A framework of map comparison methods to evaluate geosimulation models from a geospatial perspective

Alex Hagen-Zanker[1,2], Pim Martens[3]

[1] Research Institute for Knowledge Systems, PO Box 463, 6200 AL Maastricht, The Netherlands.

[2] Urban Planning Group, Technical University Eindhoven, PO Box 513, 5600 MB Eindhoven, The Netherlands.

[3] International Centre for Integrated assessment and Sustainable development, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands.

ahagen@riks.nl

Abstract Geosimulation is a form of microsimulation that seeks to understand geographical patterns and dynamics as the outcome of micro level geographical processes. Geosimulation has been applied to understand such diverse systems as lake ecology, traffic congestion and urban growth. A crucial task common to these applications is to express the agreement between model and reality and hence the confidence one can have in the model results. Such evaluation requires a geospatial perspective; it is not sufficient if the micro-level interactions are realistic. Importantly the interactions should be such that the meso and macro level patterns that emerge from the model are realistic. In recent years, a host of map comparison methods have been developed that address different aspects of the agreement between model and reality. This paper places such methods in a framework to systematically assess the breadth and width of model performance. The framework expresses agreement at the continuum of spatial scales ranging from local to the whole landscape and separately addresses agreement in structure and presence. A common reference level makes different performance metrics mutually comparable and guides the interpretation of results. The framework is applied for the evaluation of a constrained cellular automata model of the Netherlands. The case demonstrates that a performance assessment lacking either a multi-criteria and multi-scale perspective or a reference level would result in an unbalanced account and ultimately false conclusions.

Keywords: geosimulation, calibration, validation, map comparison

1 Introduction

Geosimulation is a field of geography that seeks to understand geographical patterns and dynamics as the consequence of the interactions between individual entities, like tenants, land owners, car drivers, trees, etc. Geosimulation models only prescribe the behaviour of these entities and not the resulting large scale patterns, such as segregation, urban sprawl, road congestion, forest fire, etc. These emerge as a product of the interactions between the entities. The field of geosimulation is spurred by advances in computing. As a consequence, the average desk computer can function as a virtual laboratory, where researchers can grow their own virtual cities, transport systems, rural societies, forests, etc. (Benenson and Torrens 2004)

Geosimulation models are increasingly finding region specific applications. Rather than using the models to grow geographical systems from scratch, the models are fed with a real initial situation. The models are then not just theoretical constructs, but have a more applied nature. A striking illustration is the contrast between the Schelling model (Schelling 1971) and the entity-based model of urban residential dynamics for the Yaffa area in Tel Aviv (Benenson et al. 2002). Both models are concerned with segregation and the underlying microdynamics, but the level of detail and the lessons that can be drawn differ considerably.

The constrained cellular automata land use model that is evaluated as a case in this paper is a good example of a model that has developed from a theoretical model to a practical tool. The first application of the model concerned an imaginary island with characteristics typical for Caribbean islands (Engelen et al. 1995). Later applications focused on the city of the Cincinnati Metropolitan Area on a timescale of more than 100 years (White et al. 1997). Further developments have elaborated the use of GIS data, including road network data and the dynamic integration with socio-economic land use models at multiple scales (White and Engelen 2000). The model became part of Policy Support Systems, such as the Environment Explorer (Engelen et al. 2003, de Nijs et al. 2004). Currently the model is the cornerstone of several modelling frameworks of urban and regional growth, meaning that new regional applications can be setup within hours, including the METRONAMICA (van Delden and Engelen 2006) and MOLAND (White et al. 2000, Barredo and Demicheli 2003) frameworks.

The problem that confronts the new generation of geosimulation modellers is to assess how well their virtual worlds correspond to reality. ‘Good modelling practice’ (Refsgaard and Henriksen 2004) prescribes different analytical steps. Of these, calibration and validation require an expression of goodness-of-fit of the model. Since the results of the models are typically maps, it makes perfect sense to address this question by map comparison; however the nature of geosimulation models provides some particular challenges that need to be considered.

One problem is that the resolution at which the model is defined is not equal to the scale at which the results are interpreted. The interest in the models lies in the geographical structures that unroll as a consequence of the interactions of the individual entities. The concept of complexity is relevant here. Typically the elements in the model are mutually dependent, causing feedback processes and self-organization, but also making the models sensitive to small deviations to the extent that they are chaotic or unpredictable. The consequence is that even a geosimulation model that perfectly captures the dynamics of a geographical system cannot be expected to produce result maps that correspond perfectly to reality. The models should therefore not be evaluated just at the location-to-location level, but in terms of the patterns that emerge. On the other hand; when applying a model for a particular region, one is not just interested in the global patterns, but also how the patterns are distributed in space.

A balance is sought between finding realistic patterns and finding them at the right location; the geosimulation model should create spatial configurations that are similar to reality and place them in approximately the right locations. Existing comparison methods do not strike such a balance. With few exceptions, they are either local and based on cell-by-cell overlap, or global and based on metrics summarizing the whole landscape in a single value. Despite the lack of formal methods, an expert can make this kind of balanced comparison by just looking at the map. It is therefore not surprising that in practice geosimulation models are often evaluated on the basis of such face validation. (Batty and Torrens 2005)

There are several problems associated to face validation, most pressingly the lack of objective reproducibility. A practical concern of face validation is that for some tasks, such as calibration, large numbers of consistent assessments are required. Depending on a human judge of map similarity may be too time-consuming, costly and prone to inconsistencies.

Another challenge is posed by the dynamical nature of the simulation models and the relatively small number of changes that may occur over a simulation period. It is for instance not uncommon for land use models to attain percentages of agreement between reality and model above 95%. This ‘good performance’ is then invariably due to the fact that land use patterns at the beginning of the simulation period are highly similar to those at the end (Pontius Jr. et al. 2008). This causes a real risk of misinterpretation and false confidence in the results of geosimulation models.

