Application of GIS and logistic regression to fossil pollen data in modelling present and past spatial distribution of the Colombian savanna

Flantua, S.G.A.; van Boxel, J.H.; Hooghiemstra, H.; van Smaalen, J.

DOI
10.1007/s00382-007-0276-3

Publication date
2007

Document Version
Final published version

Published in
Climate Dynamics

Clim Dyn (2007) 29:697–712
DOI 10.1007/s00382-007-0276-3

Application of GIS and logistic regression to fossil pollen data in modelling present and past spatial distribution of the Colombian savanna

Suzette G. A. Flantua · John H. van Boxel · Henry Hooghiemstra · John van Smaalen

Received: 20 December 2006 / Accepted: 28 April 2007 / Published online: 15 June 2007

Abstract Climate changes affect the abundance, geographic extent, and floral composition of vegetation, which are reflected in the pollen rain. Sediment cores taken from lakes and peat bogs can be analysed for their pollen content. The fossil pollen records provide information on the temporal changes in climate and palaeo-environments. Although the complexity of the variables influencing vegetation distribution requires a multi-dimensional approach, only a few research projects have used GIS to analyse pollen data. This paper presents a new approach to palynological data analysis by combining GIS and spatial modelling. Eastern Colombia was chosen as a study area owing to the migration of the forest–savanna boundary since the last glacial maximum, and the availability of pollen records. Logistic regression has been used to identify the climatic variables that determine the distribution of savanna and forest in eastern Colombia. These variables were used to create a predictive land-cover model, which was subsequently implemented into a GIS to perform spatial analysis on the results. The palynological data from the study area were incorporated into the GIS. Reconstructed maps of past vegetation distribution by interpolation showed a new approach of regional multi-site data synthesis related to climatic parameters. The logistic regression model resulted in a map with 85.7% predictive accuracy, which is considered useful for the reconstruction of future and past land-cover distributions. The suitability of palynological GIS application depends on the number of pollen sites, the distribution of the pollen sites over the area of interest, and the degree of overlap of the age ranges of the pollen records.

Keywords Climate change · Pollen data · Geographic information system (GIS) · Savanna · Logistic regression · Predictive modelling · Land-cover distribution · Interpolation maps

S. G. A. Flantua · J. H. van Boxel (✉) · H. Hooghiemstra · J. van Smaalen
Faculty of Science, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands
e-mail: J.H.Boxel@science.uva.nl

1 Introduction

Climate change at glacial-interglacial cycle time scales has had an impact on the vegetation in many parts of the world. Vegetation change is reflected by changes in the abundance, geographic extent, location of source areas, and floral composition of plant populations. The pollen grains from these changing plant populations are preserved in lakes and peat bogs. Sediment cores can be obtained which show temporal changes in the fossil pollen assemblages. Palynologists present these data in pollen diagrams and interpret the downcore changes in pollen spectra, and the variation in pollen representation of individual pollen taxa in terms of past vegetation change and inferred environmental conditions.

One area of palynological research has been the tropical lowlands of northern South-America. Here, savanna ecosystems occur north of a vast region of tropical rainforest. These savannas, located in Colombia and Venezuela, extend from the Eastern Cordillera to the eastern coast of Venezuela. The southern boundary of the savanna vegetation, which is transitional to tropical rainforest, has migrated in the past (Behling and Hooghiemstra 1998). Shifts of this savanna–forest transition depend heavily on annual precipitation values and the length of the dry season

while temperature change has little impact. Such a clear relationship between climate parameters and environmental setting is attractive to explore for changes in the past. Although these pollen records have revealed temporal changes in vegetation dynamics, the degree of environmental change has only been expressed in general qualitative terms, such as ''drier'' or ''wetter'' conditions, or suggestions about changes in the seasonality, such as ''shorter'' or ''longer dry period'' (Behling and Hooghiemstra 1998; Berrío 2002).

So far, little research has been carried out where palynological data has been analysed by software specially designed for spatial analysis, such as geographical information systems (GIS) (Paez et al. 2001; Davis et al. 2003). Most palynological publications, which include pollen mapping, use isopollen or isochrone maps (e.g. Birks 1989; Yu et al. 2001) and mapped pollen percentages (Brubacker et al. 2005). The integration of GIS in palynological research seems to be in an explorative stage where the applications of GIS are diverse but scarce, e.g., mapping plant-distributions (Jago and Boyd 2003; Giesecke and Bennett 2004); habitat suitability analysis (e.g., Lyford et al. 2003); and the reconstruction of past vegetation (e.g., Ray and Adams 2001; Bickford and Mackey 2004; Veski et al. 2005). Due to the complexity and the spatial heterogeneity of the variables influencing the spatial distribution of vegetation, palynological analysis thus far has mainly been limited to non-spatial methods: the search for structure in multidimensional data sets using 1D tools. There are several reasons to implement palynological datasets into GIS: palynological datasets in general are large and complex to interpret; the data consists of changes which have occurred over an area (2D surface), and over time (the 3D-variable); and frequently, data from different sites must be compared by the researcher to make an interpretation of a complete area rather than of one single site only. For this reason it has become necessary to introduce GIS as a new analytic tool for palynological research.

The aim of this paper is to detail a new approach to pollen data analysis by combining GIS and predictive modelling into a new potential palynological GIS application. A methodology is explained that can be exploited by palynologists to explore their area of research and capture it in a predictive model, which can be used to make reconstructions of past and future land-cover distributions under changing climatic conditions. In the same database, the palynological data can be implemented to use GIS to evaluate patterns of land-cover changes based on pollen counts. Furthermore, this study aims to provide a better understanding of the dynamics of the savanna distribution in Colombia, and so allow further insight into future vegetation responses to global climate change.

The questions approached by this study are: (a) Can GIS and logistic regression be used to model the spatial distribution of the Colombian savanna ecosystem? and (b) Is data from palynological site studies suitable for implementation into GIS, where it is synthesised for selected time windows?

In this paper, the employment of GIS is divided into two different but related applications. The first application is to construct a predictive model, in which the climatic variables are determined that influence the spatial distribution of the savanna ecosystem. The statistical model, derived from logistic regression, is subsequently introduced into a GIS and re-run to create land-cover maps, which are compared to the actual land-cover distribution in order to assess the model accuracy. The second application implements data from pollen records of the Colombian savanna into GIS to create land-cover maps for the last 10,000 radiocarbon years before present.

2 Setting of the study area

The Colombian savannas are vast plains stretching from the Guaviare River (an eastern tributary of the Orinoco River) to the Venezuelan border (Fig. 1). They lie in the Orinoco Basin and cover approximately 500,000 km^2 (Sarmiento 1983). This area of low lying savannas forms a level plain between 200 and 600 m altitude (Blydenstein 1967). Most of this area is covered with grass, with ribbons of gallery forests along the creeks and rivers, and patches of forest scattered on the plains. The typical combination of an open tree layer and a continuous herbaceous layer is characteristic of the Colombian savanna, although the vegetation physiognomy varies across the area from treeless savanna grassland to savanna woodland with up to

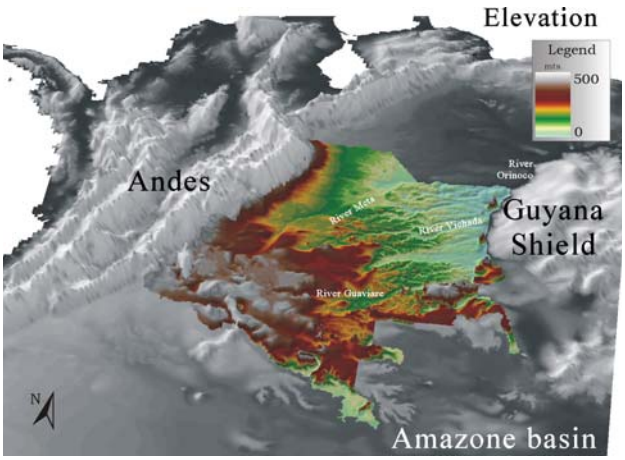

Fig. 1 Topography of the study area. Altitude in the study area is exaggerated compared to true elevation levels outside the study area (*grey colours*)

80% tree cover, and gallery forest (Sarmiento 1984). Rivers are numerous: the main ones are the Meta, Vichada, Guaviare and Inírida rivers, which are tributaries of the Orinoco River.

The climate in the savannas of the Llanos Orientales is characterized by a warm and humid climate during the rainy season from April to November (Mistry 2001). There is a gradient of higher precipitation towards the Colombian Amazon region in a southern and southwestern direction, and lower amounts in the northern part towards the Venezuelan border (Botero 1999). At the same time, the length of the dry season increases from 2 to 5 months (San José et al. 1998). The small annual temperature amplitude of <3°C contrasts with the daily variation of 10–15°C (Blydenstein 1967).

