1949

ANALYSIS OF LAND CHANGE TENDENCIES BASED ON REMOTELY SENSED DATA

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PhD)

Varga Orsolya Gyöngyi
Supervisor: Szabó Szilárd DSc

UNIVERSITY OF DEBRECEN
Doctoral Council of Natural Sciences and Information Technology
Doctoral School of Earth Sciences
Debrecen, 2020

Hereby I declare that I prepared this thesis within the Doctoral Council of Natural Sciences and Information Technology, Doctoral School of Earth Sciences, University of Debrecen in order to obtain a PhD Degree in Natural Sciences at Debrecen University.

The results published in the thesis are not reported in any other PhD theses.

Debrecen, 9 June 2020

signature of the candidate

Hereby I confirm that Varga Orsolya Gyöngyi candidate conducted his/her studies with my supervision within the Natural and Anthropogenous Processes of Lithosphere and Hydrosphere Doctoral Program of the Doctoral School of Earth Sciences between 2014 and 2020. The independent studies and research work of the candidate significantly contributed to the results published in the thesis.
I also declare that the results published in the thesis are not reported in any other theses.
I support the acceptance of the thesis.

Debrecen, 9 June 2020

signature of the supervisor

ANALYSIS OF LAND CHANGE TENDENCIES BASED ON REMOTELY SENSED DATA

Dissertation submitted in partial fulfillment of the requirements for the doctoral (PhD) degree in Geography

Written by Varga Orsolya Gyöngyi certified Geographer (MSc)

in the framework of the Doctoral School of Earth Sciences of the University of Debrecen (Natural and Anthropogenous Processes of Litosphere and Hydrosphere programme)

Dissertation advisor: Dr. Szabó Szilárd

The comprehensive examination board:

chairperson:	Dr. Posta József
members:	Dr. Kerényi Attila
	Dr. Juhász Attila

The date of the comprehensive examination: 24 May 2018

The official opponents of the dissertation:

Dr. Deák Balázs
Dr. Kristóf Dániel

The evaluation committee:

chairperson:	Dr. .....
members:	Dr. .....
	Dr. .....
	Dr. .....
	Dr. .....

The date of the dissertation defence: …… 20…

"It looks good" is not a valid accuracy statement.

(Congalton, 1991)

TABLE OF CONTENTS

1. INTRODUCTION

Land change and land change modelling have key importance in a constantly changing world where the activity of mankind results in enormous transformations and an accelerating modification of natural environment. All these changes have a diverse range of purposes – either disadvantageous damages, like illegal logging or beneficial changes in favor of natural habitat, like creating a landscape corridor for certain species. It is essential to be able to monitor changes and to project these changes forward as precisely as possible in order to reveal scenarios that also provide realistic visions of the future landscape. Land change modelling is a practical and abstract approach of the real land changes where the success of a model depends on an enormous number of possible parameters. Even if the model matches the main land characteristics of reality, the validation process may substantially distort the interpretation of results. Therefore, the modeler may be misled by unrealistic validation results and may support further erroneous land management decisions based on a wrong model.

In my dissertation I aim to reveal (1) how exactly wrong practices, which are still widely used among scientists and are frequently published, may have a bad impact on model performance interpretation; (2) what good practices there are in literature and how their appropriateness could be confirmed in a large set of land change models; (3) how results can vary with some additional circumstances apart from the parameters of the model itself, like aggregation of land categories and real change dynamics in the landscape. In this research three sets of study sites were applied, where three different approaches were illustrated based on the same cellular automaton-Markov (CA-Markov) model.

In the first set, a CA-Markov model was run in one specific study site and intensity analysis was applied for analyzing changes in reference and simulation data. Intensity analysis is a framework for land change monitoring. Along with this detailed change monitoring, Figure of Merit (FOM) and its components were calculated to have an insight to model performance. FOM is a metric that mainly focuses on the comparison of reference and simulated changes in a landscape and FOM components reveal detailed information about correctly and erroneously simulated pixels. In this case, the effect of the consistency of real landscape change dynamics on validation results was illustrated.

In the second set of study sites, 114 CA-Markov models were run with the same model parameters and the same input data, but with various sizes of study sites and various manners of aggregations of land

categories. In this case, the effect of various aggregation methods on model performance was investigated, while illustrations and findings derived from the comparison of bad and good practices of model performance validation were presented.

In the third set of study sites, 6 CA-Markov simulation models were run with the same model parameters in two study sites, focusing on sprawl-like change dynamics specifically. In this case, the differences in FOM and FOM component values related to the second set of study sites were enhanced, and an investigation on the purport of stationarity of land changes across time was presented.

The research uses remotely-sensed data either directly by processing Landsat satellite images or indirectly by using Corine Land Cover data that is also produced based on various remotely-sensed datasets.

Based on the preliminary literature study, I hypothesized the followings:

- intensity analysis could help the validation process by giving a deeper insight into changes in the landscape;
- wrong model performance approaches (Kappa Index of Agreement and Overall Accuracy) mislead the interpretation and result in high correlation with persistence in the data;
- aggregation of land use/land cover categories does not data affect model performance;
- the temporal stability in the reference and simulated data affect model performance.

The innovations of my research are the followings:

- I use intensity analysis in the model validation process;
- I investigate the possible effects of aggregation methods on model performance;
- I use a large set of model runs to present the ideas above and to prove some specific results concerning land change modeling published in scientific literature before.

I have published partial results of this research concerning the application of intensity analysis in model validation process (Varga et al., 2019) and the effect of aggregation methods on model performance (Varga et al., 2020) as part of my Ph.D. publication requirements.

Based on the results of research conducted in the three different sets of study sites, I developed my theses. My general purpose was to provide expressive cases that enlighten a deeper correspondence in validation process and help land change modelers to choose correct and suitable methods. I hope for a better understanding of possible mistakes throughout model validation process.

2. LITERATURE REVIEW

2.1. Definitions of Land

It is important to review definitions and approaches of landscape in order to define the study design appropriately and clearly. There are several definitions of landscape that have developed with the time passing by. Alexander von Humboldt was the first to think of a unique character related to landscapes and characterized landscape as *"the total character of a region"* (Farina, 2013). In modern landscape ecology, Turner et al. (2015) defined landscape as "*an area that is spatially heterogeneous in at least one factor of interest*". According to the definition in the European Landscape Convention (Council of Europe, 2000),*"landscape means an area, as perceived by people, whose character is the result of the action and interaction of natural and/or human factors".* In Hungarian literature, Kerényi (2007) defined landscape as an *individuum,* a unique part of the geosphere and a spatial unit whose basic character and boundaries were results of natural processes, but were modified as a result of anthropogenic activities in various measures. These definitions point to the fact that landscape has its own character which helps to discriminate it from other landscapes and this character is a result of a combination of natural and anthropogenic processes.

