Empirically derived suitability maps to downscale aggregated land use data

Dendoncker N.,* Bogaert P., and Rounsevell M.

45th Congress of the European Regional Science Association
23-27 August 2005, Vrije Universiteit Amsterdam
Land use and Water Management in a Sustainable Network Society

Abstract

This paper aims to derive optimal representations of land use patterns based on novel statistical techniques. In order to do so, several cross-sectionnal logistic regression models are compared by studying land use change drivers in Belgium. It is shown that a purely spatial logistic regression model gives the best statistical fit. Based on this model and on an iterative procedure using Bayes' theorem, a set of land use suitability maps are developed at two time steps. These can serve as a basis for the development of a new land use allocation procedure to downscale aggregated land use data. The method's simplicity as well as its low data requirements (only land use datasets are used) makes it easily replicable allowing application over a wide geographic area. An example is shown using the CORINE land cover dataset to downscale land use scenarios for a small area in Belgium. The results from the multinomial logistic regression as well as from the iterative procedure based on Bayes' theorem are conclusive with the resulting land use maps giving appropriate representation of land use patterns despite a tendency to further aggregate existing patterns. The method also proved useful in removing potential artificial border effects, which often arise when downscaling adjacent land use units. The resulting suitability maps could be used for a variety of applications ranging from soil science to biogeographical studies.

keywords

downscaling, suitability maps, neighbourhood, logistic regression, Bayes' theorem

*Corresponding author : Nicolas Dendoncker, Université Catholique de Louvain, UCL/SC/GEOG, Place Louis Pasteur 3, 1348 Louvain-la-Neuve (dendoncker@geog.ucl.ac.be)

1 Introduction

Empirical analyses of land use patterns have previously been used for two distinct objectives. The first is to analyse a range of variables (the so-called 'drivers' of land use change) in order to identify which are important in understanding and explaining land use patterns. This is a way of linking patterns to processes responsible for their development. Studies by Serneels and Lambin (2001), Veldkamp and Fresco (1997) and Verburg et al. (2004a) have all pursued such a goal. A second objective is to obtain the statistically optimal representation of actual or future land use patterns based on appropriate suitability maps. These maps do not necessarily have strong explanatory power because their goal is to provide adequate visualization of land use patterns rather than explaining them. The resulting maps are usually created as a step towards further analysis of, for example, the effect of land use change on biogeography (Peppler-lisbach, 2003), hydrology (Sullivan et al., 1995) or soil science (Lettens et al., 2004). For these applications, there is greater concerns for the quality of the resulting map rather than for the underlying causal relationships that explain the observed land use patterns.

Verburg et al. (2004b) state that as land use is the result of multiple processes acting over different scales, there is a need for models to integrate different scales of analysis, especially if their aim is to model land use change over a wide study area. However, most existing models only consider a single scale. An exception to the rule are the regional level integrated simulation models Briassoulis (2000) which are, by definition, multi-scalar. For example, the CLUE (Conversion of Land Use and its Effects) model (Veldkamp and Fresco, 1996), the ATEAM (Advanced Terrestrial Ecosystem Analysis and Modelling) modelling framework (Rounsevell et al., 2005a) and the Environment Explorer cellular automata (CA) based model (Engelen, 2002) all have three distinct spatial levels of analysis. It seems that providing a pan European finer resolution land use change model could be useful for various purposes. For example, land use patterns have been shown to affect numerous ecological processes (Parker and Meretsky, 2004). Since land use is a decisive factor in modelling community and species distribution, it is nowadays also being integrated in ecological modelling approaches (Peppler-lisbach, 2003). Animal ecologists want to know precisely where the different changes in land use will occur. This is especially important for migratory bird species whose population dynamics can be strongly influenced by land use change over wide areas (Gauthier et al., 2005). Downscaling is also essential to better assess the land use change impacts on biodiverisity of natural areas, which depend not only on the quantity but also the spatial configuration of natural areas, determining the relative connectivity or isolation of species and habitat (Wimberly and Ohmann, 2004). Soil scientists study changes in soil organic carbon stocks which directly depend on land use (Lettens et al., 2004).

A simple land use allocation procedure disaggregating coarse resolution land use data would be useful in such studies. Therefore, the main objective of this paper is to set the basis for a statistically consistent downscaling procedure based on: 1) logistic regression models to generate maps of probability

of land use presence based on European wide raster land use datasets and 2) a methodology to update these probability maps using Bayes' theorem and the ATEAM land use change scenarios (Rounsevell et al., 2005a). The resulting maps can then be used to downscale these land use scenario datasets. Whilst the work presented here is based on an application that uses land use change scenarios, the methodology is appropriate to the downscaling of any type of aggregated land use data, e.g. for administrative units.

2 Context of the study

2.1 land use drivers, spatial autocorrelation and neighbourhoods

Broadly speaking, two main types of variables are used within empirical land use/land use change models, i.e: the non-neighbourhood based variables and the neighbourhood-based variables. While the former category is included in virtually all land use studies, only a limited number explicitly deal with neighbourhood effects and spatial autocorrelation. Neighbourhood effects, reflecting centripetal forces, are known to have a role in the spatial structuring of land use and the landscape (Verburg et al., 2004c; White and Engelen, 1993). When regression is performed on spatial data, it is likely that spatial autocorrelation will remain in the residuals (Anselin, 2002). This will occur unless the regression is performed on a non-autocorrelated data sample , e.g. Serneels and Lambin (2001) or Peppler-lisbach (2003), but this results in a loss of information. Alternatively, in spatial models, a part of the variance can be explained by neighbouring values (Overmars et al., 2003).

2.2 logistic regression in land use modelling

Binomial multiple logistic regression (MLR) has been widely used in the field of land use and land use change modelling to quantify the relationships between driving factors and land use (change) patterns and to derive land use suitability maps (Veldkamp and Fresco, 1996; Mertens and Lambin, 1997; Hilferink and Rietveld, 1998; Serneels and Lambin, 2001; Peppler-lisbach, 2003; Verburg et al., 2002). However, it is a time consuming process that requires the acquisition and treatment of a large number of datasets. These are not always available, *a fortiori* when the extent of the study area is as large as, for example, Europe. In this case, national or even regional datasets would have to be used in order to account for regional variations in the relationships between land use and its drivers. Using only neighbourhood-based variables to build suitability maps allows other variables or drivers to be ignored. Indeed, the fact that models such as CLUE (Veldkamp and Fresco, 1996), CLUE-S (Verburg et al., 2002), the Land Use Scanner (Schotten et al., 2001), and most CA type models are 'data-hungry' and are usually unable to simulate land use dynamics in areas without a land use change

history, is considered to be their main drawback (Verburg et al., 2002). In contrast, neighbourhood based variables can be derived using a single land use dataset for the whole region studied. This makes the statistical basis of the land use allocation procedure simple and easily replicable. Considering this, a hypothesis to be tested in this study is that using neighbourhood variables alone will produce the best statistical representation of land use patterns. This hypothesis will be verified through a series of binomial regression models: 1) a model including only non-neighbourhood based land use variables; 2) a mixed model including both neighbourhood based and non-neighbourhood based variables; and 3) a purely autoregressive model including solely neighbourhood-based variables. The goodness of fit of these models will be compared. If, as hypothesized, the purely spatial model gives the best fit, a multinomial logit model will be used instead of the binomial logit model. Because they treat each land use individually, binomial MLR can provide insight into the driving factors explaining land use patterns. However, multinomial logistic regression models are probably better if the aim is to obtain the best fitting probability maps of land use presence. This is because it considers each land use as a possible alternative amongst the whole set of land use categories. Indeed, the multinomial model extends the logit to more than two states. Instead of the simple (0,1) dichotomy, there are k possible states with index $i = 1, 2, ..., k$ (Cramer, 1991). In the present study, k represents the different land use categories. Multinomial logistic regression models generate consistent probability maps of land use presence; the sum of probabilities for a given pixel being equal to 1, which is not the case in binomial MLR in which probabilities are computed for each land use independently of other land uses. This allows discrimination between land uses on a non-biased basis.