The study area was chosen so that the savanna–tropical rainforest transition zone would be centred in the defined region, providing a comparable surface area of savanna and forest. The Andes forms the western boundary of the study area, the Orinoco River the eastern limit, and the rivers Meta and Arauca outline the study area in the north (Fig. 1).

3 Modelling methods

3.1 Logistic regression basics and formula

Logistic regression is a variation of ordinary regression, which is a method used to determine the impact of independent variables on a dependent variable (Hosmer and Lemeshow 1989). In binary logistical regression, the dependent variable is an event occurrence. The observed outcome is restricted to two values, representing the presence or absence of a specific event. It produces a formula that predicts the probability of the occurrence as a function of the independent variables. The attractive aspect of logistic regression is that the impact of multiple variables can be measured at the same time, the relative importance of independents can be ranked, and the interaction effects can be evaluated.

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous dependent (response) variable and the set of independent (predictor) variables of the training data. A single formula is built, which calculates the logistic (LP) as a linear combination of the predictive variables. The inverse logistic transformation

$$p(y) = \exp(\mathrm{LP})/(1 + \exp(\mathrm{LP}))$$

is then applied to acquire response values between 0 and 1. Depending on a chosen threshold probability value, everything above this threshold indicates one condition of the binomial outcome (i.e., the presence of savanna), while everything below equals the other condition of the variable (i.e., absence of savanna; in this case, the presence of forest). This threshold is usually chosen at the default value of [0.5], but this will depend on the aims of the research project, as it can be more important to predict either the absence or the presence of any given characteristic. All these computations were performed in SPSS (Edition 11.0 for Windows). By implementing the formula and the calculated coefficients directly into GIS, probability values are obtained for every cell of the GIS grid, which results in a probability map of savanna occurrence.

3.2 Fitting the predictive model

To find the structure in the relationships between variables, Principal component analysis (PCA) was used. Using a multi-dimensional coordinate system, it groups the variables with the highest correlation into distinctive components and serves to make a first selection of model components. Variables can be entered into the logistic model in the order specified by the researcher, or logistic regression can test the fit of the model after each variable is added or deleted, called ''stepwise regression'' (Hosmer and Lemeshow 1989). Backward stepwise regression is a method, where the analysis begins with a full model and variables are eliminated from the model in an iterative process (Hosmer and Lemeshow 1989). The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be added or removed from the model, owing to reduced model accuracy, the analysis is then complete.

3.3 Measures of model evaluation

In ecological modelling literature, the different accuracy measurements are subject to debate (e.g., Fielding and Bell 1997; Foody 2002; McPherson et al. 2004; Guisan and Thuiller 2005). Therefore, various methods of accuracy assessment have been used during the process of selecting the best model.

The predictive success of binary models can be described in terms of false positive and false negative prediction errors. A false positive means that the model predicted presence when absence of savanna was determined. A false negative means that the model predicted absence when actually savanna is present. Sensitivity is the percentage of correctly predicted savanna occurrences to the total number of savanna pixels. Conversely, specificity is defined as the proportion of correctly predicted absences of savanna to the total number of absences. The overall accuracy is defined as the percentage of correctly predicted

raster-pixels to the total number of pixels in the area of interest (Fielding and Bell 1997).

Cohen's Kappa coefficient is designed to reflect the models performance in absence and presence simultaneously. This coefficient has been used extensively as an index to classify accuracy (e.g., Foody 2002), and as an inter-comparison between models (Manel et al. 2001). Although the effect of prevalence—the number of occurrences in relation to the number of samples—has been judged negligible among ecologists (e.g., Fielding and Bell 1997; Manel et al. 2001), Kappa has been criticised due to its sensitivity towards a variation in prevalence. As we do not have presumptions about the prevalence in the data set, we also include an alternative measure for the kappa statistic proposed by Allouche et al. (2006), the so-called True skill statistic (TSS). This measure is suitable for models that generate presence–absence predictions. TSS is stated to compensate for the supposedly shortcomings of the Kappa coefficient by possessing the same advantages but at the same time being independent of prevalence (Allouche et al. 2006). We used different evaluation methods to decide upon the predictive capacity of the created models.

From a confusion matrix (Table 1) that records the number of (a) true presence, (b) false presence, (c) false absence and (d) true absence, all accuracy measures are calculated (Table 2). To interpret the power of the model calculated by the Kappa and the TSS measurement, the classification by Landis and Koch (1977) is used that proposes that a [0.0–0.19] represents poor agreement, [0.2–0.39] fair agreement, [0.4–0.59] moderate agreement, [0.6–0.79] substantial agreement, and a [0.8–1.00] designates an almost perfect agreement.

3.4 Model validation

Model validation requires checking the model against independent data to see how well it predicts (Fielding and Bell 1997). The outcome of the model (logit equation) is introduced into the GIS to let the model run with the related climate layers as a data source. A map of the predicted land-cover distribution is the result of this implementation. This map is subsequently compared to the actually observed spatial distribution of savanna to evaluate how well the model performs. By subtracting the layers of the observed and predicted spatial distribution of savanna from one another in the GIS, the accuracy of the model can be illustrated, and a differentiation made between false presence and false absence predictions of savanna. According to Manel et al. (2001) ''many users of ecological models are assuming good performance because their data fits well statistically and because they can predict many occurrences correctly''. But as explained by the same author, when models are required to predict occurrence by the use of independent data, the weakness of the model becomes apparent. Therefore, implementing the logistic regression outcome into the GIS does not only test an independent data set for the models predictive capacity, but also provides insight into the models weaknesses.

Table 1 The performance of the model is specified in an error matrix

Predicted distribution	Observed distribution	
	Presence	Absence
Presence	a	b
Absence	c	d

a, the model correctly predicted the presence of the characteristic (savanna); *b*, the false predictions of presence; *c* means absence was falsely predicted as presence; and *d*, the correctly predicted absence cells. The total of all values ($a + b + c + d$) are summed in *N*

Table 2 Overview of measures for evaluating the predictive performance of the model calculated from the error matrix

Measure	Formula
Prevalence	$\frac{a+c}{N}$
Sensitivity	$\frac{a}{a+c}$
Specificity	$\frac{d}{b+d}$
Overall accuracy	$\frac{a+d}{N}$
Kappa Statistics	$\frac{(a+d)-[((a+c)(a+b))+((b+d)(c+d))]/N}{N-((a+c)(a+b)(b+d)(c+d))/N}$
TSS	$\frac{(axd)-(bxc)}{(a+c)(b+d)}$ = Sensitivity + Specificity − 1

The probability that the model will correctly classify presence of savanna is indicated by ''sensitivity''. ''Specificity'' is the probability that the model correctly classifies an absence. The rate of correctly classified cells is specified in the ''overall accuracy''. The Kappa coefficient and the true skill statistic (TSS) comprise the effect of chance in the calculation of the overall accuracy

3.5 Model assumptions

This section specifies the most important assumptions and simplifications in the model.

1. All species have different climatic tolerances and will respond independently to change (Eeley et al. 1999). In this study, the focus of attention is on the distribution of savanna and forest vegetation. Although present-day plant communities may temporally represent transient association, in the created model the forest and the savanna are considered as completely separate units and stable systems in the present-day situation.
2. The model includes only climatological and physical geographical parameters, which are assumed to be the most important factors influencing the distribution of the vegetation in the study area. Anthropogenic factors or fire are not taken into account, although both are important in the Colombian savanna, and have played

a considerable role in the recent past. Nevertheless, in this model the climatic variables are considered as the most dominant influence on patterns of plant distribution.

3. The model is one of an equilibrium state and is developed on the basis of present day climatic data. The process of change is not considered, for example the impact of C_3 and C_4-dominated vegetation through time, only representations of the current conditions of climate and vegetation.
4. The model is a representation of the present-day climate–vegetation relationship.
5. Although generally considered an important influence, soil data are not further considered for this study, because of the interdependency between soil characteristics and land-cover. A study in the Gran Sabana in Venezuela by Dezzeo et al. (2004) even failed to show a significant relation between edaphic conditions and the distribution of savanna and forest.

4 Datasets and maps

4.1 Land-cover and elevation layers

The land-cover dataset is derived from the global land-cover characteristics (GLCC) Data Base Version 2.0 (Loveland et al. 2000a). Advanced very high resolution radiometer (AVHRR) was used to achieve the high 1-km resolution (Loveland et al. 2000b). The classification, known as the international geosphere biosphere program (IGBP) land-cover classification (Belward 1996), embraces 17 classes of land-cover (Table 3). However, in this study a binary data set is required, which should comprise of either savanna or forest. In the GIS, a selection has been made excluding all redundant land-cover categories, and grouping the forest-, and savanna-categories. The following classes have been grouped to form the two categories: [Forest] = 2 + 5 and [Savanna] = 6 – 10. The classes 3, 11–17 are removed from the land-cover layer. In addition the land-cover data above 500 m altitude has been excluded (owing to the variable environmental conditions). The forest adjacent to water courses has also been omitted, since gallery forest occurrence can only be explained by the presence of water and is therefore not relevant to the characterisation of the savanna–forest transition zone. The elevation data (Belward 1996) have a 1-km horizontal resolution, which matches to that of the land-cover data.

