Planning and Land Use/Cover Scenarios: The Role of Probabilistic Algorithms

Jorge Rocha[1], Catarina Rodrigo[1], Cláudia Viana[1] and Ângela Barbosa[1]

[1]Centre for Geographical Studies, IGOT (Institute of Geography and Spatial Planning), Universidade de Lisboa, jorge.rocha@campus.ul.pt; crodrigo@campus.ul.pt; claudiaviana@campus.ul.pt; angelabarbosa@campus.ul.pt

Keywords: Probabilistic models, Prediction, Land Use/cover

1. Introduction

A cellular automata (CA) model is characterized by phase transitions that can generate complex patterns through simple transition rules. As such, this technique seems suited to model the complexity of urban systems (Clarke and Gaydos, 1998; Batty, 1995). Unlike most conventional urban models that focus more or less on the spatial patterns of urban growth, cellular automata based urban models usually pay more attention to simulating the dynamic process of urban development and defining the factors or rules driving the development. By applying different transition rules, a model based on cellular automata seeks to explore how the urban system has been developing and how this system changes under certain rules or forces.

The central component of a CA model is the transition rules that represent the logic of the process being modelled and, thus, determine the spatial dynamics of the system (White and Engelen, 2000). The transition rules define how changes the state of a cell in response to its current state and the states of its neighbours. This is the key component of CA because these rules represent the process of the system being modelled, and thus are essential to the success of a good modelling practice (White, 1998). For a restricted CA, the transitional rules are uniform and applied synchronously to all cells within the system. However, it has been pointed out a large number of different ways to define the transition rules.

The several approaches used to define transition rules, based on the understanding of the urban system and its evolution from different perspectives, resulted in different types of urban CA models. These approaches range from very simple to extremely complex. For example, in the diffusion limited aggregation (DLA) model developed by Batty a vacant cell just changes state (to occupied) if in its neighbourhood there is a occupied cell (Batty, Longley and Fortheringham, 1989). However, other urban CA based models combine different rules in order to simulate the complex behaviour of the system.

A popular approach used in cellular automata (CA) based urban modelling is to incorporate other modelling methodology into the model, especially for defining the transition rules for it. In the effort of using various mathematical approaches to configure the transition rules of a cellular automaton, it is obvious that the primary purpose of employing mathematical approaches in the modelling practice is to evaluate the adequability or probability of land for development. Therefore, this type of model is also regarded as a adequability based CA model (Li and Yeh, 2000).

In our CA urban model, we simulate different scenarios of urban growth to 2021. Our approach is to use different trends (last 2, 8, 10 and 13 years) with and without restrictions and, most important, to apply several different methods for transition rules definition. The methods test site was the Almada municipality in the South bank of Tagus River, near Lisbon (Portugal).

The mathematical models, i.e. submodels, used to define transition rules include artificial neural networks (supervised [MLP] and unsupervised [SOFM and ArtMap]), genetic algorithms (GARP and GARPOM), supported vector machines (SVM), decision trees (CRT and CHAID), generalized linear models (GLM), generalized additive models (GAM), maximum entropy (MaxEnt) and geometric distance models (Euclidian, Manhattan and Chebychev).

2. Material and methods

The study area, use to test our methodology, is the Almada municipality (Figure 1), belonging to the Lisbon Metropolitan Area. The choice of this municipality is due to diverse factors of great importance, such as: is near the capital (Lisbon); enjoys of increased accessibility since the construction of the Tagus bridge; as the possibility of urban expansion and rare natural potentialities and, between 1967 and 1986, the territory attended to a unmeasured urban pressure, resulting in a irreversible ground occupation in areas of strong physical restrictions. Between 1991, 2001 and 2011 (dates corresponding to the population census) its population did not stop to grow, even that in differentiated rhythms. The outward job dependence is a reality that the in-out movements between Lisbon and Almada corroborate.

Figure 1. Geographic location of Almada Municipality.

The proposed evaluation method combines cellular automata, with Markov chain analysis and a wide set of submodels, relying on six phases: i) data pre-selection; ii) data normalization; iii) spatial autocorrelation and multicollinearity analysis; iv) transition rules integration; v) Markov time series trends allowing predict future uses; and vi) simulation of future scenarios. In this point we will discuss all points except vi), that will be further analysed in the discussions of the submodels simulation performance.

2.1 Data pre-selection

One of the first things to consider in a model is the selection of the variables to be used. The Geographic Information systems (GIS) allow generating and manipulating information in a way that would not be possible otherwise. As the acquisition methods used, 3 types of variables are produced: (i) variables that result from vectorization of simple themes, e.g. qualitative variables such as hydrological network, road network and land use/cover; (ii) variables resulting from spatial interpolation techniques, e.g. digital terrain model; and (iii) variables resulting from processes of local, focal and zonal operations e.g. slopes and densities. The first type is typically handled in a vector environment, whereas the latter are exclusively handled in raster environments. The proposed model is based on a raster structure, adjusting to the operational characteristics of classic CA models.

There have been developed important efforts to improve the understanding of the mechanisms of production and reproduction of the urban space, as in the case of differentiated development-related positions (Harvey, 1975; 1989), auto-organization (Holland, 1998; Portugali, 1997; 2000) and the study of complex systems (Allen, 1997; O'Sullivan, 2001). Following these experiments and the work of Rocha et al. (2007) we identify the factors of greater importance on explaining urban growth in Almada (Table 1).

Table 1. Pre-selected variables.

Variable	Type
Slope	Quantitative (%)
Road Network Intersections	Quantitative (n"/km2)
Road Networksit Den	Quantitative (/km2
Distance to Road Network	Quantitative (m)
Distance to Urban Areas	(Quantitative
Accessibility to Lisbon	Quantitative (m)
Accessibility to Almada	Quantitative (m)
Accessibility to Caparica	Quantitative (m)
Territorial Attraction Index	Quantitative (adimensional)
Territorial Valorisation Index	Quantitative (adimensional)
Built Up Dynamics	Quantitative (construction licences/ km2)
National Ecological Reserve (REN)	Qualitative
National Agricultural Reserve (RAN)	Qualitative
Planning Master Plan (PDM)	Qualitative

2.2 Data normalization

The input data is developed using a set of transition rules that quantify the spatial effects that prediction cells have on land use/cover changes (Pijanowski et al., 2000). During this conversion process, the cells are encoded to represent constraints (restrictions) or probability of occurrence (factors). In the first case represent binary masks, where the zero (0) represents the absence (or denial) and one (1) presence (or acceptance), and in the second case are continuous variables, normally ranging from 0 (0% of adequability) to 1 (100% of adequability).