Turner et al. (2015) created a synthetic review of landscape ecology definitions and applications where the authors summarize the thoughts of main representatives of this field. According to this synthesis, Forman (1983) described landscape ecology as dealing with the relationships and dynamics – like the movement or flow of species, energy and mineral nutrients – among elements or ecosystems of the landscape. Risser et al (1984) determined landscape ecology as focusing on the aspects of spatial heterogeneity of the landscape, mainly the dynamics, spatial and temporal interactions, management of spatial heterogeneity, moreover its effects on biotic and abiotic processes. Forman (1995) published the patch-matrix-corridor model, which introduced essential terms in landscape science up to this day. This work determines the following definitions:

- a *patch* was defined as an area differing from its surroundings in nature or appearance;
- a *corridor* was defined as a narrow strip of a particular type which connects patches and is different from its neighboring areas;
- a *matrix* was defined as the background land cover type of a landscape which embraces and involves the other elements in the landscape.

Turner et al. (2015) also described the landscape ecology definition of Urban et al (1987) as it states that the motivation of landscape ecology is a need to comprehend ecological processes and phenomena in terms of dynamics, spatial scales, temporal scales and role of disturbance.

According to McGarigal (2002), one of the founders of landscape metrics theory, land cover types are relevant examples of a certain basic data type of landscape pattern analysis. This data type is categorical map pattern, named also as thematic or choropleth map, where the subject is represented as a mosaic consisting of discrete patches. This character is in accordance with the ecological approach of patches where the patches are discrete areas of homogeneous conditions from an ecological aspect (McGarigal, Kevin, 2002).

It is important to summarize the definitions and units which may occur in the analysis context. Scientists examine land change in a pretty wide range of researches and publications (Abd El-Kawy et al., 2011; Kim, 2016; Mallinis et al., 2014; Mallupattu and Reddy, 2013) which suggests that land change monitoring and land change analysis are really popular topics. We can find several examples which describe Land Use (LU) change analyses, Land Cover (LC) change analyses, but more often these terms are used interchangeably in literature, as land use / land cover (LULC) change analyses. Even there is abbreviation focusing on specifically the change of Land Use / Land Cover, which is LULCC meaning Land Use / Land Cover Change (Näschen et al., 2019; Ozsahin et al., 2018) and LUCC meaning Land Use/Cover Change (Mas et al., 2014).

However, there is fundamental difference between definitions of land use and land cover. DiGregorio and Jansen (2000) defined land cover as *„the observed (bio)physical cover on the earth's surface"*. DiGregorio and Jansen (2000) defined land use as it „*is characterized by the arrangements, activities and inputs people undertake in a certain land cover type to produce, change or maintain it"*. Soesbergen (2016) stated that land cover characterize the physical surface, e.g. presence of vegetation, and this character is directly observable, but land use characterize the economic and social functions of land or the purposes of human exploitation. These definitions all point to the fact that land cover refers to natural units of the surface which can be visually observed, while land use is determined by the purpose that the land is utilized for, and there is a direct relationship between them. Land use and land cover may even show different characteristics in a given unit of land. For instance, a residential area is homogeneous in a sense of land use category, since it is used as mainly permanent residence of the population, but it is heterogeneous in a sense of land cover category,

because it may consist of either buildings, roads or green areas (Verőné Wojtaszek, 2010).

In this dissertation, land change is in focus in a manner that different categories of LULC are simulated to a future state by a LULC change model. The investigation is mainly based on a ready-to-use LULC database (Corine Land Cover, henceforth referred to as CLC) that has a well-defined category scheme. This category scheme is consistent through different versions of Corine databases and this consistency has substantial importance in the modelling phase of the research. The Corine category scheme itself has possible shortcomings due to the various national methods of production (Martínez-Fernández et al., 2019) or problems when applied in local scale analyses (Diaz-Pacheco and Gutiérrez, 2014). These shortcomings are not in the focus of this research, because CLC Level 3 datasets were aggregated according to various category aggregation methods, and the possible general shortcomings of the circumstances of CLC data production may affect the study design uniformly, if any.

Within this research, there was an individual group of two study sites where specific land cover categories were determined via segmentation of remotely-sensed images (*Section 3.1.3.*). In these cases, classes were determined based on specific characteristics of the examined phenomena and visually interpretable objects, which latter condition is in accordance with the cited definitions of land cover. Therefore, these cases can be considered more specifically as land cover (LC) change models, instead of LULC change models.

2.2. Land Change Analysis

Land change monitoring has the opportunity for revealing the patterns of change and dynamics of change in the landscape (Lambin, Eric F. et al., 2003; Madrigal-Martínez and García, 2019). Some of these opportunities are based on crosstabulation matrices of different land cover maps. Post-classification comparison of remote-sensed land cover data follows this logic, since this method overlays independently classified maps originated from remotely sensed data, and creates a crosstabulation matrix based on this comparison. It can provide a basis for calculations of LULC changes from one time period to another, and help to determine the changing areas and what category they turned into (Jensen, 1996). Many scientists used post-classification comparison for the change detection analysis of remotely sensed data from various sources, such as historical aerial photographs or Landsat and ASTER satellite images (Alo and Pontius Jr, 2008; Alphan et al., 2009; El-Hattab, 2016; Halls and Kraatz, 2006). This method of establishing a

crosstabulation matrix and calculating changes of LULC classes can not only be used in case of remotely sensed data but raster land cover data derived from any sources, e.g. results of field measurements or maps generated via visual interpretation, after rasterization.

Intensity Analysis is another approach of describing land cover change, also based on crosstabulation matrices of maps from initial and end dates of a time interval. Intensity Analysis is a quantitative framework to characterize change among categories through time and to characterize patterns of changes in more and more detailed levels. It has been used recently in an increasing number of researches worldwide, for the purpose of analyzing changes in landscape through more time intervals and even through time intervals with different durations (Aabeyir et al., 2017; Castro and Rocha, 2015; Quan et al., 2017; Raphael John et al., 2014; Rocha et al., 2017; Teixeira et al., 2016; Yang, Y. et al., 2017). It could have been used for the analysis of other phenomena as well, such as dynamics of solar radiation (Li et al., 2017). It has not been widely used in simulation model evaluation issues so far, although there were papers including the analysis of gains, persistence and losses of urban and non-urban classes in various urban growth scenarios (Liu and Feng, 2016). Another example of application was the examination of the temporal pattern of urban land changes across time intervals in order to get insight to the dynamics of the study area, right before setting up a predictive model for urban land changes (Subasinghe et al., 2016). These examples point to the fact that Intensity Analysis has started to become a widely used method for monitoring landscape changes, but it has not been used for monitoring the landscape changes simulated by a land change model. The first example was our publication, which is a basis for my dissertation and a practical application of this theory (Varga et al., 2019).

