forests

MDPI

Article

Generalized Models: An Application to Identify Environmental Variables That Significantly Affect the Abundance of Three Tree Species

Pablo Antúnez [1,*], José Ciro Hernández-Díaz [2], Christian Wehenkel [2] and Ricardo Clark-Tapia [1]

[1] División de Estudios de Posgrado-Instituto de Estudios Ambientales, Universidad de la Sierra Juárez, Avenida Universidad S/N, Ixtlán de Juárez, 68725 Oaxaca, Mexico; rclark@unsij.edu.mx

[2] Instituto de Silvicultura e Industria de la Madera, Universidad Juárez del Estado de Durango, Km 5.5 Carretera Mazatlán, 34120 Durango, Mexico; jciroh@ujed.mx (J.C.H.-D.); wehenkel@ujed.mx (C.W)

* Correspondence: pantunez4@gmail.com; Tel.: +52-951-553-6362

Academic Editors: Damian Adams and Timothy A. Martin
Received: 2 December 2016; Accepted: 22 February 2017; Published: 27 February 2017

Abstract: In defining the environmental preferences of plant species, statistical models are part of the essential tools in the field of modern ecology. However, conventional linear models require compliance with some parametric assumptions and if these requirements are not met, imply a serious limitation of the applied model. In this study, the effectiveness of linear and nonlinear generalized models was examined to identify the unitary effect of the principal environmental variables on the abundance of three tree species growing in the natural temperate forests of Oaxaca, Mexico. The covariates that showed a significant effect on the distribution of tree species were the maximum and minimum temperatures and the precipitation during specific periods. Results suggest that the generalized models, particularly smoothed models, were able to detect the increase or decrease of the abundance against changes in an environmental variable; they also revealed the inflection of the regression. In addition, these models allow partial characterization of the realized niche of a given species according to some specific variables, regardless of the type of relationship.

Keywords: climate influence; temperate forest; Mexican flora; climatic niche

1. Introduction

In order to identify potential geographic or environmental space, where organisms can be established and successfully developed, researchers have used different modeling methods. Among them are correlative-type models, which aim to relate the presence or abundance of a species with different predictor variables via a mathematical function [1]. The multiple linear model has been one of the most preferred analysis techniques to explain the stochastic relationship between descriptive variables and distribution of species [2], often providing satisfactory results [3]. However, it is also common that data violate some parametric assumptions. For example, in the field of forestry and forest ecology, predictors usually have high interdependence; moreover, the data do not follow a theoretical normal distribution and there is heteroscedasticity. Sometimes there is a linear relationship between a variable of interest and some covariates, while with the remaining covariates, a curvilinear relationship may exist [4,5].

Several techniques have been explored as options to deal with those dilemmas in ecology, such as Chi-squared Automatic Interaction Detector (CHAID) [6], Machine Learning Techniques (MLTs) [7] or the generalized additive models (GAMs) proposed by Hastie and Tibshirani [8]. GAMs have become one of the most used techniques, in which non-linearity and non-parametric regression are incorporated [1]. These types of model are an extension of Generalized Linear Models (GLMs),

Forests **2017**, *8*, 59; doi:10.3390/f8030059 www.mdpi.com/journal/forests

constructed by the sum of smooth functions of predictor variables, where it is common to use defined polynomials forn intervals known as "splines" [8–10]. For this reason, a function of this type loses its purely parametric nature and becomes a semi-parametric or non-parametric model [11]. In addition, GAMs can be applied without the compliance of independent regressors or a specific normal distribution shape of the sample [10]. Furthermore, with a GAM, the algorithms enable the introduction of any distributions (e.g., Binomial, Poisson or Gamma) that enable the researcher to identify and select the best representation of the data, converting the GAM technique into a viable alternative for forestry and ecology.

GLM and GAM are common tools in ecology to model the response of living organisms to environmental factors using discrete or continuous variables [12–14]. In essence, these models do not have substantial discrepancies between them, sometimes a GLM is more accurate, and in others, a GAM is preferable. When a GAM is better, this is often attributed to its semi-parametric structure [3,15].

Many studies have been carried out to understand the relationship and interactions between environmental variables and the presence of a species in a given locality [15,16], to model habitat distribution [1] or to evaluate the effects of environmental variables [17]. However, in forestry, our knowledge of the relationship between the abundance of forest species and their habitat conditions is scarce, but is essential when we want to find efficient ways to reforest or restore specific areas.

The main purposes of this study were: (i) to find variables that better explain the behavior of the abundance of three tree species that grow naturally in the temperate forests of Oaxaca, Mexico, using 18 environmental variables; and (ii) to identify which type of model (linear or nonlinear) best describes the relationship between the abundance of each species and the analysed covariates. For addressing these objectives, generalized lineal models (GLMs) and nonlinear generalized additive models (GAMs) were tested.

2. Materials and Methods

2.1. Study Area

The study area was the forest of Santiago Comaltepec, which is a portion of the "Sierra Juárez" in the Oaxaca region, in southwestern Mexico, located at 17° 33′ 35″ north latitude and 99° 26′ 32″ west longitude, covering an area of 26.5 km^2 (Figure 1). The elevation above sea level goes from 1800 to 3000 m. The annual average maximum temperature is 13.4 °C; the annual average minimum is 4.7 °C; the summer rainfall from June to September ranges between 600 mm and 1200 mm, while the dry season lasts from December to May with an average precipitation of 225.5 mm and an average temperature of 16.7 °C [18,19].

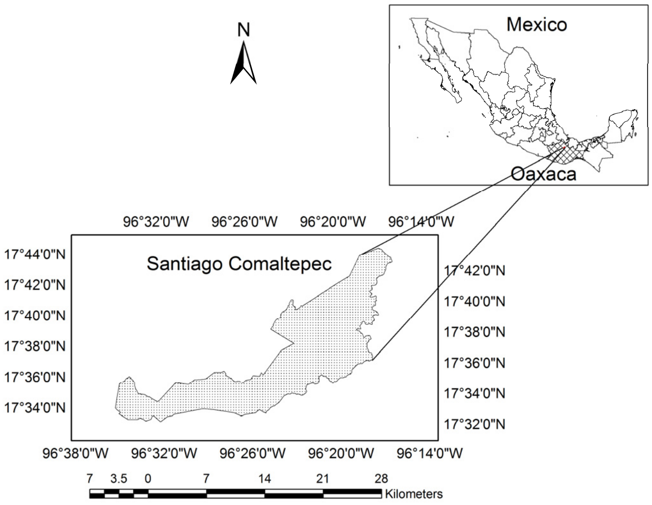

Figure 1. Location of the study area.

