also allow further improvements in future versions of this working paper, since there is a flourishing literature on optimizing Metropolis-Hastings algorithms (e.g. Maclaurin and Adams, 2014; Korattikara, Chen and Welling, 2014; Angelino et al., 2014).

Within a certain range of values of τ the usual regularity conditions for the estimator of this parameter do not hold (Smith, 1985). Note that model (4.9) is only defined for those observations for which $1+\tau\mathbf{q}>0$ (see, e.g., Calabrese and Osmetti, 2013).

To assign the priors of the SGEV model, we use LeSage (2000)'s assumption that the priors are independent

$$\nu(\rho,\boldsymbol{\beta},\sigma,\mathbf{V})=\nu(\rho)\nu(\boldsymbol{\beta})\nu(\sigma)\nu(\mathbf{V}),$$

where we follow Wang and Dey (2010) in assigning a relatively uninformative prior distribution of $\pi\sim N(0,100)$ to all parameters. For the estimation, we relabel Y, such that we estimate a model for rare zeros. Using the estimates of a non-spatial log-log model as the initial values $\boldsymbol{\beta}_0$, the correlation between $\mathbf{y}$ and $\mathbf{Wy}$ as the initial value ρ_0, and $\tau_0=0$, all estimates are updated in each MCMC iteration through the Metropolis-Hastings algorithm. $\sigma^2_{\boldsymbol{\varepsilon}}$ is kept fixed, because of lack of identifiability when simultaneously estimating the error variance and the linear regression coefficients in a latent variable model—similar to probit and logit regressions.

Let $\boldsymbol{\theta}$ be $\boldsymbol{\theta}=[\boldsymbol{\beta}',\tau,\rho]'$. The log-likelihood function is

$$\ell(\boldsymbol{\theta})=\sum_{i=1}^{N}\left(\log\pi_i(\boldsymbol{\theta})^{y_i}+\log(1-\pi_i(\boldsymbol{\theta}))^{1-y_i}\right),$$

where $\pi(\cdot)$ is defined in equation (4.9). The Metropolis-Hastings algorithm (Hastings, 1970) proceeds as follows. For each θ_j, let the value $\theta^*_{j,t+1}=\theta_{j,t}+cZ$ be generated, where Z is a draw from a standard normal distribution, c is a known constant, and t refers to the sampling iteration. The acceptance probability

$$a=\min\left\{1,\frac{\ell(\boldsymbol{\theta}^*_{t+1})\nu_\theta}{\ell(\boldsymbol{\theta}_{t+1})\nu_\theta}\right\},$$

where $\boldsymbol{\theta}^*_{t+1}=\boldsymbol{\theta}_{t+1}$, except for parameter $\theta^*_{j,t+1}$. A value m is drawn from a continuous uniform distribution with support $[0,1]$. If $m<a$, the next draw from the density function (4) is given by $\theta_{j,t+1}=\theta^*_{j,t+1}$, otherwise the draw is taken to be the current value $\theta_{j,t+1}=\theta_{j,t}$. Where the parameter is constrained, Z is drawn from a truncated standard normal distribution—in our application this holds for ρ, which is constrained to the $[-1,1]$ interval. For computational efficiency reasons and following Thomas (2007), we dynamically adapt c throughout the chain for each parameter in $\boldsymbol{\theta}$, such that the acceptance rate of the Metropolis-Hastings algorithm is approximately 60%.[4]

5 Monte Carlo simulations

In order to evaluate the performance of the proposed estimator, we perform Monte Carlo analyses whereby the estimator is applied to data of which the underlying data generation mechanism, and its associated parameters, are known. While the Monte Carlo analysis in Calabrese and Elkink (2014) focuses primarily on the estimation of the intensity of the spatial autocorrelation, ρ, we focus primarily on the prediction quality of the estimator. In particular, we are interested in the accuracy of predicting the actual events, the ones. We compare our SGEV estimator with the predictive performance of the regular logistic regression and the BGEVA estimator proposed by Calabrese and Osmetti (2013).