Table 3 Units of land-cover according to the IGPB land-cover classification

Land-cover	
Code	Description
1	Evergreen Needleleaf Forest
2	Evergreen broadleaf forest
3	Deciduous Needleleaf forest
4	Deciduous broadleaf forest
5	Mixed forest
6	Closed Shrublands
7	Open Shrublands
8	Woody savannas
9	Savannas
10	Grasslands
11	Permanent wetlands
12	Croplands
13	Urban and Built-up
14	Cropland/natural mosaic
15	Snow and Ice
16	Barren or sparsely vegetated
17	Water bodies

4.2 Climate layers

The monthly values of precipitation, temperature and potential evapotranspiration form the basis for the climatic component of the database. Legates and Willmott (1990a, b) acquired the data of the precipitation and temperature using traditional land-based gauge measurements and shipboard estimates spanning the period from 1920 to 1980. The values were then interpolated to a 0.5° lat./long. grid using an enhanced distance-weighting interpolation procedure (Legates and Willmott 1990a, b). Ahn and Tateishi (1994) produced the potential evapotranspiration data set by applying the Priestley–Taylor formula to a global data set of air temperature, albedo, cloudiness and elevation. To provide insight in the seasonality of the 12-month data series, several descriptive calculations were performed for each cell within the raster and each climate variable. This resulted in new maps, e.g., a map showing the minimum/maximum values or the range of values of each cell. An additional calculation was made for the precipitation- and potential evapotranspiration values, namely the sum of the 12-month values (the total precipitation and evapotranspiration over a year). To derive information about the dry period, we calculated during which months a water deficit would occur. As potential evapotranspiration (PET) approaches higher values during the warmer months of the year, precipitation (PREC) falls off. By the time evapotranspiration reaches the maximum values, it has exceeded precipitation. This results in a water deficiency for the vegetation. In the GIS the precipitation-layers were subtracted from the evapotranspiration-layers (PET–PREC); any surplus (PET > PREC) in the resulting layers indicates a water deficiency. To characterise the dry

period of the year, the degree of dryness (water deficit categorization) and the duration of water deficiency during a continuous period of the year (''duration long period'') were also calculated in the GIS. To provide a continuous data set visually comparable to higher resolution variables, all climate layers were interpolated and smoothed by trend surface analysis in the GIS. The same points from which data is extracted for model creation, serve as interpolation points.

4.3 Preparation of the data base: retrieving the variables

To understand how the variables spatially relate to each other, all layers were overlaid in the GIS. A GIS layer was created in which a random point raster was built. For every point, the underlying data of the variable layers was extracted into a DBF-format document. To investigate the influence of randomly distributed and evenly allocated sample points (motivated by a paper by Hirzel and Guisan 2002), a second point layer of evenly spaced grid points was constructed, and the underlying data was extracted in a similar way (''regular sampling'', Hirzel and Guisan 2002). To derive at least 300 data points for both savanna and forest, the random point raster consisted of a layer of 2000 points scattered out over the area of interest, whereas the regular sample layer consisted of a square of 18 × 18 data points.

The variables as used for the predictive modelling are listed in Table 4. The land-cover layer has been converted to a presence/absence layer of savanna and the outcomes of the descriptive statistics each form a separate GIS layer.

5 Logistic regression results and discussion

The results of the factor analysis in which all climatic variables are included are shown in Table 5. The matrix shows how significantly variables belong to which component, and the order of importance. Three components were extracted, which together explain 89.1% of the total variance (41.7; 29.3; 18.1, respectively). The strongest component consisted of the total water deficit (PET > PREC), dry period in months (DRY_PER), annual precipitation (PREC_SUM), water deficit balance (PET–PREC) and the driest month in a year (PREC_MIN). The second component was composed of temperature variables in combination with the total evapotranspiration in mm/year. The third and weakest component included the range of the precipitation values (PREC_RNG), the value of the wettest month (PREC_MAX) and the temperature range (TEMP_RNG). Based on the backward stepwise regression methodology, the best model for the logistic predictor was:

$$\begin{aligned} LP = {} & 0.822 - 0.0162\,PET\_SUM + 0.585\,DRY\_PER \\ & + 0.0136\,DRY\_SUM + 0.00100\,PREC\_SUM \\ & + 0.0107\,PREC\_MIN + 0.653\,TEMP\_MAX \\ & - 0.539\,TEMP\_RNG \end{aligned}$$

PET_SUM, total potential evapotranspiration [mm/year]; DRY_PER, duration dry period [months/year]; DRY_SUM, total water deficit [mm/year]; PREC_SUM, total precipitation [mm/year]; PREC_MIN, precipitation of driest month [mm]; TEMP_MAX, maximum temperature [°C]; TEMP_RNG, temperature range [°C].

This equation was implemented into the GIS, with the different data layers as input for the calculation. Subsequently, the application of the inverse logistic transformation (Sect. 3.2) resulted in a layer showing the probability of savanna occurrence in values ranging from 0 to 1. The actual land-cover distribution in Fig. 2a was compared to the probability surfaces in Fig. 2c, d, that show the logistic regression model output of respectively randomly and evenly distributed points. First the default [0.5] threshold was used to differentiate between the presence and absence of savanna (yellow lines in Fig. 2c, d). However, after comparing the map of the savanna probability (Fig. 2c, d) to the actual land-cover distribution

Table 4 Overview of all variables implemented into the predictive modelling

Variable	Abb.	Data type	Function	Unit	Distribution	Skewness	Kurtosis	Derived layers
Savanna		Binary (1,0)	Dependent	–	Binomial	–	–	Presence/absence
Elevation	ELEV	Numeric	Independent	Meters	App. normal	0.779	0.292	Elevation, slope, aspect
Precipitation	PREC	Numeric	Independent	mm/m	App. normal	0.313	–0.581	Descriptive statistics[a]
Temperature	TEMP	Numeric	Independent	°C/m	No normality	–1.284	2.139	Descriptive statistics[a]
Potential evapotranspiration	PET	Numeric	Independent	mm/m	Normal	0.09	0.172	Descriptive statistics[a]
Water deficit	PET > PREC	Numeric	Independent	mm/m	App. normal	–0.412	–0.275	Duration dry period, annual deficit

[a] Descriptive statistics performed on variables: mode, mean, range, variance, maximum, minimum values

Table 5 Rotated component matrix from PCA, with three explanatory components extracted

	Component		
	1	2	3
PET > PREC	0.946		
DRY_PER	0.935		
PREC_SUM	–0.915		
PET–PREC	0.913		
PREC_MIN	–0.906		
TEMP_MN		0.972	
TEMP_MED		0.964	
TEMP_MIN		0.947	
TEMP_MAX		0.916	
PET_SUM		0.744	
PREC_RNG			0.965
PREC_MAX	–0.584		0.769
TEMP_RNG			0.697

(Fig. 2a), it became appealing to increase the 0.5 threshold to 0.6 (red lines in Fig. 2c, d), which means that the boundary between savanna and forest ''moves up''. Figure 2e, f show the difference between the predicted land-cover distribution and the observed values at a 0.5 threshold, when samples where taken from random points (Fig. 2e) and evenly sampled points (Fig. 2f). When the threshold of 0.5 was changed to 0.6, the predicted land-cover (Fig. 2e) for the random sampling changed (Fig. 2g). Figure 2h shows the outcome of the 0.6 threshold for regular sampling. The accuracy measurements derived from both the threshold values are shown in Table 6.

5.1 Predictive variables in model

The variables indicated by the predictive model as determinants of the savanna distribution, correspond to earlier publications (Sarmiento 1983; San Jose et al. 1998; Rippstein et al. 2001; Hooghiemstra et al. 2002).

The components of the factor analysis assemble the variables, which are closely correlated to each other. The first component consists of just the precipitation and water deficit variables. It explained 41.7% of the total variance found in the data base, indicating a strong influence of the precipitation gradient and the dry period in the study area. The second component, composed of temperature related variables and the potential evapotranspiration, explained 29.3% of the total variance. The inclusion of potential evapotranspiration in this component made sense as high air temperatures generally increase the loss of moisture from soil and plants. The third component explained only 18% of the total variance, but was still considered important enough to include in the logistic regression formula.

In the logistic regression, the important predictors of savanna distribution were the same key variables as before: the precipitation and water deficit variable were best represented in the predictive model.