Landscape and land change analysis are also fundamental and popular research topics in Hungarian scientific literature or Hungarian study sites, and the science of the background of land changes has long traditions in Hungarian science. Landscape in general (Lóczy, 2015), landscape ecology, human transformation (Csorba and Szabó, 2009) or examination of driving factors (Deák et al., 2016; Ladányi et al., 2016) of land change are all research topics of interest among Hungarian scientists as well.

2.3. Category Aggregation

When establishing a process of land change analysis, it is a relevant need to aggregate land classes that the scientist intends to analyze. Handling too many categories may make the analysis complicated to

perform and interpret. Moreover, the change analysis of too many categories may distract the focus from outstandingly important changes in the landscape. The categories in a land change analysis also need to be comparable, because we cannot analyze appropriately the land change in a time interval where the definition and membership rules of categories of the initial and final dates are different. The problem of category aggregation dates back to the definition of Modifiable Areal Unit Problem, frequently referred to as MAUP. The problem had been partly discovered before by Gehlke and Biehl (1934), and was later thoroughly analyzed and published by Openshaw and Taylor (1979). The MAUP has two sub-problems (Wong, 2004):

(1) the *zoning effect* which means whether the number of zones in a given area is constant and new boundaries are drawn in order to set up a new zoning system, then the analytical result of the different datasets gained based on the different zoning systems will be also inconsistent;
(2) the *scale effect* is present when spatial aggregation or disaggregation of data occurs, or the spatial resolution of the data changes and at least one of these effects leads to inconsistent analytical results (Wong, 2004).

A typical example of the zoning effect is the phenomenon of gerrymandering, which is a certain way of drawing the boundaries of constituencies in order to gain particular political advantages (Johnston, 2002). More papers investigated the scale effect in connection with its importance in land change monitoring and land change modelling applications (Jelinski and Wu, 1996). Category aggregation is an important factor in land change modelling as well, since usually LULC maps are used as inputs in land change models. The aggregation of LULC map categories affects if a specific change is present or hidden in the map, i.e. aggregation of two categories can relevantly change the pattern of the land mosaic; therefore, the outcome will be biased by the method itself. According to Olmedo et al. (2018), MAUP is relevant in land change modelling in a manner that vector-to-raster conversions or resampling operations have a significant effect on the initial map of the simulation model, since it influences the cell neighborhood. Mas et al. (2015) examined deforestation in a case where they aggregated the information concerning driving factors based on spatial units. They found that MAUP produced variation, but did not have substantial effect in most cases, except for some variable pairs and specific cases where the effect was substantial. Pham (2005) stated that not many researches focused on the effects of grid size and aggregation on simulation models despite MAUP's known effects. Moreover, evidence of MAUP in grid-

based modelling approach, even theoretical or empirical, is missing (Pham, 2005).

Pontius and Malizia (2004) introduced a theory called category aggregation problem (CAP) which states that the category definition is important due to the fact it has a substantial influence on change in the map. However, it seems to be obvious that category aggregation affects the changes in the map somehow, they also introduced 5 principles that drive the effects of category aggregation and proved them mathematically. Pontius and Malizia (2004) proved that category aggregation has a tremendous effect on the confusion matrix used for accuracy assessment. It means that the accuracy results we interpret, based on traditional accuracy assessment methods, can vary with the change of actual categorization. Aldwaik and Pontius (2015) delineated a possible method for aggregation, the behaviour-based category aggregation, which intend to aggregate two categories in each step in order to maintain change as much as possible. They provided a Visual Basic for Applications (VBA) macro for performing the analysis in Microsoft Excel environment.

Generalized nomenclatures cannot express the conditions of reality in details, thus, their suitability is questionable during practical applications and actions concerning land use. The multidimensional approach of land use classification dates back to the middle of the twentieth century and emerged from urban planning due to providing an opportunity on more project-specific classifications (Guttenberg, 2002). There are different schemes for aggregating LULC categories in order to reduce the influence of LULC change examinations on the results as least as possible. According to Congalton and Green (1999), a classification scheme should be mutually exclusive and totally exhaustive. Anderson (1976) suggested the usage of a uniform classification framework with two levels for LULC data interpreted based on remote-sensing techniques. He aimed to establish a classification system which can be a basis for a uniform categorization for satellite and aerial images. There are even several other classification systems in literature – and practice – so as to represent land cover data by assigning appropriate grouping/alignment for land cover objects and types, either at global, continental or local scales (Di Gregorio and Jansen, 2000; Fosberg, 1961; Herold et al., 2009; Küchler and Zonneveld, 1988; UNESCO, 1973). While there are many available category schemes in literature, there is not any uniformly accepted category scheme, partly because they are inappropriate for uniform purposes or that the schemes are based on outdated information (Di Gregorio and Jansen, 2000). However, in many cases, the scientists perform classification in accordance with a specific

purpose, because they want to investigate a certain land cover, and a process like this does not need a comprehensive, but a focused category scheme (Abriha et al., 2018; Burai et al., 2015; Deák, M. et al., 2017; Gulácsi and Kovács, 2018; Kristóf et al., 2002).

As CLC data is a Pan-European LULC map and had five releases with a nomenclature consistent over time, it is a popular source for land change monitoring issues. This research used the most detailed CLC Level 3 data with its original classes and the classes were aggregated into various datasets according to various aggregation methods. The methods of category aggregation are described in *Section 3.4.*

2.4. Measurement of Stationarity

Usage of the words "pattern" and "dynamics" in context of land change are not necessarily related to uniformly accepted definitions. These words are used in literature characterizing various approaches of land change issues, like using a dedicated spatial index for the analysis of land change patterns (Dadashpoor et al., 2019) or studying the determinants of changes (Verburg et al., 2004b). It is important to have an insight into the patterns of land change in order to understand the changes that occur in the landscape. In case of land change, spatial and temporal considerations are equally important, since land change is a transition located in a definite place and change process has a beginning and end in time. There are a few measurements which address land change pattern analysis. According to Aldwaik and Pontius (2012), if the change in a landscape is stationary, then the changes in a given time interval show the same pattern as the pattern in another time interval (Aldwaik and Pontius Jr, 2012). They published calculations for determining stationarity in this sense. In their concept, the definition of stationarity depends on the level of analysis, because it has different conditions in case of the whole spatial extent, in case of the categories' gains and losses and in case of transitions between the categories. Sang et al. (2019) applied this method to analyze stationarity and change intensity throughout 20 years based on Landsat TM and OLI images. Runfola and Pontius Jr (2013) used the term temporal stability, which they describe as the measurement of stationarity, and define as "*the degree to which the rate of land change is consistent over a given temporal extent".* Markov models predict based on transitional probabilities and according to Mertens and Lambin (2000) (Mertens and Lambin, 2000; Runfola and Pontius Jr, 2013) if a Markov model has to deal with land change process that is not stationary, then it loses its predictive ability, unless the transitional probabilities are modified. Pontius Jr and Neeti (2010) stated that it is a good chance for uncertainty

in land change processes that these processes include human decisions, which increases the presence of non-stationary changes, while the model tries to extrapolate stationary changes. Runfola introduced Runfola's R value (Runfola and Pontius Jr, 2013) for measuring temporal stability, also known as stationarity.