Some recent work has incorporated spatial dependencies into qualitative dependent variables and discrete choice models. Augustin et al. (2001) used a spatial multinomial logit model to explore vegetation dynamics. Bhat and Guo (2003) applied this framework to the field of residential modelling while Ben-Akiva and Lerman (1985) and Mohammadian and Kanaroglou (2003) applied it to transportation planning. However, spatial multinomial logistic modelling has hardly been used in the field of land use (change) modelling. McMillen (2001) applied this framework to model land use in an urban fringe area of Chicago considering three land use classes. These included neighbourhood variables such as the characteristics of the quarter-section in which properties were located.

3 Data and methods

3.1 overview

The analysis presented here consists of three main parts:

1. Deriving the logistic regression model that gives the best statistical fit by testing several models

in a cross-sectional analysis of land use drivers in Belgium.

2. Using this model to derive baseline probability maps of land use presence.

3. Updating these probability maps with an iterative procedure based on the Bayes' theorem and on land use change scenario data.

The PELCOM (Pan European Land Use and Land Cover Monitoring) (Mücher et al., 2000) and CORINE (Coordinated Information on the European Environment) land cover map (European Commission, 1993) will serve as baseline dataset to derive suitability maps of land use presence while ATEAM scenario data (Rounsevell et al., 2005b; Erhard et al., 2005) will serve as quantitative constraints to update the probability maps.

3.2 data

3.2.1 dependent variable: land use categories

There are very few European wide land use datasets. The two main datasets are the PELCOM land cover dataset (Mücher et al., 2000) and the CORINE land cover [1] dataset (European Commission, 1993). Both were originally derived from satellite data. PELCOM is a geographical map that covers the whole of Europe at a spatial resolution of 1.1 km x 1.1 km. It was obtained from earth observation satellite images around 1996 and was validated using the CORINE land cover dataset. CORINE exists in several versions. It was first made available as a grid database at a spatial resolution of 250m, aggregated from the original vector data at a scale of 1:100000, limited to the EU15 countries and the year 1990. Despite CORINE's finer spatial resolution and generally assumed better quality over PELCOM (see e.g. Schmit et al., 2005), the PELCOM dataset was used for the cross-sectional analysis of land use drivers. This was done because a) the computation time is greatly reduced as there are over 19 times fewer PELCOM cells (24932) than CORINE cells (489401) in Belgium and b) the spatial resolution of the independent variables presented in table 1 is always coarser than that of CORINE. The following PELCOM land use/land cover classes are present in Belgium: deciduous forest, coniferous forests, mixed forests, cropland, grassland, built-up (urbanized) areas and inland water (ignored in this study). The three forest classes were aggregated into a single class, so the regression analyses were performed on four land use classes: built-up, cropland, grassland and forests. It should be noted that, at this resolution, the classification of satellite images allows only the dominant land use of each cell to be represented.

At the time the study was undertaken, no updated version of the PELCOM raster dataset was available. Therefore, a cross sectional analysis of the actual land use pattern was implemented using

[1] Although Land Use and Land Cover are two different concepts, they will be used interchangeably in this study to avoid over-complication

the 1996 version. A well-known limitation to this type of analysis is the uncertainty with respect to the causality of the supposed relationships between land use drivers and land use patterns (Verburg et al., 2004b). However, the primary goal of the present study is not to determine such relationships, but rather to find the best fit statistical model to derive land use suitability maps.

3.2.2 independent variables: non-neighbourhood based versus neighbourhood based.

Table 1 summarizes the data used in this analysis. The selection of independent variables was based on a literature review and expert judgment. Most variables are quantitative and continuous, but some are categorical or binary. The combination of different types of variables is not a problem in logistic regression modelling, however, the independence of variables can be. For example, over long time-scales, accessibility and urban expansion are closely linked, with complex cause-effect relationships. More roads can create urbanization, which in turn, can generate new roads. This excludes variables such as distance to the transport network from the analysis. The same type of reasoning makes other socio-economic factors, such as population pressure and labour availability, equally unsuitable for explaining the land use patterns within a cross-sectional analysis. In practice, many factors that are commonly used to explain land use change patterns are endogenous to the processes studied over long time scales (Verburg et al., 2004b) and this can lead to biased regression coefficients (Anselin and Kelejian, 1997; Gujarati, 1995).

A simple way of building a neighbourhood variable is to take the percentage of the land use itself within the 8 immediately surrounding cells (or less for edge located cells), that is, without taking into account the possible influence of cells that are located further away. Neighbourhood variables that take account of effects over larger distances were also derived using the enrichment factor developed by Verburg et al. (2004c) and by looking at spatial autocorrelation in the regressions' residuals.

3.2.3 land use change scenarios

The ATEAM project developed land use scenario maps of Europe for the years 2020, 2050 and 2080 (Rounsevell et al., 2005a; Erhard et al., 2005). The ATEAM land use maps give land use shares (in percent) for each cell over a 10' longitude/latitude grid. This resolution does not allow the identification of land use change effects at the landscape level and is insufficient to establish a link with local case-studies (Verburg et al., 2005). The land use classes are built-up, cropland, grassland, forest, biofuels (liquid, non-woody and woody) and surplus (i.e. abandoned land). The latter two are absent from the baseline land cover maps (i.e. PELCOM and CORINE). To match the baseline land cover categories, liquid biofuels (e.g. oilseed rape) were grouped with cropland while non-woody and woody biofuels (e.g. willow plantations) were grouped with forest.