To make our data match this 0-1 scale we need normalize (standardise) it. We use fuzzy membership functions (Zadeh, 1965). on the standardisation of the variables. Identify exactly which function to use, depends on the knowledge and understanding of the relationship between the criteria and the set of decisions, as well as the availability of information to infer the degree of membership. In the vast majority of cases, the sigmoidal or linear functions are appropriate to the modelling of reality. The sigmoidal function was applied to the factors which constitute continuous (quantitative) variables and linear function to which correspond to discrete (qualitative or categorical) values.

2.3. Spatial autocorrelation and multicollinearity

Some of the common problems that models contain are the autocorrelation and multicollinearity. Statistically, the autocorrelation is a measure that reflects how much the value of a random variable is able to influence its neighbours. For example, how the existence of a higher value influences also high values in is neighbourhood. On the other hand, multicollinearity is a common problem in models where the independent variables have exact (or approximately exact) linear relationships.

The spatial autocorrelation can be measured in different ways. The Moran index is the most widespread statistic and measures the spatial autocorrelation from the product of the standard deviations. This index is a measure of global autocorrelation, as it indicates the degree of spatial association present in the data set.

If the covariance between contiguous cells is null, the neighbourhood has no influence on the values of the variable. In this case, there is no spatial autocorrelation. If the neighbours resemble each other more than the distant cells, then there is a positive spatial autocorrelation, i.e. clustering. If the distant cells resemble more than the neighbours then there is a negative spatial autocorrelation, i.e. sprawl.

In the present we choose to use two types of neighbourhood in the spatial autocorrelation calculation: Moore and Von Neuman (Table 2). Of this analysis, we conclude that all factors have acceptable values of autocorrelation, regardless of the type of neighbourhood adopted in calculations.

Table 2. Spatial autocorrelation.

Layer	Moore	Von Neuman
Accessibility to Almada	0.9822	1.0133
Accessibility to Caparica	0.9844	1.0313
Accessibility to Lisbon	1.0055	1.0450
Territorial Attraction Index	0.9195	0.9228
Slope	0.8570	0.8736
Distance to Urban Areas	0.9817	1.0227
Distance to Road Network	0.9836	1.0194
Territorial Valorisation Index	0.9532	1.0340
Road Network Intersections	0.9949	1.0294
Road Network Density	0.9937	1.0206
Built-up Dynamics	0.9841	1.0043
Planning Master Plan (PDM)	1.0049	1.0358
National Agricultural Reserve (RAN)	1.0049	1.0382
National Ecological Reserve (REN)	1.0049	1.0382

The clearest indication of the existence of multicollinearity is when the R2 is quite high, but none of the coefficients are statistically significant according to conventional t-statistics. The consequence of multicollinearity is high standard errors in the case of moderate or severe multicollinearity and even the impossibility of any estimation if the multicollinearity is perfect. The absence of multicollinearity is one of the premises to establish a correct model. However, some authors state that it is not a serious problem if the objective of the model is the prediction.

Thus, the choice of the variables to use in the model should take into consideration their possible multicollinearity (Guisan and Thuiller, 2005). This can be detected, among other ways, through the correlation matrix. Other techniques used are the tolerance value or its inverse, called the variance inflation factor (VIF), whose high values indicate the existence of collinearity:

$$VIF \quad \frac{1}{1 \quad R_j^2} \tag{1}$$

where R_j comes from x_j regression with the other variables, i.e. the coefficient of determination of the regression equation.

It is suggested, when collinearity occurs, that those variables are removed from the analysis. In Table 3 we can see that all the factors have values within acceptable parameters, and the only one that stands out for having a higher value is Accessibility to Lisbon. Crossing the values of R2 and VIF, we left out two factors: Accessibility to Almada and Road Network Density. In both cases there is a strong correlation with Accessibility to Lisbon and the Road Network Intersections, respectively. However,

despite the latter having higher values in relation to the first, they are chosen in virtue of the analysis of in out trips still emphasize the importance of movements to Lisbon and the nodes (intersections) represent a real access to road network which does not happen with the routes, e.g. highway access is only possible in some intersections. In addition, although with higher values, both parameters are within acceptable parameters.

Table 3. Multicollinearity values.

	Factor	R2	VIF
1	Accessibility to Almada	0.376923	1.604938
2	Accessibility to Caparica	0.245495	1.325373
3	Accessibility to Lisbon	0.761179	4.187239
4	Territorial Attraction Index	0.077739	1.084292
5	Slope	0.197143	1.245552
6	Distance to Urban Areas	0.473995	1.901123
7	Distance to Road Network	0.287069	1.402661
8	Territorial Valorisation Index	0.262954	1.356768
9	Road Network Intersections	0.588384	2.429452
10	Road Network Density	0.321538	1.473923
11	Built Up Dynamics	0.564097	2.294087
12	Planning Master Plan (PDM)	0.234892	1.307005
13	National Agricultural Reserve (RAN)	0.141907	1.165375
14	National Ecological Reserve (REN)	0.445719	1.804140

2.4. Transition rules integration

A problem of CA models is determining the weights to assign to each factor. The solo simulation involves the use of several spatial variables. The contribution of each of these variables for the simulation is quantified by a weight, or parameter. The value of these parameters has a important effect on the results of the simulation, i.e. different combinations of values lead to totally different urban forms.

In most situations it is necessary to calibrate the CA model in order to ensure that the simulation generates results close to reality. This calibration is extremely difficult to accomplish. There are two main types of calibration processes: those based on trial-and-error approaches and those based on mathematical/statistical methods. In the first case it is not necessary to use rigid mathematical methods. A simple method is comparing visually the results of simulations made with different combinations of weights. However, when there are many variables it is difficult to define the combinations and access the results visually, since the generated patterns can be very complex.