According to direct observations from field data and database information that originated from the last forest inventory (2014–2015), there are different types of vegetation in the study area. Oak forest is the main vegetation, followed by tropical deciduous forest and cloud forest.

2.2. Variables and Sample

The frequency and density of a species are common indicators of its abundance [20]. This work refers to density, which is defined as the number of trees of a species per sampling plot. We counted only trees with diameter greater than or equal to 7.5 cm at 1.3 m above the ground level. Data were captured from a total of 634 circular 1000 m^2 plots. These plots were established by a local forestry and technical services office (Union Zapoteca Chinanteca—UZACHI). The plots and the registration of data were developed following the method of the National Forestry Commission, concerning the planning of the temperate forest in Mexico [21].

The species selected for this study, were: *Pinus pseudostrobus* Lindl. var. *apulcensis* (Lindl.) Shaw, *Pinus patula* Schl. et Cham. and *Quercus macdougallii* Martínez. The first is characterized by its fast growth and is commonly used to reforest degraded areas or forest sites without vegetation; the second is in high demand by sawmills, furniture factories, and pulp and paper industries [22]. *Quercus macdougallii*, besides fulfilling ecological functions, is a rare species in the study area; it was found in just 33 out of the 634 plots; moreover, it is a vulnerable species, included in the red list of the International Union for Conservation of Nature (IUCN).

For this study, 18 climatic and physiographic variables were considered, given that in several studies, it has been shown that they affect, in different ways and scales, the geographical distribution of species [23–26]. Table 1 shows the variable acronyms, their meanings and some descriptive statistics. The climatic variables for each site were obtained from the server of the Forest Service, Department of Agriculture of the United States (see Algorithms). These data were computed using methods proposed by Rehfeldt et al. [27] with climate data from 1976 to 1990, originating from about 4000 weather stations in Mexico, the southern part of the United States of America, Guatemala, Belize and Cuba [27–29]. The physiographic variables and elevation above sea level were directly obtained in the field (Table 1).

In developing this study, we did not find quantifiable information useful to evaluate the effect of human activities on the abundance of species in the study area; therefore, we did not analyse this factor, but it is worth to say that, in Santiago Comaltepec, the area under exploitation is incorporated into a certified Management Plan and that the current forest management schemes focus on sustainable development which takes into account biodiversity conservation [30].

2.3. Data Analysis

In order to identify the variables that significantly affect the abundance of the species and to detect the type of relationship (linear or nonlinear), parametric models (GLM) and non-parametric models (GAM) were tested. These two model types were chosen because they have demonstrated satisfactory results for these types of studies [15,31,32] in some cases, surpass other analysis techniques [3,33].

A total of 18 explanatory variables were used that, in previous studies, have demonstrated significant effects on forest species (Table 1). An exploratory data analysis using Pearson´s linear correlation coefficients detected a high correlation between the covariates, mainly among temperature and precipitation variables. For this reason, the modeling was conducted using two approaches: (i) taking into account collinearity, to reduce as much as possible the spatial autocorrelation [34]; and (ii) in a scenario with an absence of collinearity.

In the first case, the variables that were independent of each other, but at the same time were significantly correlated with the abundance of the species, were filtered. In this way, the chosen variables included: one precipitation variable, an index whose estimate includes temperature and precipitation and two physiographic variables, they were: mean precipitation in the growing season (GSP), annual aridity index (AI), average slope (AST) and dominant aspect (ASP).

Table 1. Descriptive statistics of the variables used in this study.

		AR	AST	ASP	ELEV	AI	GSP	MTCM	MTWM	FFP	SMRSPRPB	SPRP
Quercus macdougallii	Maximum	48	90	9	2968	0.047	2144	13.5	18.5	364	5.9	168
	Minimum	1	10	2	2001	0.014	1024	7.8	12.1	208	5.2	84
	SD	14.4	28.1	2.4	530.3	0.009	432	2.3	3	65.5	1	33.6
	Mean	17.6	32.3	6.9	2649.2	0.024	1615.5	9.6	13.8	264.7	5.6	129.8
		AR	AST	ASP	ELEV	AI	MAP	GSP	MMIN	DD5	D100	SMRP
Pinus patula	Maximum	185	95	9	3002	0.047	3063	2220	7.4	3906	32	1019
	Minimum	1	10	1	2001	0.013	1318	1022	2.9	1699	12	447
	SD	29.7	23.5	2.5	224.4	0.007	441.1	304.2	0.9	403.3	4.5	145.6
	Mean	21.8	38.4	5.9	2635.9	0.024	2140.1	1581.8	4.5	2322.8	22.1	709.7
		AR	AST	ASP	ELEV	AI	MAT	GSP	MTWM	MMAX	D100	WINP
Pinus pseudostrobus	Maximum	174	95	360	3002	0.048	15.9	2220	18.7	25.3	32	441
	Minimum	1	10	1	1983	0.013	9.5	1016	12	17.4	12	157
	SD	26.7	24	116	226.2	0.007	1.113	302.7	1.0977	1.329	4.535329	69.54
	Mean	16.9	44	207	2613	0.024	11.39	1564	13.666	19.55	21.79208	291.6