This paper introduces a framework to evaluate model performance. It applies a number of statistics that can be categorized according to two axes. The first axis is typically recognized in geographical information science and is based on the spatial unit of the analysis; it ranges from local, via focal to global. The second axis is more commonly applied in landscape ecological applications and discerns whether the presence of certain classes is considered or their spatial structure. The interpretation of the results is guided by a neutral model of landscape change that provides a common reference for all metrics (Hagen-Zanker and Lajoie 2008). This reference model is subject to the same initial situation and constraints but does not represent the processes that characterize the evaluated model. The difference in performance between the two models can thus be attributed to those processes.

The purpose of this framework is in first instance to provide an order in the large variety of performance metrics. Secondly, the framework can be a guide towards a comprehensive approach of performance assessment in the calibration and validation process of geosimulation models.

2 Method

2.1 The axis: Local, Global, Focal

Geographical Information Systems (GIS) operations can be classified as being local, focal or global in nature (Takeyama and Couclelis 1997). Local statistics relate a particular location on the map, in the case of raster maps: a cell. Analysis based on local operations is often called overlay analysis. Focal statistics have the focus on one location, but also take the neighbourhood of the location into account. Typical focal operations are spatial smoothing functions and density estimation. Focal operations are also called moving window, filter (typically for small neighbourhoods) or kernel operations (typically for large neighbourhoods). Global operations are based on aggregates over the whole map.

In the past, map comparison methods that are used as performance criteria for spatial models have been either local e.g. (Monserud and Leemans 1992, Pontius Jr. 2000) or global. Global analysis includes fractal dimension (Batty and Longley 1994, White 2006), cluster size distributions (Dungan 2006) and landscape metrics (Turner et al. 1989, Barredo and Demicheli 2003). The disadvantage of the methods is that either spatial structure (local) or spatial specificity (global) is ignored. The continuum from local to global has been investigated by multi-scale analysis on the basis of step-wise aggregations of model results and data (Costanza 1989, Kok et al. 2001, Pontius Jr. 2002) and wavelet decomposition (Briggs and Levine 1997). Both aggregation and wavelet based approaches however, suffer under the rather arbitrary positioning of the coarse scale grid relative to the original grid.

The methods in this paper consider the whole spectrum of local to global operations. The emphasis, however, is on focal operations. Not just because these have received little attention in the past, but primarily because focal operations are the ideal means to simultaneously investigate similarity in structure and location-to-location correspondence.

The use of local or focal statistics does not imply that map comparison results are only presented at the local level, i.e. as a comparison map. Local agreement can be aggregated to global statistics. The crucial distinction is whether a metric compares local, focal or global attributes.

2.2 The axis: presence and structure

The field of landscape ecology (Turner 2005) studies the relationship between landscape structure and ecological processes. One of the major considerations is that inference about spatial structure is only possible if it is objectively quantified. Consequently many metrics of landscape structure have been introduced and analyzed.

In particular two types of spatial structure are recognized: composition and configuration. Metrics of composition are based on the fraction of occurrence of land use / land cover classes at the global level. Examples of composition metrics are diversity indicators. Metrics of configuration relate to the spatial positioning of land use classes relative to each other. Configuration metrics are often calculated at the level of patches. Patches are contiguous areas of a class. In a geographical context patches are often called clusters. Examples of patch level structure metrics are size, perimeter, shape index and fractal dimension. Other configuration metrics that are not based on patches include the edge and contagion index. Configuration metrics are common in landscape ecology (McGarigal et al. 2002) and they are sporadically used as performance criteria for geosimulation models of urban dynamics (Turner et al. 1989, Barredo and Demicheli 2003).

In landscape ecological studies, metrics are typically applied and analyzed at the landscape scale, i.e. global. When landscape metrics are applied at the focal level, the distinction between composition and configuration blurs, since the focal composition is dependent of the global configuration. For example, focal assessments of patch size will, depending on the size of the focal window, correlate strongly with focal assessments of entropy, since large patches lead to homogenous areas. Therefore, this paper re-emphasizes the distinction between configuration and composition to that between structure and presence. Similarity in structure is achieved when structure metrics (either of configuration or composition) describing two maps, focal windows or locations, are similar. Similarity in presence is location specific and achieved when the composition of two locations or focal windows are similar.

2.3 Comparison methods

Considering the two axes, six classes of map comparison methods could possibly be identified. Locally, at the level of a single cell, it is not possible to recognize spatial structure however. Otherwise, at the global level all notion of location is lost and therefore the notion of presence is obsolete. Thus, four classes of map comparison methods remain. Table 1 presents the comparison methods of this paper and places them in the framework of the two axes.

Table 1. Overview of comparison methods applied in this paper

	Local	Focal	Global
Presence	Kappa	Moving Window Euclidean	-
Structure	-	Moving Window Patch size	Cluster size distribution

2.3.1. Local presence: Kappa statistics

Cell-by-cell map comparison methods consider two compared maps as a number of paired observations. Each paired observation consists of the classes found at one cell in the compared maps. Apart from direct overlap, spatial structure is not considered and the full information available to cell-by-cell methods can therefore be tabulated in the contingency table. This matrix tabulates for each pair of classes how often it occurs. Table 2 gives the generic form. It is well established as a cornerstone of accuracy assessment (Foody 2002).

An obvious metric of map correspondence is the fraction of agreement, which is the fraction of all observed pairs where the first and second class are identical. The cell-by-cell metric that is more often used however is Kappa. This statistic corrects the fraction of agreement for the fraction of agreement that can be expected given the number of cells of each class. For instance, consider two maps that are both 80% forest and 20% desert. A random spatial distribution of these quantities over the map would be at least 60% identical and the expected agreement of these maps is $0.8^2 + 0.2^2 = 0.68$.