[4] We constrain c to remain in the $[0.005,10]$ interval.

The data generation process generates one continuous independent variable $X \sim N(0, 4\sigma_\varepsilon)$, whereby σ_ε is the standard deviation of the randomly generated error term. The error term itself follows a Generalized Extreme Value distribution, with location parameter $a = 0$ and scale parameter $b = 1$, while the shape parameter is set to $\tau \in \{-0.25, 0.25\}$ to generate positively skewed data. With the inclusion of the intercept in the model, the linear parameters of the data generation process are $\boldsymbol{\beta} = [0, 1]'$. We performed the simulations using relatively small sample sizes, $N = 500$. The level of spatial autocorrelation is varied from entirely absent to moderate autocorrelation, $\rho \in \{0, 0.5\}$. Finally, the dependent variable y is constructed by applying a threshold different from the zero in (2.1), such that the proportion of ones is predetermined, in these simulations to 20%.[5] We thus end up with four different configurations of the data generation parameters. We perform estimations on 60 replications of each parameter configuration.[6]

To evaluate the predictive quality of the regression model a wide range of statistics are available. We are primarily focused on the correct estimation of the ones in the sample, since these are rare—it would be easy to predict the zeros by simply predicting zeros in all cases. In the context of our empirical example, the prediction of state failure, it is also reasonable to assume that it is more important not to miss a potential future failed state than that it is to slightly overestimate the probability of state failure for a state that is less at risk. A similar rationale applies for example to the risk of bank defaults, currency defaults or default on credits, where (central) banks will be more concerned with correctly identifying those cases that are truly at risk than with incorrectly identifying some cases as at risk while they are not. Applying a similar logic to credit defaults, Calabrese and Osmetti (2013) therefore calculate the Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the subset of cases where $Y = 1$—the error being defined as $y - \hat{\pi}(\mathbf{x})$. Because the SGEV model as defined in (4.9) is only valid for $1 + \tau\mathbf{q} > 0$, we calculate and evaluate predicted probabilities only for those cases where, given the estimates of β and ρ, this is the case.

The left plot in Figure 1 provides a graphical depiction of the MSE for the predicted values, for those cases where we observe $y = 1$. This plot shows that under the absence of spatial autocorrelation ($\rho = 0$), existing models such as the logistic regression or BGEVA outperform the SGEV model when it comes to correctly identifying the ones in the data. This is of course not surprising, given that the complication our estimator is designed to address—the presence of spatial interdependence—is not relevant. When ρ is increased to a moderately high level of autocorrelation, however ($\rho = 0.5$), the accuracy of identifying the ones in the data set is notably better for the SGEV estimator.

While we emphasise the importance of correctly predicting high probabilities for the rare outcome of $Y = 1$, the estimator still needs to classify the cases for both positive and negative outcomes. In other words, while we can prioritize a low true positive rate over a low false positive rate, we still need to make sure that our classifier identifies both negative and positive cases. The right plot in Figure 2 therefore provides the overall Mean Squared Error, including both $y_i = 1$ and $y_i = 0$ cases. It is clear from this plot that, while positive cases are generally better identified by the SGEV estimator, this is combined with a relatively high false positive

[5] This affects the estimation of the intercept, but the performance of the estimation of that particular parameter is not our current concern.

[6] This is an unusually low number of variations and low number of replications as Monte Carlo studies go and this will be increased in future iterations of the paper. Future iterations will include larger sample sizes, level of ρ comparable to those in Calabrese and Elkink (2014) and different proportions of ones in the dependent variable, evaluating each combination of parameters over 1000 replications.

