Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 164 (2003) 257–270

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Spatial autocorrelation in multi-scale land use models

K.P. Overmars [a,b,*], G.H.J. de Koning [c], A. Veldkamp [b]

[a] *Centre of Environmental Science (CML), Leiden University, P.O. Box 9518, 2300RA Leiden, The Netherlands*
[b] *Laboratory of Soil Science and Geology, Wageningen University, P.O. Box 37, 6700AA Wageningen, The Netherlands*
[c] *Institute of Soil Science and Forest Nutrition, University of Goettingen, Büsgenweg 2, 37770 Goettingen, Germany*

Received 14 January 2002; received in revised form 8 January 2003; accepted 17 February 2003

Abstract

In several land use models statistical methods are being used to analyse spatial data. Land use drivers that best describe land use patterns quantitatively are often selected through (logistic) regression analysis. A problem using conventional statistical methods, like (logistic) regression, in spatial land use analysis is that these methods assume the data to be statistically independent. But, spatial land use data have the tendency to be dependent, a phenomenon known as spatial autocorrelation. Values over distance are more similar or less similar than expected for randomly associated pairs of observations. In this paper correlograms of the Moran's *I* are used to describe spatial autocorrelation for a data set of Ecuador. Positive spatial autocorrelation was detected in both dependent and independent variables, and it is shown that the occurrence of spatial autocorrelation is highly dependent on the aggregation level. The residuals of the original regression model also show positive autocorrelation, which indicates that the standard multiple linear regression model cannot capture all spatial dependency in the land use data. To overcome this, mixed regressive–spatial autoregressive models, which incorporate both regression and spatial autocorrelation, were constructed. These models yield residuals without spatial autocorrelation and have a better goodness-of-fit. The mixed regressive–spatial autoregressive model is statistically sound in the presence of spatially dependent data, in contrast with the standard linear model which is not. By using spatial models a part of the variance is explained by neighbouring values. This is a way to incorporate spatial interactions that cannot be captured by the independent variables. These interactions are caused by unknown spatial processes such as social relations and market effects.

Keywords: Land use model; Spatial autocorrelation; Autoregressive model; Regression; Multi-scale; Land use drivers

1. Introduction

Various modelling approaches exist for the simulation and exploration of land use change. Land use change modelling, especially if done in a spatially-explicit, integrated and multi-scale manner, is an important technique for the projection of alternative pathways into the future, for conducting experiments that test our understanding of key processes, and for describing the latter in quantitative terms (Lambin et al., 2000; Veldkamp and Lambin, 2001). Land use change models represent part of the complexity of land use systems. They offer the possibility to test the sensitivity of land use patterns to changes in selected variables. They also allow testing of the stability of linked social and ecological systems, through scenario building. While, by definition, any model falls short of incorporating all aspects of reality, it provides valuable information on the system's behaviour under a range of conditions. Different modelling

* Corresponding author. Tel.: +31-71-527-7475; fax: +31-71-527-4796.
E-mail address: overmars@cml.leidenuniv.nl (K.P. Overmars).

0304-3800/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0304-3800(03)00070-X

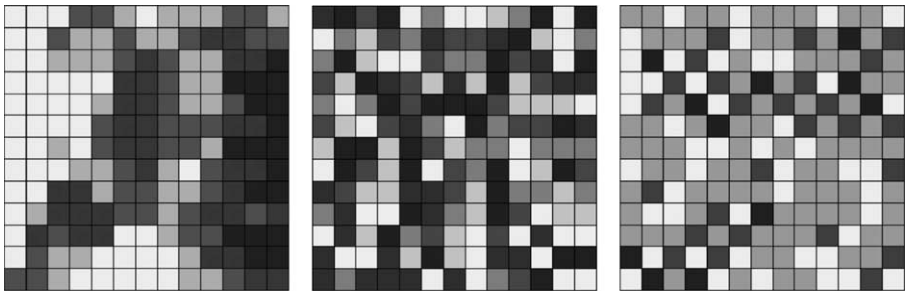

Fig. 1. Visualisation of positive spatial autocorrelation (left), no spatial autocorrelation (middle), and negative spatial autocorrelation (right) in an imaginary 13 × 13 grid. The different tones of grey indicate different values of a variable.

approaches have been adopted in the study of land use/land-cover change (see reviews by Sklar and Costanza, 1991; Lambin, 1994; Riebsame and Parton, 1994; Kaimowitz and Angelsen, 1998; Lambin et al., 2000; Veldkamp and Lambin, 2001). In several land use models statistical tools are being used to analyse spatial data, for example data organised as polygons or grid cells. Examples of statistical approaches in gridded spatial data are CLUE (Veldkamp et al., 2001; Verburg et al., 2002) and GEOMOD (Pontius et al., 2001). In these approaches the study area is sub-divided into grid cells and described by a pre-determined set of biophysical and socio-economic variables. Through (logistic) regression analysis, those variables are selected that best describe land use patterns quantitatively, the so-called land use drivers (Verburg et al., 1999).

Land use modelling often involves substantial amounts of data with a spatial component. Theories about spatial processes are constructed and hypotheses tested. Much of this testing is done with conventional statistical methods. The problem of using conventional statistical methods in spatial land use analysis, like linear regression based on ordinary least squares (OLS) (in this study indicated with 'standard linear model') and logistic regression using ROC, is that these methods assume the data to be statistically independent and identically distributed (iid) (Cliff and Ord, 1981). But, spatial land use data have the tendency to be dependent, a phenomenon known as spatial autocorrelation, which can be defined as the property of random variables to take values over distance that are more similar or less similar than expected for randomly associated pairs of observations, due to geographic proximity (Fig. 1) (Legendre and Legendre, 1998).

Spatial dependency could be seen as a methodological disadvantage, but on the other hand it is exactly what gives us information on spatial pattern, structure and processes (Gould, 1970). Spatial dependency contains useful information but the appropriate statistical methods have to be used to deal with it.

The effects of spatial dependence on conventional statistical methods are various, for example, biased estimation of error variance, *t*-test significance levels, and overestimation of R^2 (Anselin and Griffith, 1988). All the usual statistical tests have the same behaviour: in the presence of positive autocorrelation, computed test statistics are too often declared significant under the null-hypothesis. Negative autocorrelation may produce the opposite effect (Legendre and Legendre, 1998). This is caused by the fact that an observation carries less information then an independent observation, since it is partly predictable from its neighbours and a new sample point does not bring with it one full degree of freedom (Cliff and Ord, 1981; Legendre and Legendre, 1998).

Until recently (see, e.g. Nelson, 2002) often ordinary statistics were used in studies dealing with spatial data, although several techniques are available to deal with spatial autocorrelation (Anselin and Griffith, 1988). A reason for this could be that the spatial modelling techniques originate in the econometric sciences and are not easy applicable to land use models.