This research applies Runfola's R value for measuring the temporal stability between time intervals which are used for calibration and validation of Markov models. This research considers also the stationarity of calibration changes and changes simulated by the Markov model and how the difference between these stationarity values addresses model performance in a large set of simulation models.

2.5. Land Change Models

There is a really wide range of simulation model types in literature. It is important to position the model used in this study design in order to have an insight to the purposes and logic of the model.

Lambin et al. (2000) published a paper in the topic of agricultural land-use models in which they stated that land change processes should have the purpose of addressing the following questions, at least one of them:

- *WHY?* – the question addresses environmental and cultural variables which explain land change the most;
- *WHERE?* – the question addresses the locations affected by land change;
- *WHEN?* – the question addresses the rate of land change.

Lambin et al. (2000) also published a classic grouping of land change models where they grouped the models based on addressing these questions, as follows:

- *Empirical-statistical models:* these models aim to identify the causes of changes via mainly multiple linear regression analyses. These models are able to predict changes which are represented in the training data and had been measured through a long period before.
- *Stochastic models:* these models are based on transitional probability information which is statistically estimated from transitions that have been observed in the past.
- *Optimisation models:* these models are specific for agriculture, since they are based on land rent theories and the models aim to approach a status where the land earns the highest rent.

- *Dynamic simulation models:* these models are based on biophysical and socio-economic processes and their interaction, while aiming to simulate these processes. Therefore it is a system-focused approach and demands the a priori understanding of the driving forces.

Soesbergen groups models into the following categories based on the work of Heistermann et al. (2006), as follows:

- *Geographic models*: these models use local characteristics and suitability to allocate land. The availability of geographic information systems (GIS) made it possible to develop geographic models, and they are capable of simulating phenomena mainly at regional or local scales (van Soesbergen, 2016).
- *Economic models:* these models focus on demand and supply of land-intensive commodities to allocate land. Computable General Equilibrium and Partial Equilibrium models (De Rosa et al., 2016) are examples of this approach, since in case of these models the allocation is based on market conditions.
- *Integrated models* combine the features of the former two model types. It combines geographic approach, where geographic analysis determine the allocation of land, with economic approach, where world market analysis determines demand and supply characteristics (van Soesbergen, 2016)

Brown et al. (2013) determined five types of modeling approaches which are grouped according to both if they emphasize process or pattern and if they emphasize projection or explanation, as follows:

- *machine learning:* this approach focuses on patterns of change. The approach uses algorithms for finding relationships between changes and characteristics of locations where the changes are observed and derive this information from spatial variables. Brown et al. (2013) mentions artificial neural network, CART (classification and regression trees) and logistic regression as examples of methods used for variable selection in this approach;
- *cellular approach:* this approach focuses on either process or pattern, since it simulates changes based on combining likelihood maps with spatial interactions;
- *sector-based economic models:* this approach is purely economical and addresses demand for land while focusing on the equilibrium of the market based on demand and supply relations;
- *spatially disaggregated economic models:* this approach is about to understand land changes as a result of individual decisions in accordance with microeconomic theories;

- *agent-based approach:* this approach focuses on establishing observed land change via design and content determined by the user, based on interactions by which the user assumes to influence the processes.

Van Schrojenstein Lantman et al. (2011) identified further concepts of land use change in literature in their review, based on practical considerations: *cellular automata, statistical analysis, Markov chains, artificial neural networks, economics-based models and agent-based systems.* This grouping is a result of a slightly different approach, but it has a substantial overlap with the ideas of the groupings above.

Models can belong to a combination of groups according to the cited grouping approaches. In this research, I used CA-Markov model that simulates transitions among categories and combines the features of cellular automaton approach and Markov approach. They are also used separately. Many land change models use Markov extrapolation, like Idrisi's Land Change Modeller, Idrisi's CA-Markov (Eastman, 2012a) and DINAMICA model (Filho et al., 2002). Cellular automaton is integrated into model applications individually as well, like it is used in the SLEUTH model (Clarke et al., 1997; Silva and Clarke, 2002). CA-Markov belongs to the group of stochastic models and answers the question of *WHEN?,* according to Lambin et al (2000), in a manner that it focuses on the rate of land change based on the past status of land while does not necessarily consider the reasons for the change. This latter feature depends on the exact model in which this approach is integrated, e.g. Idrisi's Land Change Modeler is capable of involving spatial variables. The CA-Markov model can implement various weighting factors (El-Hallaq and Habboub, 2015; Myint and Wang, 2006) and has been applied to specific fields of land change, such as urban growth (Jalerajabi and Ahmadian, 2013; Sang et al., 2011) and historical land use research (Clarke et al., 1997; Iacono et al., 2015). Previous studies dealt with the behavior of land change models in terms of quantity and allocation of land changes, such as in GEOMOD and TerrSet's Land Change Modeller applications (Olmedo et al., 2015; Pontius and Malanson, 2005).

CA-Markov is a cellular approach according to the grouping of Brown et al (2013) and the authors warn about that "*these models are limited in their ability to represent decision making processes"* due to their logic behind the modelling process. In general, land change models and scenarios are useful inputs for landscape planning and management and there are researches for the possibilities and circumstances of utilization in practice (Convertino and Valverde Jr, 2013; Lippe et al., 2017). I use this model in a large set of model runs in various study areas

across Europe, North and South America. The driving forces could be so diverse that managing various forces implemented into spatial variables would distract the focus of the research. In this research mainly the metrics and influencing factors of simulation performance are relevant and the driving forces of land change are irrelevant. The utilization of the results is relevant in practical aspects, i.e. the validation of a land change model. The CA-Markov model makes it possible to run the models without determining driving factors of change in the study areas and simulate future land changes based on purely the characteristics of land changes in the past. Furthermore, the model can be run with the exact same variables throughout the study design, except for the cases of American study areas, therefore these latter examples were interpreted separately.

2.6. Land change model validation

When running a simulation model, it is a fundamental need to characterize the agreement between reference and simulated change. Turner et al. (1989) published a paper of possible evaluation methods for spatial simulation models. The author examined metrics for spatial pattern, spatial predictability and goodness-of fit analyses (Turner et al., 1989). There are various approaches for simulating a model, depending on the model itself as well. It is an an extremely widely used method to compare the simulated LULC map to the map representing the reference LULC of the same date, and the agreement is characterized by an index used for accuracy assessment in remote sensing applications, like the Kappa Index of Agreement or overall accuracy (Grigorescu et al., 2011; Halmy et al., 2015; Keshtkar and Voigt, 2016; Mishra and Rai, 2016; Popovici et al., 2018; Singh et al., 2015; Yang et al., 2014). In these cases, the metrics of agreement between the two maps were used to evaluate model performance, but these indices evaluate persistent areas as agreement, and they are capable of returning high agreement values even if the agreement between reference and simulated changes is low. Another metric in literature, the Figure of Merit – also referred to as critical success index (Jollife and Stephenson, 2003; Klug et al., 1992; Perica and Foufoula-Georgiou, 1996; Pontius Jr, R. G. et al., 2011), focuses on the intersection of reference and simulated change, making it possible to approach model performance based on the success of simulating the changes, not the persistence. Pontius Jr et al. (2011) examined simulation models from cases published in scientific literature where the authors derived and presented possible combinations of comparisons between the relevant maps. Figure of Merit has components that characterize pixels according to being simulated erroneously or

correctly, based on a three-map comparison approach of the reference map, simulated map and a reference map from the previous date which the simulation was based on. This concept also appears in other scientific fields, like behavioral analysis (Lerman et al., 2010).