In the second case, we refer to a fairly common approach in urban modelling with cellular automata, which is the incorporation of other modelling methodologies in the model, with a special focus on the integration of their transition rules, i.e. predictive submodels. The most used mathematical models include the multicriteria evaluation (Wu and Webster, 1998; Wu, 1995), in particular together with Analytic Hierarchy Process (AHP) developed by Saaty (1980, 1986), multiple regression (Sui and Zeng, 2001) and principal component analysis (Li and Yeh, 2002).

In an effort to use various mathematical approaches to configure CA transition rules, normally becomes obvious that the primary purpose of the application of these approaches is the evaluation of their

adjustment to urban develops adequability. Therefore, this type of model is also known as adequability CA model (Li and Yeh, 2000). From here we can see that there is a high and diverse number of statistical integration models can be used in modelling (Guisan and Zimermann, 2000). Their role is to obtain response values that translate the adequability through the combination of the independent variables values relating to the known distribution of the phenomenon.

Taking into account this large variability, the choice of a single method implies the loss of several potentialities present in others. In addition, the holding of several methods simultaneously is suggested on modelling as a way to decrease the uncertainty of the process (Pearson et al., 2006). Considering these prerogatives and the potential gain in predictive performance, we opted for the use of several different models.

The applied methods include maximum entropy (MAXENT) (Jaynes, 1957; Philips et al., 2006), generalized linear model (GLM) (Guisan et al. 2002), generalized additive model (GAM) (Hastie and Tibshirani, 1990), support vector machine (SVM) (Vapnik, 1995), Chi-squared Automatic Interaction Detection (CHAID) (Kass, 1980; Thuiller, 2003) and Classification and Regression Trees (CRT/CART) (Breiman et al., 1984; Thuiller, 2003), artificial neural networks (supervised [multilayer perceptron MLP ((Rumelhart, Hinton and Williams, 1986)] and unsupervised [self organizing feature maps SOFM (KOHONEN, 1993) and adaptive resonance theory ARTMAP (Carpenter and Grossberg, 1987)]), genetic algorithms (genetic algorithm for rule set production - [GARP] and genetic algorithm for rule Set production open modeller [GARPOM]) (Stockwell and Peters, 1999) and geometric distance models (Euclidian, Manhattan and Chebychev) (Tsoar et al., 2007).

2.5. Time series trends

One of the key components of CA models is time, i.e. the change of a cell state at a given time is controlled by the state of the cell itself and the state of his neighbours in the previous step. Thus, the model needs to be configured not only spatially as well as temporally. However, the current practices of urban CA based modelling focus essentially the spatial dimension of urban land-use change. There is little concern with the temporal dimension of the model.

The Markov chain is a mathematical model that describes a certain type of process that moves in a sequence of steps and through a set of States (Baker, 1989). The attractiveness of the Markov chain lies in the fact that the model parameters are easily estimated. The transition probabilities can be estimated statistically from a sample of transitions that occurred during a certain time interval. In this way, the Markov chain only requires the establishment of a finite number of states and that transition probabilities are known. Time steps correspond to land use/cover binary maps (urban/non-urban) obtained by photo-interpretation (Figure 2). Depict having a database from 1960 to 2008 we chose to use just the data from 1995 and later, because this maps were made over ortophotomaps and have a higher positional accuracy than the other (based on aerial photograph).

The Markov chain analysis generated two tables with, respectively, the transition land use areas (Table 4) and the transition probabilities (Table 5), for 4 periods.

The 4 periods considered were of 13 (1995-2008), 10 (1998-2008), 8 (2000-2008) and 2 (2006-2008) years. Based on the values obtained we can then predict the probability of land use transitions to 2021 based on different growth trends, i.e. different periods (4).

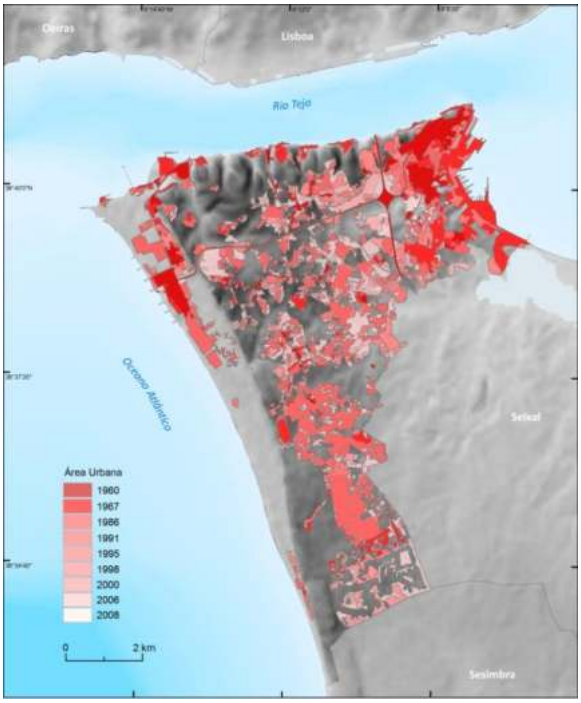

Figure 2. Land use/cover from 1960 to 2008.

Table 4. Transition areas (km2).

	1995 – 2008		1998 – 2008		2000 – 2008		2006 – 20	
	Non-urban	Urban	Non-urban	Urban	Non-urban	Urban	Non-urban	Ur
Non-urban	25.9464	8.1779	25.2250	8.8993	25.0907	9.0335	32.1285	
Urban	0.0000	34.4108	0.0000	34.4108	0.0000	34.4108	0.0000	

Table 5. Transition probabilities.

	1995 – 2008		1998 – 2008		2000 – 2008		2006 – 2008	
	Non-urban	Urban	Non-urban	Urban	Non-urban	Urban	Non-urban	Urb
Non-urban	0.7604	0.2396	0.7392	0.2608	0.7353	0.2647	0.9415	
Urban	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1

3. Submodels Validation

A receiver operating characteristic (ROC), or simply ROC curve, provides tools to select possibly optimal models and to discard suboptimal ones. Because ROC graphical plots dissociate classifiers performances through the asymmetry of class and cost errors, they have advantages over other evaluation measures.