The objectives of this paper are the following: (i) to demonstrate the presence of spatial autocorrelation in a case study for Ecuador at different spatial scales (de Koning et al., 1998). Spatial autocorrelation is

tested for different data sets at different aggregation levels to get insight in the behaviour and extent of the spatial patterns. In relation to spatial autocorrelation it is important to consider the scale of the data. Both resolution and the extent of spatial data influence the pattern that can be observed. (ii) To evaluate a spatial regression model and its scale dependency for this case study.

2. Methods

2.1. Data and study area

The data set used for this study consists of spatial explicit land use data, bio-geophysical data and socio-economic data of Ecuador. This data set was originally created by de Koning et al. (1998) to determine the factors, so-called land use drivers, that best describe land use patterns and is based on maps (biophysical factors) and census data (land use data and socio-economic factors). In the study of de Koning et al. (1998) significant driving forces ($P < 0.05$) were selected using stepwise regression from a set of 23 potential drivers (Table 1) for four selected land use types at three different aggregation levels. The highest resolution in the data set are 5×5 min (9.25×9.25 km) cells. By averaging data of 2×2 cells, 3×3 cells, up to 6×6 cells, higher aggregation levels were created. The different aggregation levels were used to simulate different scales. The area of Ecuador was spatially stratified on the basis of altitude defining three main eco-regions called: Coast, Andes and Amazon (Fig. 2).

Table 1
Variables included in the stepwise regression analysis (de Koning et al., 1998)

Variable	Unit
Land use data	
Percentage permanent crops	–
Percentage temporary crops	–
Percentage grassland	–
Percentage natural vegetation	–
Land use drivers	
Percentage soils with texture class 1 (<35% clay)	–
Percentage soils with texture class 2 (35–55% clay)	–
Percentage soils with texture class 3 (>55% clay)	–
Percentage soils with slope class 1 (<8% clay)	–
Percentage soils with slope class 2 (8–16% clay)	–
Percentage soils with slope class 1 (>16% clay)	–
Percentage low fertility soils	–
Percentage medium fertility soils	–
Percentage high fertility soils	–
Altitude	masl
Total annual precipitation	mm
Distance to nearest urban centre	km
Distance to nearest road	km
Distance to nearest river	km
Total population per surface area	km^{-2}
Rural population per surface area	km^{-2}
Urban population per surface area	km^{-2}
Percentage of total population living in poverty	–
Percentage of rural population living in poverty	–
Percentage of total population that is illiterate	–
Percentage of rural population that is illiterate	–
Percentage of total population working in agriculture	–
Percentage of rural population working in agriculture	–

2.2. Detection of spatial autocorrelation

Spatial structures, like spatial dependency, can be described through structure functions. The most commonly used structure functions are correlograms, variograms and periodograms. They can be used to quantify the spatial dependency per distance class, a so called lag. In correlograms autocorrelation values are plotted against distance classes. This can be computed for both univariate (Moran's I (Moran, 1950) or Geary's c (Geary, 1954)) and multivariate data (Mantel correlogram) (Legendre and Legendre, 1998). This study uses correlograms of the Moran's I (Eq. (1)). Correlograms are preferable over, for example, semi-variograms for two reasons. First, the significance of the correlation coefficient (in this case the Moran's I) can be tested and second, correlograms are standardised, so different cases can be compared (Legendre and Fortin, 1989; Meisel and Turner, 1998).

$$\text{Moran's } I: \quad \text{for } h \neq i$$
$$I(d) = \frac{(1/W)\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \bar{y})(y_i - \bar{y})}{(1/n)\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (1)$$

In Eq. (1) the y_h's and y_i's are the values of the observed variable at sites h and i. The values of w_{hi} are the weights. The weights w_{ij} are written in a $(n \times n)$ weight matrix W. W is the sum of the weights w_{hi} for a given distance class (Legendre and Legendre, 1998).

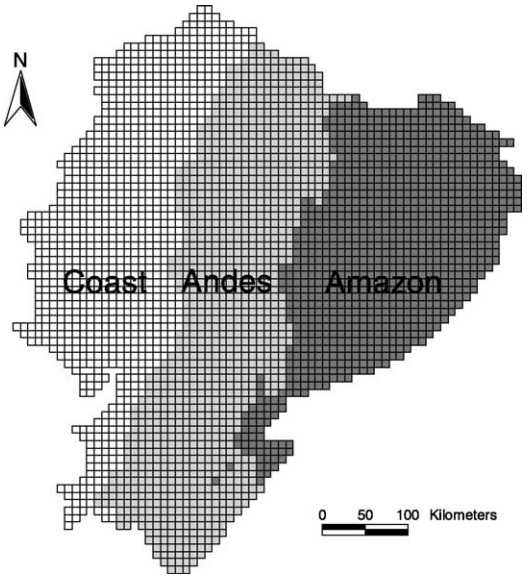

Fig. 2. Location of the three eco-regions in Ecuador (for the 1×1 grid size).

The weight matrix depicts the relation between an element and its surrounding elements. Weight can be based, for example, on contiguity relations or distance. In a weight matrix based on contiguity, a value unequal to zero in the matrix represents pairs of elements with a certain contiguity relation and a zero represents pairs without contiguity relation. Two examples of contiguity relations are rook case and queen case. The first takes only full neighbours into account and the latter all eight surrounding cells. The complete matrices contain the contiguity relations of all pairs of points. Besides this contiguity principle it is also imaginable to make weight matrices based on geographic distances, like inverse distance. To compute the outcome of spatial regression models the spatial weight matrix should be row-standardised, instead of equal weights, to yield a meaningful interpretation of the results (Anselin, 1992). In a row-standardised matrix the values are represented as fractions to accomplish that sum of all values in a row of the weight matrix equals one. The row-standardised matrices are also used to calculate the Moran's *I*.

The value of Moran's *I* generally varies between 1 and -1, although values lower than -1 or higher than $+1$ may occasionally be obtained. Positive autocorrelation in the data translates into positive values of *I*; negative autocorrelation produces negative values. No autocorrelation results in a value close to zero (Legendre and Legendre, 1998).

Spatial autocorrelation can be analysed on unmodified data or on the residuals of a regression analysis. If autocorrelation is detected on the regression residuals, this can imply that the regression model should have an autoregressive structure or that non-linear relationships between the dependent and the independent variables (trend surface analyses) are present or that one or more important regressor variables are missing (Cliff and Ord, 1981; Griffith, 1992; Miron, 1986; Long, 1998).

2.3. Analysis of the presence of spatial autocorrelation

Generally spoken, two main causes of spatial structure exist. First, spatial structure can be caused by a dependence of *y* upon one or several variables *x* which are spatially structured. The pattern is a reaction to another variable. This is also called a trend or gradient. Second, spatial structure can appear when the process that has produced the values of *y* is spatial in itself, and reflects interaction between sites (Legendre and Legendre, 1998; Cliff and Ord, 1981). Of course, in reality both reaction and interaction might affect the spatial structure.