AR: Abundance rate of species (units); AST: average slope of the terrain (%) (obtained with a clinometer Suunto®); ASP: dominant aspect of the site (zenith = 1, northeast = 2, east = 3, southeast = 4, south = 5, southwest = 6, west = 7, northwest = 8 and north = 9); ELEV: elevation above sea level (m) (registered with a Garmin satellite navigation receiver; AI: annual aridity index ($DD5^{0.05}$/MAP), see Sáenz-Romero et al. [29]); GSP: mean precipitation in the growing season corresponding to the mean cumulative precipitation from April to September (mm); MTCM: mean temperature in the coldest month (°C); MTWM: mean temperature in the warmest month (°C); FFP: average length of frost-free period (days); SMRSPRPB: summer/spring precipitation balance: (July + August)/(April + May) (mm); SPRP: spring precipitation (April + May, mm); MAP: mean annual precipitation (mm); MMIN: mean minimum temperature in the coldest month (January) (°C); DD5: degree-days above 5 °C (based on mean monthly temperature) (degree-days); D100: Julian date, the sum reaching 100 of degree-days above 5 °C (degree-days); SMRP: summer precipitation (July + August) (mm); MAT: Mean annual temperature (°C); MMAX: mean maximum temperature in the warmest month (June) (°C); WINP: winter precipitation (November + December + January + February) (mm); SD: Standard deviation.

In the other approach of the analysis, we intended to adjust non-parametric models (GAMs) assuming a scenario with an absence of collinearity, initially including the 18 variables. They were then eliminated one by one, leaving in the models only the variables that showed a significant contribution ($p < 0.01$) to explain the global deviance of each model (DE) and, therfore, that allowed reduced mean squared error of prediction (SME). For this purpose, simple models (including just one variable) and multiple models (including several variables at a time) were tested. In the multiple type models, ASP and AST were always included because these variables can change drastically in short distances within the study area.

Measures of the prediction error (SME) in both GLM and GAM were obtained by cross validation using for this purpose the *cv.glm* function in the "*boot*" R package for GLM [35] and "*CVgam*" function in the "*gamclass*" R package for GAM models [36].

In addition to SME, other indicators such as the values of global deviance statistic (DE), which express the percentage of variance explained by the model, the Akaike Information Criterion (AIC) and the analysis of residual values by quantile-quantile plot (Q-Q-plot), were also used to support the filtering of variables and choose the degrees of freedom in the GAM models.

Mathematical expressions, as well as the theoretical framework about GLM, can be consulted in Nelder [9], McCullagh and Nelder [12], McCulloch [37] among others. A general linear model (GLM) has a random part (linear predictor), a systematic component and a link function that specifies the links between the variable of interest and the systematic part of the model [9].

Any coefficient of an explanatory variable was considered statistically significant at 5%, if its p-value was less than 0.0028 (after the Bonferroni correction) [38].

A GAM is constructed by the sum of smoothed functions of the predictor variables, which can identify the types of effects and nonlinear relationships between variables. For this purpose, it is common to use polynomials defined based on intervals known as splines [8,10].

Here, we reproduce the general expression of a GAM according to Wood [10]:

$$\mathrm{g}(\mu_\mathrm{i}) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \quad (1)$$

where $\mu_\mathrm{i} \equiv \mathrm{E}(Y_i)$ and $Y_i \sim$ some exponential family distribution.

Y_i is a variable of interest, X_i^* is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_i are smooth functions of the covariates.

GLM and GAM allow the inclusion of a specific theoretical distribution in their initial structure. In this study, the Poisson distribution was integrated in the functions since they are count data and only have positive integer values [39,40].

Due to the flexibility of GAMs, there is a risk of over-fitting of these models [41]. In order to integrate the optimal smoothing parameters for *Pinus patula* and *Pinus pseudostrobus*, Generalized Cross Validation (GCV) described by Craven and Wahba [42] was used employing the "*mgcv*" package in *R* [43], while for *Quercus macdougallii*, degrees of freedom were manually estimated by trial and error, which was feasible since there were only 33 observations. In the case of parametric models, a reference threshold of $\alpha = 0.01$ (after the Bonferroni correction) was used to test the evidence against the null hypothesis of no significant contribution of each covariate.

3. Results

The coefficients associated with each environmental variable whose contribution was significant for the parametric models and the degrees of freedom for non-parametric models are shown in Table 2. The reader may appreciate the magnitude of the evidence against the null hypothesis (no significant contribution of a variable) by observing the p-values associated with each covariate. P-values with an asterisk were significant ($p < 0.01$, after a Bonferroni correction), using an initial $p = 0.05$.

Table 2. Coefficients and adjustment indicators for GLMs and GAMs.

Predictors	*Quercus macdougallii*				*Pinus patula*				*Pinus pseudostrobus*			
	GLM		**GAM**		**GLM**		**GAM**		**GLM**		**GAM**	
	Parms	***p*-Values**	**EDF**	***p*-Values**	**Parms**	***p*-Values**	**EDF**	***p*-Values**	**Parms**	***p*-Values**	**EDF**	***p*-Values**
Intercept	7.724	<0.001 *	2.6	<0.001 *	10.21	<0.001 *	2.8	<0.001 *	15.89	<0.001 *	2.6	<0.001 *
AST	−0.002	0.430	2.8	<0.001 *	−0.002	<0.001 *	4.0	<0.001 *	0.004	<0.001 *	7.6	<0.001 *
ASP	0.032	0.125	2.4	<0.001 *	0.068	<0.001 *	2.0	<0.001 *	0.001	0.489	4.9	<0.001 *
GSP	−0.002	0.001 *	1.9	<0.001 *	−0.002	<0.001 *	4.0	<0.001 *	−0.004	<0.001 *	4.9	<0.001 *
AI	−84.66	<0.001 *	1.0	<0.001 *	−168	<0.001 *	5.0	<0.001 *	−265.3	<0.001 *	4.9	<0.001 *
SME		250.29		322.59		810.15		774.45		686.91		706.6
DE		8.8		46.9		14.08		25.6		11.45		20.3
AIC		469.9		346.5		9825.7		8705.3		7540.3		7725.2

AST: Average slope of the terrain; ASP: dominant aspect or geographic orientation of the ground; GSP: mean precipitation in the growing season (April–September); AI: annual aridity index; SME: Square mean error estimated by cross validation; DE: Global deviance, expressing the percentage of variance explained by the model; AIC: Akaike Information Criterion values; Parms: parameter values in GLM models; EDF: Estimated freedom degrees in the GAM models; * indicates significant *p*-values (less than 0.01), after the Bonferroni correction at an initial significance level of 0.05.