Table 2. Contingency table

Map A \ Map B	1	2	$\cdots$	c	Sum
1	t_{11}	t_{12}		t_{1c}	t_{1+}
2	t_{21}	t_{22}	$\cdots$	t_{2c}	t_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c	t_{c1}	t_{c2}	$\cdots$	t_{cc}	t_{c+}
Sum	t_{+1}	t_{+2}	$\cdots$	t_{+c}	t_{++}

t_{ij} is the number of cells of class i in map A and class j in map B. t_{i+} is the number of cells of class i in map A. t_{+i} is the number of cells of class j in map B. t_{++} is the total number of cells.

The following equations express how Kappa is calculated from the contingency table:

$$\mathrm{K} = \frac{\mathrm{P}(\mathrm{A}) - \mathrm{P}(\mathrm{E})}{1 - \mathrm{P}(\mathrm{E})} \tag{1}$$

$$P(A) = \frac{1}{t_{++}} \sum_{i=1}^{c} t_{ii} \tag{2}$$

$$P(E) = \frac{1}{t_{++}} \sum_{i=1}^{c} t_{i+} t_{+i} \tag{3}$$

where *K* is Kappa, *P(A)* is fraction of agreement and *P(E)* is the expected fraction of agreement. A *K* of 0 corresponds to the expected level of agreement, identical maps get value 1 and the lowest possible score is -1. The Kappa statistic as originally introduced (Cohen 1960) is not intended as a map comparison metric, but as a general method of evaluating agreement of paired observations. Interestingly the same method was already introduced decades earlier for the purpose of comparing weather forecast maps (Heidke 1926). The metric is therefore also known as the Heidke Skill Score.

Further developments have extended the use of Kappa metrics; Monserud and Leemans (1992) calculate Kappa values for individual classes by temporary reclassifying the maps to binary maps. Fig. 1 shows two maps; the corresponding contingency table and Kappa statistics are given in tables 3 and 4. The Kappa results indicate that the two maps are most similar in terms of the class city and least similar for the class park, despite the fact that the contribution to the fraction of agreement of that class is very small. The fact that all cells of the class park are contained within the same (bottom right) region of the map is not recognized by the Kappa statistic, but will be by the focal statistics discussed in the following sections.

Table 3. Example contingency table

Map 1 \Map 2	Open	River	City	Park	Total Map 1
Open	1767	61	119	48	1995
City	32	92	18	8	150
River	154	1	96	0	251
Park	74	11	0	19	104
Total Map 2	2027	165	233	75	2500

Table 4. Example Kappa statistics

	Overall	Open	City	River	Park
P(A)	0.79	0.71	0.037	0.038	0.0076
P(E)	0.66	0.65	0.0040	0.0094	0.0012
Kappa	0.94	0.94	0.97	0.96	0.72

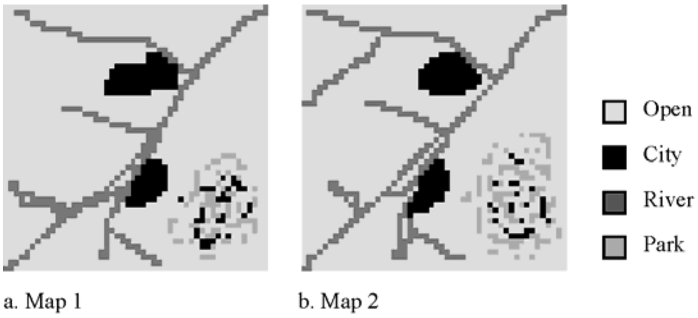


Fig. 1. Example pair of land use maps

2.3.2 Focal presence: Moving window Euclidean distance

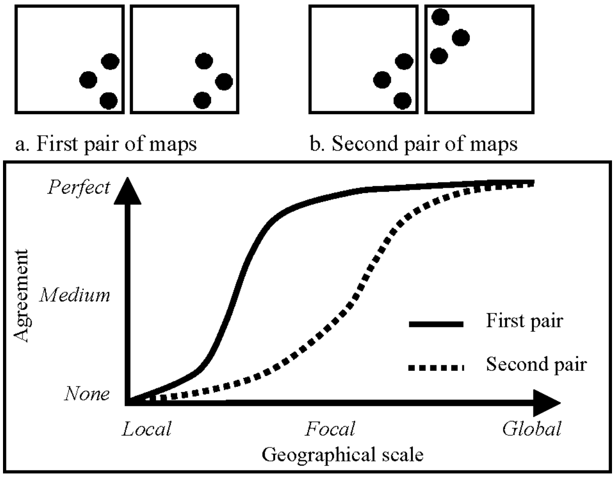
As the previous section explains, cell-by-cell comparison methods do not consider spatial structure except for direct overlap. Many geosimulation models however, including the model evaluated in this paper, are not expected to achieve such precision in their predictions. If the model achieves to place land use classes approximately in the right location, it can already be considered to perform well. With the purpose of considering such near agreement a moving window approach is followed that compares the composition of the window in both maps.

The idea of this method is count differences that are mitigated in the close neighbourhood only as small errors. In other words, an over-prediction of a class at one location can partially compensate the under-prediction of that same class at a close-by location. A larger window to detect such mitigating errors will lead to smaller errors. Fig. 2 shows two pairs of maps that in a cell-by-cell approach would be considered fully distinct (Kappa = -1), but with increasing window sizes more of the similarity between the two maps is recognized. When the moving window covers the whole map, the two maps are considered identical. The crucial point is that with increasing window size, the similarity of the first map is recognized earlier (or is stronger) than that of the second map. It is apparent that focal statistics provide an insight in the nature of the agreement of both pairs that neither local nor global statistics can provide.

The moving window Euclidean distance metric is calculated as the mean Euclidean distance on the basis of the proportions of each land use class over all cells in the map, as follows:

$$\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{j=1}^{c}\left(p_{i,j}^{A} - p_{i,j}^{B}\right)^2} \qquad (4)$$

where i iterates over all n cells on the map and j iterates over all c classes in the legend. E_i is the Euclidean distance of a moving window centered on cell i. $\bar{E}$ is the mean Euclidean distance. $p^{A}_{i,j}$ is the fraction of cells within a circular window around cell *i* that is in class *j* in map *A* (likewise for map *B*).