It is useful to assess whether the dominant effects are caused by reaction to external forces or by interaction between neighbouring individuals. When reaction is the major influence, a regression model is appropriate, whereas interactive effects suggest the need for a model with a spatially dependent covariance structure. In order to decide for the most appropriate model, it is useful to examine the residuals of a regression model for spatial dependence. When the presence of spatial autocorrelation has been demonstrated, a possibility to deal with it is to draw a random sample that is not autocorrelated and then apply conventional statistical tests (Verburg and Chen, 2000). In fact, this is a loss of information.

2.4. Spatial autoregressive models

The most general formulation of a spatial autoregressive model is Eq. (2) (Anselin, 1988; LeSage, 1999). From the general model we can derive specific

models by imposing restrictions (for interpretation of the different models see Anselin, 2002). Setting $X = 0$ and $W_2 = 0$ produces a first-order spatial autoregressive model, explaining variation in y as a linear combination of contiguous or neighbouring units with no other explanatory variables. Setting $W_2 = 0$ produces a mixed regressive–spatial autoregressive model. This model has additional explanatory variables in the matrix X to explain variation in y over the spatial sample of observations. This model is also called the simultaneous model (Anselin, 1988) or simultaneous spatial autoregression (Kaluzny et al., 1997) and is originally based on the work of Whittle (1954). Setting $W_1 = 0$ results in a regression model with spatial autocorrelation in the disturbances. Anselin (2002) describes this model as a standard regression model with spatially filtered variables. A model known as the spatial Durban model contains a spatial lag in both the dependent variable and the independent variables.

$$y = \rho W_1 y + X\beta + u \qquad u = \lambda W_2 u + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2 I_n) \tag{2}$$

In Eq. (2), y contains a $n \times 1$ vector of cross-sectional dependent variables, X represents an $n \times k$ matrix of explanatory variables, and W_1 and W_2 are known $n \times n$ spatial weight matrices. Rho (ρ) is a coefficient on the spatially lagged dependent variable and λ is a coefficient on the spatially correlated errors (LeSage, 1999). Beta (β) is $k \times 1$ vector with linear regression coefficients like in a standard linear regression model. The error term (ε) is an $n \times 1$ vector of independent identically normally distributed variables with zero mean and variance σ^2.

In case of a row-standardised W_1, the spatial part of the mixed regressive autoregressive model functions as an extra variable equal to the (weighted) mean of observations from contiguous cells. If we assume spatial dependence between the observations in the data set y, some part of the total variation in y across the spatial sample is explained by each observation's dependence on its neighbours. "The parameter ρ would reflect that in the typical sense of regression" (LeSage, 1999).

In the case study presented here the mixed regressive–spatial autoregressive model is used. To estimate the parameters of this model Maximum Likelihood estimation is applied, since OLS estimation for spatial autoregressive models is biased (Anselin, 1988).

2.5. Measures of fit in spatial models

In the presence of spatial autocorrelation there is not much meaning to giving each observation equal weight in a measure of fit. Thus, the traditional R^2 measure of fit, based on the decomposition of the total sum of squares into explained and residual sums of squares, is not applicable to the spatial lag model. Instead, a number of so-called pseudo R^2 measures can be computed. The pseudo R^2 that is used in this study is defined as the ratio of the variance of the predicted values over the variance of the observed values for the dependent variable. In the standard regression model, this variance ratio is equivalent to the R^2, but in the spatial lag model it is not (Anselin, 1992). The pseudo R^2 is a general guide to assess fit, but does not have the type of meaning R^2 has in the standard regression model (SpaceStat support, 2000). So, the traditional R^2 and the pseudo R^2 cannot be compared, but it is possible to compare the pseudo R^2 of different spatial models.

The proper measures for goodness-of-fit for the spatial model are based on the likelihood function. These include the value of the maximised log likelihood (LIK), the Akaike Information Criterion (AIC) and the Schwartz Criterion (SC). The likelihood-based measures are directly comparable with those achieved for the standard regression model. The model with the highest LIK, or with the lowest AIC or SC has the best goodness-of-fit (Anselin, 1992). The LIK is not a standardised indicator like R^2, and therefore cannot be interpreted as an absolute value.

2.6. Analysis

We analysed the data in two different ways. (1) We tested the data for spatial autocorrelation at multiple resolutions; (2) we constructed spatial regression models that incorporate spatial autocorrelation at multiple resolutions.

To calculate the Moran's I statistic, SpaceStat (Anselin, 1998, 1999) is used, in combination with an extension of SpaceStat for Arcview (Anselin and Smirnov, 1999). The weight matrices are calculated based on distance (Anselin, 1999). All pairs of centroids (distances between the centres of gravity) of the grid cells are classified into lags. The lag size chosen is 10 km, resulting in lag distances of 0–10, 10–20 and 20–30 km, etc. Within a lag the weights

are equal. For higher aggregation levels the same lag distribution was used. In those cases the first few lags can be empty since the cell size exceeds the boundaries of the first lags. By using the same lag distribution it is possible to compare the Moran's *I* of different aggregation levels for exactly the same lag.

Correlograms are constructed and compared for all combinations of the three eco-regions, three aggregation levels and the four land use types. The residuals of the regression models constructed by de Koning et al. (1998) are evaluated as well.

To calculate mixed autoregressive-regressive models SpaceStat (Anselin, 1998, 1999) is used as well. For every combination of land use, aggregation level and eco-region a mixed regressive autoregressive model is calculated. First, a spatial model is constructed using the independent variables that were selected by de Koning et al. (1998) in combination with a spatial part using a weight matrix based on contiguity relations of the first lag. Other weight matrices were tested, but the first lag weight matrix proved to be most powerful. By applying this method some of the originally selected variables turned out to be insignificant. Insignificant variables are removed one by one until a model with solely significant variables remained.

3. Results

3.1. Results of the testing for spatial autocorrelation

3.1.1. Correlograms of the 1 × 1 grid level

In Fig. 3 correlograms of the Moran's *I* (distance as the upper distance of a lag) of the surface percentage within cells of the four land use types for the 1 × 1 aggregation level are presented for the three eco-regions. All three land use types show positive spatial autocorrelation in all eco-regions, which decreases gradually with distance. The correlograms of eco-regions Coast and Amazon show small differences in Moran's *I* between land use types, while in eco-region Andes somewhat larger differences occur. Fig. 4 shows the correlograms of the selected driving factors for permanent crops (1 × 1 grid) in Coast and Andes. As the land use data, all driving factors show positive spatial autocorrelation, which decreases gradually with distance.