Table 1: Independent Variables

Variable	Type	Description	Source
Biophysical Variables			
Slope	Continuous	1km resolution. The slope data set describes the maximum change in the elevations between each cell and its eight neihbours. The slope is expressed in integer degrees of slope between 0 and 90	USGS - Hydro1k derivative elevation database
Elevation	Continuous	1km resolution DEM of Europe. Units: meters	USGS - Hydro1k derivative elevation database
Soil characteristics			
Silt content of topsoil	Continuous	Silt content (%) of the 30 first centimetres of the soil corresponding to the root zone, which is important for agricultural land uses. Mean value per soil association	Aardewerk database (Van Orshoven et al., 1993) + digital version of the soil association map of Belgium (Tavernier and Marechal, 1962)
Sand content of topsoil	Continuous	Sand content (%) of the 30 first centimetres of the soil corresponding to the root zone, which is important for agricultural land uses. Mean value per soil association	Ditto
Clay content of topsoil	Continuous	Clay content (%) of the 30 first centimetres of the soil corresponding to the root zone, which is important for agricultural land uses. Mean value per soil association	Ditto
Soil depth	Binary	Soil associations with a B-textural horizon are defined as deep (1) while associations without a B-textural horizon were classified as not deep (0)	Digital version of the soil association map of Belgium (Tavernier and Marechal, 1962)
Wetness index	Ordinal	Soil associations were defined as 'dry', 'normal' and 'wet'	
Climate Variables			
Temperature	Continuous	Yearly mean temperature between 1960 and 1997 interpolated to the PELCOM grid (Mücher et al., 2000)	(Mitchell et al., 2004)
Temperature inxex	Continuous	Inxex representing the cumulated growth potential for winter cereals based on criteria of the FAO	(Mitchell et al., 2004)
Precipitation	Continuous	Yearly mean temperature between 1960 and 1997 interpolated to the PELCOM grid (Mücher et al., 2000)	(Mitchell et al., 2004)
Accessibility variables			
Distance to open water	Continuous	Minimum distance to the coast or to the nearest main river	ESRI data
Distance to urban cores	Continuous	Minimum distance to the nearest historic urban centre - base on the distance to the core (in terms of population density) of the 53 communes with high functional urbanisation	Merenne-Schoumaker et al. (1998)
Neighbourhood variables			
Percentage of the surrounding land use	Continuous	cf. text	Mücher et al. (2000) and European Commission (1993)

3.3 logistic regression

The general approach followed here was to first perform standard binomial logistic regressions, assuming that non-neighbourhood related variables would lead to a good statistical representation of the baseline land use pattern. For such a logistic regression the model form is:

$$logit(p(Y = 1)) = \alpha + \beta x \tag{1}$$

where Y is the dependent variable, x is the vector of covariates and α and β are the model parameters. This model is suqsequently referred to as the '*purely regressive model*'.

If autocorrelation remains in the residuals, this can either mean that some variables are missing, and/or that there is a need to include a spatial part in the model. Whilst it is hard to estimate which variables are missing and perhaps even harder to find the appropriate data, including a spatial variable is always possible. The model form becomes:

$$logit(p(Y = 1)) = \alpha + \beta x + \gamma y \tag{2}$$

where γy is the neighbourhood variable and associated weights. This model is subsequently referred to as the '*mixed model*' (including both neighbourhood and non-neighbourhood variables).

Alternatively, a strictly '*autoregressive model*' can be developed:

$$logit(p(Y = 1) = \gamma y \tag{3}$$

These binomial models will be applied to each of the four PELCOM land use categories (i.e. urban, cropland, grassland and built-up). If the '*autoregressive model*' gives the best statistical fit, a multinomial model will be used for the reasons described in section 2.2. Since no ancillary data is needed in a purely autoregressive model, the resolution of the baseline land use dataset is no longer considered to be a restriction in the choice of baseline land use data. Of course, spatial autocorrelation in land use patterns is scale dependent (Overmars et al., 2003). At an aggregate level, urban areas are clustered and have positive autocorrelation (Arai and Akiyama, 2004). However, at the scale of the individual parcel, negative autocorrelation resulting from negative externalities (e.g. congestion effects) has been found among urban parcels (Irwin and Geogheghan, 2001; Parker and Meretsky, 2004). For agricultural land uses, positive externalities result from e.g. economies of scales (Munroe et al., 2001) or imitation strategies (Rounsevell et al., 2003). Although this has not been verified (due to e.g. lack data representing potential land use drivers at appropriate scales), it is assumed that a purely autoregressive model would also give the best fit at the CORINE resolution.

Therefore, the multinomial model was applied to the CORINE land cover dataset (250m resolution) as well as the PELCOM dataset (1.1km). This gain in spatial resolution is combined with a gain in

quality as the official classification accuracy of CORINE is about 87% (European Commission DGXII-D, 2000) while it ranges from 50.6% to 73.3% (depending on the region) for PELCOM (Mücher et al., 2000).

The general multinomial logistic regression and the methodology to update the conditional probabilities of land use presence is described in the following sections. A subset of CORINE was chosen to demonstrate these methods, corresponding to a small area (9 ATEAM cells - about 3000km^2) of Southern Belgium (figure 1). As mentionned before, the various land cover classes present in the area were reclassified into 'built-up areas', 'cropland', 'pastures', and 'forests'. All minor land cover classes ($< 1\%$ of total area) that did not match any of these categories (e.g. water courses, bare rocks...) were reclassified as 'others'. This was done mainly for computational purposes and to best correspond to the ATEAM scenarios of land use change.

Multinomial logistic regression models belong to the category of discrete choice models. This category of models originates from the economic sciences and are based on random utility theory, which assumes that the decision-maker's preference for an alternative is captured by the value of an index, called utility. A decision-maker selects the alternative from the choice set that has the highest utility value (Mohammadian and Kanaroglou, 2003). In the case of land use modelling, a cell (or pixel) can be conceptualised as an aggregated group of decision makers. The general form of the multinomial logit model can be written as:

$$p_{ij} = \frac{exp(\mu V_{ij})}{\sum_{k \in C_k} exp(\mu V_{kj})} \qquad k \in C_k \tag{4}$$

where p_{ij} is the probability that a CORINE cell j would take land use i, rather than any other land use k in the choice set C_k, conditional on knowing the utility function V_{kj} for all k land uses in the choice set (μ is non-negative scale parameter). In the case of a purely spatial logit model, the utility function can be written as:

$$V_{kj} = \sum_{s=1}^{S} \rho_{ijs} y_{is} \tag{5}$$

where ρ is the weight factor given to the S neighbours and y_{is} takes value 1 if cell j is occupied by land use i, 0 otherwise. For a more complete discussion about discrete choice modelling principles and methods, see Ben-Akiva and Lerman (1985).

A Multinomial Discrete Choice (MDC) procedure was implemented using the SAS software. The dependent variable 'decision' takes value 1 when a specific alternative is chosen, otherwise it takes value 0. Each cell is allowed to choose one and only one of the five possible alternatives (land use categories - the 'others' category were included for completeness). In order to differentiate between land use categories, a set of four dummy independent variables x_i (taking the value 1 when alternative i is chosen, 0 otherwise) were added to the model together with the neighbourhood variable.