A further approach of model validation is the $Kappa_{simulation}$ published by Hagen-Zanker (Hagen-Zanker, 2006), which was defined as „*the coefficient of agreement between the simulated land-use transitions and the actual land-use transitions".* This index focuses on whether the changes are explained more by the simulation than they would be explained by a random distribution. Pontius Jr (2000) revised the shortcomings of Kappa and advised using further variations of the index. Pontius Jr et al (2011) later discouraged using Kappa and its variations due to its baseline of randomness and warned about misleading results when interpreting this metric while comparing two maps. They presented a new idea of map comparison via crosstabulation matrix, by introducing alternatives for measuring disagreement between the maps, namely quantity and allocation disagreement. While Hagen-Zanker's validation method accounts for the transitions in reference and simulated data, this approach possibly involves a baseline of randomness as well, due to applying Kappa. Among other indices, Kappa variations are available in multi-purpose Map Comparison Toolkit software as well (Visser and de Nijs, 2006).

There have already appeared more complex methods for the assessment of land cover change simulation models in literature, which mainly serve the purpose of validating models that involve spatial variables. One of them is Total Operating Characteristic, as known as TOC, which monitors the results in term of location and quantity, as it compares a Boolean variable versus a rank variable and assesses prediction accuracy at several diverse threshold levels (Pontius Jr, R. G. and Si, 2014). The TOC shows more extended information compared to the Relative Operating Characteristic, as known as ROC (Jamal, 2012; Pontius Jr, R. G. and Batchu, 2003; Pontius Jr, R. G. and Parmentier, 2014). Sensitivity analysis is another method widely used in model assessment that aims to answer which of the input factors can be relatively helpful in reducing the uncertainty of the output and which of them should be eliminated in order to reduce the variance of the output (Saltelli et al., 2004). Sensitivity Analysis was used in a wide range of practical applications, such as for parametrization of logistic regression equations (Van Dessel et al., 2011) and sensitivity analysis of Markovian models (Cao, X. R. and Wan, 1998; Chan and Darwiche, 2005; Charitos and van der Gaag, 2006; Renooij, 2010).

By presenting trends of model validation in literature I aimed to highlight either the wrong approaches or underline the reasons for their failures. I also revised the alternatives, by which the modeler can get a more appropriate insight to model performance. In the following section, I present the methods I used in this research, along with descriptions focusing on a methodological aspect, therefore enabling reproducibility as well.

3. METHODS AND STUDY DESIGN

3.1. Dataset

3.1.1. Corine Land Cover data

In this research Corine Land Cover data (Coordination of Information on the Environment, henceforth referred to as CLC) was used, which is a LULC database produced by the European Environment Agency, managed by the Copernicus Land Monitoring Service recently (Feranec et al., 2016). In the frame of CLC program, a geographic information system was established that contains information about land cover status of years 1990, 2000, 2006, 2012 and 2018. It was produced at a 1:100 000 scale based on the interpretation of various data sources by the time and available technological opportunities passing by, e.g. Landsat-TM, Landsat-MSS, SPOT (HRV XS), IRS, RapidEye and Sentinel-2 images (Büttner and Kosztra, 2017). A minimal mapping unit of 25 hectares and 100 m width (latter in case of linear objects) was applied. The databases were reported as having a thematic accuracy of 85% at least (Büttner et al., 2004; Büttner, 2014; Büttner and Kosztra, 2017). CLC data is a frequently used dataset for various landscape analysis purposes and land monitoring issues, such as hemeroby studies or landscape pattern analysis, also in Hungarian study areas (Csorba and Szabó, 2009; Túri, 2010). Corine Land Cover data is an extremely widely used data source for a range of subfields of environmental monitoring (Bielecka and Jenerowicz, 2019; Stathopoulou et al., 2007; Yague and Garcia, 2004).

CLC datasets were used as input data in study site groups 1 and 2, as described in *Sections 3.2.1 and 3.2.2.*, CLC has the advantage that it has a nomenclature consistent over time and the manner of data acquisition represents an approximately regular sampling over time, since CLC is published every 6 years. However, the images used for data processing showed a slight deviation from 6 years' time interval. CLC nomenclature consists of 3 standard levels in a nested hierarchical order and the most detailed third level assigns 44 LULC categories (Kosztra et

al., 2019). Standard level 2 and 1 has a maximum of fewer categories, therefore aggregating classes into ones with broader definition. This standard level system was used as an approach for category aggregation. This approach is described in *Section 3.3.* in detail. CLC nomenclature is presented in *Appendix 1* in detail.

3.2. Study sites

3.2.1. Study Site Group 1

Study site group 1 (*Figure 1*) consists of one specific study site located around Tokaj city and the estuary of Tisza and Bodrog rivers in the west neighborhood of the settlement. It is a junction of 5 microregions (Tokaji-hegy, Bodrogköz, Löszös-Nyírség, Hajdúhát, Taktaköz) and is located on the common administrative boundaries of two counties (Borsod-Abaúj-Zemplén, Hajdú-Bihar) and two NUTS2 regions (Northern Hungary, Northern Great Plain). Tokaj Wine Region Historic Cultural Landscape, a UNESCO World Heritage site, and its traditional vineyards also intersect the study area (Kerényi, 2015; Varga et al., 2019) This intersection and presence of Natura 2000 sites also contributes to the protected status of particular parts of the study area. Lowland chernozem and alluvial meadow soils are dominant, based on loess coverage, which are appropriate for arable and pasture land use as well as viniculture. Latter could have been cultivated for centuries due to favorable local aspect features, however, the area is charged with intense erosion (Kerényi, 2015). Deciduous forest coverage is typical mainly in areas with brown forest soil and relatively higher altitude, alongside the rivers or as afforestation patches sparsely within the S and SE part of the study site (Dövényi, 2010). Therefore, the land cover structure is quite heterogeneous, even related to a nationwide scale, because either forest coverage, extended built-up areas, water bodies, arable and pasture lands appear together. However, the partly protected status is an obvious limit for possible land cover changes, and under this circumstance, the study area shows an extremely low ratio of changing areas throughout the 12 year-long study time interval.