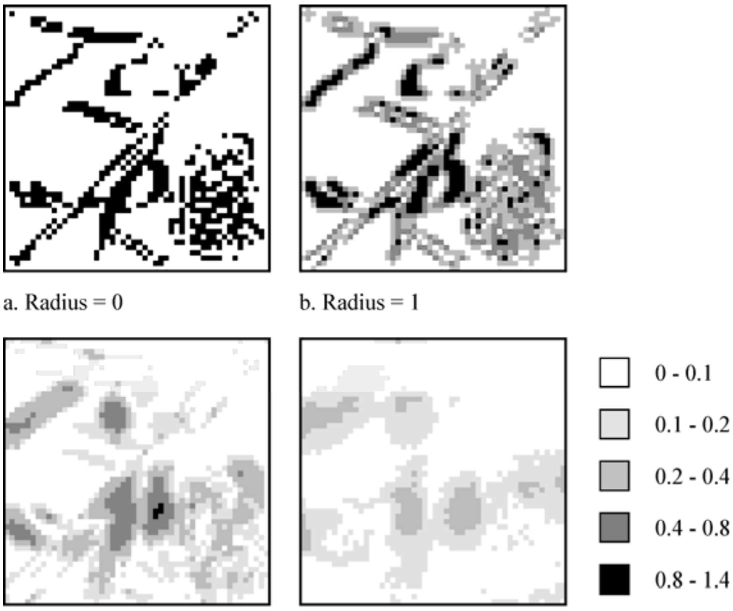


a. First pair of maps

b. Second pair of maps

c. Agreement as a function of scale

Fig. 2. Two pairs of maps that are both considered fully distinct from the local perspective of and identical from the global. Balanced analysis should find that the first pair has stronger correspondence (source: Hagen-Zanker 2006)



a. Radius = 0

b. Radius = 1

c. Radius = 3

d. Radius = 6

Fig. 3. Visualization of the Euclidean distance metric for the example maps. With increasing radius, Euclidean distance decreases and images blur more.

The values E_i can be visualized as a map, and like the Kappa metrics presented in the previous section these metrics can be calculated for individual classes as well as for all classes combined. Fig. 3 and table 5 present the results for the example introduced in fig. 1. The results illustrate how with increasing window sizes the errors reduce. Note that for the class park the error reduces most strongly with increasing radius, because that class is dominated by small errors.

Table 5. Example Euclidean distance statistics

Radius	Overall	Open	City	River	Park
0	0.30	0.28	0.074	0.16	0.080
1	0.21	0.19	0.065	0.12	0.049
3	0.13	0.12	0.050	0.074	0.028
6	0.081	0.072	0.035	0.044	0.020

2.3.3 Focal structure: Moving window patch size

Land use models are not only expected to place land use classes at (approximately) the right location, but also to arrange the classes in the right structure. As the strength of these models is to capture the processes underlying spatial morphology, they may even be expected to perform better in terms of structure than presence. There are many indicators of spatial structure. A commonly used metric, because of its straightforward interpretation, is patch size. A patch is also called a cluster and consists of all contiguous raster cells of one and the same class.

The focal comparison of spatial structure that is applied in this paper is based on two additional spatial (raster) layers that are derived from the original categorical maps. The value of a cell in the first layer is the size of the cluster that the cell belongs to. The second layer contains the weight for each cell, the weights are chosen such that only one category is considered at a time (i.e. cells of other categories have a weight of zero) and the total weight of all cells in a cluster is 1. The focal comparison of spatial structure is made by comparing the weighted average cluster size of the focal window of both maps, according to the following equations:

$$\bar{D} = \sqrt{\frac{\sum_{i=1}^{n} \left(D_i\right)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} \left(S_i^A - S_i^B\right)^2}{n}} \qquad (5)$$

$$S_i^A = \frac{\sum_{j=1}^{s} w_{i,j}^A s_{i,j}^A}{\sum_{j=1}^{s} w_{i,j}^A} \tag{6}$$

where i iterates over all cells in the map and j iterates over all cells in the window. $s^A{}_{i,j}$ is the size of the cluster that includes the j-th cell in the window centred around cell i in map A. $w^A{}_{i,j}$ is the weight associated to that cell, which is the inverse of the cluster size or zero if the cell is not taken in by the class for which the comparison is performed. D_i is the difference in cluster sizes at location i and $\overline{D}$ is the root mean squared difference over the whole map. D_i can be used to visualize the spatial distribution of differences over the map.

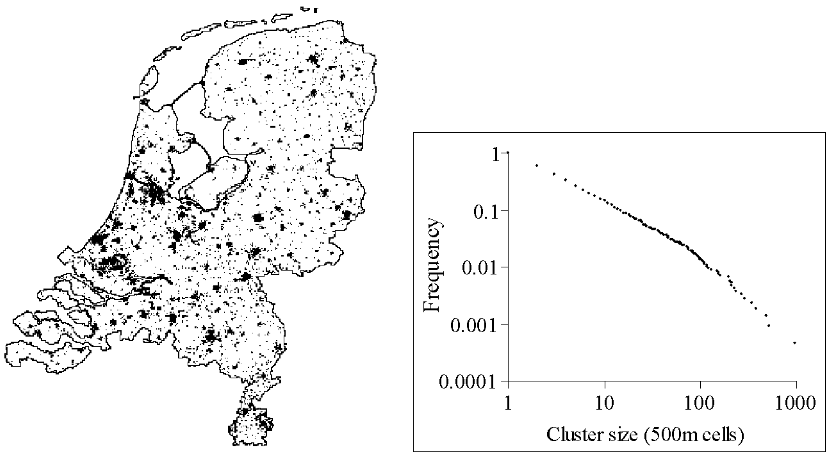
2.3.4 Global structure: Cluster size frequency distribution