The usual approach is to establish a threshold or cut-off value in order to split the response variable in dichotomous classes. The value of the threshold is usually set equal to the a priori occurrence probabilities of the phenomenon under study. However, the definition of a threshold is more a matter of using the model than a feature of the model itself. For this reason, it is better to use threshold-independent methods.

Usually ROC curves are used to generate summary statistics. One of the most commons is the area under the ROC curve (AUROC), or just area under curve (AUC). When using normalized fuzzy maps, the AUC is equal to the probability that a classifier will rank a randomly chosen high instance (e.g. 0.8) above a lower randomly chosen one (e.g. 0.2) (Fawcett, 2006).

Table 6 shows the AUC values for all submodels tested against 2008 land use. The results are globally satisfactory ranging from 0.61 to 0.85. It is interesting to note that distance based models, despite being conceptually simpler, are among those who get the best results. This data can be important in a future dissemination by non-specialized users, where the simplification of procedures takes on greater importance.

Understanding that all models have their virtues but also some drawbacks, but being unable to work with all of them, we select those who demonstrated a performance rated as good. So, henceforth our analysis will focus the generalized linear model (GLM), the generalized additive model (GAM), the model of maximum entropy (MaxEnt), the two genetic algorithms (GARP and GARPOM) and the model based on Manhattan distance (Manha).

We should bear in mind that all considerations concerning the parameters obtained by the models are valid only for the range of values (presences and absences) used. As such, these relationships between variables can be widely different, arising from the use of different samples, e.g. for a more extensive study area. All selected submodels, including less performance ones (GARP and Manha), can be used in conjunction, in order to better characterize the potential existence of urban areas. This method consists in the addition of consistent areas of all the six models, using as a limit of adequability the AUC value that maximizes the results of each one, i.e. weighted average. This procedure is relatively similar to that proposed by Araujo and New (2006), for the reduction of uncertainty in predictive modelling.

Table 6. Submodels AUC values.

Submodel	AUC	Accuracy	Submodel	AUC	Accuracy
GLM	0.851	Good	MLP 1 hidden layer 12 neurons	0.787	Reasonable
MAXENT	0.838		MLP 2 hidden layers 12 neurons	0.783	
GARPOM	0.820		MLP 2 hidden layers 4 neurons	0.782	
GAM	0.816		MLP 1 hidden layer – 8 neurons	0.779	
GARP	0.811		ARTMAP	0.771	
MANHATTAN (MANHA)	0.802		CHAID	0.768	
CHEBYCHEV	0.792	Reasonable	CRT	0.767	
EUCLIDIAN	0.792		SVM100	0.757	
MLP 2 hidden layers – 8 neurons	0.788		SVM200	0.725	
MLP 1 hidden layer 4 neurons	0.787		SOFM	0.613	Weak

4. Submodels Comparison

Analysing submodel-by-submodel (Table 7), it may be noted that, on average, the GLM, GAM and Manha, favour the probability of occurrence of urban use. However, the Manha has a minimum adequability of 0.4941, which does not correspond to reality. This is due to the use of pseudo absences that may not be true negatives. The GAM and the genetic algorithms (GARP and GARPOM) have the higher standard deviation values.

Table 7. Submodels descriptive statistics.

	N	Average	Minimum	Maximum	Standard deviation
MANHA	2729908	0.8295	0.4941	0.9412	0.0653
GAM	2729908	0.8336	0.0000	1.0000	0.2369
GARP	2729908	0.5824	0.0000	1.0000	0.3683
GARPOM	2729908	0.6716	0.0000	1.0000	0.3977
GLM	2729908	0.8557	0.0000	1.0000	0.1870
MAXENT	2729908	0.3292	0.0039	0.8510	0.1968
AVG6M	2729908	0.6836	0.0000	0.9412	0.2180

Comparing the submodels with each other (Table 8), we can find that they are relatively related, presenting correlation values that oscillate between 0.60 (GLM GAM) and 0.96 (AVG6M GARPOM). The AVG6M has the greater correlation with all the others (between 0.96 and 0.79), which goes alongside the idea of its use.

To compare the variances of the different submodels we used F-ratio statistic. The F-ratio is used to determine whether the variances in two independent submodels are equal. If the F-ratio is not statistically significant, we can assume that there is homogeneity of variance. This value is obtained by dividing the largest variance for the smallest. If the F-ratio exceeds the critical value, i.e. there is no difference between the variance of the submodels, the null hypothesis is rejected.

Table 8. Pearson correlation coefficient between submodels.

	MANHA	GAM	GARP	GARPOM	GLM	MAXENT	AVG6M
MANHA	1.00						
GAM	0.76	1.00					
GARP	0.81	0.70	1.00				
GARPOM	0.87	0.66	0.86	1.00			
GLM	0.84	0.60	0.68	0.85	1.00		
MAXENT	0.77	0.61	0.80	0.83	0.79	1.00	
AVG6M	0.92	0.79	0.93	0.96	0.86	0.89	1.00

In our case, because the degrees of freedom are always greater than 1000 (>>120), the sample is statistically regarded as . Thus, regardless of the degree of confidence selected (0.10; 0.05; 0.25) the threshold corresponds to 1. The F-ratio values (Table 9) always exceed this value (1), so there is a statistically significant difference between the variances of the models. Despite all, the AVG6M is the one that presents values closer to the threshold, namely 1.36 (with GLM), 1.23 (with MaxEnt) and 1.18 (with GAM). A visual comparison can give the notion of areas where the one submodel (column) has higher (green), equal (yellow), or lower (red) prediction values than another submodel (line).

Table 9. F-ratio statistic for inter-model variances.