Temperature plays an important role as predictor variable in the form of the warmest month value, and also more surprisingly, as the annual temperature range. The area of the Colombian savanna has a small amplitude of annual temperature change which should not influence the vegetation distribution significantly (Rippstein et al. 2001). However, in logistic regression the variables are combined to achieve a certain degree of predictive accuracy, i.e., variables form a model when in combination, not individually. The temperature range value may indeed be considered as a relevant predictor variable, because it is related to the precipitation or evapotranspiration values. The temperature variable when considered separately does not play a significant part in the predictive capacity of the model. However, when included in the logistic regression model, it increased the overall accuracy of the predictions.

5.2 Model evaluation: overall accuracy, sensitivity and specificity

When the logistic regression procedure was completed within the statistical programming, the model generated an overall accuracy of 80.7%, for the random sampling method at the 0.5 threshold. Based on only this statistical outcome, it could be concluded that this model is a sufficiently accurate representation of the real land-cover distribution. According to Manel et al. (2001), about 36% of the users of presence–absence models in ecological publications during the period of 1989–1999 confined their model evaluation to this general prediction success value, while many more did not carry out any evaluation (55%). In our study we also used the Kappa's coefficient and the TSS to evaluate the models performance. By running the logistic model in the GIS, the different patterns of the predicted and the actual land-cover distribution could be better understood.

Based on the general accuracy measurements—overall accuracy, sensitivity and specificity—all created models achieved an acceptable predictive power with an overall accuracy range of 81–86%. The deviating values of sensitivity and specificity are noticeable, where there exists a certain trade-off of predictive capability in achieving a higher overall accuracy. In the random sampling method, the threshold of 0.6 reaches a higher overall accuracy (84.5%) than the 0.5 cut-point (80.7%). This can be explained by the shifting of the savanna-boundary further north in Fig. 2c (the savanna boundary shifts from the yellow to red line), which resulted in a larger area correctly predicted as forest, with an increase in specificity (Table 6,

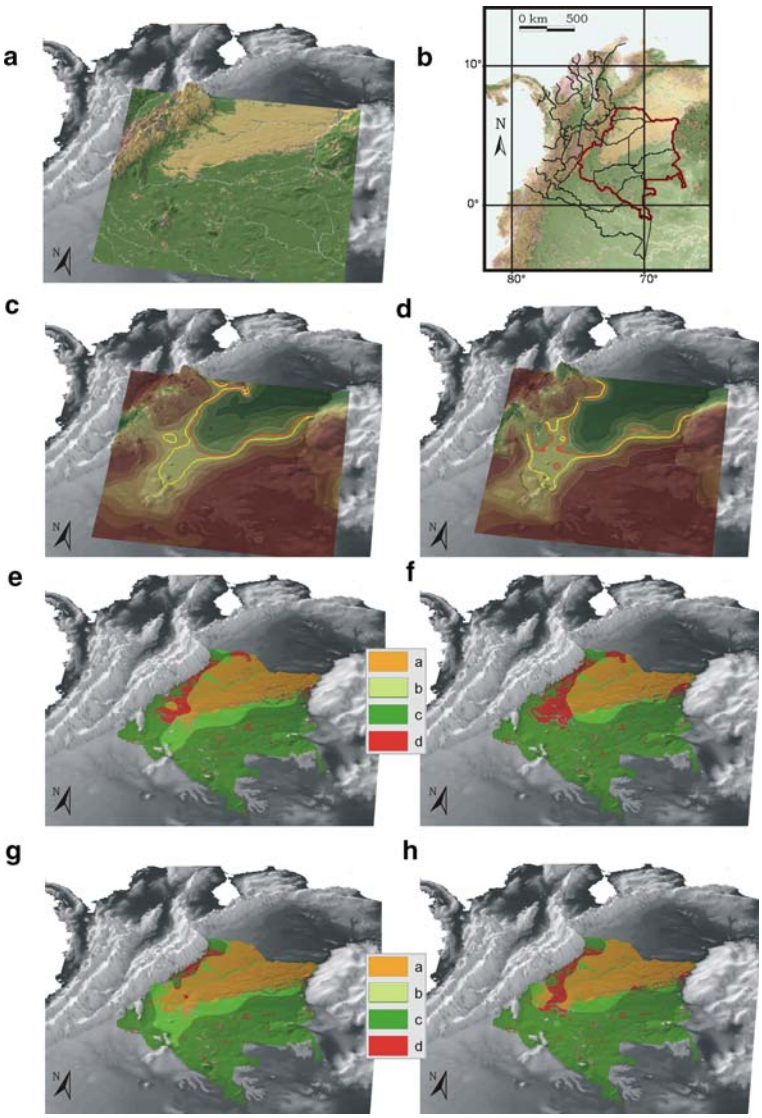

Fig. 2 Study area and outcomes of predicted land-cover distribution by logistic model: **a** South-America indicating the location of the study area and the actual land-cover distribution, *green* is forest and *yellow* is savanna; **b** location of the Colombian savanna biome between the Andes and the Guyana Shield (03–07°N, 68–71°W); **c** probability map of savanna occurrence based on random data sampling; **d** Probability map of savanna occurrence based on regular sampling. *Yellow lines* indicate the 0.5 threshold, *red lines* delineate the 0.6 cut-point; **e**–**h** Differences between observed land-cover distribution and model prediction; **e** based on random sampling at 0.5 threshold; **f** random sampling at 0.6 threshold; **g** regular sampling at 0.5 threshold; **h** regular sampling at 0.6 threshold. The *letters* correspond to the letters of Table 1: (*a*) indicates correctly predicted savanna [*Dark yellow*]; and (*b*) represents where the model falsely predicted savanna [*Bright green*]; (*c*) indicates correctly predicted forest [*Dark green*]; while (*d*) shows where the model failed to predict savanna [*Red*]. The statistical accuracy of these outcomes is detailed in Table 6

81.0–91.7%). However, the trade-off exists in the decrease of the sensitivity which fell from a 80.0 to 70.9% correctly predicted savanna area. When the aim is to create a model representing the savanna distribution, the power to predict the occurrence of savanna (sensitivity) is evidently thought to be of more value than to predict its absence (the specificity). In the regular sampling methodology, the evaluation of the different accuracy measurements is clearly important. At a 0.5 threshold, the overall accuracy is at a good predictive capacity of 81.7% of the cases. However, when considering the large area that the model included within the 0.5 threshold for savanna, it is surprising that the model performed well in predicting savanna presence (90.4%).

Table 6 Predictive accuracy of the created models (using either a random or a regular sampling method, and setting the threshold at a 0.5 and a 0.6 cut-point)

	Random sampling distribution		Regular sampling distribution	
Threshold value	0.5	0.6	0.5	0.6
Prevalence (%)	34.7	34.5	34.5	34.9
Overall accuracy (%)	80.7	84.5	81.7	85.7
Sensitivity (%)	80.0	70.9	90.4	79.0
Specificity (%)	81.0	91.7	77.0	89.3
Kappa	0.588	0.647	0.624	0.685
TSS	0.610	0.626	0.674	0.683

Simple accuracy measures have been expressed as percentages

When random sampling and regular sampling were compared (Table 6), the values of Kappa and TSS are both higher for regular sampling. Regular sampling on a raster of 18 by 18 grid points therefore gives more robust results than random sampling, despite the fact that less points were used (324 data points as opposed to 2,000 points used for the random sampling). The model that seems to fit all requirements is the regular sampled model at a 0.6 cut-

point (Fig. 2h). The overall accuracy is highest, predicting 85.7% of the cases correctly, while both Kappa and TSS values range up to a value of 0.68, which is stated to be within ''substantial agreement'' (Landis and Koch 1977). In both sampling methods the Kappa and TSS values show an increase of the models agreement with the true distribution land-cover, when changed from a 0.5 to 0.6 threshold.

When comparing all models in Fig. 2, the random sampling resulted in more smoothed outcomes, while the evenly distributed samples yielded more irregular shapes. Taking into account that the original climate database has a resolution of 0.5° (55 km), from which it was interpolated into more detailed climatic patterns, the regular distribution of points resulted in some unrealistic variation. However, the regular sampling still provided more powerful predictive models, which concurs with the findings of Hirzel and Guisan (2002), who showed that systematic sampling is more accurate and robust than random sampling strategies.

5.3 False prediction of absence (forest)

The red areas in Fig. 2e indicate that the model predicted an absence of savanna at 7.9% of the total area of interest, when actually the savanna is present. This suggests that according to the model, this specific area differs markedly from the savanna region northward, due to differences in climatic conditions. Figure 3 shows an evaluation of the data in the GIS, showed for one climatic variable, the total annual precipitation, a diverging distribution. The encircled area shows precipitation values up to 3,800 mm/year, which compares to the values at the south eastern precipitation front—rainforest-area of the shown map (Fig. 3, pointer). To consider the possibility that these interpolations overestimated the precipitation values, the precipitation maps of the Geographic Institute Agustin Codazzi (IGAC) were consulted. Focusing on high precipitation areas in the Colombian savanna, similar high rates were found in the corresponding region with values ranging from 4,000 to 5,000 mm per year (IGAC 2002) and as early as 1967, Blydenstein observed that in the Llanos Orientales ''total annual rainfall ranged between 1,700 and 2,000 mm, increasing sharply near the base of the mountains to over 4,000 mm at Villavicencio'' (Blydenstein 1967). It is therefore possible that in this region with steep precipitation gradients and few meteorological stations, the high precipitation rates at Villavicencio have been extrapolated over a region that is too large in area. Comparison of our precipitation map with Latorre (1977) supports this hypothesis.