At the global level many characteristics can describe spatial structure. For urban systems fractal metrics have been considered adequate (Batty and Longley 1994, de Keersmaecker et al. 2003). One particular aspect of spatial structure in which a fractal ordering becomes clear is the cluster size frequency distribution. In many urban systems a power law can be observed between the size of clusters and frequency of occurrence of clusters that size or smaller. This regularity has been utilized to describe urban patterns (Benguigui et al. 2006), simulate urban change patterns (Schweitzer and Steinbink 1997) and compare maps (Dungan 2006, White 2006). The global structure comparison applied in this paper compares cluster size frequency distributions of both the simulated and the real map. Fig. 4 illustrates the occurrence of a power law distribution for the case of urban clusters in the Netherlands.

Notwithstanding this example, cluster sizes of land use classes do not always adhere to power law distributions. The Kolmogorov-Smirnov (KS) distance which can be used to express the similarity between series of sampled data without assuming any distribution is therefore of great practical use. The KS-distance is the maximum difference in cumulative frequency between two sampled distributions:

$$D_{A,B} = \sup_x \left| F_A(X) - F_B(X) \right| \tag{7}$$

where $D_{A,B}$ is the KS distance of the two sampled distibutions A and B of variable x. In other words, the KS distance is the maximal vertical distance between two cumulative frequency distributions.



a. Urban areas in the Netherlands

b. Cluster size distribution

Fig. 4. Cluster size frequency distribution of urban clusters in the Netherlands. The straight line in the log-log plot indicates a power law distribution.

2.4 Reference levels

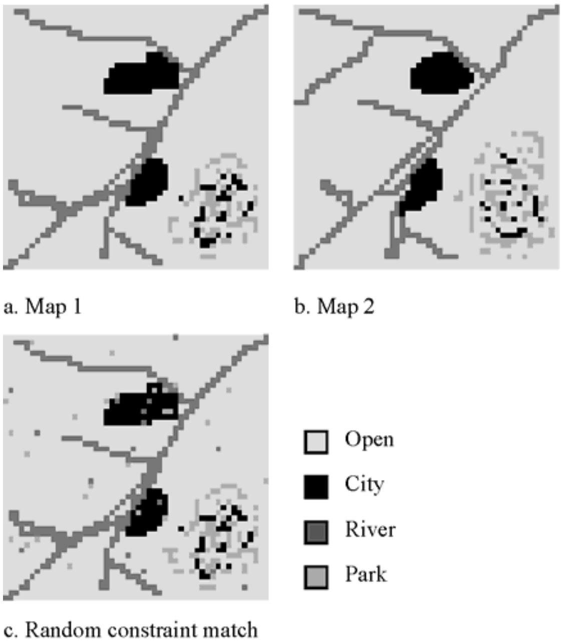
The methods described in the previous sections present a comprehensive overview of similarity between model and reality. It is not sufficient however to identify strengths and weaknesses of the model on the basis of these results, since the metrics are expressed at various scales. Moreover, not all of the registered similarity is a consequence of the performance of the model. In practice it appears that much of the similarity must be attributed to boundary conditions and constraints that are exogenously imposed on the model (Hagen 2003, Pontius Jr. et al. 2004, Hagen-Zanker and Lajoie 2008), most strikingly this is the case for models with an exogenously determined initial situation.

These two problems of interpretation are mitigated by introducing a reference model. This model is subject to the same constraints and boundary conditions as the tested model, but otherwise represents as little process as possible. The difference in performance between the tested model and the reference model, can then be attributed to the processes that are present in the tested, but not in the reference model. The reference model provides a common standard and individual results become mutually comparable.

The boundary conditions and constraints in the current case are formed by the initial situation and the total area constraints, i.e. the total area of each land use class is an input to the model. The reference model to observe these conditions and meet the constraints is the ‘random constraint match’ model.

The random constraint match model compares the number cells on the initial map to the area required by the constraints. Of those classes that are over-presented in the initial map, the surplus cells are selected randomly to change to another class. The classes that are under-represented are distributed randomly over the cells selected in the previous step.

Consider the earlier example whereby map A is now used as the initial situation and map B for the total area constraints. In map A there are 2027 cells of class open, in map B there are 1995. Therefore the class is over-represented in the initial situation by 2027 – 1995 = 32 cells. Likewise the class city is over-represented by 15 cells. The classes river and park are under-represented by respectively 18 and 29 cells. The random constraint match model initializes with map A. It then randomly selects 32 cells of class open and 15 cells of class city. These 47 selected cells are then, in random order, replaced by 18 cells of class river and 29 of class park. Fig. 5 shows the two maps and the derived results of the random constraint match.



a. Map 1

b. Map 2

c. Random constraint match

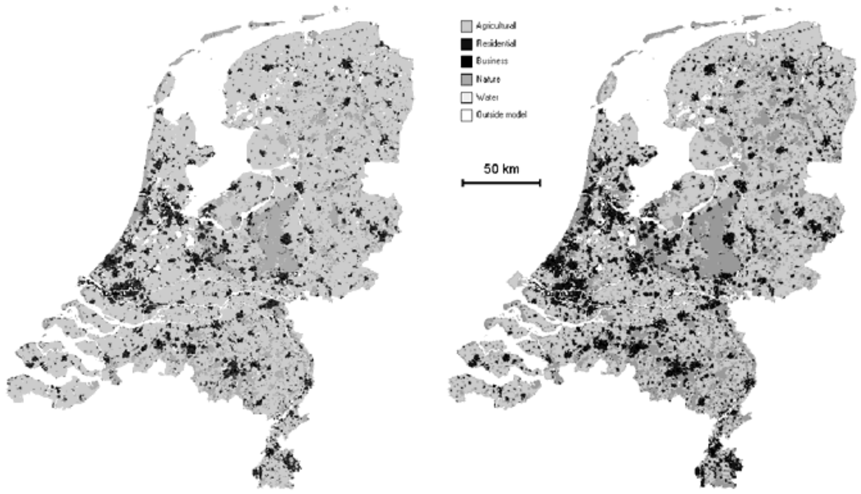
Fig. 5. Application of the random constraint match model where the total area constraints are posed by map 2 and the initial situation is that of map 1.