In spite of the fact that most p-values were less than 0.01, adjustment indicators suggested a poor linear relationship (GLM) between the abundance of trees and the predictors compared. Including four variables, the global variances (DE) explained by the linear models were just 8.8% for *Quercus macdougallii*, 14.08% for *Pinus patula* and 11.45% for *P. pseudostrobus*, which are clearly lower than the percentages explained by the GAMs, although, the cross-validation values of the prediction error (SME) and the AIC values resulted similarly in both model types, except for *Q. macdougallii* (Table 2). Comparing the adjustment among species, *Q. macdougallii* showed the lowest values of AIC in both model types, meaning that the models' adjustment for this species was better than for the other two. The square mean error value was also remarkably lower in both *Q. macdougallii* models (Table 2).

AST and ASP in the GAM were significant for *Q. macdougallii*, but were not significant in the corresponding GLM, suggesting that there could be a relationship other than linear between these covariates and the abundance of this species.

GLMs suggested that precipitation from April to September and the annual aridity index, linearly explain the variability of species abundance since their corrected p-values were significant ($p < 0.01$) in the three species. However, due to a poor fit of all the models (see the SME, DE and AIC values), it is not possible to make a robust prediction of abundance by a parametric model using only these predictors.

The simple GAMs detected the variables that showed a greater effect on the abundance of each species by observing the value of individual deviance (DE) of each variable (Table 3). The three variables that separately explain the higher percentage of variability of the abundance of *Quercus macdougallii* were MTCM, MTWM and SPRP. For *Pinus patula* FFP, SMRSPRPB and ELEV showed high values of deviance, while MMAX, GSP, WINP and MTWM showed the largest percentage of variability of *Pinus pseudostrobus*.

The results for multiple models revealed another perspective; in this case, all variables included for *Pinus patula* and *Pinus pseudostrobus* were significant while AST, ASP, ELEV and FFP were significant in *Quercus macdougallii* GAM model (Table 3).

The resulting polynomial functions for a GAM include a constant term in its first component (Intercept), for example, for *Pinus patula*, (in a simple form) would be as follows:

$$Y = 2.8 + f_1(x_{1=AST}) + f_2(x_{2=ASP}) + f_3(x_{3=ELEV}) + f_3(x_{4=MAP}) + \ldots \quad (2)$$

where Y is the abundance of the species. Figure 2a,b show the results of the non-parametric GAM using the Poisson distribution. These results reveal two behaviors: on one hand, the type of relationship between the variables studied and on the other, the behavior of the abundance of each species at different gradients of the covariates.

Table 3. Estimated degrees of freedom and adjustment indicators of the GAM models.

Quercus macdougallii					*Pinus patula*				*Pinus pseudostrobus*			
COV	**Model Type**	**DE**	**EDF**	***p*-Values**	**COV**	**DE**	**EDF**	***p*-Values**	**COV**	**DE**	**EDF**	***p*-Values**
AST	M		3.3	0.0019 *	AST		3.8	<0.001 *	AST		7.7	<0.001 *
	S	14.2	3.7			1.3	4.6			3.5	4.8	
ASP	M		2.3	0.0003 *	ASP		1.9	<0.001 *	ASP		5	<0.001 *
	S	3.8	1.9			6.3	4.9			4.4	4.9	
ELEV	M		3.8	0.0015 *	ELEV		4.5	<0.001 *	ELEV		4.8	<0.001 *
	S	15.5	3.6			18.4	4.9			10.7	4.8	
GSP	M		1	0.4694	MAP		3.0	<0.001 *	WINP		4.9	<0.001 *
	S	23.6	4.9			17.5	4.9			11.0	3.9	
SPRP	M		1.9	0.0291	GSP		4.0	<0.001 *	D100		4.1	<0.001 *
	S	23.8	4.9			17.8	5.0			9.7	4.6	
SMRSPRPB	M		1.2	0.5672	SMRP		1.9	<0.001 *	GSP		5	<0.001 *
	S	8.3	4.8			18.1	4.9			11.1	4.9	
MTCM	M		3.9	0.0221	DD5		1.6	<0.001 *	MMAX		4.9	<0.001 *
	S	24.7	4.9			17.6	4.9			11.2	4.8	
MTWM	M		2	0.0400	D100		1.9	<0.001 *	MAT		4.1	<0.001 *
	S	24.1	4.9			18.0				10.8	4.7	
FFP	M		1.9	<0.001 *	MMIN		4.8	<0.001 *	MTWM		4.9	<0.001 *
	S	18.9	4.3			18.6	4.9			11.0	4.7	
INTERCEPT (M)		2.5				2.8				2.5		
AIC (M)		266				7940.5				6821		
SME (M)		3462				7596				7008		
DE (M)		76.8				33.6				31.9		
UBRE (M)		2.84				19.70				17.69		

AST: average slope of the terrain; ASP: dominant aspect or geographic orientation of the ground; ELEV: elevation above sea level; GSP: mean precipitation in the growing season; SPRP: spring precipitation (April + May); SMRSPRPB: summer/spring precipitation balance: (July + August)/(April + May); MTCM: mean temperature in the coldest month; MTWM: mean temperature in the warmest month; FFP: average length of frost-free period; MAP: mean annual precipitation; SMRP: summer precipitation (July + August); DD5: degree-days > 5 °C (based on mean monthly temperature); D100: julian date the sum to 100 of degree-days above 5 °C; MMIN: mean minimum temperature in the coldest month (January); WINP: winter precipitation (November + December + January + February); MMAX: mean maximum temperature in the warmest month (June); MAT: Mean annual temperature; EDF: Estimated degrees of freedom associated with each independent variable estimated by the multiple model; AIC: Akaike information criterion; SME: Square mean error estimated by cross validation; Model type: M = Multiple model, S = Simple model; DE: proportion of the null deviance explained by the model; UBRE: Un-Biased Risk Estimator criterion; * *p*-values < 0.0028, after the Bonferroni correction at an initial significance level of 0.05.