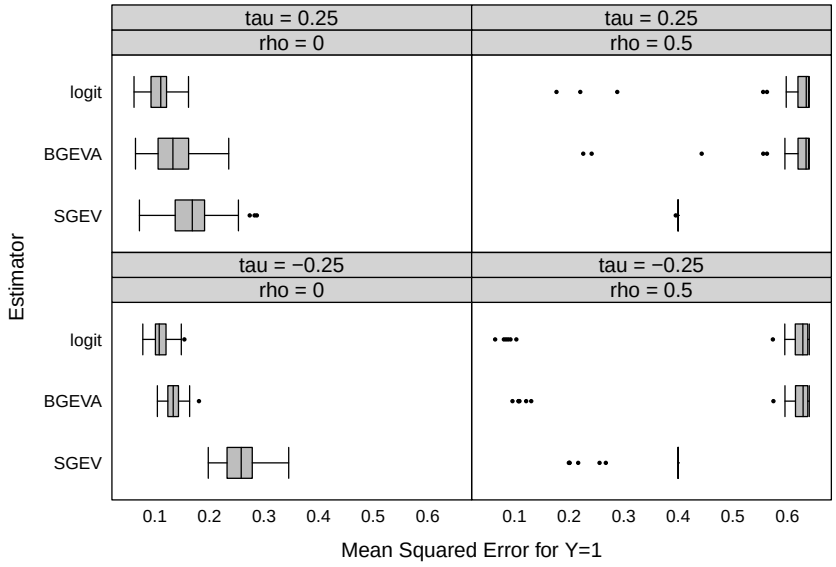

Figure 1: *Distribution of the Mean Squared Error for the prediction of ones ($MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\boldsymbol{\pi}}(\mathbf{x}))^2$, for all $y_i = 1$), for different configurations of the parameters, across 60 replications.*

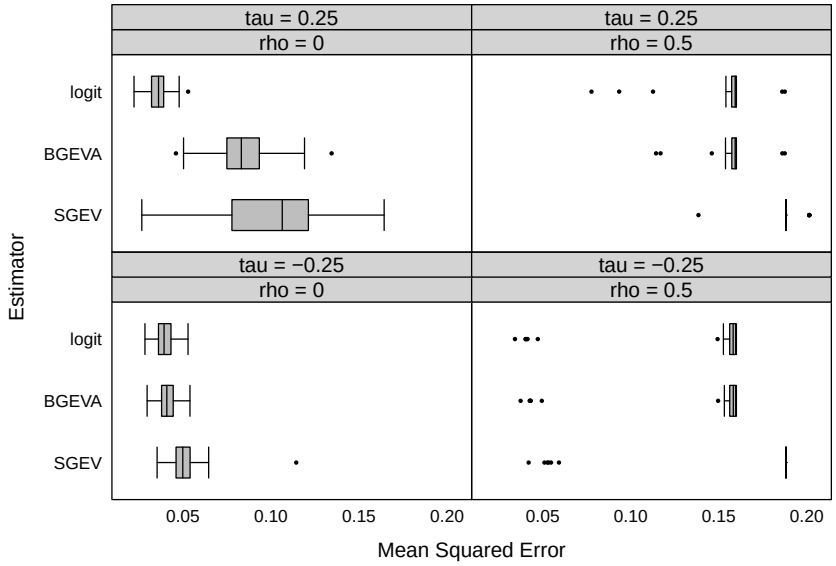

Figure 2: *Distribution of the Mean Squared Error for the prediction of y ($MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\boldsymbol{\pi}}(\mathbf{x}))^2$, for the full sample), for different configurations of the parameters, across 60 replications.*

rate.

Since the prediction of ones and zeros can be sensitive to the choice of the prediction threshold—for what predicted probability $P(Y = 1)$ do we conclude that this is case is probably a 1?—alternative statistics are available that consider the entire range of potential threshold values. A common procedure is to sort all predicted values $\hat{y}^*$, then move the decision threshold along the range of values, and plot the proportion of correctly classified ones against the proportion of incorrectly classified ones. This is referred to as the Receiver Operating Charateristic curve (see, e.g., Hand and Anagnostopoulos, 2014) and is often used to compare the classification quality of different estimators or algorithms. The curve will generally be above the 45 degree diagonal line, since otherwise we could simply create a better classifier by swapping the labelling of the predictions. The further the ROC curve from the diagonal line, the better the classifier. A numerical indicator for this prediction quality is the Area Under Curve (AUC) statistic, which is the area under the ROC curve, and thus ranges typically from 0.5 to 1—an AUC below 0.5 indicates a classifier performing worse than random classification of cases.