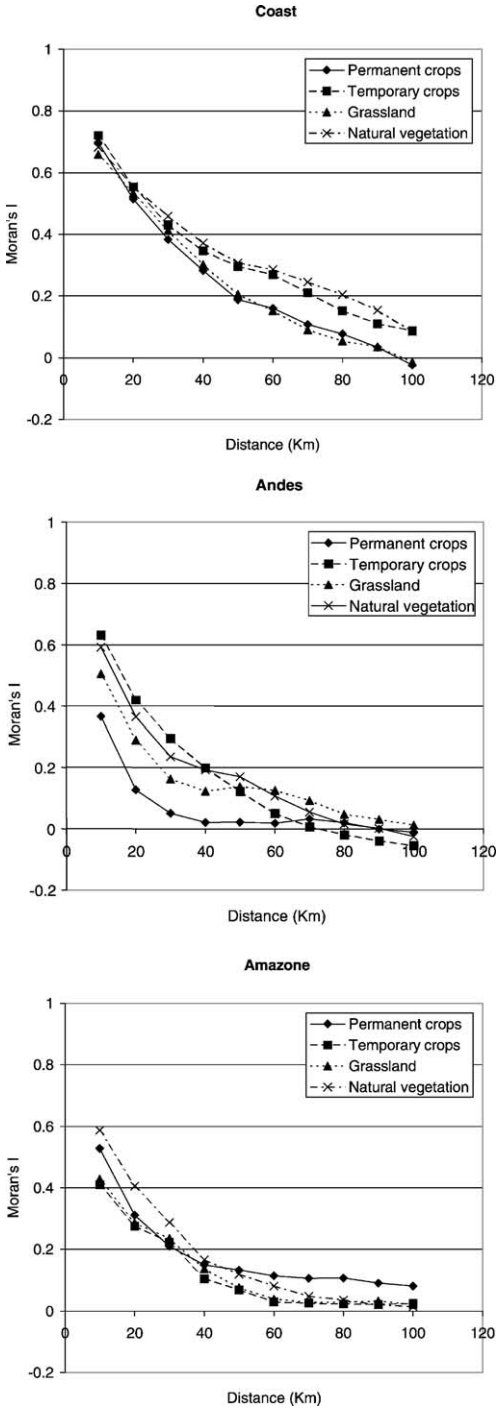

Fig. 3. Correlograms with Moran's *I* comparing different land use types in three eco-regions, grid size 1 × 1.

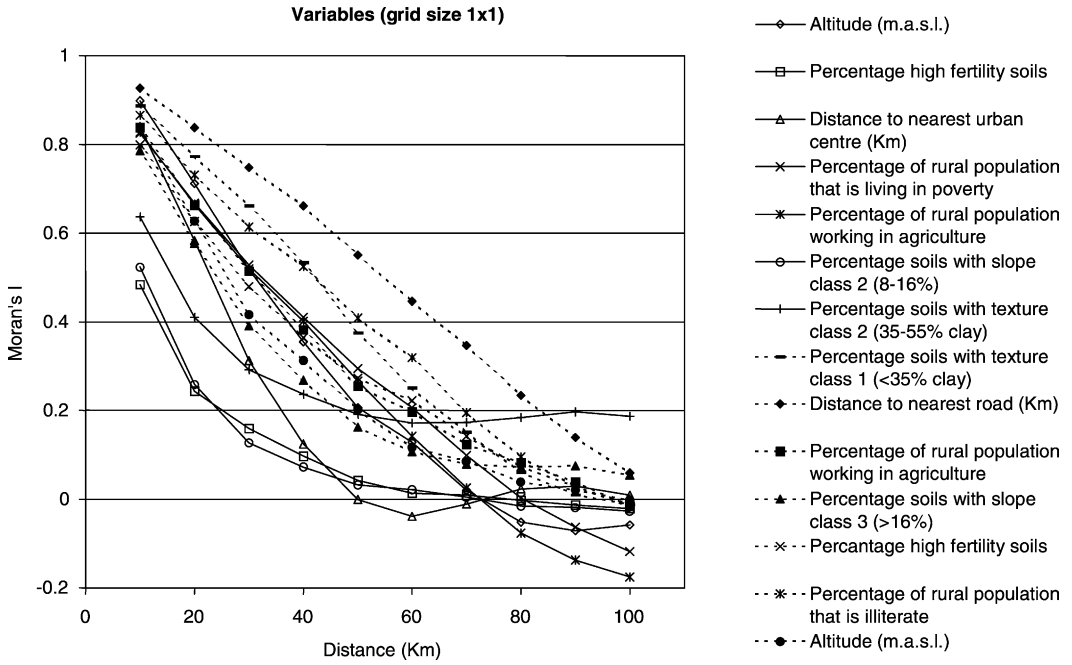

Fig. 4. Spatial autocorrelation in the variables used in the standard linear model of permanent crops, grid size 1 × 1, of Coast (full line) and Andes (dashed).

3.1.2. Correlograms of the three different aggregation levels

The effect of spatial scale on spatial autocorrelation for the different land use types is analysed by comparing three aggregation levels. The correlograms of the different aggregation levels (two examples are presented in Fig. 5) show clear differences in Moran's *I* between the aggregation levels. At higher aggregation levels, temporary crops in the coastal area have higher Moran's *I* indices (Fig. 5). In the example for permanent crops in the Andes (Fig. 5) the same phenomenon can be observed, but this disappears for distances over 40 km, because over 40 km the patch size is smaller than the cell size and the values of Moran's *I* are low and insignificant ($P < 0.001$). At the 5 × 5 aggregation level a different pattern appears with the highest Moran's *I* at a distance of 90 km, but this value is statistically not significant ($P < 0.001$).

3.1.3. Spatial autocorrelation in the residuals

The residuals of the standard linear regression models for different land use types constructed by de Koning et al. (1998) are tested for spatial autocorrelation. Fig. 6 shows the Moran's *I* of the residuals together with the Moran's *I* of the land use types as presented in Fig. 3 "Coast". The spatial autocorrelation in the residuals is less than in the original data, though still significant autocorrelation is present.

Spatial autocorrelation in the first lag of the residuals (Table 2), which can be used as an indicator to use the standard linear model, of aggregation level 1 is significant for all land use types in both eco-regions.

Table 2
Moran's *I* of the first lag

Coast	1 × 1	3 × 3	5 × 5
Residuals of permanent crops	0.5399	0.2160	**0.0035**
Residuals of temporary crops	0.5232	0.2617	**0.0129**
Residuals of grassland	0.5179	0.5005	0.2867
Residuals of natural vegetation	0.3823	0.3358	**0.1831**
Andes			
Residuals of permanent crops	0.2653	0.1744	**0.0125**
Residuals of temporary crops	0.3482	0.2396	**−0.0028**
Residuals of grassland	0.3369	**0.0908**	**0.2059**
Residuals of natural vegetation	0.3571	**0.0717**	**0.0743**

Numbers in bold are not significant ($P < 0.05$).