By simulating the world's distribution of ecosystems in the absence of fire, Bond et al. (2005) discovered that large areas in South America are dominated by grasslands and

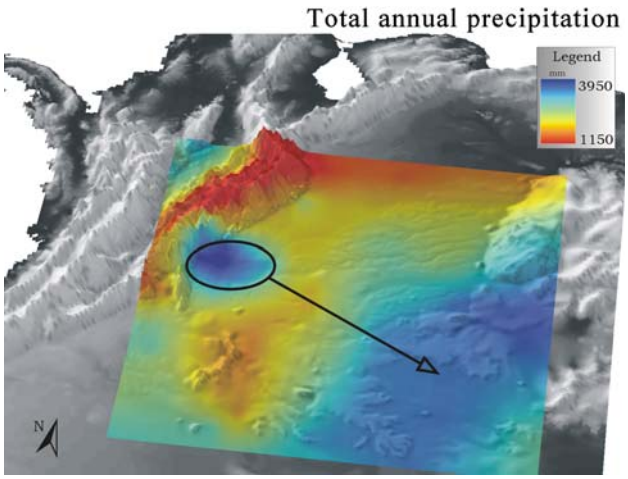

Fig. 3 Distribution of total annual precipitation

savannas, when the climatic conditions could support woodlands and forests. Although confined to a very coarse resolution, as the simulations are set for a global scale, it becomes evident from the reconstructed maps that forest and woodlands would replace large areas of the Colombian savanna in the absence of frequent burns. This partly explains the model's ''inaccuracy'' of predicting forest, when the current land-cover is savanna.

5.4 False prediction of presence (savanna)

The bright green area in Fig. 2e indicates that the model predicted presence of savanna for 14.1% of the total area of interest, when actually savanna is absent. This suggests that the model identified this specific area as dissimilar to the forest-region southward owing to the differences in climatic conditions.

An evaluation of the data in the GIS, showed that the area of false presence resembles the spatial limits of the total annual water deficit. In Fig. 4, ''water deficit'' (PET > PREC) is shown, with the yellow line indicating the boundary of savanna distribution according to the model. The southern border of the false prediction area appears to be strongly related to the isolines of the water deficit. The layer of the duration of the dry period (not shown here) has similarly situated isolines, indicating a strong influence of the degree of dryness.

5.5 Best fit model and possible improvements

Owing to the general increase of the model-accuracy, the 0.6 cut point is preferred for the predictive model. The 85.7% overall accuracy of this model is considered as an acceptable performance. Nevertheless, the inaccuracy of the model is greatest, not surprisingly, in the zone where the transition of forest to savanna takes place. This tran-

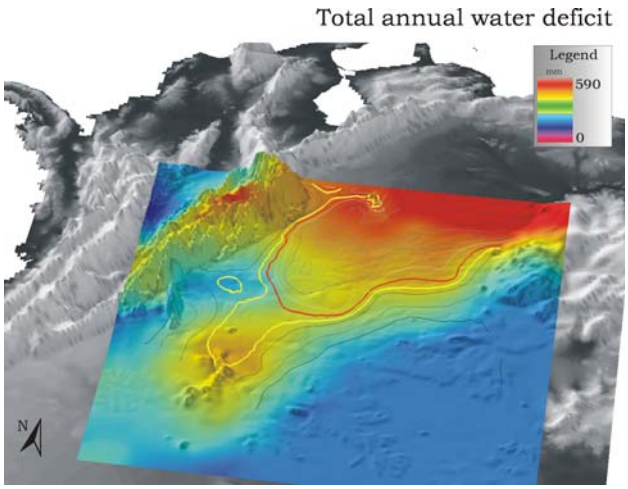

Fig. 4 Distribution of the annual water deficit layer (mm), with delineation of the 0.5 threshold by the *yellow line* and the 0.6 in *red*

sition zone is of considerable relevance to this study as it is the focus of attention of palynological research, as most of the pollen sites are located in this zone. Although the model performs well, the characterisation of the transition zone model therefore remains difficult.

Several factors may have influenced the predictive performance of the model. The first factor is within the model itself: the absence of one (or more) explanatory variable(s), which could increase the ability to differentiate between savanna presence and absence—a general goal to improve the accuracy of the model, in particular, the effect of fire. To further develop the model, consisting only of climate variables, the introduction of a more complex form of logistic regression modelling with more discriminant components may be useful.

The second factor is the precision of the data. The readily available surface layers of the climatic variables were interpolated from local measurements at meteorological stations to produce maps. However, if there is a lack of evenly distributed stations, especially in areas with strong gradients, the interpolation is inaccurate. The created datasets still depend greatly on the expert knowledge of the area to correct erroneous interpolation values. Furthermore, the model assumes the vegetation is in pseudo-equilibrium with its environment. Nonetheless, this area of vegetation could well be in a phase of shifting towards a new equilibrium after an environmental change. Part of the deficiency in the model's predictive capacity may actually be adequately explained by ecological theory and historical events, if the temporal and stochastic dimensions of population dynamics are taken into account (Guisan and Thuiller 2005).

By implementing the logistic regression model into GIS, the weaknesses of the model have become evident. It is in this part of the ecological modelling process that the utility of GIS is demonstrated, showing that the spatial patterns of the models are directly comparable to the true land-cover patterns. Not only can the interpretation of the pattern of errors contribute to an improvement of the model, but can aid the understanding of the responsiveness of land-cover to different environmental conditions, and therefore to the system as a whole.

6 Palynological GIS application

Fossil pollen spectra are translated into palaeo-environmental and palaeo-climatological conditions following the papers by, e.g., Behling and Hooghiemstra (1998, 1999). Berrio et al. (2000) discuss how pollen from savanna vegetation may be over/under represented in the pollen record. In previous studies pollen data of the study area were assessed in the vegetation model Biome-3 (Marchant et al. 2004, 2006). In this study, pollen data of the savannas of the Colombian Llanos Orientales were introduced into the GIS through pollen percentages implementation and interpolation methods, to evaluate the application of palynological data in GIS. An assessment is made of the suitability of the pollen data for a GIS analysis, in which both limitations of the present approach and recommendations for future work are discussed.

6.1 Pollen sites and time series

The ten pollen records available in our study area are positioned in an east to west transect as can be seen in Fig. 5. Site-specific data are listed in Table 7. The transect covers a distance of approximately 480 km. The data of the pollen records are organized in pollen diagrams, which display the percentages of pollen taxa found. The pollen taxa are classified into the following ecological groups, according to Behling and Hooghiemstra (1999, 2000): (1) trees of forest and gallery forest; (2) shrubs and trees of savannas; (3) savanna herbs; (4) aquatics; (5) ferns. In order to make a dichotomous land-cover layer (savanna/forest) in the GIS, all ecological groups, which are considered as representative for savanna vegetation (groups 2 and 3), are grouped together. Forest vegetation is based on the pollen from group 1. The data from the aquatics and ferns are not taken into account.

To obtain a reconstruction of temporal land-cover changes, pollen spectra at successive time slices have been compared. The time slices of interest were selected based on the amount of available data and the degree of change compared with earlier time slices to make meaningful temporal intervals (Fig. 6). To be able to use the interpolation tool of the GIS—Geostatistical Analyst, more than nine points are required (ESRI 2001). The age ranges of the

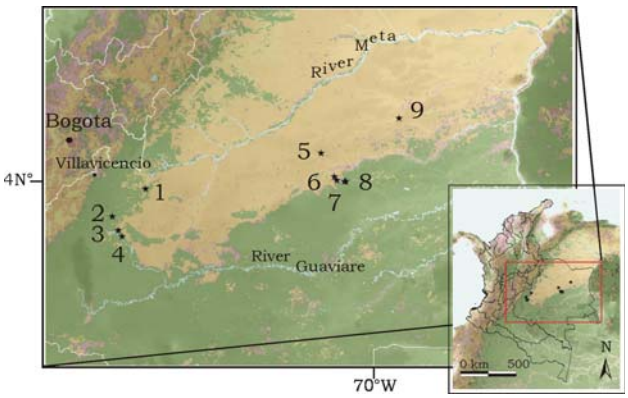

Fig. 5 Location of pollen sites. The numbers are according to Table 7. *Green* reflects forest and *yellow* reflects savanna

available pollen records of the Colombian savanna are shown in Fig. 7. The chart shows already that interpolation-attempts are not useful for time slices older than 8,000 ^{14}C year BP, because they include data from only four or less sites. To overcome this problem of point interpolation, more data points were added in the far north and south of the area of interest, which were assumed to have 90 and 0% pollen of savanna taxa, respectively (Table 7). The pollen sites in the north correspond to places where lakes have been cored, but where the cores did not contain enough pollen to construct a reliable time series. The presence of gallery forest in a savanna area explains why the proportion of savanna taxa never reaches values close to 100% (e.g., Berrio et al. 2000), and a lower value reflects the savanna–forest transition.