The AUC statistic suffers from a number of deficiencies as a measure to evaluate the performance of different estimators. A particular feature of the statistic that is of concern to our Monte Carlo study is the fact that this statistic amounts to a weighted average minimum loss measure. This average loss is determined by the relative cost of misclassifying zeros or ones, whereby the cases are weighted depending on the cumulative density function of the two classes across the range of scores (that is, $F_0(t) = Pr[\hat{y}^*(\mathbf{x}) < t|Y = 0]$ and $F_1(t) = Pr[\hat{y}^*(\mathbf{x}) < t|Y = 1]$, with score $\hat{y}^*(\mathbf{x})$ the prediction of the classifier and t the threshold value) (Hand, 2009, 108–111). In other words, the relative cost of misclassifying ones or zeros—and thus the evaluation of the relative performance of the classifier—are dependent on the classifier used, as opposed to being dependent on the underlying application, which is incoherent. Hand (2009) and Hand and Anagnostopoulos (2013, 2014) outline this complication and provide an alternative to the AUC score, the so-called H-index, which addresses this deficiency. Similar to the AUC statistics, the H-index has a range of 0 to 1, with high values indicating improved performance. We use the `hmeasure` package in R (Anagnostopoulos and Hand, 2012) to calculate the H-index statistics, the results of which are summarized in Figure 3.

Under the absence of spatial autocorrelation ($\rho = 0$) we again see a relatively worse performance for the SGEV estimator compared to the more conventional logistic regression. Again, while this implies a warning for using this type of model in a context where there is no strong reason to expect spatial autocorrelation, this is not the type of application we have in mind when developing the estimator. Under spatial autocorrelation, however, the right half in Figure 3, the H-index does show better results for the SGEV estimator than it does for the logistic or BGEVA regression models. The discrepancy between this finding and the only marginally improved true positive rate—combined with a high false positive rate—can be explained by the choice of threshold used for the classification. While the convention is to take $\pi^* = 0.5$ as the threshold value, this is by no means the obvious choice in the context of rare events, where the identification of the ones is inherently more important than the identification of the zeros (Calabrese, 2014).

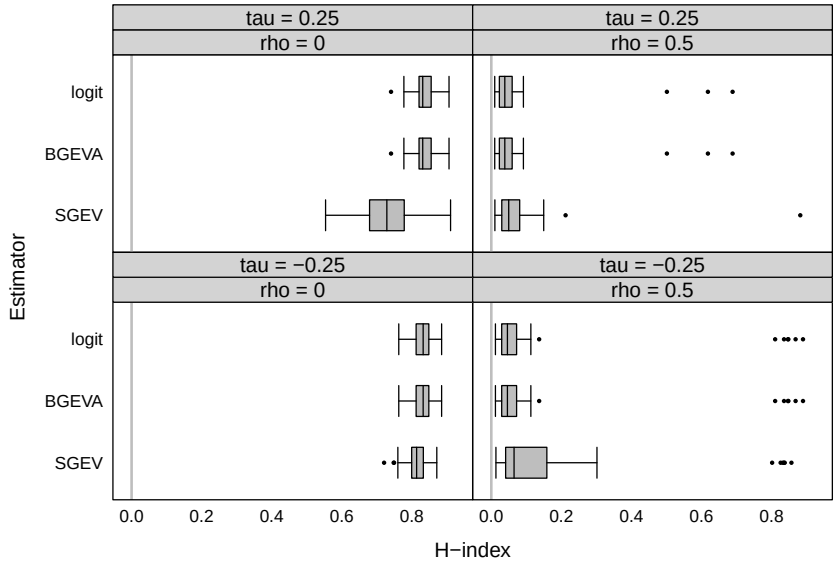

Figure 3: *Distribution of the H-index, for different configurations of the parameters, across 60 replications.*