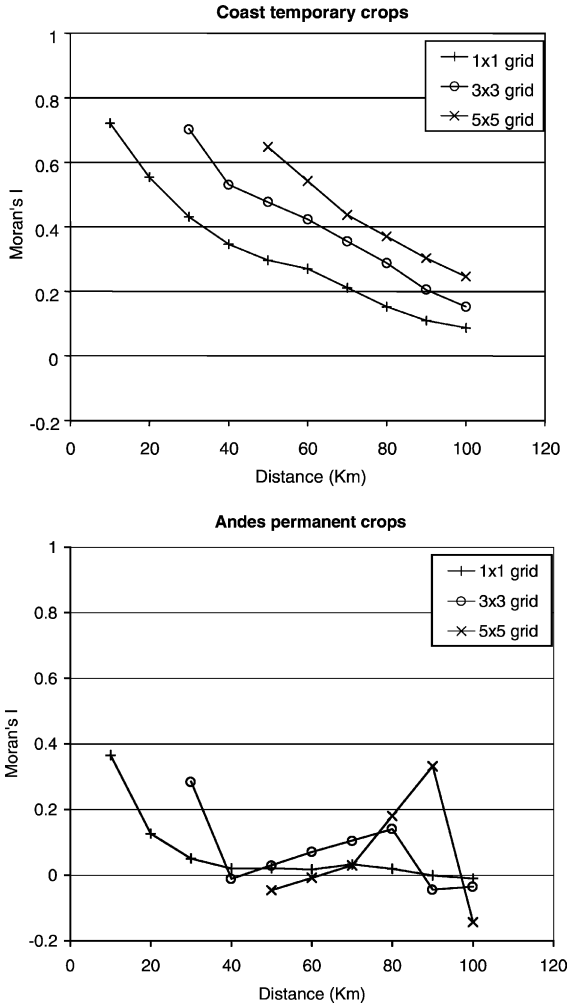

Fig. 5. Comparison of the Moran's I for different aggregation levels.

Coast shows higher autocorrelation than Andes. The Moran's I of the residuals for the 5×5 grid is not significant for the first lag for any of the cases, except for grassland in eco-region coast. Spatial autocorrelation in the residuals extents to 40–50 km (Fig. 6) and a cell of the 5×5 grid is of the same extent, therefore autocorrelation at this aggregation level disappears.

3.2. Results spatial autoregressive models

In this paragraph some examples are presented to illustrate the difference between the standard (multiple) linear model based on OLS and the mixed regressive autoregressive model. Table 3 shows the results for eco-region coast, grid size 1×1, for permanent crops.

Table 3, (A) shows the original standard linear model, with variables that were selected by de Koning et al. (1998). The output contains the measure of fit (R^2), coefficient estimate, standard error, t-test value and associated probability. The LIK is given for comparison with the spatial models. Applying the mixed autoregressive regressive model (Table 3, (B)), using the same variables as the standard linear model, results, as expected, in smaller values of the estimated regression coefficients. This is because a part of the prediction is now based on the autoregressive term. The significance of the parameters also decreases and one variable "percentage of rural population that is illiterate" is not longer significant ($P < 0.05$). The LIK of the mixed autoregressive regressive model is higher than the LIK of the standard linear model indicating a better goodness-of-fit.

From the mixed autoregressive regressive model 1, the insignificant variable is dropped to construct model 2. This results in a model with only significant ($P < 0.05$) variable estimates (Table 3, (C)). The pseudo R^2 drops from 0.5412 to 0.5394 and the LIK from -3551.6 to -3552.3. So, the goodness-of-fit drops only slightly by excluding this variable.

In the mixed autoregressive regressive model the autocorrelation in the residuals disappeared (Fig. 7). The difference in spatial autocorrelation between the two different spatial models is negligible. The maps of the residuals of the standard linear model and the mixed regressive autoregressive model 2 (Fig. 8) give an impression of the functioning of the spatial model. The standard linear model (left) shows large residuals with a clear pattern (spatial autocorrelation). The spatial model (middle) shows small residuals with no clear pattern (no spatial autocorrelation). A total of 17% of the residuals is switched from positive to negative or the other way around. Overall, the residuals are considerably lower when using the spatial model.

The residuals of the standard linear model for the case of grassland at the 3×3 aggregation level in the Andes do not show significant autocorrelation (Moran's $I = 0.0908$, see Table 2). However, applying a spatial model results in a significant spatial component ($P < 0.05$). So, the Moran's I gives no ground to apply a spatial model, but the spatial model

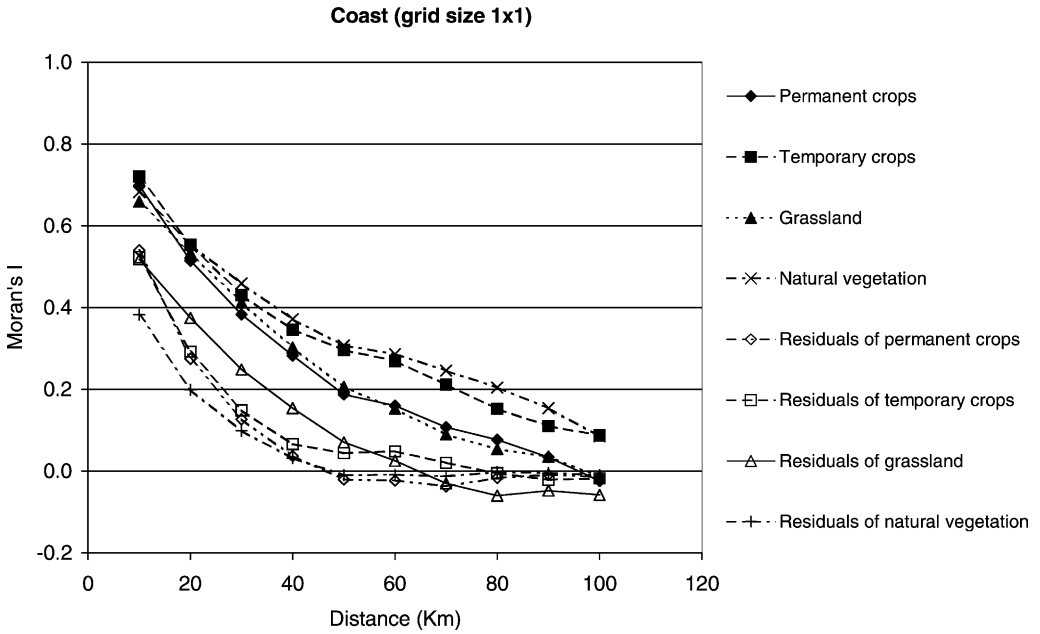

Fig. 6. Correlograms with Moran's I comparing different land use types and their residuals after standard linear regression (eco-regions Coast, grid size 1×1).

itself yields a significant spatial component. The two criteria can both be used to make a decision to use a spatial model or not. In another case with insignificant spatial autocorrelation (Moran's $I = -0.0028$) in the first lag (temporary crops, 5×5, Andes), it is not necessary to consider a spatial model, because the coefficient of the spatial part is not significant (see Table 4).