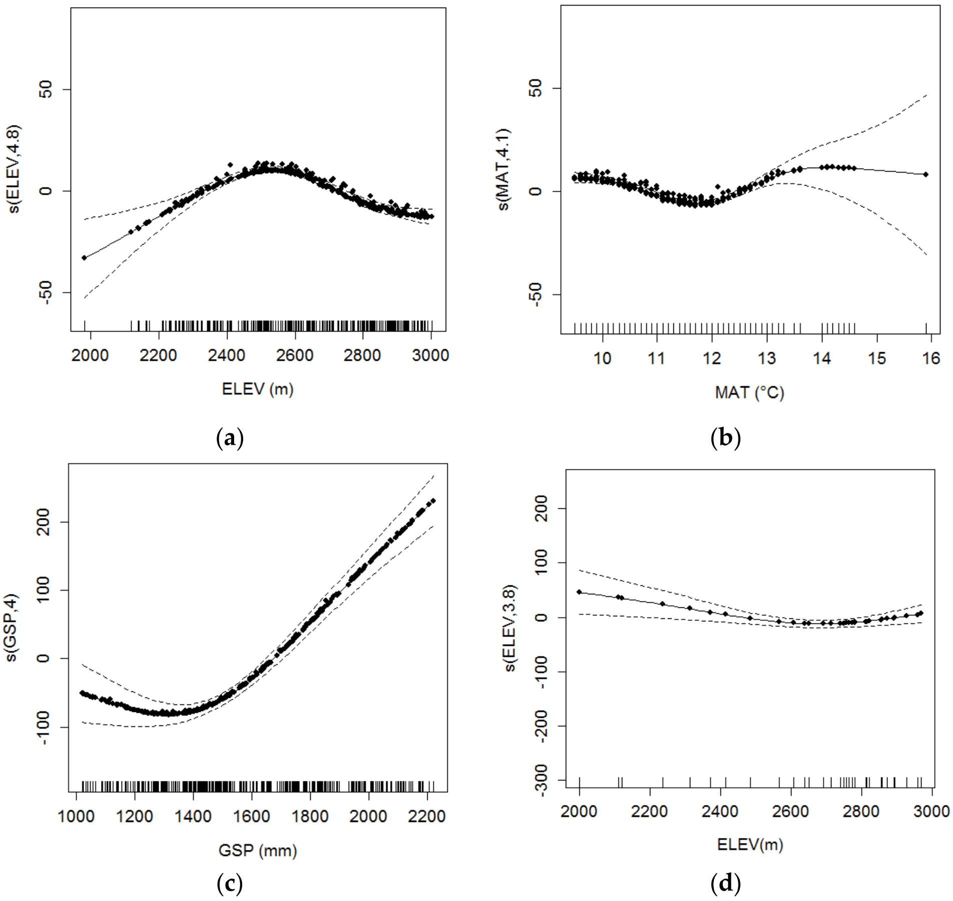

Figure 2. Response of abundance rate of: (**a**) *Pinus pseudostrobus* to the smooth function *f(ELEV)*; (**b**) *Pinus pseudostrobus* to the *f*(*MAT*); (**c**) *Pinus patula* to the *f*(*GSP*) *and* (**d**) *Quercus macdougallii* to the *f*(*ELEV*). Dashed lines indicate the point-wise standard errors for each smoothing term with a 95% confidence interval (CI). The values on the vertical axis are estimated degrees of freedom.

Mostly nonlinear relationships between abundance of species and the covariates were identified. In some cases they were clearer, such as in *Pinus pseudostrobus* against the elevation above sea level (Figure 2a) or *Pinus patula* in relation to the precipitation from April to September (Figure 2c). In addition, signs of linear effects were observed for some variables, such as the slope and the exposure, which influenced the abundance of both pine species, and still more evident was the effect of ASP, AST, ELEV or FFP on the abundance of *Pinus patula*.

Figure 2d showed a linear trend of *Quercus macdougallii* against ELEV values. No drastic changes are perceived as ELEV values increase; however, the strip of uncertainty is more noticeable when the elevation above sea level takes values less than 2500 m.

Using the multiple GAM, the *Quercus macdougallii* model showed a high value of the global deviance (76.8%), followed by the *Pinus patula* model with 33.6% and the *P. pseudostrubus* model with 31.9%. The good fit of *Quercus macdougallii* was reflected in the graphs of the residuals, which were mostly distributed within the range of confidence of Q-Q plot (gray shaded area) at 99% of CI (Figure 3b), and there was moderate dispersion between the adjusted values and response values (Figure 3a).

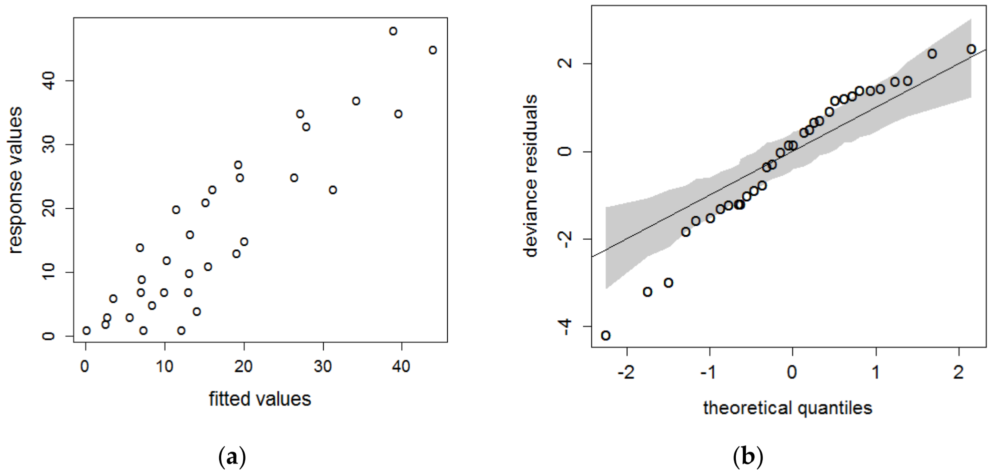

Figure 3. Fitted against response values of *Quercus macdougallii* smooth function (**a**), and distribution of its residuals in a Q-Q plot, showing a 99% confidence interval (**b**).