6 Empirical application to state failure

To test the model on data related to state failure, we use a definition of W that is block-diagonal, with

$$W_{ijt} = \begin{cases} 1, & R(i) = R(j) \text{ and } i \neq j \\ 0, & \text{otherwise,} \end{cases} \tag{6.10}$$

with $R(i)$ the region of observation i, which is constant over time.[7] This W is subsequently normalized such that all rows add up to one. The model specification we use is inspired by, but does not closely follow, Goldstone et al. (2010).[8] In their replication data, specific matched samples are used to address the rare events nature of the data, while we estimate our models on a full data set of all countries and years where we have data on the relevant variables. All independent variables are lagged by one year relative to the dependent variable. To account for world-wide shocks, year dummies are included. The independent variables are sourced from the *Nations, Development, and Democracy* data set by Wejnert (2007) and the data used is from 1975 onwards.

Table 1 provides the regression estimates for the four models on identical data, with in addition the H-index measure of prediction quality. Strikingly, while for the Monte Carlo analysis the SGEV estimator performed well in the presence of spatial autocorrelation, in this empirical data the results appear weaker—for the within-sample prediction of the outcome variable, the H-index suggests a substantially weaker classifier than the existing models. When we fix τ in the estimation, we obtain what appear to be much better results and indeed in these

[7] Future iterations of the paper will use a contiguity matrix based on immediate adjacency of land borders.

[8] This is a very crude version of the model—improvements to the model specification itself is left to future iterations of the paper.

	Logistic	Gibbs	BGEVA	SGEV*	SGEV
intercept	-17.08	-17.87	-2.910	-19.80	-17.70
Polity IV	-0.034	-0.008	-0.037	-0.041	-68.34
Polity IV squared	-0.010	0.001	-0.001	-0.007	-36.93
Log of infant mortality	0.168	0.295	0.116	-0.064	-1.277
Log of GDP per capita	-0.506	-0.724	-0.281	-0.903	-73.55
ρ		-0.22		-0.10	-0.23
τ			0.25	0.25	6.28
N	1653	1653	1653	1653	1653
H-index	0.467	0.545	0.480	0.559	0.237

Table 1: *Regressions explaining state failure as classified in Goldstone et al. (2010), using regional membership to define the adjacency matrix. Models include year fixed effects and the dependent variable is observed at $t+1$. For the BGEVA and SGEV* models, τ is fixed* a priori.

circumstances, the H-index for the SGEV estimator is the highest, while the regular spatial Gibbs sampler by LeSage (2000) generates good predictions as well.

The estimate for the autoregression parameter ρ by the regular Gibbs sampler, which it is generally good at estimating (Calabrese and Elkink, 2014), suggests an absence of, or even negative spatial autocorrelation—the 95% Highest Posterior Density interval is $[-0.41, -0.05]$ and the same negative coefficient is identified by the SGEV estimator. The weak predictive results of the SGEV estimator might therefore be due to lack of actual spatial autocorrelation in this data, at least when the spatial contiguity matrix is defined as a block-diagonal matrix of six world regions. Alternatively, given the high performance of the SGEV* model, where τ is fixed, the estimation of the shape parameter τ itself might be problematic.

While the paper is primarily concerned with the accurate prediction of the rare event, the empirical results do raise the question whether the estimation of τ in particular is problematic. Figure 4 provides some insight into the performance of the SGEV estimator when it comes to correctly identifying the shape parameter τ. Strikingly, under the absence of any spatial autocorrelation, the estimates of τ are reasonably accurate, albeit with high variance. Under spatial autocorrelation, however ($\rho = 0.5$), the estimates of τ are nearly always zero, which is the initial value τ_0.[9]

7 Conclusion

This paper proposes an estimator designed for rare events, through the application of an asymmetric link function that weighs the positive cases more strongly in the estimation than the negative cases. We provide initial Monte Carlo simulations to evaluate the performance of this estimator. The current implementation provides a high true positive rate—the rare events are generally correctly identified—but also a relatively high false positive rate. The H-index however, which is used for the evaluation of the overall prediction quality with arbitrary decision threshold, shows a slightly better performance of our estimator than existing estimators, in the

[9] Investigation of this pattern is left for the next iteration of the working paper.