Furthermore, to facilitate extrapolation outside the geographical area of the pollen sites, four extra outlier points were created to enlarge the area covered by the interpolations (Table 7c). These additional data points were set at 0% savanna pollen, given the fact that these additional points are complementary to the palynological pollen sites and are located far outside the savanna area.

Solving the problem of insufficient point data also gives a useful trend direction to the interpolated surface, i.e., a

Table 7 Site-specific data of the pollen records used in this study

a.	Name of Pollen site	Coordinates	Elevation m.a.s.l.	Age range (^{14}C year BP)	^{14}C No.	References
1	Mozambique	3°58′N,73°03′W	175	0–3,450	7	Berrio et al. (2002)
2	Agua Sucia	3°35′N,73°31′W	300	0–5,500	4	Wijmstra and Van der Hammen (1966)
3	Las Margaritas	3°23′N,73°26′W	290	850–9,760	9	Wille et al. (2003)
4	Loma Linda	3°18′N,73°23′W	310	0–8,720	8	Behling and Hooghiemstra (2000)
5	El Angel	4°28′N,70°34′W	200	2,000–10,030	5	Behling and Hooghiemstra (1998)
6	El Piñal	4°08′N,71°23′W	180	0–18,000	6	Behling and Hooghiemstra (1999)
7	Chenevo	4°05′N,70°21′W	150	0–7,260	6	Berrio et al. (2002)
8	Carimagua	4°04′N,70°14′W	180	0–8,270	6	Behling and Hooghiemstra (1999)
8	Carimagua-Bosque	4°04′N,70°13′W	180	0–1,300	9	Berrio et al. (2000)
9	Sardinas	4°58′N,69°28′W	180	0–11,600	6	Behling and Hooghiemstra (1998)

b.	Name of extra point	Coordinates	Extra point	Pollen %	References
1	Caño La Mata	6°90′N,70°45′W	North	90	Hooghiemstra, personal info
2	La Maporita	6°93′N,70°47′W	North	90	Hooghiemstra, personal info
3	Grimonero	7°03′N,72°00′W	North	90	Hooghiemstra, personal info
4	Las Tres Marias	6°98′N,70°58′W	North	90	Hooghiemstra, personal info
5	La Porfira	6°92′N,70°50′W	North	90	Hooghiemstra, personal info
6	La Viga-Porfira	6°95′N,70°48′W	North	90	Hooghiemstra, personal info
7	Pantano de Monica	0°42′S,72°04′W	South	0	Behling et al. (1999)

c.	Outliers points	
1	North-east corner	7°05′N,66°50′W
2	North-west corner	7°05′N,75°00′W
3	South-east corner	1°50′S,66°50′W
4	South-west corner	1°50′S,75°00′W

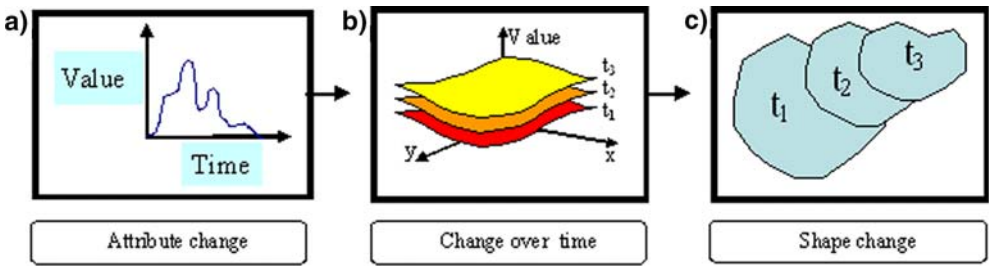

Fig. 6 Steps to create a time series of a variable: **a** attribute change; **b** change over time; **c** shape change

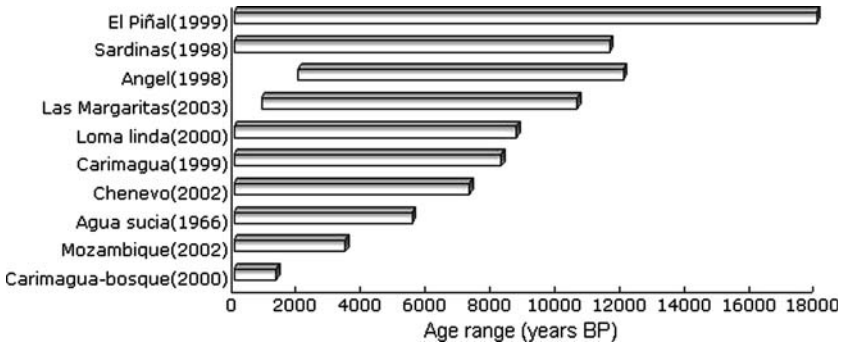

Fig. 7 Age ranges (^{14}C year BP) of the pollen records used in this study

high probability of savanna in the north compared to a low probability in the south and periphery of the study area.

6.2 Incorporating the pollen data into the GIS

The pollen percentages of the sites come from the original Excel files on which the original pollen diagrams were based. Pollen percentages per site and per sample have been inserted into dbf-format documents, which form the basis of datasets for selected time slices. The data of a time slice were interpolated to form maps of savanna pollen percentages, which are indicative of the land-cover distribution.

6.3 Interpolation methods

Choosing a proper interpolation method depends on the specific database. Some methods demand hardly any specifications, while others, like co-kriging, require an experienced GIS practitioner. There have been a number of comparisons of interpolation methods enabling the creation of guidelines for researchers to choose the best interpolation method. This issue is out of the focus of this paper. The usefulness of any interpolation method depends on the characteristics of the data set. However, we aim to evaluate how the distribution of pollen sites affects the capacity to make interpretations of the complete area.

A variety of interpolation techniques are available, which all have their own characteristics (Erdogan et al. 2005). In this study, two different interpolation methods were used, (1) Local Polynomial: this is a quick deterministic extrapolator that is smooth and therefore less exact. Few decisions are required to make the interpolation. There is no assessment of prediction errors, and there are no assumptions required of the data. (2) Radial Basis Functions: this technique is used for creating surfaces from measured points based on the degree of smoothing. The surface must go through each measured point location. It is a moderately quick deterministic interpolation technique, which is more robust and thus more exact than Local Polynomial. However there are more parameter decisions, which allow a variety of map outputs. This flexibility of the different applications within the interpolation functions requires decision-making. There is no assessment of prediction errors and no assumptions about the data. There are five different Radial Basis functions, from which the ''multiquadratic function'' is considered the best (Erdogan et al. 2005), and therefore used in this study.

These interpolation methods are applied using the Geostatistical Analyst (ESRI 2001) within ArcGIS. In the Local Polynomial interpolation, a power of two is used and the weight distance optimised. Interpolated values less than zero were omitted from the analysis.

7 Results and discussion

The layers created by the Local Polynomial interpolation method are shown in Fig. 8. The selected interpolation method resulted in general delineations of pollen percentages. From Berrio et al. (2000), we estimated the boundary

between forest and savanna at approximately 40% savanna pollen. This was taking into account the over-representation of arboreal pollen owing to: (1) the amount of pollen produced by trees; (2) the lakes cored were sometimes fringed by trees, even in savanna conditions. This boundary is indicated by the black line in Fig. 8. Although these maps should be interpreted cautiously as they are constructed from a few points, several trends are obvious. Circa 8,000 years ago the savanna covered a greater area. After this period, the savanna retreated, reaching a minimum coverage at ca. 3,000 year BP. The savanna then extended, and ca. 1,000 year BP the savanna–forest boundary seemed to change to a more southeast–northwest orientation that can still be observed today (Fig. 2a). The strip of forest at the foot of the Andes, resulting from orographic rainfall, is not represented in these interpolated maps, as no cores from that area have yet been taken.

The Local Polynomial method (Fig. 8) is a very general extrapolator, which connects the points of similar values. The alternative Radial Basis Functions interpolation method (Fig. 9) demands more parameter decisions. This method requires repetitive runs to adjust the selected parameters to the outcomes. If not, the interpolated maps show unrealistic distribution of pollen percentages. However, a continuous adjustment of the outcome to meet the expected distribution of land-cover seems arbitrary and subjective. Selecting a proper method requires a trial-and-error application to see which method is suitable for each dataset. The GIS offers several different interpolation methods that demand specifications of the distribution or number of data points.