The maps of the study site were derived from CLC vector data concerning years 2000, 2006, and 2012, were rasterized into 25 m resolution maps and the categories were aggregated according to CLC Level 1 nomenclature (5 categories). The CLC data was available at the website of the Institute of Geodesy, Cartography and Remote Sensing (FÖMI), Hungary. The maps of 2000 and 2006 were inputs for calibrating the model, while 2012 served as an input only for validation. The model simulated a LULC map for 2012, and by calculating FOM

and its components, furthermore performing intensity analysis, a comprehensive study of land changes in the study site was conducted. FOM, FOM components, and intensity analysis are described in Sections *3.6.2.1 and 3.5.2* in detail.

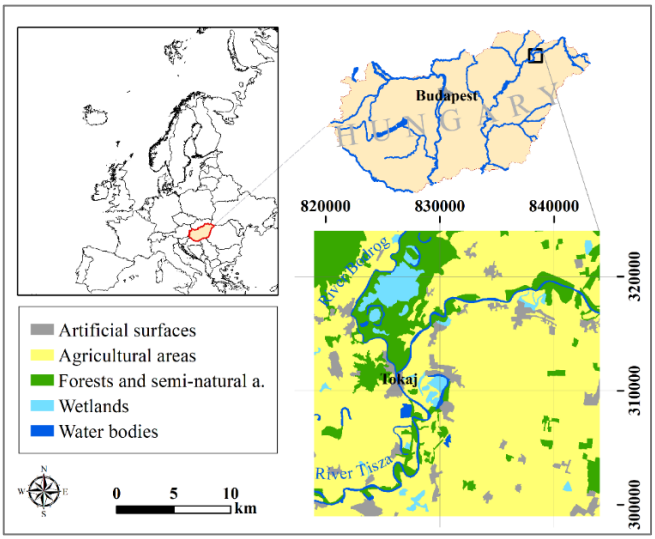

Figure 1. *Location of study site group 1, consisting of one study site, an area located around Tokaj, NE Hungary.*

3.2.2. Study Site Group 2

In this study site group, models were set up in eight different landscapes across Europe. The landscapes were selected solely based on the quantity of changing areas according to CLC change layers for 2000-2006 and 2006-2012. CLC change layers have been produced with a finer minimal mapping unit (5 ha; Büttner, 2014)), therefore change layers contain more detailed information related to a simple comparison of CLC LULC map layers from different dates. The selected landscapes were subject to as high ratio of changes as possible in at least one of the time intervals of the analysis (2000-2006 or 2006-2012). Further condition for selection was that the study sites must have had 20 categories as a maximum in each relevant date (2000, 2006 and 2012) according to CLC Level 3 classification, and must have had exactly the same number of categories in at least the first two dates (2000 and 2006). The CA-Markov model can handle only the cases where the category numbers are equal in the calibration interval (interval between 2000 and 2006), therefore the study areas must not have change concerning the number of categories in the calibration interval. It points to the fact that the model cannot handle newly appearing or vanishing categories.

Under the described conditions 24 areas were selected in 8 different landscapes, because in each landscape three zoom levels were applied. Large (L) zoom level consisted of the whole selected landscape and two further zoom level were assigned completely within the large area: medium (M) and small (S) subareas. Therefore, the small subarea was always a subset of the medium subarea, and the medium subarea was always a subset of the large subarea. Selected subareas were clipped from CLC's 100 m resolution raster layers. The 8 landscapes were named after the closest cities (Figure 1) in order to identify them more easily. Finally, all the selected subareas had the following characteristics:

- the subareas had the exact same area by zoom level (L = 62500 ha, M = 15625 ha, S = 2500 ha), therefore the subareas had the exact same pixel number by zoom level;
- all the subareas had the exact same 100 m pixel resolution, independently from zoom levels;
- the subareas had the exact same category numbers in 2000 and 2006;
- the subareas had the largest ratio of changing area possible.

Classes of all the 24 areas were aggregated according to various aggregation methods described in *Section 3.3.* and *Figure 5* in detail. Therefore, five LULC maps were created in all the 24 areas – original data and further four ways of aggregation – which increased the number of observations to 120 (= 8 landscapes * 3 zoom levels * 5 aggregation methods). There were 6 exceptions in case of one aggregation method where the aggregation did not make sense – reasons detailed in *Section 3.3.* – which resulted in 114 cases altogether. For all these 114 cases, CA-Markov models were run, and further variables were calculated concerning model performance (FOM, FOM components, quantity and allocation disagreement of simulation), comparison of reference and simulated 2012 maps (Overall Agreement, Kappa index of Agreement), simple metrics of changing areas and temporal stability.