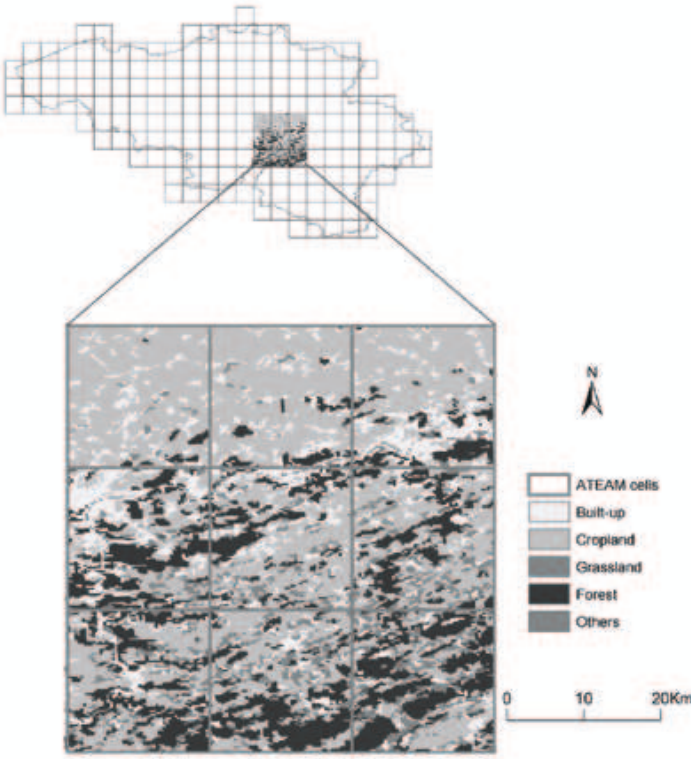

Figure 1:
Study area - 9 ATEAM cells

3.4 a method to update the probabilities of land use presence

Assume that there are k potentially observable land use categories within any CORINE cell over the study area, these categories being denoted as $c_1, \ldots, c_k$. Without loss of generality, let us consider that for any given CORINE cell, the 8 immediate neighbours have been taken into account when estimating the probability of observing each possible category for the central cell at an initial time step t (see Figure 2). If we denote $\mathcal{N}_\alpha$ as a specific configuration of these neighbours categories (among the set of k^8 possible configuration of categories), the multinomial logit procedure provides us with the conditional probabilities $p(c_i|\mathcal{N}_\alpha)$ (where $i = 1, \ldots, k$), i.e. the probabilities of each possible category at the central cell given the neighbours categories configuration $\mathcal{N}_\alpha$. Moreover, from the CORINE database, we are also able to obtain good estimates of the global theoretical frequencies (marginal probabilities) for each land use, that will be denoted as $p(c_i)$. From Bayes' theorem, we can also write that

$$p(c_i|\mathcal{N}_\alpha) = \frac{p(\mathcal{N}_\alpha|c_i)p(c_i)}{\sum_{j=1}^{k} p(\mathcal{N}_\alpha|c_j)p(c_j)} = \frac{1}{A}p(\mathcal{N}_\alpha|c_i)p(c_i) \tag{6}$$

where the denominator plays the role of a normalization constant, so that we always have $\sum_i p(c_i|\mathcal{N}_\alpha) = 1$. Based on this result we can state that, up to a multiplicative constant, the conditional probabilities can be computed as the product of marginal probabilities $p(c_i)$ – that may change according to possible global modifications of the land use proportions over time – and conditional probabilities $p(\mathcal{N}_\alpha|c_i)$ that characterize the local spatial organization of the various land uses, assumed to be largely invariant over time. As we have explicitly specified the $p(c_i|\mathcal{N}_\alpha)$ values from the multinomial logit model as well as the $p(c_i)$ values from the whole CORINE database, these various $p(\mathcal{N}_\alpha|c_i)$ values can be computed from eq. (6), with

$$p(\mathcal{N}_\alpha|c_i) = A\frac{p(c_i|\mathcal{N}_\alpha)}{p(c_i)} \tag{7}$$

This way of presenting the problem is particularly useful as it allows us to take into account the effect of the various ATEAM scenarios. Indeed, let us consider that at time t' after t the theoretical proportions of land use become equal to new values $p'(c_1), \ldots, p'(c_k)$. Assuming that the $p(\mathcal{N}_\alpha|c_i)$ values remain unchanged when moving from time t to time t', rewriting eq. (7) for this time instant t' gives us the result

$$A\frac{p(c_i|\mathcal{N}_\alpha)}{p(c_i)} = p(\mathcal{N}_\alpha|c_i) = A'\frac{p'(c_i|\mathcal{N}_\alpha)}{p'(c_i)} \Longleftrightarrow p'(c_i|\mathcal{N}_\alpha) \propto \frac{p'(c_i)}{p(c_i)}p(c_i|\mathcal{N}_\alpha) \tag{8}$$

subject to the constraint $\sum_i p'(c_i|\mathcal{N}_\alpha) = 1$. The last relation in eq. (8) is thus an updating rule for the computation of new conditional probabilities $p'(c_i|\mathcal{N}_\alpha)$ when the marginal probabilities are changed from the set of values $\{p(c_1), \ldots, p(c_k)\}$ that apply at time t to the new set of values $\{p'(c_1), \ldots, p'(c_k)\}$ that apply later at time t'. In practice, these new marginal probabilities are the land use frequencies of the ATEAM cell in which the CORINE cell is included (see Figure 1).

In summary, for all of the n CORINE cells that belong to the same ATEAM cell :

1. for each land use c_i, the marginal probability $p(c_i)$ at the initial time step t is considered to be identical for all CORINE cells and can be estimated from the corresponding observed frequency computed from the whole CORINE database (or solely from the CORINE cells that belong to this ATEAM cell if proportions appear to vary in a significant way from one ATEAM cell to another);

2. for each land use c_i, the marginal probability $p'(c_i)$ at time t' is considered to be identical too for all CORINE cells and is considered to be equal to the corresponding land use frequency f_i' specified for this ATEAM cell;

3. for each CORINE cell, the updated conditional probabilities $p'(c_i|\mathcal{N}_\alpha)$ $(i = 1, \ldots, k)$ are computed (up to a normalization constant) using eq. (8), where the initial conditional probabilities $p(c_i|\mathcal{N}_\alpha)$ are coming from the previously fitted logit multinomial model.

4. for each CORINE cell, the outputs of eq. (8) are normalized so that $\sum_i p'(c_i|\mathcal{N}_\alpha) = 1$, as required for a valid probability distribution.

Though the above procedure provides us with a simple method for computing new conditional probabilities for any CORINE cell included within a given ATEAM cell, there is still a consistency issue. The marginal probabilities as directly specified by the corresponding ATEAM frequencies $\{f'_1, \ldots, f'_k\}$ may differ somewhat from the marginal probabilities $\{p'(c_1), \ldots, p'(c_k)\}$ that can be estimated afterwards from the new set of conditional probabilities using the formula

$$\widehat{p}\,'(c_i) = \frac{1}{n}\sum_{j=1}^{n} p'(c_i|\mathcal{N}_{[j]}) \tag{9}$$

where $p'(c_i|\mathcal{N}_{[j]})$ refers to the estimated conditional probability of category c_i for the jth CORINE cell. This consistency problem is an indirect consequence of the normalization step 4 that involves a different normalization constant for each CORINE cell in general. Indeed, using eq. (9) with the ouputs of Step 3 will lead to $\widehat{p}\,'(c_i)$ values that are identical to the specified f_i ATEAM frequencies, but these outputs do not provide valid distributions (as in general these probabilities will not sum up to one for each CORINE cell), whereas the ouputs of Step 4 will respect this validity condition but using eq. (9) will then no longer garantee that $\widehat{p}\,'(c_i) = f_i$ in general. Stated in other words, the above procedure does not allow the user to derive conditional probability values that respect at the same time the validity conditions $\sum_i p'(c_i|\mathcal{N}_\alpha) = 1$ and the specified frequencies f_i for the corresponding ATEAM cell.