The pollen dataset used in this study exhibits several limitations. The orientation of the transect of sites follows a west to east direction, when in the past the savanna–forest transition zone mainly had shifted in a north to south direction. Interpretation of the pollen graphs only allows conclusions to be made about relative changes instead of the desired geographic extent of the vegetation. Pollen sites distributed more evenly over the study area would allow an effective interpolation. A transect of data points only results in a 1D representation of the data (e.g., north–south), as the points give information about the local circumstances and the degree of change along the line of interpolated points. A large area will be subject to extrapolations with a higher uncertainty due to missing reference points. When the setting of the landscape remains constant over the length of a transect, the interpolation of the point data will be limited. The influx of pollen from edaphically determined gallery forest along the rivers in the savanna area caused another bias. Climatic conditions could not sustain forest. According to Berrío et al. (2000), the pollen signal of savanna is under-represented in lake sediments when the lake is totally surrounded by gallery forest. Once interpolated pollen maps have been satisfactorily produced, it remains to be decided on which percentage margin a differentiation is made between savanna and forest. Based on the influences of gallery forest, one could easily underestimate the presence of savanna. An

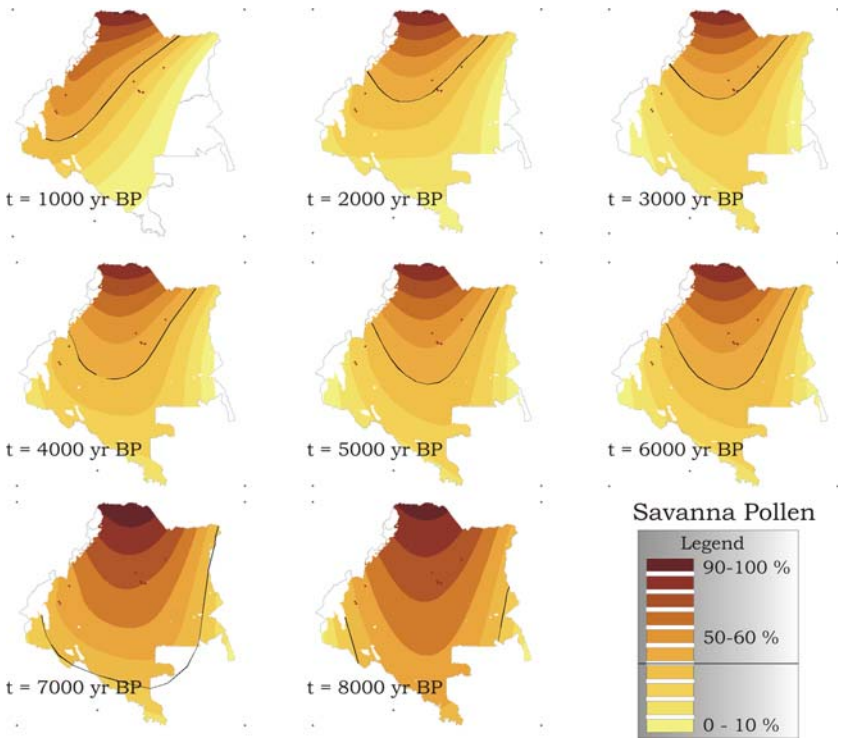

Fig. 8 Interpolated pollen percentages of taxa reflecting savanna vegetation based on Local Polynomial Interpolation method. The interpolated area corresponds to the area delineated in Fig. 2b. Selected time slices range from 1,000 to 8,000 ^{14}C year BP. (Interpolation specification: power = 2, Ideal weight distance activated). The *black line* indicates the estimate for the savanna–forest boundary

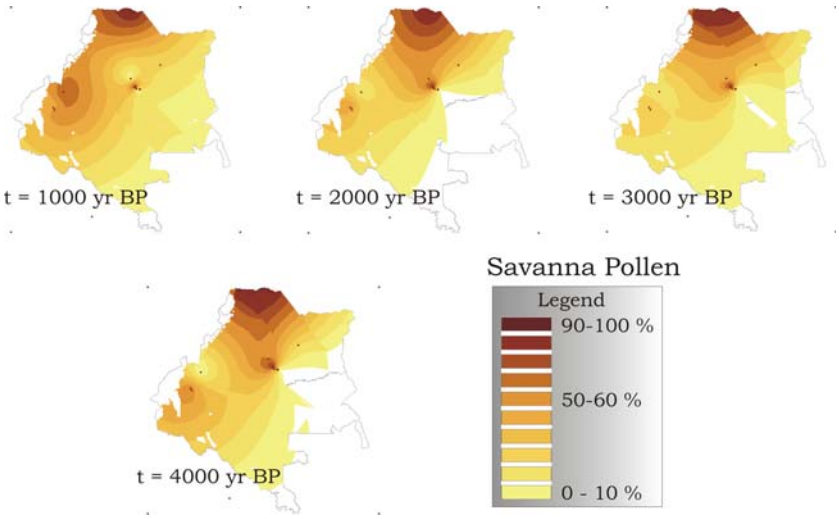

Fig. 9 Maps of interpolated pollen percentages of taxa reflecting savanna vegetation based on Radial Based Functions interpolation method. The interpolated area corresponds to the area delineated in Fig. 2b. Selected time slices range from 1,000 to 4,000 ^{14}C year BP

appropriate data set of modern pollen rain data is therefore necessary. Based on the pollen dataset used in this study, the transition zone between savanna and forest is cannot be accurately located.

8 Conclusions: GIS, statistics and palynology

The combination of GIS and logistic regression used in this study is a novel approach to modelling the spatial distribution of savanna vegetation and the incorporation of palynological data into GIS. Logistic regression has been chosen since the relationship between species distribution and predictor environmental variables is made obvious. The model resulting from the logistic regression, is run in a GIS, giving a visual representation of the predictive capacity of the model, outlining limitations, and so facilitating the improvement of the model.

To further improve the performance of the model, the logistic regression procedure as well as the data set can be developed further. The introduction of a more complex form of logistic regression modelling with more discriminant components (not only climate) may result in a higher precision of predictions. In addition to improving the resolution of the data set, it would be functional to add one (or more) explanatory variable(s) to the model, which were not considered in this study, such as the effect of fire. This would most probably increase the capacity of the model to differentiate between presence and absence of savanna vegetation. Other interesting modifications in the model would be the incorporation of inter-species competition effects, migration processes, and the effects of human disturbance. The incorporation of these factors would provide greater insight in the dynamic interface of savanna and forest.

When the model is considered to adequately represent the vegetation distribution, the model can be further employed to improve the interpretation pollen database. As subsequent step the manipulation of one or more climatic variable(s) by increasing or decreasing the overall climatic values with a certain percentage is proposed. The logistic model should then be re-run and introduced into the GIS according to the same methodology. The degree of vegetation change compared to the relative change of a climatic variable can then be defined and compared with the relative changes seen in the pollen data of the area of interest, to understand the response of the vegetation to changing environmental conditions in both past and future context.

To use this model to reconstruct past and future vegetation distributions, the effect of fluctuating atmospheric pCO_2 levels on the vegetation must be taken into account. As plant species are affected differently by variations in pCO_2 levels, this climatic component has contributed to the plant distribution in the past (Boom et al. 2002).

A number of recommendations for future modelling work stem from the results of this study: more complex forms of logistic regression modelling should be explored; more advanced interpolation methods should be implemented; one or more predictor variables, such as fire frequency, biotic interaction, and human disturbance should be implemented. Locations of pollen sites should be more evenly distributed over the study area to aid the understanding of the geographical migration of land cover boundaries in space and time, with the help of interpolation methods in GIS.

We conclude that analysis of pollen data in GIS offers new possibilities to evaluate multi-site data. A regional synthesis is not merely descriptive or embedded in a vegetation model (Biome-3) but the impact of particular climatic variables and their geographical gradients can be assessed,

while also sensitivity experiments may be carried out with the dataset.

Acknowledgments The present study was made possible by a cooperative effort between the Palaeoecology department and the GIS Studio. We thank: Guido van Reenen for additional advice as head of the GIS Studio at the University of Amsterdam; Willem Bouten (University of Amsterdam), for his interest in exploring new fields of collaboration; Juan Carlos Berrio (now at University of Leicester, UK) and Hermann Behling (now at University of Göttingen, Germany) for making available the raw data sets of the pollen records. Dan Yeloff (University of Amsterdam) is thanked for improving the English of the manuscript. We thank two anonymous reviewers for constructive comment on an earlier version of this paper.