4. Discussion

GAMs, both in their simple and multiple structure, successfully described the abundance of a species in response to unitary changes of a climatic variable. Graphical results of GAMs not only revealed the trend or turning points of regression, but also showed the possible limits within which the optimum abundance of each species could occur (Figure 2a–d). For example, Figure 2b shown that between 9 and 13 °C, there is a greater distribution of data, the trend of the curve is relatively stable and there is a marked reduction in the uncertainty region (95%). This means that the range of maximum abundance of *Quercus macdougallii* in relation to mean annual temperature would be between these values.

In this study, we did not find data to include the effect of human activities; neither did we find any study addressing such effect on the distribution and abundance of species in the study area. However, since 1983, the communities of the Sierra Juárez have adopted projects aimed to recover the forest quality, focusing also on the regeneration and conservation of the native pine species, such as: *Pinus patula*, *P. pseudostrobus* and *P. ayacahuite*, which are the most abundant in the study area. Furthermore, the forest management scheme of the timber harvesting area in Santiago Comaltepec has the certification of good practices, awarded by the Forest Stewardship Council (FSC). Nonetheless, the effect of human activities ought to be addressed in future studies, because management certification by itself does not guarantee the conservation of biodiversity [44], and anthropogenic activity has had increased influence on natural ecosystems in recent years [45,46].

Knowing the variables that have a significant effect on the abundance of a species is important because these variables could condition higher or lower presence of that species in a locality. This information could be vital for decision making. The models tested in this research suggest that the aridity index, extreme temperature (the maximum of the warmest month) and the rainfall during specific periods (winter precipitation, summer precipitation, and the precipitation from April to September), significantly affect the abundance of species (Tables 2 and 3). The aridity index has been reported as a variable that is correlated to the altitudinal gradient, which is useful for a possible assisted migration of species to different climate change scenarios [47], whereas extreme temperatures and specific rainfall have also been observed as key variables for determining the density of pine and oak in the northwest of Mexico [25].

Environmental preferences and geographical distribution of a given species are inherent aspects. Therefore, a review is needed of some limitations of mathematical models such as the ones used in this study: Firstly, a simple regression model permits only the testing of the hypothesis that there is a high probability that the tested covariate, and no other influences (or due to chance) explain the predicted variable [48]. That is, a good prediction model only increases the probability of finding a species of interest in a given locality where that species really exists, but there will always be a degree of uncertainty as well.

Hence, possible errors in modeling must be taken into account, such as: (i) Predicting the abundance of a species based on a list of variables; the model implicitly assumes that the predicted range or potential space is fully occupied by the modeled species, which does not always happen [49]; furthermore, the spatial distribution of organisms has adopted a dynamic behavior over time [49], so that a potential site may or may not be sparsely vegetated by certain species for a certain period (e.g., during sampling) due to progressive succession of plants; or a temporary absence could be found due to natural causes, such as season of the year, attack of pests or diseases or inter-species competition; (ii) The current presence of a species reflects the contexts of the past [49] which in turn could give rise to uncertainty (though on a smaller scale) in predicting habitats; (iii) The global environmental conditions follow changing trends of different duration [50,51], so it is possible that in a certain case, an observed species may be declining or at risk of extinction in a locality or it could be suffering displacement, dispersion or fragmentation of its habitat [52,53], but the prediction model does not detect this dynamic behavior.

Despite the limitations of predictive models [1,54], several studies have demonstrated the usefulness of correlative models [55,56]. Our results reported here could contribute to the identification of environmental preferences of the species studied, as in the case of *Quercus macdougallii*, which is included on the red list of the IUCN. In general, the proven models served the purpose of the study, but we do not rule out exploring other tools in the future, which will complement these results, such as spatial autocorrelation regression models [57,58] or the Chi-squared Automatic Interaction Detector (CHAID) together with Regression Trees [6], as well as probability functions which would be useful to find the climatic values in which the maximum likelihood of abundance occurs and also, generate maps of maximum abundance.

5. Conclusions

In this study, the GAMs better served our purpose than GLMs, considering that their results not only revealed the trend of the response variables or the inflection of the regression, but also served to detect when species abundance increased or decreased. GAMs also allow partial characterization of the realized niche of a given species, according to some specific variables, regardless of the type of relationship (Figure 2a–d), which represents a substantial advantage for these types of studies. Furthermore, GAMs are useful in studies with multiple variables in instances when the observations do not follow a specific (normal) distribution. Particularly, smoothed models, both in their simple and multiple forms, are useful tools that help answer questions like those raised in this report: Which variables have a greater and more significant effect on the abundance of each studied species? Or, is there a linear or a curvilinear relationship between the variable of interest and its covariates? The results obtained with GAMs could also be useful to find potential areas for endemic or threatened *taxa*, because through adequate models it is possible to obtain specific values for each variable that might affect where each species could grow with greater chances of success.

Acknowledgments: This research was funded by the Professional Development Program for Teachers (PRODEP) and Universidad de la Sierra Juarez (UNSIJ) in Oaxaca, Mexico. We are very grateful to the local forestry services unit (UZACHI) who provided valuable information. We would also like to thank the reviewers, known and unknown, for their useful suggestions to improve this report.