A way to avoid this consistency problem is to use an iterative procedure that aims at modifying the set of conditional probabilities in such a way that both conditions can be fulfilled at the same time. Let us rearrange the set of all conditional probabilities from Step 4 in a $k \times n$ two-dimensional table $T^{[0]}$, such that

$$T^{[0]} = \begin{bmatrix} p'(c_1|\mathcal{N}_{[1]}) & p'(c_1|\mathcal{N}_{[2]}) & \cdots & p'(c_1|\mathcal{N}_{[n]}) \\ p'(c_2|\mathcal{N}_{[1]}) & p'(c_2|\mathcal{N}_{[2]}) & \cdots & p'(c_2|\mathcal{N}_{[n]}) \\ \vdots & \vdots & \ddots & \vdots \\ p'(c_k|\mathcal{N}_{[1]}) & p'(c_k|\mathcal{N}_{[2]}) & \cdots & p'(c_k|\mathcal{N}_{[n]}) \end{bmatrix} \tag{10}$$

where for the sum along the columns of $T^{[0]}$ we have

$$\sum_{i=1}^{k} T^{[0]}_{ij} = \sum_{i=1}^{k} p'(c_i|\mathcal{N}_{[j]}) = 1 \qquad \forall j = 1, \ldots, n \tag{11}$$

whereas for the sum along the lines we have

$$\sum_{j=1}^{n} T^{[0]}_{ij} = \sum_{j=1}^{n} p'(c_i|\mathcal{N}_{[j]}) = n\widehat{p}\,'(c_i) \qquad \forall i = 1, \ldots, n \tag{12}$$

with $\widehat{p}\,'(c_i) \neq f_i$ in general. Let us now define a new table $T^{[1]}$ obtained using the relations

$$\begin{aligned} S_{ij} &= T_{ij}^{[0]} \frac{f_i}{\sum_{j=1}^{n} T_{ij}^{[0]}} \quad \forall i,j \quad \text{so that } \textstyle\sum_j S_{ij} = f_i \quad \forall i \\ T_{ij}^{[1]} &= S_{ij} \frac{1}{\sum_{i=1}^{k} S_{ij}} \quad \forall i,j \quad \text{so that } \textstyle\sum_i T_{ij}^{[1]} = 1 \quad \forall j \end{aligned} \tag{13}$$

Clearly, applying such a transformation does not modify the sum of each line and we still have $\sum_j T_{ij}^{[1]} \neq n f_i$ in general. However, it can be proved that after a repeated use of eq. (13) in order to get $T^{[2]}, T^{[3]}, \ldots$, we will obtain

$$\sum_{j=1}^{n} T_{ij}^{[\infty]} = n f_i \quad \forall i = 1, \ldots, k$$

which is the result that was sought. In practice, convergence is relatively fast so that few iterations are needed for getting $\widehat{p}\,'(c_i) \simeq f_i$. This algorithm is a straightforward variation around the well-known iterative rescaling procedure, widely applied in statistics for the fitting of a multidimensional contingency table subject to contraints on its marginal values.

When downscaling aggregated land use data from adjacent land use units, it is likely that the borders between these units remain visible at the finer resolution. In the present case, it makes no sense to assume that two adjacent CORINE cells belonging to two different ATEAM cells would have completely different sets of marginal probabilities (i.e. different sets of land use ratios at the ATEAM resolution). However, the marginal probabilities values from the ATEAM cells constrain the updating of the conditional probabilities (see eqn. 5). Therefore, the original marginal probabilities at both time steps t (given by CORINE) and t' (given by the ATEAM scenarios) have to be recomputed for all CORINE cells using inverse distance weighted based smoothing algorithms within the iterative procedure.

This iterative procedure giving new conditional probabilities was encoded and applied using the MATLAB software. To avoid confusion, the 'others' category from the CORINE map (i.e. 'fixed land covers' such as water or bare rocks) and the 'surplus' categories (i.e. land that no longer has an economic value and becomes abandoned) from the ATEAM scenarios were excluded from the vectors of marginal probabilities at all time steps.

4 Results

4.1 binomial logistic regression

For all of the land uses studied, the approaches based on the enrichment factor or on the regression's residuals were not significantly better than the regressions that included a simple Moore neighbourhood variable. On the contrary, they tended to give slightly poorer fits. For more details about these various

neighbourhood sizes and models the reader is referred to Dendoncker et al. (2005). Therefore, the spatial variable based only on the Moore neighbourhood was retained. Table 2 summarizes the results of the different models for each land use studied. Results are compared using the Receiver Operating Characteristic (ROC) (Pontius and Schneider, 2000) and a classification table giving correctly classified cells at a 0.5 probability threshold. By itself, a classification table can only give a relative indication of how well the model performs, but it remains a very useful tool when comparing the performance of different models. The results are discussed at length by Dendoncker et al. (2005). A general conclusion that can be drawn is that stricly *autoregressive models* always outperform *purely regressive models* and give similar fit to the *mixed models*. Therefore, it can be concluded that a purely autoregressive approach is entirely appropriate in order to obtain the best statistical model of land use distribution at a relatively coarse resolution in Belgium.

Table 2: binomial logistic regression

Land use	goodness of fit	non-spatial	mixed	spatial
BUILT-UP	ROC stat	0.75	0.89	0.88
	% correct	88.7	91.6	91.6
CROPLAND	ROC stat	0.83	0.94	0.94
	% correct	73.0	86.1	86.0
GRASSLAND[a]	ROC stat	0.72	0.94	0.94
	% correct	81.7	89.5	88.3
FOREST	ROC stat	0.92	0.98	0.98
	% correct	88.3	94.2	94.3

[a] In order to account for differences in management, the dataset was divided in two. The given results of the non-spatial and mixed models concern the Northern region of Belgium. For more detail see Dendoncker et al. (2005). Applying a strictly autoregressive model avoids this division.

4.2 multinomial autologistic regression

The results of the MDC procedure are summarized in table 3. The general fit of the model is very good (Mc Fadden's likelihood ratio index = 0.63) and all variables are highly significant (p value < 0.0001). This confirms that using solely neighbourhood variables to derive probability maps of land use presence at the CORINE resolution was appropriate. The associated probability table is of greater importance for the rest of this study, as the main goal of this model is to derive baseline suitability maps. This set of probabilities will be the reference database for the updating methodology described in section 3.4.

Table 4 gives the contingency table (also called error or confusion matrix) comparing the original CORINE data with the predicted land use (i.e. the land use category with the highest probability in the choice set). The overall accuracy (0.77) and the kappa coefficient (0.63) (Richards and Jia, 1999;

Pontius, 2000) are both good. To consider the similarity of location and the similarity of quantity independently, the kappa statistic has been partitioned into 'K-location' (as defined by Pontius (2000)) and 'K-histo' (as defined by Hagen (2002)). The former equals 0.65 while the latter equals 0.97. This suggests that the model predicts quantity with more accuracy than location. Not surprisingly, the model does not predict the 'others' category well. This can be explained by the under-representation of this artificially constructed category and the fact that it mainly consists of linear elements (rivers, i.e. not clustered) along with some small patches of bare rock. Furthermore, this category is not subject to change and will be masked when the allocation procedure is applied to real data. It was only included here for completeness, to have a consistent set of alternatives (ensuring that predicted probabilities add up to one) and to avoid holes in the gridded dataset. All other land use categories are fairly well predicted.