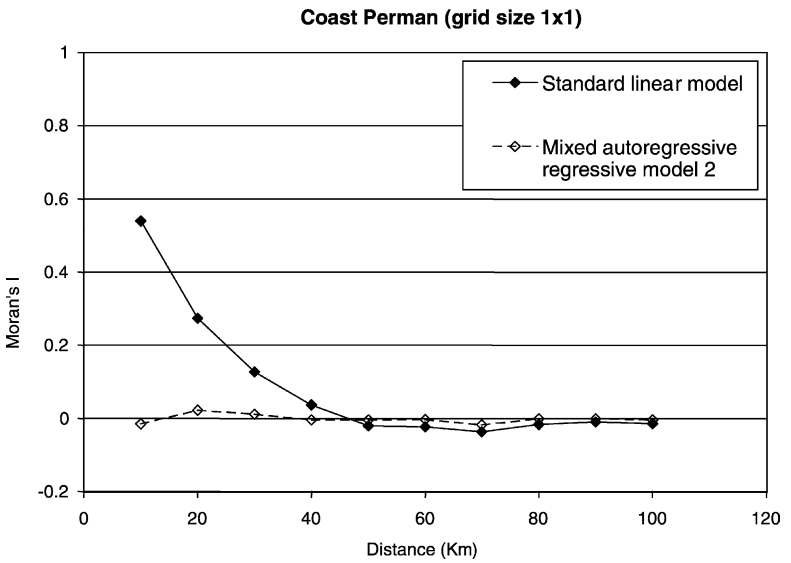

Fig. 7. Spatial autocorrelation in the residuals of two models (calculated in Table 2).

Table 3
Calculated model parameters of three different models for permanent crops in the coastal area with grid size 1 × 1

Variable	Coefficient	S.D.	*t*-value	Probability
(A) Linear model				
Constant	−1.971	2.534	−0.778	0.437
Percentage soils with texture class 1 (<35% clay)	19.997	1.409	14.188	0
Distance to nearest road (km)	−0.338	0.037	−9.127	0
Percentage of rural population working in agriculture	0.885	0.128	6.907	0
Percentage soils with slope class 3 (>16%)	10.849	1.563	6.939	0
Percentage high fertility soils	6.951	1.557	4.465	0
Percentage of rural population that is illiterate	−0.541	0.165	−3.272	0.001
Altitude (masl)	−0.024	0.003	−8.611	0
R^2	0.325			
LIK	−3821.2			
Variable	Coefficient	S.D.	*z*-value	Probability
(B) Mixed autoregressive regressive model 1				
ρ	0.731	0.024	30.441	0
Constant	−0.202	1.728	−0.117	0.907
Percentage soils with texture class 1 (<35% clay)	6.608	1.050	6.293	0
Distance to nearest road (km)	−0.110	0.026	−4.178	0
Percentage of rural population working in agriculture	0.215	0.089	2.419	0.016
Percentage soils with slope class 3 (>16%)	3.991	1.083	3.685	0
Percentage high fertility soils	2.225	1.073	2.074	0.038
Percentage of rural population that is illiterate	−0.141	0.113	−1.252	0.211
Altitude (masl)	−0.010	0.002	−4.993	0
Pseudo R^2	0.541			
LIK	−3551.6			
(C) Mixed autoregressive regressive model 2				
ρ	0.734	0.024	30.755	0
Constant	−0.926	1.624	−0.571	0.568
Percentage soils with texture class 1 (<35% clay)	6.552	1.050	6.240	0
Distance to nearest road (km)	−0.121	0.025	−4.851	0
Percentage of rural population working in agriculture	0.176	0.083	2.108	0.035
Percentage soils with slope class 3 (>16%)	3.691	1.059	3.485	0
Percentage high fertility soils	2.218	1.073	2.068	0.039
Altitude (masl)	−0.009	0.002	−4.824	0
Pseudo R^2	0.539			
LIK	−3552.3			

Table 4 lists some output of models for all possible combinations following the same procedure as before. Again it is illustrated that spatial autocorrelation in most cases is lower in eco-region Andes than in the Coast area. In general, ρ's decrease with aggregation level. Only a few ρ's are not significant. If the significance of the spatial part of the model (ρ) is taken as a criterion whether or not to apply a spatial model, 20 out of 24 models would qualify for applying a spatial model. The models of 5 × 5 permanent crops/coast, permanent crops/Andes and temporary crops/Andes do not have a significant ρ. Those are the same cases that have a patch size smaller then the cell size.

If the weight matrix is row-standardised ρ can be interpreted as the percentage of the prediction that is predicted with the spatial part. Looking at the ρ's in Table 4 it is clear that a large part, 21–74%, of the prediction is due to the spatial part. The other variables, which are considered to be independent driving factors, make up the remaining part.

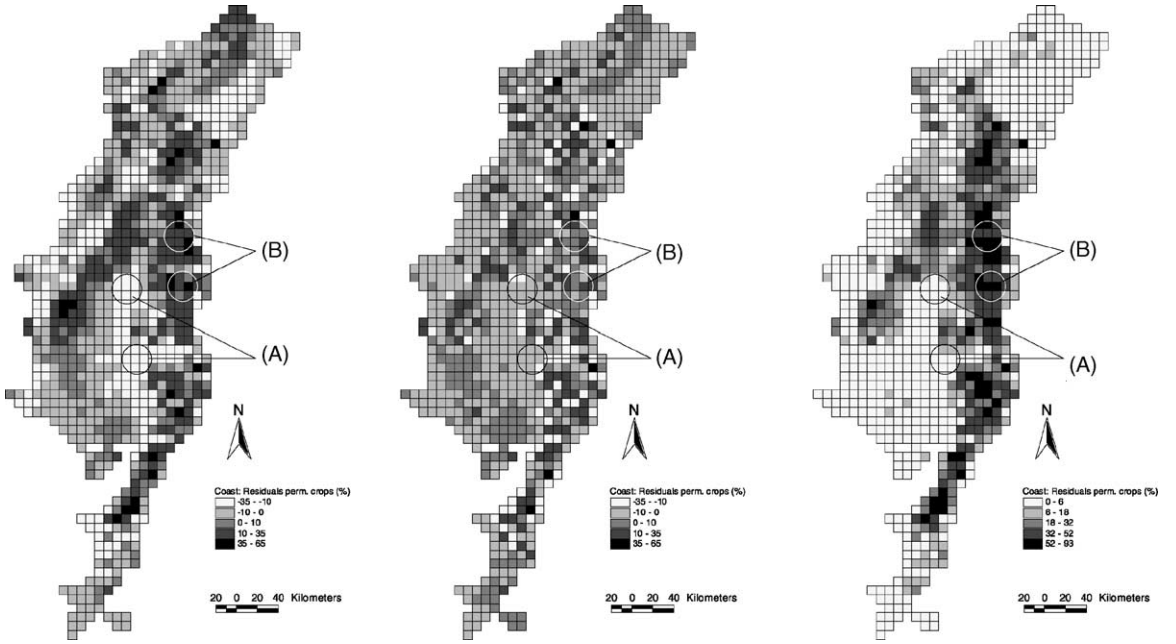

Fig. 8. Residuals of the standard linear model (left) and the mixed autoregressive regressive model 2 (middle) and the original values (right).