	MANHA	GAM	GARP	GARPOM	GLM	MAXENT
GAM	13.146					
GARP	31.765	2.416				
GARPOM	37.052	2.818	1.166			
GLM	8.186	1.606	3.880	4.526		
MAXENT	9.072	1.449	3.501	4.084	1.108	
AVG6M	11.136	1.181	2.853	3.327	1.360	1.227

To quantify the relationship of each submodel with all others at the same time, we carry out a multiple linear regression (MLR). In table 10 we may find the results of MLR calculations. The diagonal values correspond to the value of the line intersection and the other values match the relationship of each submodel with the one that is being analysed. Here we can see that the AVG6M is a valid option, showing values of R and R2, near 0.999 (higher values among all submodels tested). The residuals, i.e. difference between the observed and expected values, for all the regressions carried out. The areas where the values are higher are those where the worst model represents all others (the green represents positive values and negative red).

Table 10. Multiple linear regression parameters by comparing a submodel (dependente) with all others (independent).

	MANHA	GAM	GARP	GARPOM	GLM	MAXENT	AVG6M	R^2	R
MANHA	91.7450	0.3942	0.4051	0.4084	0.3654	0.4349	2.5813		
GAM	0.8427	8.8040	0.9466	0.9514	0.9878	0.9449	5.7254		
GARP	0.8862	0.9688	4.6950	0.9386	1.0261	0.9371	5.7939		
GARPOM	0.8965	0.9770	0.9417	9.8620	0.9590	0.9455	5.7766		
GLM	0.6389	0.8711	0.8841	0.8236	9.1759	0.8206	5.1604		
MAXENT	0.9189	0.9339	0.9050	0.9101	0.9198	1.0449	5.5463		
AVG6M	0.1623	0.1684	0.1665	0.1655	0.1721	0.1651	0.3694	0.999	

5. Results Analysis

The previously discussed 7 submodels calculated served as basis to simulate future scenarios of urban occupation using a 5x5 extended Von Neuman neighbourhood. As the largest gap between records of land use/cover is 13 years, i.e. 1995 2008, it is not advisable to create simulations beyond 13 years into the future, which based on the year of the last record (2008), limits the time horizon to 2021. However, it is common knowledge that the simulation models perform better for small periods than for more extended ones. So what can be done to decrease the risk of developing a simulation so ahead in time? The solution can be easily explained by an analogy to golf training.

The golfer shoots several times the ball toward the hole and the result will be a set of balls around the hole (assuming that it is difficult to get it right in the hole at the first time, as a model does not predict exactly a future situation). Some of those balls fall on left side of the hole, others on the right, still others ahead and so on, forming a cloud of balls. Statistically, there is a high probability of the hole location being in the geometric centre of the balls distribution. This probability increases as the number of balls raises and, consequently, their density increases.

With models the situation is the same. Despite very often the purpose of the model is not to accurately predict a situation, but rather to explain the influence of factors in a given system and make it possible to identify how the system reacts to changes in the parameters, its power of simulation can be greatly increased through the creation of several alternative scenarios. This means that creating various scenarios of evolution of urban area to 2021, assures that there is a much higher probability of find the future real situation between them.

Hence, the importance of the time horizons considered (4 in total). As inferred from the land use/cover evolution changes in Almada municipality, the evolution of urban area did not happen always in the same way in space and, especially, in time. So, it was decided to adopt a set of four situations ranging from the overall trend of the last decade, e.g. last 13 years (1995-2008) and 10 years (1998-2008), to the

latest trends, e.g. 8 years (2000-2008) and 2 years (2006-2008). From the combination of all the elements, we have established 7 predictive submodels and 4 distinct trends of urban growth, originating a 7x4 matrix.

This matrix was further subjected to two distinct approaches. First, the predictive submodels were introduced without changes, i.e. fully consisting of factors. Remember that in the elaboration of these submodels, factors such as REN and RAN, were considered as highly conditioning, but not restrictive of urban use. That leads to 28 urban occupation scenarios for 2021.In the second approach, REN and RAN were used as binary masks, where the existence of the factor would correspond to zero, i.e. inability to build and the absence of a factor, i.e. possibility to build. Therefore, these two sets of information were considered as constraints in all submodels by multiplying the firsts by the seconds. In this way, we fully prevent urban sprawl within areas of REN and/or RAN in the simulations carried out. This second approach resulted in another set of 28 scenarios of urban area in 2021.

In a more quantitative analysis (Table 11) we can observe that the simulations follow the trends of land use/cover change. So, as the last two years (2006-2008) were of urban growth strong slowdown, the simulation based on this rate of growth is the one that generates less urban area for 2021, with values ranging between 35.866 km2 (AVG6M) and 35.869 km2 (Maxent). The most fruitful period is the 8 years one, spanning from 2000 to 2008. The predicted urban are for 2021, according to the 8 years trend, oscillates between 42.74 km2 of GAM and AVG6M, and the 42.751 km2 of manha. As the simulation based on the growth of 2006 to 2008 predicts a small increase in the urban area, it is understood that the 8 years dynamic is due essentially to the situation between 2000 and 2006.

Table 11. Quantitative results of the simulations (UR means unrestricted and RE means restricted).

	2 years trend		8 years trend		10 years trend		13 years tre	
	UR	RE	UR	RE	UR	RE	UR	RE
Edmanha	35.867	35.869	42.743	42.751	42.613	42.620	41.910	41.913
GAM	35.868	35.868	42.740	42.744	42.608	42.613	41.905	41.908
GARP	35.872	35.871	42.745	42.742	42.614	42.610	41.908	41.908
GARPOM	35.869	35.869	42.744	42.746	42.612	42.615	41.908	41.912
GLM	35.867	35.867	42.742	42.746	42.613	42.616	41.909	41.913
Maxent	35.877	35.872	42.742	42.744	42.610	42.613	41.906	41.910
AVG6M	35.866	35.870	42.740	42.743	42.609	42.613	41.903	41.911

The two next simulations, give predictions of successively smaller urban area. Indeed, considering the trend of the last decade (1998-20008) the urban area in 2021 would oscillate between 42.608 km2 in the case of GAM and 42.62 km2 to the manha. Finally, following the trend of the last 13 years (1995-2008) we have a decrease in growth trend, standing just above the trend of the last two years, although with values much closer to the other two simulations (10 and 8 years) than this last (2 years). The trend of the last 13 years varies between 41.903 km2 of urban area based on AVG6M and the 41.913 km2 generated by manha. In short, we can build three groups of scenarios, one with a fairly moderate growth (trend of the last two years), a steep growth (trend of the past eight years), and an intermediate situation, although tending to be closer to the larger dynamic situations (the last ten and thirteen years).