A spatial comparison of the CORINE data with the predicted land use is given in figure 2. It clearly shows that the model tends to further compact existing land use clusters while isolated cells are lost. This leads to a more aggregated and less fragmented landscape pattern. This is confirmed by some simple spatial metrics computed with the Patch Analyst extension for Arcview (e.g. mean patch size of all land use categories - except 'others' significantly increases while total edge length diminishes). Moreover, cells located at the fringes between land use clusters are usually poorly predicted (figure 2c). This is the consequence of using a purely neighbourhood-based regression model. However, Dendoncker et al. (2005) showed that no other model would give a better fit. One has to remember that this is a mean map showing the most probable (in a statistical sense) land use pattern but not necessarily the most realistic one. In fact, whenever a decision is made to represent each cell by one hard categorical value, there is a loss of information. Real information is contained by the distribution of conditional probabilities as given by the spatial multinomial logistic regression model. A certain level of uncertainty is associated therefore with each decision made. In other words, a cell can be mapped as grassland with conditional probabilities of 0.3 or 0.99 but uncertainty is obviously much lower in the latter case. However, the uncertainty of prediction is never represented on maps, which would become over-complex and difficult to read.

Table 3: The MDC procedure

Discrete Response Profile		
LAND USE	Frequency	Percent
Built-up	4920	15.58
Cropland	16920	53.57
Grassland	2809	8.89
Forest	6737	0.63
Dependent Variable		Decision
Number of Observations		31585
Number of Cases		157925
Likelihood Ratio		64260
McFadden's LRI		0.6321
Parameter Estimates		$Pr > \|t\|$
dummy1	1.70	< 0.0001
dummy2	1.50	< 0.0001
dummy3	1.53	< 0.0001
dummy4	1.50	< 0.0001
neighbourhood	0.63	$< .0.0001$

Table 4: contingency table

		PREDICTED							
		Built-up	Cropland	Grassland	Forest	Others	Total	p(a)	0.77
OBSERVED	Built-up	0.09	0.04	0.01	0.01	0.00	0.15	p(max)	0.98
	Cropland	0.03	0.46	0.02	0.03	0.00	0.54	p(e)	0.37
	Grassland	0.01	0.02	0.05	0.01	0.00	0.09	kappa	0.63
	Forest	0.01	0.02	0.01	0.17	0.00	0.21	klocation	0.65
	Others	0.00	0.00	0.00	0.00	0.00	0.01	khisto	0.97
		0.14	0.55	0.08	0.22	0.01	1 (31585 cells)		

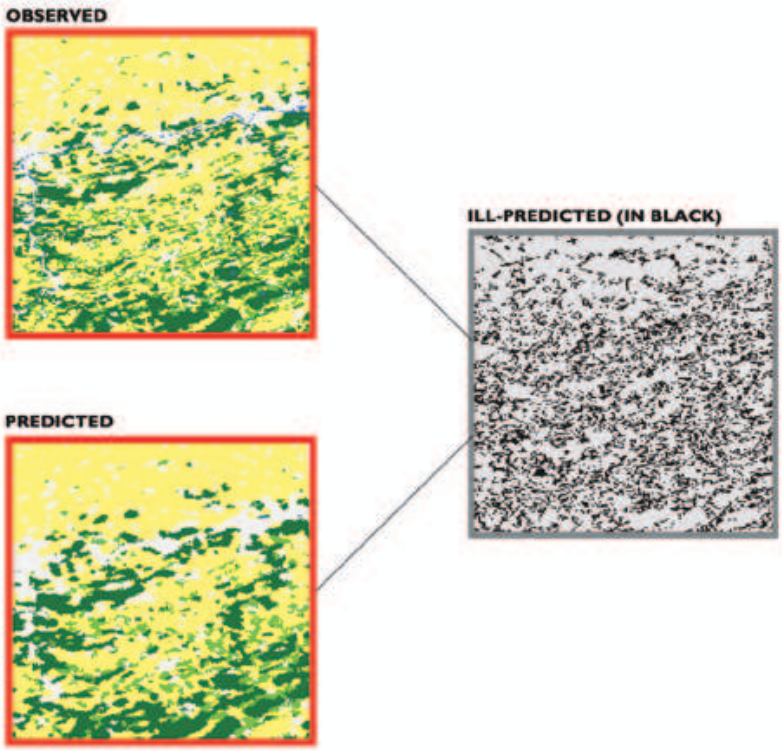

Figure 2:

The MDC procedure - observed vs predicted

4.3 updated probabilities of land use presence

Figure 3 shows the land use map resulting from the downscaling of the four ATEAM scenarios using the updated conditional probabilities for the year 2020 (time t'). The land use with the highest predicted conditional probability is represented. As expected from the baseline probability map resulting from the multinomial logit procedure (see figure 2), the landscape pattern is more aggregated and less fragmented. However, no border effects can be seen.

All river cells from the baseline CORINE dataset (see figure 1) have artificially been replaced by different land use categories, this is not considered to be a problem as the 'others' (i.e. river and bare rock) category will be restricted from change in a practical application of the method. Similarly, no indication is given as to which pixels will be abandoned from commercial management (i.e. the 'surplus' category from the ATEAM scenarios). However, other studies suggest that for agriculture these are likely to take place in less favoured areas (Verburg et al., 2005).

The main differences between scenarios that can be observed in figure 3 result from the differences in the land use frequencies given by the ATEAM scenarios rather than from the probability updating methodology itself. For example, the A2 and B1 scenarios display very similar land use patterns and quantities, both preserving most of their grassland cells although, on the whole, European grassland diminishes much more in A2 (Rounsevell et al., 2005a). In the A1 scenario grassland diminishes mainly in the Least Favoured Areas (LFA's) represented by the southern cells in this example. In general grassland is replaced by 'liquid' biofuels classified as cropland here. Finally, B2 shows the almost complete disappearance of grassland and its replacement by 'solid' biofuels here classified as forest.