3. Application and results

The evaluated model is the Constrained Cellular Automata (CCA) land use model (White et al. 1997) as it is applied in the Environment Explorer spatial planning support system (Engelen et al. 2003). The system incorporates several other models besides the CCA, but the spatial distribution of land use classes is the responsibility of the CCA and that aspect is evaluated here.

The CCA model simulates land use change by year-by-year allocating land use classes to those cells for which they have the highest potential. The allocation process is constrained such that the exogenous area demand for each land use class is met. The potential of each location for different land use types is dynamic and changes over time as a function of the land use classes that are found in the neighbourhood of each location. Hereby a reciprocal relation comes about, where the potential layers are determined by the spatial distribution of land use classes and at the same time the spatial distribution of land use classes is determined by the potential layers. This mutual relation causes the complex and self-organizing behaviour of cellular automata models and is hypothesised to underlie the formation of urban morphology.

The potential layers are not only a function of the dynamic neighbourhood effect but are a composite measure of several layers. Besides the neighbourhood effect, these are the accessibility of the road network, the zoning status of cells, and the physical suitability of the land. Finally, a stochastic perturbation is included.



a. Land use in 2000

b. Model result for 2030

Fig. 6. Land use maps of the Netherlands; model and real.

For the calibration and validation tasks only limited land use data was available. Land use maps were available for 1989, 1993, 1996 and 2000. In an early stage it was decided that the period of calibration is 1989-1996 and validation 1996-2000. The consideration was that the calibration influences the performance of the model, whereas the validation only measures the performance. Better results can therefore be expected if the longer period is used for calibration. The intended period of application is 2000-2030, therefore a considerable gap remains between the duration over which the model is calibrated / validated and is applied.

The model distinguishes 15 land use classes. For the sake of model evaluation these are reduced to 5 main classes: agricultural, residential, business, water and nature. The model runs in steps of 1 year and the cell size is 500m. Fig. 6 shows the initial land use map and the final model result.

At the local level, cell-by-cell agreement is expressed by the Kappa statistics for the whole map and for individual land use classes (Table 6). Over the calibration period the agreement of the CCA model is better than that of the reference model for the classes agriculture, residential, urban and nature. Note that the remaining two classes, water and foreign, are assumed constant in time by the model. Therefore the distribution of these classes is identical in both the CCA land use model and the RCM model. Over the validation period the CCA model does not outperform the reference model, except for the class nature, where the difference is minimal.

The results of the focal presence comparison display (fig. 7) the same pattern as the Kappa results; over the calibration period the CCA land use model outperforms the RCM reference model, but over the calibration period it does not. With increasing radius of the moving window the errors of both the CCA and the RCM model decrease, but the relative position remains the same. Again the difference in performance of both models is small.

The results for focal structure, measured on the basis of patch size are more diverse (fig. 8). Like the earlier results, the CCA outperforms the reference model over the calibration period, but now also for the larger focal windows over the validation period.

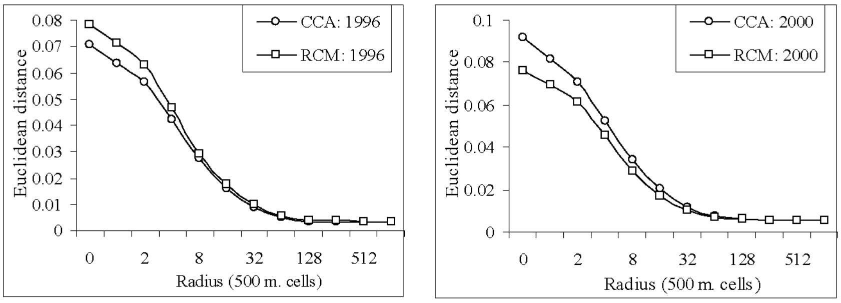
Table 6. Local presence comparison: Kappa statistics

	CCA: 1996	RCM: 1996	CCA: 2000	RCM: 2000
Overall	0.896	0.884	0.867	0.890
Agricultural	0.907	0.894	0.881	0.906
Residential	0.867	0.854	0.827	0.864
Business	0.769	0.766	0.731	0.809
Nature	0.924	0.908	0.907	0.904
Water	0.915	0.915	0.874	0.875
Foreign	0.960	0.960	0.949	0.949

CCA: Comparison between Constrained Cellular Automata model results and real data
RCM: Comparison between Random Constraint Match model results and real data

From the perspective of global structure, the CCA model proves to be more similar to reality than the reference model (fig. 9). The difference in cluster size distribution between the CCA model results and reality are substantially smaller than those of the RCM model. This is not only the case for all classes over the calibration period, but also for the classes residential and nature over the validation period. The class business shows little difference between the two models over the validation period. Agriculture is not included in these graphs because the patches of agriculture are so large that there are only a few of them on the map.