Still in a global analysis, immediate upraises the fact that, perhaps paradoxically at first glance, the simulations with restrictions predict greatest amount of urban area than the ones without restrictions (Figure 3). The only exception is the simulation based on biennial trend. This was due to the fact that with 2 years of trend, spatial relations are still dull and the time window is too small to model a complex system, as the urban. In this way, the linear functions prevail and the built-up area expands, in this context with logic, more easily than by the more restricted space.

In the simulations based on more years, the interactions between microelements increase, coming into sight non-linear behaviours and the emergence of new behaviours. This forced the urban sprawl to follow alternate routes, because of the restrictions imposed, finding new realities that encourage new behaviours, which in turn enables the discovery of new routes and so on.

In a submodel level analysis it can be seen that in the simulation based on 2 years trend the submodel that more urban area generates is the maximum entropy (MAXENT). The entropy is a measure of the amount of uncertainty in a probability distribution or a system subject to constraints. Entropy measures the number of different microstates that may arise in a particular macrostate. The term has been used as the basis for the maximization of entropy models of spatial interaction.

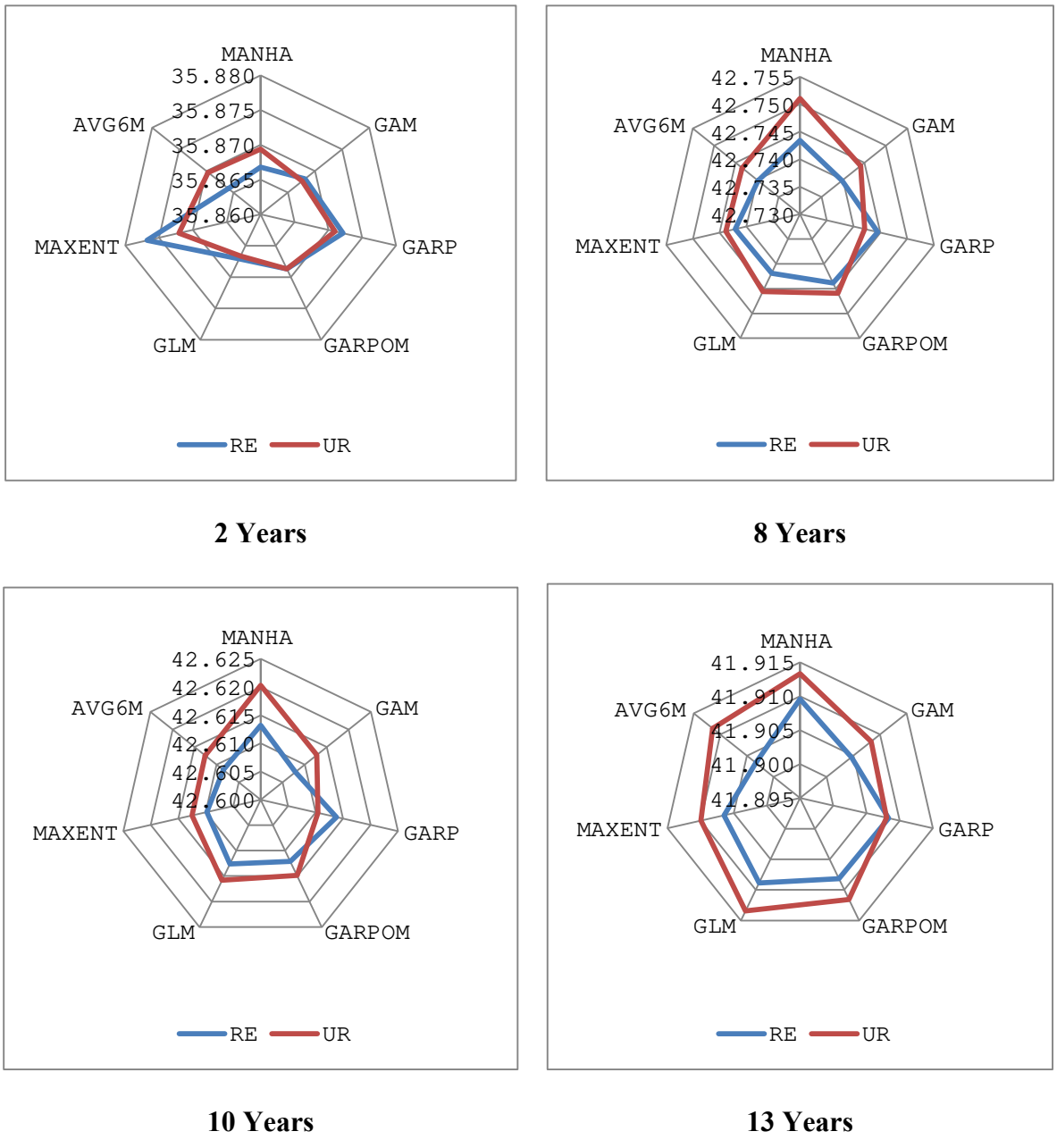

Figure 3. Quantitative results of the simulations.

The maximum entropy models are statistical models to identify the pattern of a spatial location, most likely a system subject to constraints. This approach was introduced in geographic modelling by Wilson in 1967, as the basis for a stricter interpretation of the gravity model. The model of maximum entropy make it possible to easily calculate the effects of new housing, and have been used in varied urban models, e.g. in the Lowry model. As in the situation of two years, there is a relatively small period to set the trend the system don't get auto-organized, predominating, due to the many possible behaviours, the existence of microstates.

In the remaining three simulations, based on 8, 10 and 13 years trends, the system tends to the auto-organization and the MAXENT begins to lose protagonism. In this context, the effect of distance gets predominance from what the model that stands out is the manha, which is based on Manhattan distance. At first (8 and 10 years trends), the manha is the more productive simulation in generating built space when the model is run with restrictions.