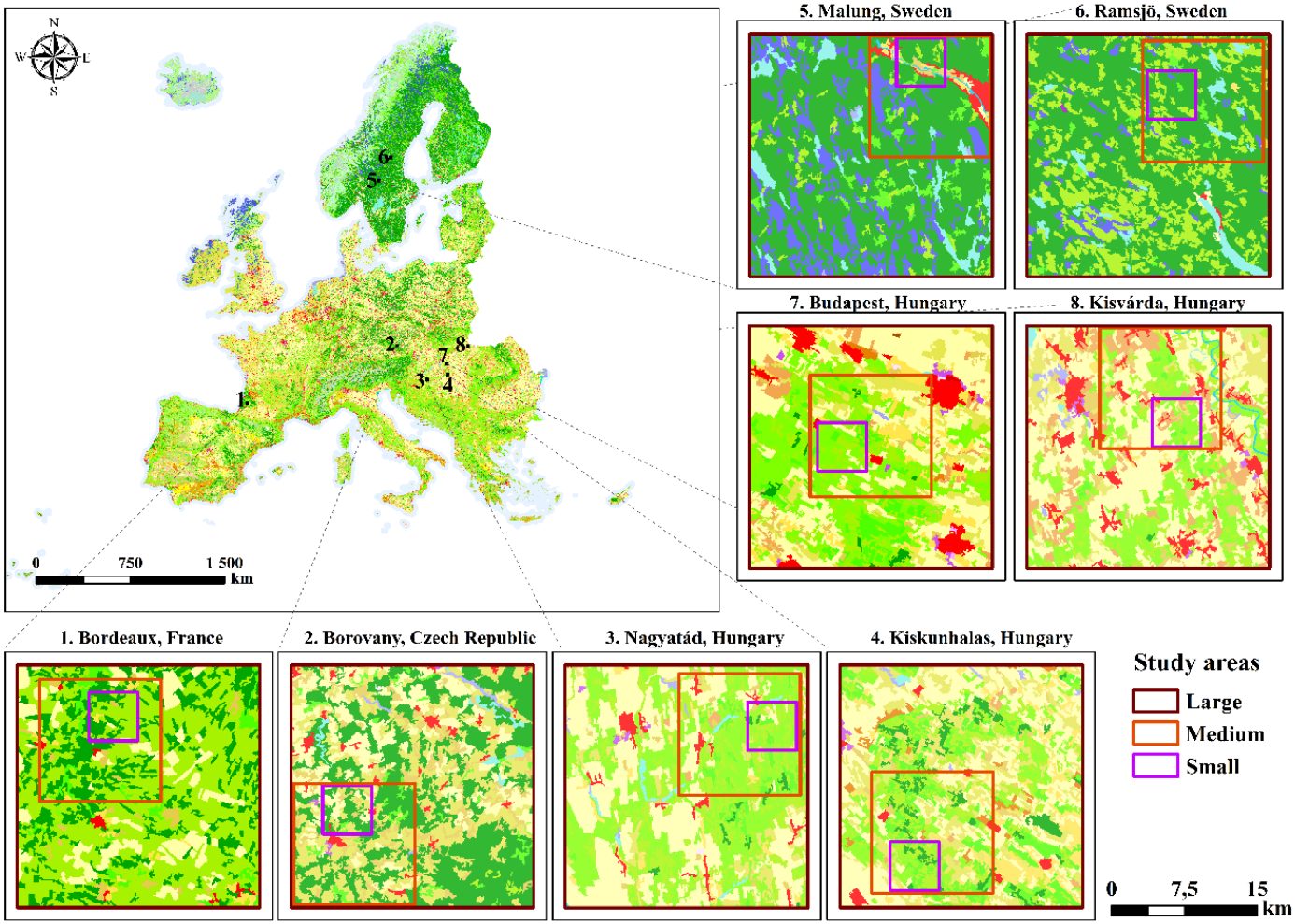

Figure 2. *Study sites of study site group 2. The sites are located across Europe, specifically across the area of EEA countries, covered by Corine Land Cover dataset. The figure also presents zoom levels (large, medium and small) assigned with squares with different colors inside the study sites, in CLC 2012 LULC maps. The original figure was*

3.2.3. Study Site Group 3

This group consists of two study extents (Atchafalaya Bay area, Amazonian area) and 4 more subareas, generated by stepwise zooming into the extents, similar to the process in study site group 2. Altogether there were six areas in this group, two large (L), two medium (M) and two small areas (S). These study extents are located in America (*Figure 3*) and they have several characteristics that discriminate them against the other two study site groups.

These two extents were assigned in order to specifically test the performance of the model in areas where the changes show a sprawl-like pattern, i.e. change that is concentrated in the areas neighboring the original categories. In Atchafalaya Bay the main accelerator of changes is delta accumulation (Atchafalaya River delta and Wax Lake Outlet). Studies about the delta accumulation dates back to the 1980's (DeLaune et al., 1987; Tye and Coleman, 1989). In the Amazon, the main accelerator of dramatic changes is deforestation, which is a long-term problem in Brazil (Carvalho et al., 2019), but the problem has renewed in the last years partly due to the new governmental attitude (Carvalho et al., 2019; de Area Leão Pereira, E. J. et al., 2019). Both described environmental changes can be regarded as clear examples of sprawl-like phenomena.

The processing scheme of these study extents was also different, because the maps were derived from Landsat Thematic Mapper images. These images were downloaded from Earth Explorer website of the United States Geological Survey, as known as USGS (U.S. Geological Survey, 2016). Landsat images are widely used in scientific researches (Almeida et al., 2016; Ruelland et al., 2008; Viana et al., 2019; Zhu and Woodcock, 2014), partly due to their long-term availability, since it has provided continuous data from the 1970's. This long-term availability and relatively dense – 16 days – revisit time makes it possible to perform long-term monitoring studies. These two study sites are not covered by CLC area of interest, thus it was impossible to use the same dataset for the analysis as in the other two study site groups. Since these two areas were also subjects for running CA-Markov models, they had to meet the requirements of running CA-Markov, e.g. equal number of categories in all maps. These two areas' analysis also had the special purpose of modelling sprawl-like phenomena. That is why the LULC maps were evolved by determining 2 categories in both study sites, enhancing the relevant phenomena. First category determines the category that would potentially sprawl in terms of the examined phenomenon, and the other category functions as the background and target of the change generated

by the phenomenon. Therefore the first category, e.g. deforestation in the Amazonian study site, is presumed to show a large amount of gain, while the other category (forest) loses area along with the spread of deforestation, thus suffering the change.

Three Landsat images were processed in each study area from three different dates with quasi-equal time intervals between them. It was a requirement when downloading the images that they must have been acquired in the same season (or with a maximum of 2 months difference within the same period of the year). The images were acquired in 1990, 2000 and 2010 with almost equally 10 years between them in case of the Amazonian site. The images were acquired in 1990, 2003 and 2016 with almost equally 13 years between them in case of the Achafalaya Bay site. They were processed in Trimble eCognition software via segmenting the images by a multiresolution segmentation ruleset and the segments were classified into two categories based on visual interpretation. The segmentation process was supported by using NDVI layer in case of forest and MDNWI layer in case of water, in order to identify forest and water land covers more effectively. These indices could be computed based on the original bands of Landsat images (Baret et al., 1989;Xu, 2006). The accuracy assessment procedure was performed in accordance with Congalton's (1991) and Cochran's (1977; Olofsson et al. 2014) recommendations. The reported overall accuracy was over 85% in each map. The same processing scheme was used in this case as in a study area in Nyírség, NE Hungary before, where we achieved high classification accuracy related to a pixel-based classification approach. Furthermore, a pixel-based classification approach frequently results in salt and pepper effect that would be disadvantageous when analyzing a sprawl-like phenomenon. We reported the accuracy results of this processing scheme in Varga et al. (2014). In order to match the resolution of study site group 2 maps, the two-class LULC maps were resampled from the original 30 m spatial resolution of Landsat to 100 m by the nearest neighbor resampling method.

Based on the CA-Markov simulations, further variables were calculated, similar to study site group 2, concerning either model performance (FOM, FOM components, quantity and allocation disagreement of simulation), comparison of relevant reference and simulated maps of 2010 and 2016 (Overall Agreement, Kappa index of Agreement), simple metrics of changing areas or stationarity (Runfola's R values).