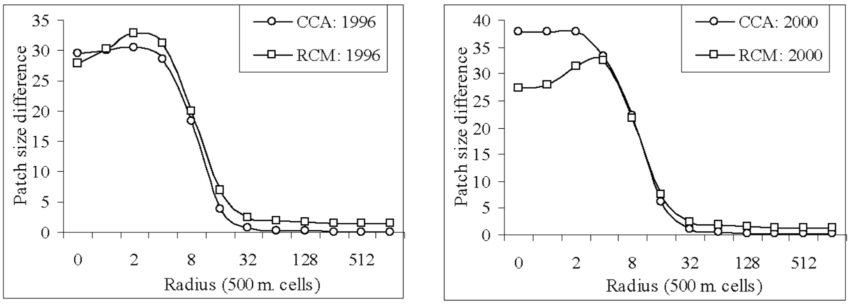
For the intended application period no data is available. The historical trend, however, is that the cluster size distribution is relatively stable in time. The CCA is better able to maintain a stable cluster size distribution.



a. Calibration period

b. Validation period

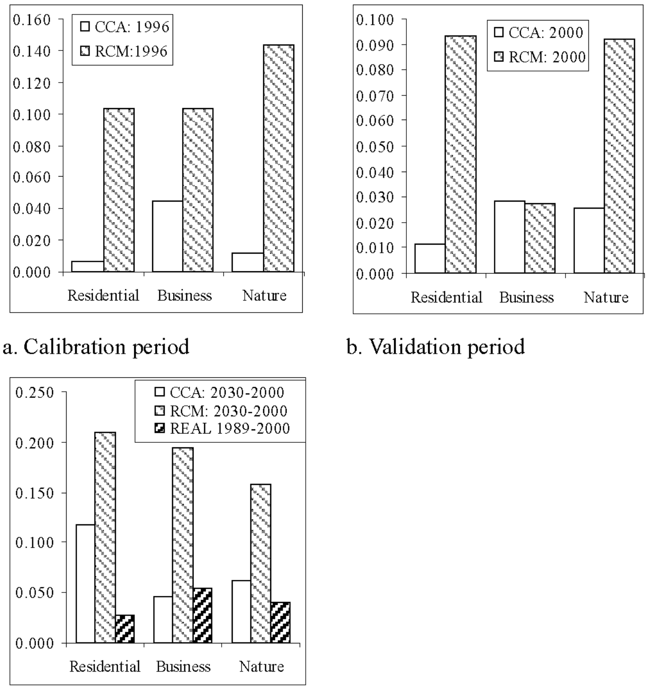
Fig. 7. Focal presence comparison: Euclidean distance



a. Calibration period

b. Validation period

Fig. 8. Focal structure comparison: Patch size



a. Calibration period

b. Validation period

c. Application period

Fig. 9. Cluster size frequency distribution, compared by Kolmogorov-Smirnov distance.

4. Discussion and conclusions

The data available for the model calibration and validation is less than ideal. Although the model is to be used for explorations of approximately 30 years into the future, data is available for a period that spans only over 11 years. Moreover, this short period is split into separate calibration and validation periods, leaving 7 years for calibration and 4 years of validation. This is of particular significance since the land use model (like most geosimulation models) is a dynamic model, and rather than simulating land use patterns it simulates land use change patterns. The short calibration period means that relatively little change takes place and thus that the calibration routine has little information to pick up on. As a result, the risk of over-calibration is very real. By over-calibration is meant that too many parameters are fitted to too little observations and rather than fitting to the general trend, the parameters are fitted to eccentricities in the data. Over the period 1989-2000 about 6% of all cells change state, it is unknown how many of these transitions are real changes and how many are data errors.

The comparison of the results between the calibration and validation period shows a decrease of performance over the validation period. This can always be expected and does not necessarily indicate over-calibration. At the local level results are better than the reference model over the calibration period, but not over the validation period. This is the strongest indication that over-calibration took place and the recommendation is made that future development of the model should aim to reduce the number of parameters and increase the amount of data by employing data available over longer time periods. There is a synergy between these two recommendations because the number of parameters can be reduced by considering less land use classes and if less land use classes are considered more data is available, also over longer periods. These recommendations have already been followed in a later project with good results (van Vliet 2006).

The verdict of 'over-calibrated' must not be too rashly made however. This conclusion is indeed suggested by the results at the local level and also at the focal level when presence is concerned. The results focal results of spatial structure, as well as the global results offer a more positive interpretation. For these results too, the performance (relative to the reference model) over the validation period is less than that of the calibration period, but over both periods the land use model outperforms the reference model. Therefore we can conclude that the strength of the model is in simulating urban structure at coarser scales rather than the precise (or approximate) land use class at particular locations. In summary, table 7 describes model performance according the two axes along which the performance metrics have been organized.

Table 7. Overview of model performance according to the two axes of performance criteria

	Local	Focal	Global
Presence	--	-	
Structure		+	++

-- poor performance, - mediocre performance, + acceptable performance, ++ good performance

Within the framework of the two axes many individual performance metrics can be applied. During this research we have applied more metrics than those reported here. In particular the Fuzzy Kappa metric (Hagen 2003) has been used to measure presence at the focal level. Structure at the focal level has also been measured in terms of edge density and fractal dimension (McGarigal et al. 2002). These results are not reported here, but do offer support for the summary presented in table 7.

All results, but most clearly, the focal metrics show a strong correlation between the performance of the evaluated model and the reference model. Apparently, a large proportion of model performance must be attributed to factors exogenous to the model. This demonstrates the need of the reference model, since without this reference any analysis of the performance criteria will be clouded by the impact of exogenous factors.

The contradiction in the results between the various performance metrics underlines the need for a multi-criteria analysis as applied in this paper. Focus on a single metric bears the risk of over-confidence in the results (in the current case if only structure metrics would be considered) or under-valuation of the model (in the current cast if only presence metrics would be considered).

We generally recommend this approach for all studies involving calibration and validation of geosimulation models. Also exciting recent developments towards rigorous sensitivity analysis of large numbers of simulations (Jantz and Goetz 2005, Kocabas and Dragicevic 2006) can gain in scope and relevance if these recommendations are followed.