Author Contributions: P.A. Conceived and coordinated the research project, statistical analysis and the primary writing of text. J.C.H.-D. and C.W. contributed to the interpretation of results and helped in writing the text. R.C.-T.: advisor and text editor.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guisan, A.; Edwards, T.C.; Hastie, T. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecol. Model.* **2002**, *157*, 89–100. [CrossRef]
2. Carrascal, L.M.; Bautista, L.M.; Lázaro, E. Geographical variation in the density of the white stork Ciconia ciconia in Spain: Influence of habitat structure and climate. *Biol. Conserv.* **1993**, *65*, 83–87. [CrossRef]
3. Pliscoff, P.; Fuentes-Castillo, T. Modelación de la distribución de especies y ecosistemas en el tiempo y en el espacio: Una revisión de las nuevas herramientas y enfoques disponibles. *Rev. Geogr. Norte gd.* **2011**, *48*, 61–79. [CrossRef]
4. Cade, B.S.; Noon, B.R. A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.* **2003**, *1*, 412–420. [CrossRef]
5. Weiher, E. Species richness along multiple gradients: Testing a general multivariate model in oak savannas. *Oikos* **2003**, *101*, 311–316. [CrossRef]
6. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *J. R. Stat. Soc. Ser. C* **1980**, *29*, 119–127. [CrossRef]
7. Olden, J.D.; Lawler, J.J.; Poff, N.L. Machine learning methods without tears: A primer for ecologists. *Q. Rev. Biol.* **2008**, *83*, 171–193. [CrossRef] [PubMed]
8. Hastie, T.; Tibshirani, R. Generalized additive models. *Stat. Sci.* **1986**, *1986*, 297–310. [CrossRef]
9. Nelder, J.; Wedderburn, R. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **1972**, *135*, 370–384. [CrossRef]
10. Wood, S. *Generalized Additive Models: An Introduction with R*; Chapman Hall/CRC: Boca Raton, FL, USA, 2006; pp. 1–33.
11. Wang, L.; Liu, X.; Liang, H.; Carroll, R.J. Estimation and variable selection for generalized additive partial linear models. *Ann. Stat.* **2011**, *39*. [CrossRef]
12. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: London, UK, 1989; p. 511.
13. Nicholls, A.O. How to make biological surveys go further with generalised linear models. *Biol Conserv.* **1989**, *50*, 51–75. [CrossRef]
14. Jaberg, C.; Guisan, A. Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. *J. Appl. Ecol.* **2001**, *38*, 1169–1181. [CrossRef]
15. Elith, J.; Graham, J.C.H.; Anderson, P.; Zimmermann, N.E. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **2006**, *29*, 129–151. [CrossRef]
16. Maravelias, C.D. Trends in abundance and geographic distribution of North Sea herring in relation to environmental factors. *Mar. Ecol. Prog. Ser.* **1997**, *159*, 151–164. [CrossRef]
17. Murase, H.; Nagashima, H.; Yonezaki, S.; Matsukura, R.; Kitakado, T. Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: A case study in Sendai Bay, Japan ICES. *Afr. J. Mar. Sci.* **2009**, *66*, 1417–1424. [CrossRef]
18. CNA, Comisión Nacional del Agua. Servicio Meteorológico Nacional. Available online: http://smn.cna.gob.mx (accessed on 7 January 2016).
19. INEGI, Instituto Nacional de Estadística y Geografía. Available online: http://www.beta.inegi.org.mx/app/mapa/espacioydatos/default.aspx (accessed on 8 January 2016).
20. Schweik, C.M. Social norms and human foraging: An Investigation into the spatial distribution of Shorea Robusta in Nepal. Forest, Trees and People Programme. Available online: http://www.treesforlife.info/fao/Docs/P/X2104E/X2104E06.htm (accessed on 17 October 2016).
21. CONAFOR, Comisión Nacional Forestal, Sistema de Planeación Forestal Para Bosque Templado (SIPLAFOR). Available online: http://fcfposgradoujedmx/spf/inicio/documentosphp (accessed on 7 September 2016).
22. Muñoz-Flores, H.J.; Sáenz-Reyes, J.; García-Sánchez, J.J.; Hernández-Máximo, E.; Anguiano Contreras, J. Áreas potenciales para establecer plantaciones forestales comerciales de *Pinus pseudostrobus* Lindl. y *Pinus greggii Engelm*. en Michoacán. *Rev. Mex. Cienc. For.* **2011**, *2*, 29–44.
23. Nadezda, M.T.; Gerald, E.R.; Elena, I.P. Impacts of climate change on the distribution of *Larix* spp. and *Pinus sylvestris* and their climatypes in Siberia. *Mitig. Adapt. Strateg. Glob.* **2006**, *11*, 861–882. [CrossRef]
24. Sáenz-Romero, C.; Rehfeldt, G.E.; Crookston, N.L.; Duval, P.; St-Amant, R.; Beaulieu, J.; Richardson, B.A. Spline models of contemporary, 2030, 2060 and 2090 climates for Mexico and their use in understanding climate-change impacts on the vegetation. *Clim. Chang.* **2010**, *102*, 595–623. [CrossRef]