With the absence of constraints, the urbanizing space available increases and, therefore, there is an increment of the possible paths to follow. This multiplicity of solutions comes from a competition for space, which favours the genetic algorithms, which duplicate by analogy from biology the competition between species (land uses in this case). Therefore, it is not surprising that when we face an unrestricted space, GARP is the simulation with highest performance.

These claims can be confirmed by comparing the relationships between the simulations with and without restrictions. In the simulation based on two years trends, the two submodels generate more urban area with no restrictions than with restrictions, namely Maxent and GARP. Then, with 8 and 10 years trends, the only submodel that can create greater urban area without restrictions than with restrictions is GARP, for the reasons already pointed out. We cannot help to note, because it is interesting and reinforces our analysis, that GARP being more productive without restrictions, behaves in the opposite way when operating with restrictions.

Finally, in the simulation based on the trend of the last 13 years, as the system is already auto-organized, neighbourhood and proximity relationships acquire particular emphasis, and the manha is the predominant submodel, both with as without restrictions. Once again, when it comes to simulate with restrictions, GARP is the one that generates least urbanized area but, at the same time, is the one whose value are closest in the two simulations (with and without restrictions). To reinforce the importance of the distance, we must also mention that with restrictions, manha performance is matched by the Generalized Linear Model (GLM).

Analyzing the agreement between the simulations results, we can easily see that this is less when these are made without restriction. This situation is logic, to the extent that there are more possible states and the system can evolve in many different directions. Even so, in 8, 10 and 13 years trends, the number of cells assigned to urban use in all the seven models amounts to 75%. This value is higher, accounting for 94% , in the biennial based simulation, which makes perfect sense, because if this is the least dynamic, when there are fewer state changes it will also be less error probability.

When there is not total agreement, predominates the result presented by just one model (2%), which reflects the specificity of each submodel algorithm. All other combinations have equal distribution among themselves and between models, with the exception of the two-year simulation, whose intra-simulation as also equitable distribution, but in an inter-simulation comparison, their percentages present lower values.

When we use restrictions, the situation is similar but with a slight change of scale. The unique solution continues to represent about 2% in all the simulations, as well as the common result to the seven submodels. However, the decline of the relative importance of the agreement of 6, 5, 4 and 3 submodels in detriment of the 7, allows expanding the percentage of perfect agreement good to 92%. As in the

previous situation, the particular case of simulation with 2 years presents spatially results slightly more consistent, with a value of 96% of total agreement.

In spatial terms it may be noted that the highest degree of disagreement arises on the South side, in every simulation of 8, 10 and 13 years trend, This happens mainly because the South is, in part, a luxury gated community, which does not generate the same rules as the rest of the territory, suggesting that it should be the subject of further studied.

The disagreement is also high in the northern part of the municipality, in the simulations with 8 and 10 years and with restrictions. Here, essentially two zones fail, located at Northwest and Northeast, being created by excess by only one of the models. These areas although theoretically good for housing, are not valued by the population. This situation should have been guarded by the use of the valuation tax of the territory (IMI), but its weak predictive potential turned out to condition the result.

One way to quantify the agreement is through the Kappa index. The Kappa Index of Agreement method ( ), KHAT or KIA is based on a simple cell-by-cell comparison, which checks for each cells pair on two maps if they are equal or not. This results in a comparison map that displays the spatial distribution of the agreement, without requiring any previous parameters. This method evaluates the agreement between the result of the simulation and a purely random assignment of classes to the cells (designated by agreement due to chance).

Another simple cell-by-cell statistic is the correct fraction, also known for overall accuracy. This statistic is calculated by dividing the number of correct cells by the total number of cells. However, the overall accuracy is considered unreliable as a general measure of similarity. The reason is that the correct fraction is a biased method, i.e. tends to consider maps with few categories and/or unevenly distributed, more similar than the ones with many categories and/or equally distributed.

The following (extreme) case illustrates the difference between the kappa statistic and the overall accuracy. Having a model to predict parks locations in an urban area, we have also a binary map of presences and absences. In reality, there is only one park per 100 cells. This means that a model that predicts wrongly parks in this place still gets an impressive result of proper fraction of 0.98 (absence on both maps). Another model predicts the presence of houses in the same area. These are found in 50% of all locations (i.e. cells). A model that predicted 80% of these locations would only achieve a overall accuracy of 0.8. Thus, even though the parks model is clearly worse than the model of the houses, the first produces a correct proportion much higher, simply because of the uneven distribution of absences and presences.

The Kappa values are generally lower than those obtained for overall accuracy. This difference is expected, since each index use different types of information from the contingencies table. Overall accuracy only uses data that lie along the diagonal of the matrix, excluding the omission and commission errors. On the other hand, the Kappa index incorporates the diagonal elements of the matrix as a product of the marginal values of rows and columns. It is impossible to unequivocally point the measure that must be applied and when. Usually, it is desirable to calculate and evaluate the two.

However, in order to obtain a more balanced similarity we often use the Kappa index. The Kappa statistic results from two types of similarities, the similarity of quantity and the similarity of location. In this case the amount refers to the total number of cells attached to each category found on the map, i.e. the histogram, and the location refers to the spatial distribution of the different categories of the map. To know to what extent the measures of of location and quantity similarity are represented in the Kappa statistic, we divided it into two statistics, the Kappa histogram (KHisto) and the Kappa location (KLoc). It can be found more information about the KHisto and KLoc in Hagen (2002). The relationship between the 3 statistics is given by:

$$\text{Kappa} \quad \text{KHisto} \quad \text{KLoc} \tag{2}$$

These statistics were calculated for the simulations carried out without restrictions (Table 12). This option was taken because this is the actual situation, i.e. there has always been urban growth in areas of RAN and REN. In addition, the new legislation of the REN is more permissive in terms of use, establishing some activities compatible with the various levels of protection. These condition, associated with the fact that the management of both the REN and RAN, is made by the local authorities, points for this scenario as the most suitable to follow. For all the reasons already pointed out, we choose as final submodel the average of all models with good performance of AUC, weighted by its predictive power, i.e. submodel AVG6M. The kappa statistics were only calculated for this final submodel.