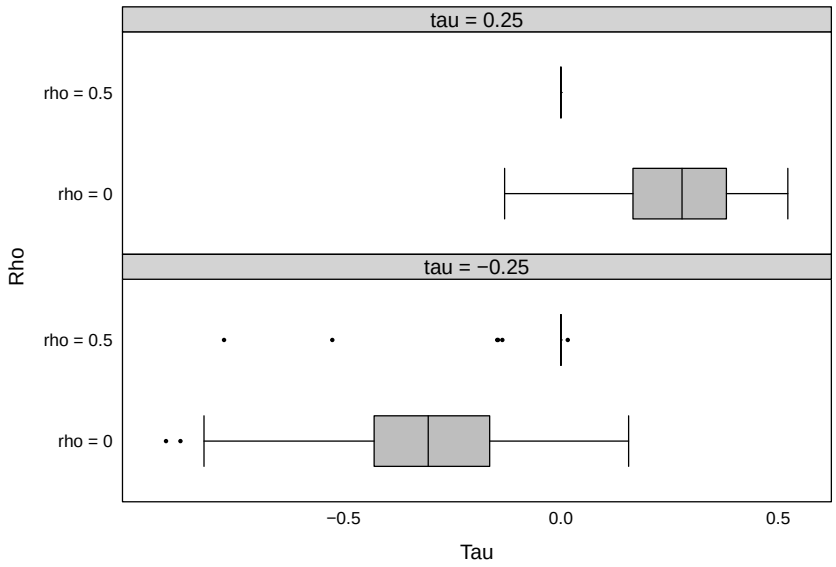

Figure 4: *Distribution of the estimates of τ, for different configurations of the parameters, across 60 replications.*

context of spatial autocorrelation. Future work will improve the quality of the predictions and apply the estimator to the study of state failures in an international context using an appropriate model specification.

References

Agresti, A. 2002. *Categorical Data Analysis*. Wiley, New York.

Anagnostopoulos, Christoforos and David J. Hand. 2012. *hmeasure: The H-measure and other scalar classification performance metrics*. R package version 1.0.

Angelino, Elaine, Eddie Kohler, Amos Waterland, Margo Seltzer and Ryan P. Adams. 2014. "Accelerating MCMC via parallel predictive prefetching." arXiv preprint arXiv:1403.7265.

Anselin, Luc. 2002. "Under the hood. Issues in the specification and interpretation of spatial regression models." *Agricultural Economics* 27:247–267.

Beron, Kurt J. and Wim P.M. Vijverberg. 2004. Probit in a spatial context: a Monte Carlo analysis. In *Advances in spatial econometrics. Methodology, tools and applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer pp. 169–195.

Bivand, Roger. 1998. "A review of spatial statistical techniques for location studies." Paper presented at the CEPR symposium on New Issues in Trade and Location (2277), Lund, Sweden, 28-30 August, 1998.

Calabrese, Raffaella. 2014. "Optimal cut-off for rare events and unbalanced misclassification costs." *Journal of Applied Statistics* 41(8):1678–1693.

Calabrese, Raffaella and Johan A. Elkink. 2014. "Estimators of binary spatial autoregressive models: A Monte Carlo study." *Journal of Regional Science* 54(4):664–687.

Calabrese, Raffaella, Johan A. Elkink and Paolo Guidici. 2014. "Interdependence of European Banks in Distress: An Application of a Binary Spatial Regression Model." Paper presented at the Annual Convention of the International Studies Association, Toronto.

Calabrese, Raffaella and Silvia A. Osmetti. 2013. "Modelling small and medium enterprise loan defaults as rare events: The generalized extreme value regression model." *Journal of Applied Statistics* 40:1172–1188.