25. Martínez-Antúnez, P.; Hernández-Díaz, J.C.; Wehenkel, C.; López-Sánchez, C.A. Estimación de la densidad de especies de coníferas a partir de variables ambientales. *Madera Bosques* **2015**, *21*, 23–33. [CrossRef]
26. Martínez-Antúnez, P.; Wehenkel, C.; Hernández-Díaz, J.C.; Corral-Rivas, J.J. Use of the Weibull function to model maximum probability of abundance of tree species in northwest Mexico. *Ann. For. Sci.* **2015**, *72*, 243–251. [CrossRef]
27. Rehfeldt, G.E.; Crookston, N.L.; Warwell, M.V.; Evans, J.S. Empirical analyses of plants climate relationships for the western United States. *Int. J. Plant Sci.* **2006**, *167*, 1123–1150. [CrossRef]
28. Crookston, N.L.; Rehfeldt, G.E.; Ferguson, D.E.; Warwell, M. FVS and Global Warming: A Prospectus for Future Development. Available online: http://www.treesearch.fs.fed.us/pubs/30963 (accessed on 1 June 2016).
29. Sáenz-Romero, C.; Rehfeldt, G.E.; Ortega-Rodríguez, J.M.; Marín-Togo, M.C.; Madrigal-Sánchez, X. *Pinus leiophylla* suitable habitat for 1961–1990 and future climate. *Bot. Sci.* **2015**, *93*, 709–718. [CrossRef]
30. Chapela, F. El Manejo Forestal Comunitario Indígena en la Sierra de Juárez, Oaxaca Los Bosques Comunitarios de México Manejo Sustentable de Paisajes Forestales. Available online: http://www2ineccgobmx/publicaciones/libros/532/cap5pdf (accessed on 17 September 2016).
31. Guisan, A.; Graham, C.H.; Elith, J.; Huettmann, F. Sensitivity of predictive species distribution models to change in grain size. *Divers. Distrib.* **2007**, *13*, 332–340. [CrossRef]
32. Austin, M. Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecol. Mod.* **2007**, *200*, 1–19. [CrossRef]
33. Meynard, C.N.; Quinn, J.F. Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *J. Biogeogr.* **2007**, *34*, 1455–1469. [CrossRef]
34. Dormann, F.C.; McPherson, J.M.; Araújo, M.B.; Kühn, I. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* **2007**, *30*, 609–628. [CrossRef]
35. Canty, A.; Ripley, B. Package 'Boot'. Available online: https://cranr-projectorg/web/packages/boot/bootpdf (accessed on 17 March 2016).
36. Maindonald, J. Package 'Gamclass'. Available online: https://cranr-projectorg/web/packages/gamclass/gamclasspdf (accessed on 7 August 2016).
37. McCulloch, C.E. Generalized linear models. *J. Am. Stat. Assoc.* **2000**, *95*, 1320–1324. [CrossRef]
38. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802. [CrossRef]
39. Bio, A.M.F.; Alkemade, R.; Barendregt, A. Determining alternative models for vegetation response analysis: A non-parametric approach. *J. Veg. Sci.* **1998**, *9*, 5–16. [CrossRef]
40. Lehmann, A. GIS modeling of submerged macrophyte distribution using Generalized Additive Models. *Plant Ecol.* **1998**, *139*, 113–124. [CrossRef]
41. Austin, M.P. Spatial prediction of species distribution: An interface between ecological theory and statistical modeling. *Ecol. Model.* **2002**, *157*, 101–118. [CrossRef]
42. Craven, P.; Wahba, G. Smoothing Noisy Data with Spline Functions Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation Numerische. *Mathematik* **1978**, *31*, 377–404. [CrossRef]
43. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B Met.* **2011**, *73*, 3–36. [CrossRef]
44. Rodríguez-Rivera, V.; Alfonso-Corrado, C.; Aguirre-Hidalgo, A.; Campos, J.E.; Venegas-Barrera, C.S.; Clark-Tapia1, R. Galls and host occurrences along a forest gradient in Sierra Juárez, Oaxaca, Mexico. *J. Environ. Biol.* **2017**, *38*, 1–7. [CrossRef]
45. Hunter, P. The human impact on biological diversity. *EMBO Rep.* **2007**, *8*, 316–318. [CrossRef] [PubMed]
46. Crowther, T.W.; Glick, H.B.; Covey, K.R.; Bettigole, C.; Maynard, D.S.; Thomas, S.M.; Smith, J.R.; Hintler, G.; Duguid, M.C.; AmatullI, G.; et al. Mapping tree density at a global scale. *Nature* **2015**, *525*, 201–205. [CrossRef] [PubMed]
47. Sáenz-Romero, C.; Martínez-Palacios, A.; Gómez-Sierra, J.M.; Pérez-Nasser, N.; Sánchez-Vargas, N.M. Estimación de la disociación de Agave cupreata a su hábitat idóneo debido al cambio climático. *Rev. Chapingo Ser. Cien.* **2012**, *18*, 291–301.
48. Oreskes, N.; Shrader-Frechette, K.; Belitz, K. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **1994**, *263*, 641–646. [CrossRef] [PubMed]

49. Seoane, J.; Bustamante, J. Modelos predictivos de la distribución de especies: Una revisión de sus limitaciones. *Ecología* **2001**, *15*, 9–21.
50. Rehfeldt, G.E.; Crookston, N.L.; Sáenz-Romero, C.; Campbell, E.M. North American vegetation model for land-use planning in a changing climate: A solution to large classification problems. *Ecol. Appl.* **2012**, *22*, 119–141. [CrossRef] [PubMed]
51. Zhu, Q.; Jiang, H.; Liu, J.; Peng, C.; Fang, X.; Yu, S.; Zhou, G.; Wei, X.; Ju, W. Forecasting carbon budget under climate change and CO2 fertilization for subtropical region in China using Integrated Biosphere Simulator (IBIS) model. *Pol. J. Ecol.* **2011**, *59*, 3–23.
52. Desai, A.R.; Normets, A.; Bolstad, P.V.; Chen, J.; Cook, B.D.; Davis, K.J.; Euskirchen, E.S.; Gough, C.; Martin, J.G.; Ricciuto, D.M.; et al. Influence of vegetation and seasonal forcing on carbon dioxide fluxes across the Upper Midwest, USA: Implications for regional scaling. *Agric. For. Meteorol.* **2008**, *148*, 288–308. [CrossRef]
53. Goparaju, L.; Jha, C.S. Spatial dynamics of species diversity in fragmented plant communities of a Vindhyan dry tropical forest in India. *Trop. Ecol.* **2010**, *51*, 55–65.
54. Austin, M.P.; Belbin, L.; Meyers, J.A.; Doherty, M.D.; Luoto, M. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecol. Mod.* **2006**, *199*, 197–216. [CrossRef]
55. Raxworthy, C.J.; Martinez-Meyer, E.; Horning, N.; Nussbaum, R.A.; Schneider, G.E.; Ortega-Huerta, M.A.; Peterson, A.T. Predicting distributions of known and unknown reptile species in Madagascar. *Nature* **2003**, *426*, 837–841. [CrossRef] [PubMed]
56. Martínez-Meyer, E.; Peterson, A.T. Conservatism of ecological niche characteristics in North American plant species over the Pleistocene to Recent transition. *J. Biogeogr.* **2006**, *33*, 1779–1789. [CrossRef]
57. Ward, M.D.; Gleditsch, K.S. Spatial Regression Models. 2008. Available online: http://us.corwin.com/sites/default/files/upm-binaries/21130_Chapter_11.pdf (accessed on 19 January 2017).
58. Le Gallo, J. Cross-section spatial regression models. In *Handbook of Regional Science*; Fischer, M.M., Nijkamp, P., Eds.; Springer: Berlin, Germany, 2014; pp. 1511–1533.