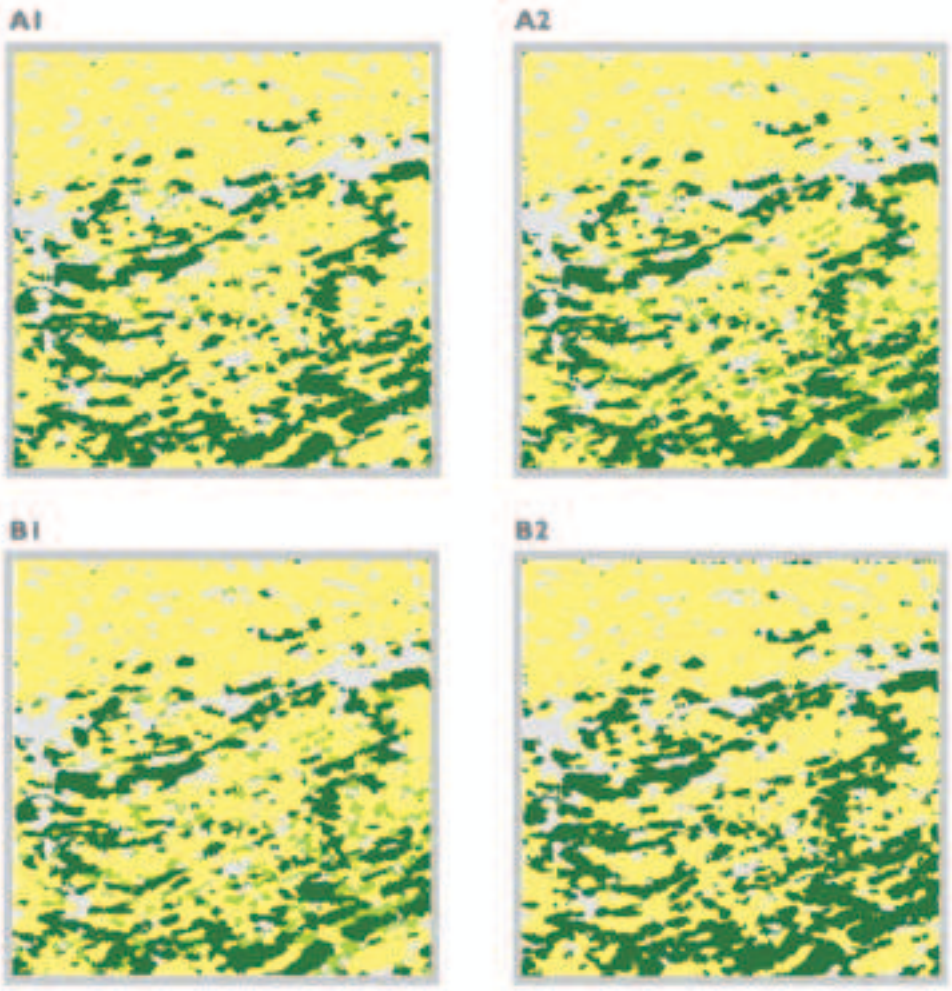

Figure 3:
A1fi 4 scenarios downscaled - 2020

5 Discussion

The main goal of this paper was to establish a statistically sound methodology to downscale aggregate land use data and to be able to visualize land use patterns from the ATEAM scenarios of land use change. Results from the spatial binomial regression, the multinomial autologistic regression as well as from the iterative procedure based on Bayes' theorem were conclusive. The resulting land use maps give appropriate representation of the land use pattern despite a tendency to increase clustering and *a fortiori* loose part of the initial fragmentation. This is a consequence of using a purely autoregressive model. However, using neighbourhood variables is considered to be a real advantage as no ancillary data are needed to derive the statistically based suitability maps. These datasets are often incomplete, are not always available at adequate resolution or spatial extent, can be of debatable quality or simply

do not exist. Thus, finding appropriate datasets to derive optimal variables that represent land use drivers is always a time and effort consuming process. The methodology developed in this paper is simpler and quicker to implement. A final advantage of the multinomial logistic regression procedure is that it allows to discriminate between land uses as complete vectors of conditional probabilities are derived (i.e. the probabilities of presence for all land use classes within each cell sum to one). The new CORINE Land Cover raster dataset giving land use in 2000 has just been made available and could serve as a basis for a quasi validation of the multinomial logit procedure.

The downscaled land use change scenario presented in figure 4 are 'mean' maps, representing the most probable (in a statistical sense) but not necessarily the most realistic patterns. In order to derive such maps, it is necessary to adapt the calculated suitability with decision rules that reflect the assumed changes in location preferences (Verburg et al., 2005). In addition to the above comments, it should be noted that downscaling is merely the last step in the multi-scale land use modelling process. The main differences between land use models (especially in land use quantities) will always arise from the interpretation of storylines and their translation into quantitative scenarios. Nevertheless, it is important to use optimal suitability maps to underpin a rule based procedure. Rules need to include elements such as policy constraints (e.g. designation of protected areas, precise delineation of LFA's...). Some conversions should be made impossible (e.g. existing urban and river cells will always remain as such), others should only be possible after certain time lags. A simple conversion matrix as proposed by Verburg et al. (2005) would account for this. Finally, some rules could be specific to certain scenarios and some *a posteriori* decisions need to be made regarding the location of biofuels and abandonned land. When this rule based approach is achieved, the resulting land use scenario maps will be useful for a large set of applications in a range of scientific disciplines. The statistical analysis presented here was a first step in this direction.

References

Anselin L. (2002), 'Under the hood. issues in the specification of spatial regression models', *Agricultural Economics* **27**, 247–267.

Anselin L. and Kelejian H.H. (1997), 'Testing for spatial error autocorrelation in the presence of endogenous regressors', *International Regional Science Review* **20**, 153–182.

Arai T. and Akiyama T. (2004), 'Empirical analysis for estimating land use transition potential functions - case in the tokyo metropolitan region', *Computers, Environment and Urban Systems* **28**, 65–84.

Augustin N.H., Cummins R.P. and French D.D. (2001), 'Exploring spatial vegetation dynamics using logistic regression and a multinomial logit model', *Journal of Applied Ecology* **38**, 991–1006.

Ben-Akiva M. and Lerman S.R. (1985), *Discrete choice analysis: theory and applications to travel demand*, Cambridge, MA.

Bhat C.R. and Guo J. (2003), A mixed spatially correlated logit model: formulation and application to residential choice modelling, *in* 'IATBR conference', Lucerne, Switzerland.

Briassoulis H. (2000), Analysis of land use change: theoretical and modeling approaches., *in* S. Loveridge, ed., 'The web book of regional science', West Virginia University, Morgantown.

Cramer J.S. (1991), *The logit model: an introduction for economists*, Arnold edn, London.

Dendoncker N., Bogaert P. and Rounsevell M. (2005), 'A spatial logistic regression model to analyse land use distribution in belgium', *Agriculture, Ecosystems and Environment* **submitted**.

DGXII-D European Commission (2000), Pelcom: Development of a consistent methodology to derive land cover information on a european scale from remote sensing for environmental modelling; final report, Technical report, Editor Mücher C.A.

Engelen, G. White R. de Nijs A.C.M. (2002), Environment explorer: spatial support system for the integrated assessment of socio-economic and environmental policies in the netherlands., *in* '1st Biennal Conference of the International environmental modelling and software society', Lugano.

Erhard M., Carter M., Mitchell T., Reginster I., Rounsevell M. and Zaehle S. (2005), 'Data and scenarios for assessing ecosystem vulnerability across europe', *Regional Environmental Change* **submitted**.

European Commission (1993), Corine land cover map and technical guide, Technical report, European Union Directorate General Environment (Nuclear Safety and Civil Protection), Luxembourg.

Gauthier G., Giroux J.F., Reed A., Bechet A. and Belanger L. (2005), 'Interactions between land use, habitat use, and population increase in greater snow geese: what are the consequences for natural wetlands?', *Global Change Biology* **11**(6), 856–868.