References

Barredo JI, Demicheli L (2003) Urban sustainability in developing countries' megacities: modelling and predicting future urban growth in Lagos. Cities 20:297-310

Batty M, Longley P (1994) Fractal cities: a geometry of form and function. Academic Press Professional, Inc. San Diego, CA, USA

Batty M, Torrens PM (2005) Modelling and prediction in a complex world. Futures 37:745-766

Benenson I, Torrens PM (2004) Geosimulation: object-based modeling of urban phenomena. Comput Environ Urban Syst 28:1-8

Benenson I, Omer I, Hatna E (2002) Entity-based modeling of urban residential dynamics: the case of Yaffo, Tel Aviv. Environ Plan B: Plan Des 29:491–512

Benguigui L, Blumenfeld-Lieberthal E, Czamanksi D (2006) The dynamics of the Tel Aviv morphology. Environ Plan B: Plan Des 33:269 – 284

Briggs WM, Levine RA (1997) Wavelets and field forecast verification. Mon Weather Rev 125:1329-1341

Cohen J (1960) Coefficient of agreement for nominal scales. Educ Psychol Meas 20:37-46

Costanza R (1989) Model goodness of fit: a multiple resolution procedure. Ecol Model 47:199-215

de Keersmaecker ML, Frankhauser P, Thomas I (2003) Using fractal dimensions for characterizing intra-urban diversity: the example of Brussels. Geogr Anal 35:310-329

de Nijs TCM, de Niet R, Crommentuijn L (2004) Constructing land-use maps of the Netherlands in 2030. J Environ Manag 72:35-42

Dungan JL (2006) Focusing on feature-based differences in map comparison. J Geogr Syst 8:131-143

Engelen G, White R, de Nijs T (2003) Environment Explorer: spatial support system for the integrated assessment of socio-economic and environmental policies in the Netherlands. Integr Assess 4:97-105

Engelen G, White R, Uljee I, Drazan P (1995) Using cellular automata for integrated modelling of socio-environmental systems. Environ Monit Assess 34:203-214

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

Hagen-Zanker A (2006) Map comparison methods that simultaneously address overlap and structure. J Geogr Syst 8:165-185

Hagen-Zanker A, Lajoie G (2008) Neutral models of landscape change as benchmarks in the assessment of model performance. Landsc Urban Plan 86:284-296

Hagen A (2003) Fuzzy set approach to assessing similarity of categorical maps. Int J Geogr Inf Sci 17:235-249

Heidke P (1926) Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. Geogr Annal 8:301–349

Jantz CA, Goetz SJ (2005) Analysis of scale dependencies in an urban land-use-change model. Int J Geogr Inf Sci 19:217-241

Kocabas V, Dragicevic S (2006) Assessing cellular automata model behaviour using a sensitivity analysis approach. Comput Environ Urban Syst 30:921-953

Kok K, Farrow A, Veldkamp A, Verburg PH (2001) A method and application of multi-scale validation in spatial land use models. Agric Ecosyst Environ 85:223-238

McGarigal K, Cushman SA, Neel MC, Ene R (2002) FRAGSTATS: Spatial Pattern Analysis Program for Categorical Maps. University of Massachusetts. www.umass.edu/landeco/-research/fragstats/fragstats.html. Accessed 14 October 2008

Monserud RA, Leemans R (1992) Comparing global vegetation maps with the Kappa statistic. Ecol Model 62:275-293

Pontius Jr. RG (2000) Quantification error versus location error in comparison of categorical maps. Photogramm Eng Remote Sens 66:1011-1016

Pontius Jr. RG (2002) Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. Photogramm Eng Remote Sens 68:1041-1049

Pontius Jr. RG, Huffaker D, Denman K (2004) Useful techniques of validation for spatially explicit land-change models. Ecol Model 179:445-461

Pontius Jr. RG, Boersma W, Castella J-C, Clarke K, de Nijs T, Dietzel C, Zengqiang D, Fotsing E, Goldstein N, Kok K, Koomen E, Lippitt CD, McConnell W, Pijanowski B, Pithadia S, Sood AM, Sweeney S, Trung TN, Veldkamp AT, Verburg PH (2008) Comparing the input, output, and validation maps for several models of land change. Ann Reg Sci 42:11-37

Refsgaard JC, Henriksen HJ (2004) Modelling guidelines: terminology and guiding principles. Adv Water Resour 27:71-82

Schelling TC (1971) Dynamic models of segregation. J Math Sociology 1:143-186

Schweitzer F, Steinbink J (1997) Urban cluster growth: Analysis and computer simulations of urban aggregations. In: Schweitzer F (ed) Self-organization of complex structures: From individual to collective dynamics. Gordon & Breach London, pp 501-518

Takeyama M, Couclelis H (1997) Map dynamics: integrating cellular automata and GIS through geo-algebra. Int J Geogr Inf Sci 11:73-91

Turner MG (2005) Landscape ecology: What is the state of the science? Annu Rev Ecol Evol Syst 36:319-344

Turner MG, Costanza R, Sklar FH (1989) Methods to evaluate the performance of spatial simulation-models. Ecol Model 48:1-18

van Delden H, Engelen G (2006) Combining participatory approaches and modelling: lessons from two practical cases of policy support. International Environmental Modelling and Software Society. http://www.iemss.org/iemss2006/sessions/all.html. Accessed 14 October 2008

van Vliet J (2006) Validation of land use change models: a case study on the Environment Explorer. Universiteit Wageningen. http://www.lumos.info/publications-en.php. Accessed 14 October 2008

White R (2006) Pattern based map comparisons. J Geogr Syst 8:145-164

White R, Engelen G (2000) High-resolution integrated modelling of the spatial dynamics of urban and regional systems. Comput Environ Urban Syst 24:383-400

White R, Engelen G, Uljee I (1997) The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. Environ Plan B: Plan Des 24:323-343

White R, Engelen G, Uljee I, Lavalle C, Ehrlich D (2000) Developing an urban land use simulator for European cities. In: Fullerton E (ed) Proceedings of the 5-th EC-GIS Workshop held in Stresa, Italy 28-30 June 1999. European Commission, Joint Research Centre Ispra, pp 179-190