In Table 12 we can observe the good agreement of the submodel AVG6M with all others, showing always values above 90% (0.9). The agreement of the histogram of the submodel AVG6M with the manha, GARP, GARPOM and the Maxent is always total. The overall accuracy and Kappa are very homogeneous, despite of using 8, 10 or 13 years trends, and increase a little when using only a 2 years trend. This demonstrates that in the first moments of expansion, transitions are more assertive. In all cases the Kappa is always less than the KHisto, due solely to lower values of Kloc, which indicates that even when evolution takes place in same quantity it not necessarily occurs in the same direction.

Table 12. Agreement of submodel AVG6M without restrictions with all the others.

AVG6M		MANHA	GAM	GARP	GARPOM	GLM	MAXENT	
2 Years Trend	Kappa	0.98107	0.98108	0.97771	0.98316	0.96754	0.96528	
	KLoc	0.98107	0.98115	0.97771	0.98316	0.96757	0.96528	
	KHisto	1.00000	0.99993	1,00.000	1.00000	0.99996	1.00000	
	Overall accuracy	0.99057	0.99058	0.98889	0.99161	0.98383	0.982	
8 Years Trend	Kappa	0.92805	0.93366	0.91759	0.95260	0.93077	0.92910	
	KLoc	0.92805	0.93410	0.91759	0.95260	0.93117	0.92910	
	KHisto	1.00000	0.99952	1.00000	1.00000	0.99957	1.00000	
	Overall accuracy	0.96661	0.96921	0.96176	0.97801	0.96787	0.967	
10 Years Trend	Kappa	0.92874	0.93394	0.91894	0.95287	0.93022	0.92948	
	KLoc	0.92874	0.93436	0.91894	0.95287	0.93061	0.92948	
	KHisto	1.00000	0.99954	1.00000	1.00000	0.99958	1.00000	
	Overall accuracy	0.96686	0.96928	0.96230	0.97803	0.96754	0.967	
13 Years Trend	Kappa	0.92655	0.93564	0.92728	0.95848	0.92842	0.93444	
	KLoc	0.92655	0.93603	0.92728	0.95848	0.92877	0.93444	
	KHisto	1.00000	0.99959	1.00000	1.00000	0.99963	1.00000	
	Overall accuracy	0.96545	0.96972	0.96579	0.98047	0.96632	0.96916	

6. Conclusions

In our CA urban model, we simulate different scenarios of urban growth to 2021. Our approach is to use different trends (last 2, 8, 10 and 13 years) with and without restrictions and, most important, to apply several different methods for transition rules definition. The mathematical models used to define transition rules include artificial neural networks, genetic algorithms, supported vector machines, decision trees, generalized linear models, generalized additive models, maximum entropy and geometric distance models.

In the second part of our work we select the 6 best performing submodels. In addition, another submodel (named AVG6M) was created through a weight mean of these six best submodels. Using the six selected methods and the newly created one, we simulate urban area growth scenarios to the year 2021. Overall, 56 simulations were carry out, using 7 submodels, 4 tends, and 2 different approaches (with and without restrictions). All the simulations were compared in relation to the predicted amount of urban area, spatial distribution of predicted urban area and geographic agreement between them.

In an effort to use different mathematical approaches to configure CA transition rules, it becomes obvious that the primary objective of the application of these approaches to modelling is the evaluation

of suitability or probability of the cell to develop, i.e. change of state. Therefore, we can see that the number of statistical integration models that can be used in correlative modelling of a phenomenon distribution of is high and diverse. Their role is to obtain response values that translate the suitability of the space by the combination of the values of the independent variables, concerning the known distribution of the phenomenon.

Possibly the most important feature of the models, which has been the subject of several studies, relates to the comparison of their predictive performance (Brotons et al., 2004; Tsoar et al., 2007), yet these models differ in several other features that often influence or potentiate their use in modelling. Taking into account this high variability, the choice of a single method implies the loss of several advantages present in others. In addition, holding several methods simultaneously is suggested on modelling the distribution of a phenomenon, as a way to reduce the uncertainty of the process and the variability of the models.

7. References

Allen, P.M. (1997). Cities and regions as evolutionary complex systems. Geographical Systems, 4, 103-130.

Araujo, M.B. & New, M. (2006). Ensemble forecasting of species distributions. Review. Trends in Ecology and Evolution, 22(1): 42-47.

Baker, W.L. (1989). A review of models of landscape change. Landscape Ecology, SPB Academic Publishing, The Hague, vol. 2 n.º 2, pp. 111-133.

Batty, M. (1995). New ways of looking at cities. Nature, Vol: 377(6550):574-574.

Batty, M., Longley, P. & Fotheringham, S. (1989). Urban Growth and Form: Scaling, Fractal Geometry and Diffusion - Limited Aggregation, Environment and Planning A. 21, 11, p. 1447-1472.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004). Presence-absence versus presence-only modeling methods for predicting bird habitat suitability. Ecography, 27: 437Ð448.

Carpenter, G.A. & Grossberg, S. (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns, Applied Optics, 26(23), 4919-4930.

Clarke. K.C. & Gaydos, L.J. (1998). Loose-coupling a cellular automata model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. International Journal of Geographical Information Science, 12, pp. 699 714.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1): 37 46.

Erlien, C.M., Tuttle, J.P., McCleary, A.L., Mena, C.F., & Walsh, S.J. (2006). Complexity theory and spatial simulations of land use/land cover dynamics: the use of what if scenarios for education, land management, and decision-making. Geocarto International, 21(4): 67-74.

Fang, S., Gertner, G.Z., Sun, Z., & Anderson, A.A. (2005). The impact of interactions in spatial simulation of the dynamics of urban sprawl. Landscape and Urban Planning, 73: 294-306

Fawcett, T. (2006). An introduction to ROC analysis, Pattern Recognition Letters, 27, 861 874.

Guisan, A. & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8, 993-1009

Guisan, A., Edwards, T.C. & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157: 89-100.