Table 4
Summary of spatial models for combinations of two different eco-regions, four land use types and three aggregation levels

Eco-region	Land use	Aggregation level	ρ	Number of variables (number of variables in the standard linear model)
Coast	Permanent crops	1	0.734	6 (7)
Coast	Temporary crops	1	0.736	6 (7)
Coast	Grassland	1	0.679	6 (7)
Coast	Natural vegetation	1	0.580	5 (7)
Andes	Permanent crops	1	0.478	4 (7)
Andes	Temporary crops	1	0.602	6 (7)
Andes	Grassland	1	0.555	4 (7)
Andes	Natural vegetation	1	0.590	7 (7)
Coast	Permanent crops	3	0.470	7 (7)
Coast	Temporary crops	3	0.568	2 (5)
Coast	Grassland	3	0.708	2 (5)
Coast	Natural vegetation	3	0.520	4 (6)
Andes	Permanent crops	3	0.292	5 (5)
Andes	Temporary crops	3	0.312	5 (5)
Andes	Grassland	3	0.286	3 (5)
Andes	Natural vegetation	3	0.186*	6 (6)
Coast	Permanent crops	5	0.079*	4 (4)
Coast	Temporary crops	5	0.212	5 (5)
Coast	Grassland	5	0.449	2 (3)
Coast	Natural vegetation	5	0.328	3 (3)
Andes	Permanent crops	5	−0.094*	3 (3)
Andes	Temporary crops	5	0.076*	4 (4)
Andes	Grassland	5	0.415	6 (6)
Andes	Natural vegetation	5	0.296	6 (7)

* Not significant ($P > 0.05$).

4. Discussion

The detected spatial autocorrelation can be used to describe and compare the spatial structure of the data. The distance at which the values of the Moran's *I* in the correlogram approach zero can be interpreted as average patch size. A comparison of patch sizes of the land use types between different eco-regions shows that Coast has patch sizes of 80–100 km, Amazon 60–100 km and Andes 40–80 km. The values of the Moran's *I* are clearly different for the three eco-regions. This means that the regions have different patterns and different spatial characteristics and that it is relevant to make the stratification for the three regions. In fact, any stratification that can be made with prior knowledge has to be made on the base of the assumption that the differences $(y_h - y_i)$ for any distance d must have zero mean and finite variance (identically distributed) over the study area, independently of the location where the differences are calculated (Legendre and Legendre, 1998). The differences between eco-regions exist in all aggregation levels. Within an eco-region, differences in Moran's *I* between land use types are less pronounced. An exception is the patch size of permanent crops in the Andes, which is clearly smaller than the patch size of the other land use types, indicating a different spatial structure.

The spatial structures of the land use types show similar characteristics as its potential driving factors. In both the land use data and the data of the driving factors eco-region Coast has high values of Moran's *I* with large patch size and Andes has moderate values with smaller patch size. So, the characteristics of the driving forces are reflected in the land use type that is explained by these variables. This is an indication that the land use is, at least partly, a reaction upon its driving factors.

According to de Koning et al. (1998) the occurrence of patterns in land use/cover can *disappear or emerge* going from one scale to the other. For example, it is possible that a certain data set has a lot of variability at a very detailed scale (low aggregation level), in which a certain pattern exists that can be seen and detected by aggregating the data. This would increase spatial autocorrelation with increasing spatial scale. On the other hand, it is possible that the aggregation level exceeds the level of the pattern. In that case no spatial autocorrelation will be detected.

In the results of this study higher aggregation levels show higher Moran's *I*. There are various interpretations possible to explain this: (1) At the low aggregation level the pattern is too noisy and through averaging the data is smoothed and a larger scale pattern becomes clear. (2) Averaging the data is a linear operation. However, at the lowest aggregation level the slope of the Moran's *I* tends to decrease with increasing distance, which is in fact a non-linear relation. So, the mean Moran's *I* of two points on the fine scale graph will always lie on the concave side along the original graph when the average is calculated. The autocorrelation of the values of aggregated cells will, therefore, be higher than the average autocorrelation of those cells. This effect increases with higher aggregation levels (Rastetter et al., 1992).

The residuals of the standard linear regression are less autocorrelated than the original data. So, the driving factors used in the regression equation capture part of the pattern, and land use is, at least partly, a reaction upon its spatially autocorrelated driving factors. Also the extent of the autocorrelation decreased from 100 to 50 km. However, there is still significant autocorrelation present in the residuals, which indicates that the regression equations are not sufficient to explain all spatial patterns. This can be the result of spatial interactions, which can be caused by, for example, social relations like imitation or market effects like the clustering of producers to gain benefits from production at a larger scale (economies of scale).

The visual presentation of the residuals of the models in Fig. 8 provides a clear insight in the differences between the models. If we consider location A in the original values (Fig. 8, right) and in the standard linear model (Fig. 8, left) low original values result in large negative residuals. The regression model predicts, based on the driving factors, a higher land use then is occurring at present. Near B (Fig. 8) the situation is the other way around. In other words, the predicted values have a smoother character (closer to the mean) than the original values. With independent data residuals would be randomly distributed.

Comparing the residuals of the spatial model (Fig. 8, centre) with the original values (Fig. 8, right) it is clear that extreme residuals (both positive and negative) of the spatial model occur only at places where original

values are high (location B). But, the residuals are randomly distributed and do not show spatial autocorrelation anymore. This is what would be expected from a regression analysis. The residuals are the outliers of the observed situation. Using the spatial model the prediction of land use area in a cell depends partly on the area of that land use in neighbouring cells. The model is not just based on independent variables (drivers), but also on the pattern of land use itself.

Basically, the test for spatial autoregression in the residuals can be used to decide to use a standard linear regression model or a model that accounts for autocorrelation (Legendre and Legendre, 1998). But, as shown in this study, there will be cases in which the application of a spatial model will lead to a significant ρ (thus a significant spatial part), while the Moran's I does not show any spatial autocorrelation.

By using spatial models, a part of the variance is explained by neighbouring values. This can be seen as unsatisfactory, because the explanation of a variable using the value of the neighbouring cell seems trivial. The aim of most land use/cover studies is to identify independent factors or proxies that explain land use (driving factors) and to make a good prediction based on those explaining factors. However, including a spatial part in the model is a way to deal with spatial interactions that cannot be captured otherwise.

Besides spatial interactions, trends or the omission of regression variables could cause the importance of the spatial part (as shown in Table 4). In that case, other driving factors, which are not yet taken into account, play a role in the occurrence of land use. If these variables have a spatial pattern, this is detected in the residuals as autocorrelation. Although omission of explaining variables and spatial interaction cannot be clearly distinguished from each other through analysis of residuals, the application a spatial model is recommended, because it is a statistically sound model for spatially dependent data. Another way to improve the model is to include new variables or transform variables until spatial autocorrelation in the residuals is not present anymore.

5. Conclusion

Moran's I can be used to identify and quantify spatial dependency (spatial autocorrelation) in spatially explicit land use studies. Positive spatial autocorrelation was detected in the land use data and in the driving factors of the case study of Ecuador. The residuals of the standard linear model were also autocorrelated, which indicates that the standard multiple linear regression model cannot capture all spatial autocorrelation in the land use data.