References

Ahn CH, Tateishi R (1994) Development of a Global 30-minute grid potential evapotranspiration data set. J Jpn Soc Photogramm Remote Sens 33:12–21

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol [OnlineEarly]. doi: 10.1111/j.1365-2664.2006.01214.x

Behling H, Hooghiemstra H (1998) Late Quaternary palaeoecology and palaeoclimatology from pollen records of the savannas of the Llanos Orientales in Colombia. Palaeogeogr Palaeoclimatol Palaeoecol 139:251–267

Behling H, Hooghiemstra H (1999) Environmental history of the Colombian savannas of the Llanos Orientales since the Last Glacial Maximum from lake records El Piñal and Carimagua. J Paleolimnol 21:461–476

Behling H, Hooghiemstra H (2000) Holocene Amazon rainforest–savanna dynamics and climatic implications: high-resolution pollen record from Laguna Loma Linda in eastern Colombia. J Quaternary Sci 15:687–695

Behling H, Berrío JC, Hooghiemstra H (1999) Late Quaternary pollen records from the middle Caquetá River basin in central Colombian Amazon. Palaeogeogr Palaeoclimatol Palaeoecol 145:193–231

Belward AS (ed) (1996) The IGBP-DIS global 1 km land cover data set (DISCover)-proposal and implementation plans: IGBP-DIS. Working Paper No. 13, Toulouse, France, p 61

Berrío JC (2002) Lateglacial and Holocene vegetation and climatic change in the lowland Colombia. PhD thesis, University of Amsterdam, p 240

Berrío JC, Hooghiemstra H, Behling H, Botero P, Van der Borg K (2002) Environmental history of the western Colombian savannas of the Llanos Orientales since the Middle Holocene from Laguna Mozambique and Chenevo: transect sysnthesis. The Holocene 12:35–48

Berrío JC, Hooghiemstra H, Behling H, Van der Borg K (2000) Late Holocene history of savanna gallery forest from Carimagua area, Colombia. Rev Palaeobot Palynol 111:295–308

Bickford S, Mackey B (2004) Reconstructing pre-impact vegetation cover in modified landscapes using historical surveys and remnant vegetation data: a case study in the Fleurieu Peninsula, South Australia. J Biogeogr 31:787–805

Birks HJB (1989) Holocene isochrone maps and patterns of tree-spreading in the British Isles. J Biogeogr 16:503–504

Blydenstein J (1967) Tropical savanna vegetation of the Llanos of Colombia. Ecology 48:1–15

Bond WJ, Woodward FI, Midgley GF (2005) The global distribution of ecosystems in a world without fire. New Phytol 165:525–538

Boom A, Marchant R, Hooghiemstra H, Sinninghe Damsté JS (2002) CO_2- and temperature-controlled altitudinal shifts of C_4- and C_3-dominated grasslands allow reconstruction of palaeoatmospheric pCO_2. Palaeogeogr Palaeoclimatol Palaeoecol 177:151–168

Botero P (1999) Paisajes fisiograficos de Orinoquia-Amazonia (ORAM) Colombia. Análisis Geográficos 27–28:361 + maps. Instituto Geográfico 'Augustin Codazzi', Bogotá

Brubacker LB, Anderson PM, Edwards ME, Lozhkin AV (2005) Beringia as a glacial refugium for boreal trees and shrubs: new perspectives from mapped pollen data. J Biogeogr 32:833–848

Davis BAS, Brewer S, Stevenson AC, Guiot J (2003) The temperature of Europe during the Holocene reconstructed from pollen data. Quat Sci Rev 22:1701–1716

Dezzeo N, Chacon N, Sanoja E, Picon G (2004) Changes in soil properties and vegetation characteristics along a forest–savanna gradient in southern Venezuela. For Ecol Manag 200:183–193

Eeley HAC, Lawes MJ, Piper SE (1999) The influence of climate change on the distribution of indigenous forest in KwaZulu-Natal, South Africa. J Biogeogr 26:595–617

Erdogan S, Sahin M, Gullu M, Baybura T, Tiryakioglu I, Yavasoglu H (2005) Comparing the performance of different interpolation techniques in obtaining DEMS. International symposium on modern technologies, education and professional practice in geodesy and related fields, Sofia, 03–04 November 2005

ESRI —Environmental Systems Research Institute (2001) Using ArcGIS Geostatistical Analyst. Redlands, California, USA, p 306

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sens Environ 80:185–201

Giesecke T, Bennett KD (2004) The Holocene spread of *Picea abies* (L.) Karst. in Fennoscandia and adjacent areas. J Biogeogr 31:1523–1548

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009

Hirzel A, Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modeling. Ecol Model 157:331–341

Hooghiemstra H, Van der Hammen T, Cleef A (2002) Evolution of forests in the Northern Andes and Amazonian lowlands during the Tertiary and Quaternary. In: Guariguata MR, Kattan GH (eds) Ecología y conservación de bosques neotropicales. Edictiones Libro Universitario Regional, Cartago, pp 43–58

Hosmer DW, Lemeshow S (1989) Applied logistic regression, Chap 1–5. Wiley, New York

IGAC —Instituto Geográfico Agustín Codazzi, Atlas de Colombia 5ª Edición. 2002. Imprenta Nacional, Bogotá

Jago LCF, Boyd WE (2003) A GIS atlas of the fossil pollen and modern records of *Ficus* and related species for Island Southeast Asia, Australia and the Western Pacific. Aust Geogr Stud 41:58–72

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Latorre EA (1977) Atlas de Colombia. Institutio Geografico ''Augustin Codazzi'', Bogotá, p 283

Legates DR, Willmott CJ (1990a) Mean seasonal and spatial variability in Gauge-corrected, global precipitation. Int J Climatol 10:111–127

Legates DR, Willmott CJ (1990b) Mean seasonal and spatial variability in global surface air temperature. Theor Appl Climatol 41:11–21

Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu J, Yang L, Merchant JW (2000a) Global land cover characteristics database (GLCCD) Version 2.0. http://www.edcdaac.usgs.gov/glcc/globdoc2_0.html

Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu J, Yang L, Merchant JW (2000b) Development of a global land cover characteristics database and IGBP DISCover from 1-km AVHRR Data. Int J Remote Sens 21:1303–1330

Lyford ME, Jackson ST, Betancourt JL, Gray ST (2003) Influence of landscape structure and climate variability on a late Holocene plant migration. Ecol Monogr 73:567–583

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Marchant R, Boom A, Behling H, Hooghiemstra H, Melief B, Van Geel B, Van der Hammen T, Wille M (2004) Colombian vegetation at the Last Glacial Maximum: a comparison of model- and pollen-based biome reconstructions. J Quaternary Sci 19:721–732

Marchant R, Berrio JC, Behling H, Boom A, Hooghiemstra H, (2006) Colombian dry moist forest transitions in the Llanos Orientales—a comparison of model and pollen-based biome reconstructions. Palaeogeogr Palaeoclimatol Palaeoecol 234:28–44

McPherson JM, Jetz W, Rogers DJ (2004) The effect of species' range sizes in the accuracy of distribution models: ecological phenomenon or statistical artefact? J Appl Ecol 41:811–823

Mistry J (2001) World Savannas. Prentice Hall, London, p 344

Paez MM, Schäbitz F, Stutz S (2001) Modern pollen-vegetation and isopoll maps in southern Argentina. J Biogeogr 28:997–1021

Ray N, Adams JM (2001) A GIS-based vegetation map of the world at the last Glacial Maximum (25,000–15,000 BP). Internet Archaeol 11. http://www.ncdc.noaa.gov/paleo/pollen.html

Rippstein G, Escobar G, Motta F (2001) Agroecología y biodiversidad de las Sabanas en Llanos orientales de Colombia. Cali: Centro internacional de Agricultura Tropical (CIAT). Publication CIAT No. 322

San Jose JJ, Montes R, Mazorra M (1998) The nature of savanna heterogeneity in the Orinoco Basin. Global Ecol Biogeogr Lett 7:441–445

Sarmiento G (1983) The savannas of tropical America. In: F Bourlière (ed) Tropical savannas. Ecosystems of the world, vol 13. Elsevier, Amsterdam, pp 245–288

Sarmiento G (1984) The ecology of neotropical savannas. Harvard University Press, Cambridge, p 235

Veski S, Koppel K, Poska A (2005) Integrated palaeoecological and historical data in the service of fine-resolution land use and ecological change assessment during the last 1000 years in Rõuge, southern Estonia. J Biogeogr 32:1473–1488

Wijmstra TA, Van der Hammen T (1966) Palynological data on the history of tropical savannas in northern South America. Leidse Geologische Mededelingen 38:71–90

Wille M, Hooghiemstra H, Van Geel B, Behling H, De Jong A, van derBorg K (2003) Submillennium-scale migrations of the rainforest–savanna boundary in Colombia: 14C wigglematching and pollen analysis of core Las Margaritas. Palaeogeogr Palaeoclimatol Palaeoecol 193:201–223

Yu G, Tang L, Yang X, Ke X, Harrison S (2001) Modern pollen samples from alpine vegetation on the Tibetan Plateau. Global Ecol Biogeogr 10:503–519