Falk, M., J. Haler and R. Reiss. 2010. *Laws of Small Numbers: Extremes and Rare Events*. Springer.

Fleming, Mark M. 2004. Techniques for estimating spatially dependent discrete choice models. In *Advances in spatial econometrics. Methodology, tools and applications*, ed. Luc Anselin, Raymond J.G.M. Florax and Sergio J. Rey. Berlin: Springer pp. 145–167.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted Robert Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder and Mark Woodward. 2010. "A global model for forecasting political instability." *American Journal of Political Science* 51(1).

Hand, David J. 2009. "Measuring classifier performance: A coherent alternative to the area under the ROC curve." *Machine Learning* 77:103–123.

Hand, David J. and Christoforos Anagnostopoulos. 2013. "When is the area under the receiver operating characteristic curve an appropriat emeasure of classifier performance?" *Pattern Recognition Letters* 13:492–495.

Hand, David J. and Christoforos Anagnostopoulos. 2014. "A better Beta for the *H* measure of classification performance." *Pattern Recognition Letters* 40:41–46.

Hastings, W. K. 1970. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika* 57:97–109.

Iqbal, Zaryab and Harvey Starr. 2008. "Bad neighbors: Failed states and their consequences." *Conflict Management and Peace Science* 25:315–331.

King, Gary and Langche Zeng. 2001*a*. "Explaining Rare Events in International Relations." *International Organization* 55:693–715.

King, Gary and Langche Zeng. 2001*b*. "Logistic Regression in Rare Events Data." *Political Analysis* 9:137–163.

Klier, Thomas and Daniel P. McMillen. 2008. "Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples." *Journal of Business & Economic Statistics* 26(4):460–471.

Korattikara, Anoop, Yutian Chen and Max Welling. 2014. "Austerity in MCMC land: Cutting the Metropolis-Hastings budget." arXiv preprint arXiv:1304.5299.

Kotz, S. and S. Nadarajah. 2000. *Extreme Value Distributions. Theory and Applications*. Imperial College Press, London.

LeSage, James P. 2000. "Bayesian estimation of limited dependent variable spatial autoregressive models." *Geographical Analysis* 32(1):19–35.

Maclaurin, Dougal and Ryan P. Adams. 2014. "Firefly Monte Carlo: Exact MCMC with subsets of data." arXiv preprint arXiv:1403.5693.

Manski, C. F. and S. R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice-based Samples." *Econometrica* 45(8).

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. Chapman Hall, New York.

McMillen, Daniel P. 1992. "Probit with spatial autocorrelation." *Journal of Regional Science* 32(3):335–348.

McMillen, Daniel P. 1995. Spatial effects in probit models: a Monte Carlo investigation. In *New directions in spatial econometrics*, ed. Luc Anselin and Raymond J.G.M. Florax. Berlin: Springer Verlag pp. 189–228.

Pinkse, Joris and Margaret E. Slade. 1998. "Contracting in space: an application of spatial statistics to discrete-choice models." *Journal of Econometrics* 85:125–154.

Smith, R. L. 1985. "Maximum likelihood estimation in a class of non-regular cases." *Biometrika* 72:67–90.

Thomas, Timothy S. 2007. "A primer for Bayesian spatial probits, with an application to deforestation in Madagascar." Companion Paper for the World Bank Policy Research Report on Forests, Environment, and Livelihoods.
URL: *http://www.timthomas.net*

Verbeek, Marno. 2008. *A guide to modern econometrics*. Chichester: John Wiley & Sons.

Wang, Xia and Dipak K. Dey. 2010. "Generalized Extreme Value regression for binary response data: An application to B2B electronic payments system adoption." *Annals of Applied Statistics* 4(4):2000–2023.

Wejnert, Barbara. 2007. "Nations, Development, and Democracy, 1800-2005.". Inter-university Consortium for Political and Social Research, Ann Arbor, NY.
URL: *http://doi.org/10.3886/ICPSR20440.v1*