Spatial autocorrelation can be very different at different aggregation levels, even though they are based on the same data. Theoretically, spatial autocorrelation can increase and decrease going from one aggregation level to another. In this study, the Moran's I increased with higher aggregation levels. This is the result of the smoothing character of averaging data and by the fact that the relation between Moran's I and distance is not linear.

Land use models that select drivers of land use patterns through regression, often overestimate their role in the presence of spatial autocorrelation. Spatial autoregressive models are suited to deal with spatial data and provide a solution that is statistically sound. This study demonstrated that different conclusions can be drawn from the same data using spatial or conventional statistics, especially with regard to the goodness-of-fit, the significance of regression coefficients and the relevance of the land use drivers. In the standard linear models used by de Koning et al. (1998) the residuals of the models still showed spatial autocorrelation, especially in low aggregation levels. To overcome this, mixed regressive–spatial autoregressive models were applied. Their residuals have no spatial autocorrelation and a better goodness-of-fit. In most cases one or two of the variables of the standard linear model turned out to be insignificant and consequently were removed from the model.

Within any spatially explicit study of land use change spatial autocorrelation will occur, depending on the spatial scale. Therefore, identification and quantification of spatial autocorrelation in land use studies using spatial data should be a standard procedure. Furthermore, the residuals of standard linear regression analyses of spatially explicit land use studies should be tested for spatial autocorrelation. If the spatial autocorrelation in the residuals cannot be excluded by adding regression variables or incorporating a trend, a spatial regression model is most appropriate.

The findings of the spatial analysis with correlograms and the application of spatial models can lead

to better insight in the data and processes that determine land use (change) and can give cause to unravel its complexity even more from a spatial perspective.

Acknowledgements

The authors would like to acknowledge the Foundation for the Advancement of Tropical Research (WOTRO) of The Netherlands Organisation for Scientific research (NWO) for the funding of the research on which this paper is based.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht, p. 284.

Anselin, L., 1992. SpaceStat tutorial. Regional Research Institute. West Virginia University, Morgantown, WV, p. 263.

Anselin, L., 1998. SpaceStat version V1.90 R26 (distributed by BioMedware, Inc.).

Anselin, L., 1999. Spatial Data Analysis with SpaceStat and ArcView, Workbook, 3rd ed. (draft). p. 90.

Anselin, L., 2002. Under the hood: issues in the specification and interpretation of spatial regression models. Agric. Econ. 27 (3), 247–267.

Anselin, L., Griffith, A.D., 1988. Do spatial effects really matter in regression analysis? Papers Reg. Sci. Assoc. 65, 11–34.

Anselin, L., Smirnov, O., 1999. The SpaceStat Extension for ArcView. Bruton Center, University of Texas, Dallas, Richardson, TX.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London, p. 266.

Geary, R.C., 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician 5, 115–145.

Gould, P.R., 1970. Is statistics inferens the geographical name for a wild goose? Econ. Geography 46, 439–448.

Griffith, D.A., 1992. What is autocorrelation? Reflections on the past 25 years of spatial statistics. l'Espace Géographique 21, 265–280.

Kaimowitz, D., Angelsen, A., 1998. Economic Models of Tropical Deforestation: A Review. Center for International Forestry Research (CIFOR), Bogor, p. 139.

Kaluzny, S.P., Vega, S.C., Cardoso, T.P., Shelly. A.A., 1997. S-Plus Spatial Stats. User's Manual for Windows and Unix. Springer, New York.

de Koning, G.H.J., Veldkamp, A., Fresco, L.O., 1998. Land use in Ecuador: a statistical analysis at different aggregation levels. Agric. Ecosyst. Environ. 70, 231–247.

Lambin, E.F., 1994. Modelling Deforestation Processes: A Review. Trees, Tropical Ecosystem Environment Observation by Satellites, Research Report No. 1. European Commission Joint Research Centre/European Space Agency, p. 113.

Lambin, E.F., Rounsevell, M.D.A., Geist, H.J., 2000. Are agricultural land-use models able to predict changes in land-use intensity? Agric. Ecosyst. Environ. 82, 321–331.

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetation 80, 107–138.

Legendre, P., Legendre L., 1998. Numerical Ecology. Developments in Environmental Modelling 20. Elsevier, Amsterdam, p. 853.

LeSage, J.P., 1999. The Theory and Practice of Spatial Econometrics. Department of Economics, University of Toledo, p. 296.

Long, D.S., 1998. Spatial autoregression modeling of site-specific wheat yield. Geoderma 85, 181–197.

Meisel, J.E., Turner, M.G., 1998. Scale detection in real and artificial landscapes using semivariance analysis. Landscape Ecol. 13, 347–362.

Miron, J., 1986. Spatial autocorrelation in regression analysis: a beginner's guide. In: Gaile, G.L., Wilmot, C.J. (Eds.), Spatial Statistics and Models. Dordrecht, pp. 201–222.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Nelson, G.C., 2002. Introduction to the special issue in spatial analysis for agricultural economics. Agric. Econ. 27, 197–200.

Pontius, R.G., Cornell, J.D., Hall, C.A.S., 2001. Modeling the spatial pattern of land-use change with GEOMOD2: application and validation for Costa Rica. Agric. Ecosyst. Environ. 85, 191–203.

Rastetter, E.B., King, A.W., Cosby, B.J., Hornberger, G.M., O'Neill, R.V., Hobbie, J.E., 1992. Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. Ecol. Appl. 2 (1), 55–70.

Riebsame, W.E., Parton, W.J., 1994. Integrated modeling of land use and cover change. Bioscience 44, 350–357.

SpaceStat support: e-mail correspondence with support@spacestat.com, the helpdesk from SpaceStat.

Sklar, F.H., Costanza, R., 1991. The development of dynamic spatial models for landscape ecology: a review and prognosis. In: Turner, M.G., Gardner, R.H. (Eds.), Quantitative Methods in Landscape Ecology. Ecological Studies 82. Springer, Berlin, pp. 239–288.

Veldkamp, A., Lambin, E.F., 2001. Editorial: predicting land-use change. Agric. Ecosyst. Environ. 85 (1–6), 379–386.

Veldkamp, A., Kok, K., de Koning, G.H.J., Verburg, P.H., Priess, J., Bergsma, A.R., 2001. The need for multi-scale approaches in spatial specific land use change modelling. Environ. Model. Assess. 6 (2), 111–121.

Verburg, P.H., Chen, Y., 2000. Multiscale characterization of land-use patterns in China. Ecosystems 3, 369–385.

Verburg, P.H., de Koning, G.H.J., Kok, K., Veldkamp, A., Bouma, J., 1999. A spatial explicit allocation procedure for modelling the pattern of land use based upon actual land use. Ecol. Model. 116, 45–61.

Verburg, P.H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., Sharifah Mastura, S.A., 2002. Modeling the spatial dynamics of regional land use: the CLUE-S model. Environ. Manage. 30 (3), 391–405.

Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–439.