Gujarati D.N. (1995), *Basic Econometrics (3rd ed.)*, Mcgraw-Hill edn, New York.

Hagen A. (2002), Map comparison - methods, Technical report, RIKS, Maastricht.

Hilferink M. and Rietveld P. (1998), Land use scanner: an integrated gis based model for long term projections of land use in urban and rural areas, Technical report, Tinbergen Institute.

Irwin E.G. and Geogheghan J. (2001), 'Theory, data, methods: Developing spatially explicit economic models of land use change', *Agriculture, Ecosystems and Environment* **85**, 7–23.

Lettens S., Van Orshoven J., van Wesemael B., Perrin D. and Roelandt C. (2004), 'The inventory-based approach for prediction of soc change following land use change', *Biotechnology, Agronomy, Society and Environment* **8**(2), 141–146.

Mücher C.A., Steinnocher K.T., Kressler F.P. and Heunks C. (2000), 'Land cover characterization and change detection for environmental monitoring of pan-europe.', *International Journal of Remote Sensing* **21**(6), 1159–1181.

McMillen D. (2001), 'An empirical model of urban fringe land use', *Land Economics* **65**(2), 138–145.

Merenne-Schoumaker B., Van der Haegen H. and Van Heck E. (1998), *Recensement général de la population et des logements au 1er mars 1991: urbanisation*, Cheruy C. edn, Bruxelles.

Mertens B. and Lambin E.F. (1997), 'Spatial modelling of deforestation in southern cameroon. spatial disaggregation of diverse deforestation processes.', *Applied Geography* **17**, 143–162.

Mitchell T.D., Carter T.R., Jones P.D., Hulme M. and New M.G. (2004), A comprehensive set of high-resolution grids of monthly climate for europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100) . tyndall centre working paper no. 55, Technical report, Tyndall Centre, Norwich, UK.

Mohammadian A. and Kanaroglou P.S. (2003), Applications of spatial multinomial logit model to transportation planning, *in* 'Moving through nets: the physical and social dimensions of travel', Lucerne.

Munroe D., Southworth J. and C.M. Tucker; (2001), 'The dynamics of land-cover change in western honduras: spatial autocorrelation and temporal variabtion', *Agricultural Economics* **27**(3), 355–369.

Overmars K.P., de Koning G.H.J. and Veldkamp A. (2003), 'Spatial autocorrelation in multi-scale land use models', *Ecological Modelling* **164**, 257–270.

Parker D. C. and Meretsky V. (2004), 'Measuring pattern outcomes in an agent-based model of edge-effect externalities using spatial metrics', *Agriculture, Ecosystems and Environment* **101**(2-3), 233–250.

Peppler-lisbach C. (2003), 'Predictive modelling of historical and recent land-use patterns', *Phytocoenologia* **33**(4), 565–590.

Pontius R.G. (2000), 'Quantification error versus location error in comparison of categorical maps', *Photogrammetric Engineering and Remote Sensing* **66**(8), 1011–1016.

Pontius R.G. and Schneider L.C. (2000), 'Land use change model validation by a roc method', *Agriculture, Ecosystems and Environment* **85**, 269–280.

Richards J.A. and Jia X. (1999), *Remote sensing digital image analysis, an introduction*, Springer edn, Berlin.

Rounsevell M., Ewert F., Reginster I., Leemans R. and Carter T.R. (2005b), 'Future scenarios of european agricultural land use, ii: Estimating change in land use and regional allocation', *Agriculture, Ecosystems and Environment* **107**(2-3), 101–116.

Rounsevell M.D.A., Annetts J.E., Audsley E., Mayr T. and Reginster I. (2003), 'Modelling the spatial distribution of agricultural land use at the regional scale', *Agriculture, Ecosystems and Environment* **95**, 465–479.

Rounsevell M.D.A., Reginster I., Araujo M.B., Carter T.R., Dendoncker N., Ewert F., House J.I., Kankaanpää S., Leemans R., Metzger M., Schmit C., Smith P. and Tuck G. (2005a), 'A coherent set of future land use change scenarios for europe', *Agriculture, Ecosystems and environment* **In press**.

Schmit C., Rounsevell M.D.A. and La Jeunesse I. (2005), 'The limitations of spatial land use data in environmental analysis and policy', *Environmental Science and Policy* **submitted**.

Schotten K., Goetgeluk R., Hilferink M. and Scholten H. (2001), 'Residential construction, land use and the environment. simulations for the netherland using a gis-based land use model', *Environmental Monitoring and Assessment* **6**, 133–143.

Serneels S. and Lambin E.F. (2001), 'Proximate causes of land-use change in narok district, kenya: a spatial statistical model', *Agriculture, Ecosystems and Environment* **85**, 65–81.

Sullivan T.J, McMartin B. and Charles D.F. (1995), 'Re-examination of the role of landscape change in the acidification of lakes in the adirondack mountains, new york', *Science of the Total Environment* **183**(3), 231–248.

Tavernier R. and Marechal R. (1962), Soil survey and soil classification in belgium, *in* 'Transactions of a joint meeting of the International Society of Soil Science', Palmerston North, New Zealand, pp. 298–307.

Van Orshoven J., Deckers J.A., Vandenbroucke D. and Fewen J. (1993), 'The complete database of belgian soil profile data and its applicability in planning and management of rural land', *Bulletin de Recherche Agronomique de Gembloux* **28**, 197–222.

Veldkamp A and Fresco L.O. (1996), 'Clue-cr: an integrated multi-scale model to simulate land use change scenarios in costa rica.', *Ecological Modelling* **91**, 231–248.

Veldkamp A. and Fresco L.O. (1997), 'Reconstructing land use drivers and their spatial scale dependence for costa rica (1973 and 1984)', *Agricultural Systems* **55**(1), 19–43.

Verburg P.H, de Nijs T., van Eck J.R., Visser H. and de Jong K. (2004c), 'A method to analyse neighbourhood characteristics of land use patterns', *Computers, Environment and Urban Systems* **28**(6), 667–690.

Verburg P.H, Ritsema van Eck J.R., de Nijs T.C.M., Dijst M.J. and Schot P. (2004a), 'Determinants of land use change patterns in the netherlands', *Environment and Planning B* **31**(1), 125–150.

Verburg P.H, Schot P., Dijst M. and Veldkamp A. (2004b), 'Land use change modelling: current practice and research priorities', *Geojournal* **61**, 309–324.

Verburg P.H, Schulp C.J.E., Witte N. and Veldkamp A. (2005), 'Downscaling of land use change scenarios to assess the dynamics of european landscapes', *Agriculture, Ecosystems and Environment* **submitted**.

Verburg P.H., Soepboer W., Limpiada R., Espaldon M.V.O. and Sharifa M.A. (2002), 'Modeling the spatial dynamics of regional land use: the clue-s model', *Environmental Management* **3**, 391–405.

White R. and Engelen G. (1993), 'Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns', *Environment and Planning A* **25**, 1175–1199.

Wimberly M.C. and Ohmann J.L. (2004), 'A multi-scale assessment of humand and environmental constraints on forest land cover change on the oregon (usa) coast range', *Landscape Ecology* **19**, 631–646.