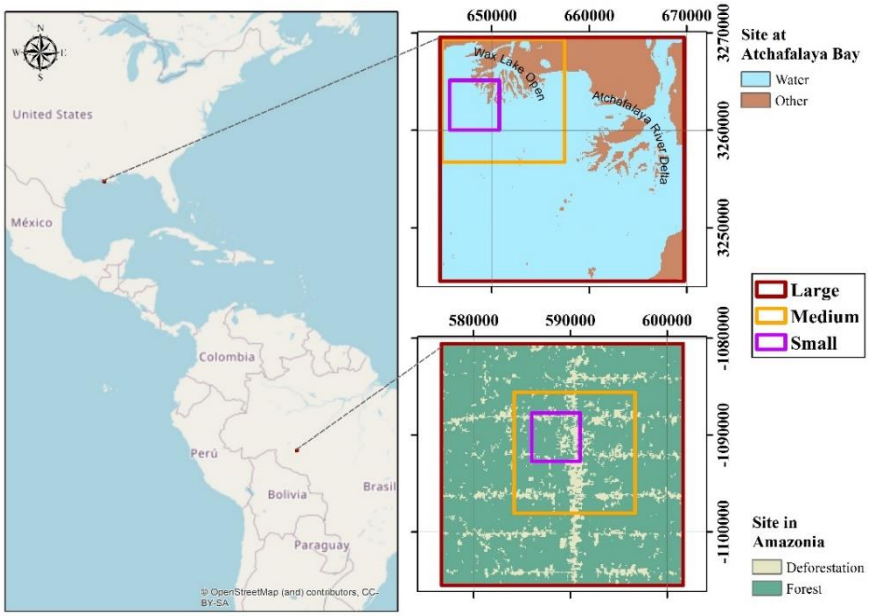

Figure 3. *Study sites of study site group 3. One site is located around Atchafalaya Bay and Wax Lake Outlet, one site is located in Amazonia, Brazil. The overview map is presented based on an Open Street Map layer (CC-BY-SA).*

3.3. Aggregation methods

In this section, the summary of aggregation methods used in study site group 2 are described. To provide a clear overview of these aggregation methods, an example of the classification schemes applied in a specific study site is presented in *Figure 5* at the end of *Section 3.3.*

3.3.1. Corine Standard Levels

CLC classification scheme has a nested hierarchical nomenclature with 3 standard levels. The CLC dataset basically classify all the areas of EEA countries into 44 categories, processed with respect to match a thorough technical guideline (Büttner et al., 2004). When updating CLC datasets, new remotely-sensed datasets and technologies with more developed features were involved in the processing workflow, keeping up the pace with continuously developing technological innovations (Büttner and Kosztra, 2017). Although the processing methods and the base data varied in case of different CLC datasets over time, there are uniform characteristics that remained consistent over time, namely the 25 ha minimal mapping unit and the guaranteed 85% thematic accuracy. These uniform characteristics ensure a relatively common basis for analyses of the datasets (Büttner, 2014).

In this research, all the three hierarchical levels of CLC datasets were used. The areas determined by study site group 2 were clipped from CLC Level 3 dataset, with respect to equal category numbers in 2000 and 2006, and with a maximum of 20 categories in all study sites. CLC Level 3 is a category scheme that was used in the analysis as a basis and no aggregation was performed in the data at all. However, all the aggregations were based on this data. Throughout the analysis, this scheme is referred to as *CLC Level 3 (L3)* method.

CLC Standard Level 2 is a superior hierarchy level related to Level 3, and classifies Level 3 categories into a maximum of 15 categories. The categories were aggregated based on the hierarchical nomenclature scheme by a reclass procedure. Throughout the analysis, this aggregation is referred to as *CLC Level 2 (L2)* aggregation.

CLC Standard Level 1 is a superior hierarchy level related to Levels 2 and 3, and classifies all categories into a maximum of 5 categories. The categories were aggregated based on the hierarchical nomenclature scheme by a reclass procedure. Throughout the analysis, this aggregation is referred to as *CLC Level 1 (L1)* aggregation.

3.3.2. Behavior-based category aggregation

The main aim of this type of aggregation method is to maintain net change, which is the change originating from quantity differences between two dates (Aldwaik et al., 2015). This information can be derived from an error matrix set up between the maps of the two dates. There is a Visual Basic for Applications (VBA) macro published by the authors of the concept for extracting this information and to follow step-by-step whether various types of change starts to decrease when aggregating a pair of classes. The macro advises pairs of classes to aggregate while shows the actual net and swap change for the user (*Figure 4 and Table 1*).

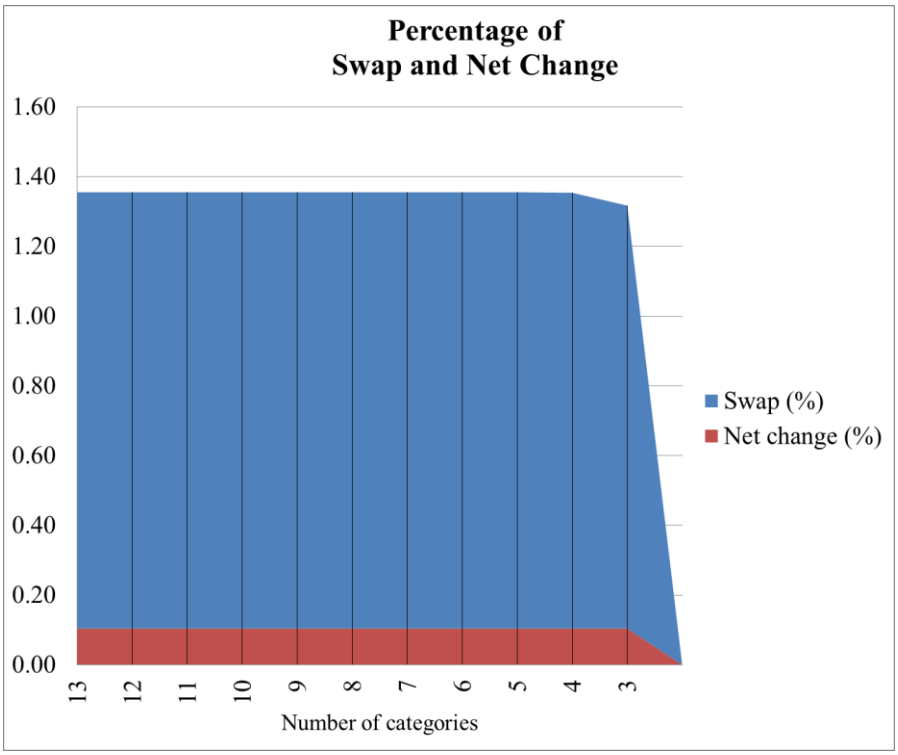

Figure 4. *Diagram produced by behavior-based category aggregation VBA macro when performing aggregation of study site Malung, zoom level L. The diagram shows different types of change metrics and the decrease in change metrics when reaching different numbers of categories. Net change is change originating from quantity differences between two dates, swap change is a type of change originating from location.*

Swap change is a type of change attributed to location (Pontius Jr et al., 2004). The user can decide in each step if they want to continue with aggregation. The macro does not perform the aggregation, just show a possible scenario for aggregations and their consequences regarding net change. This macro was used for executing the aggregation for all the study sites and all zoom levels, 24 areas altogether. However, this aggregation concept has not been capable of managing more than one time interval simultaneously in a manner that considers the changes in both time intervals. That is why the L3 classes of the calibration interval were aggregated dictated by the behavior-based aggregation method, then the classes of validation interval were aggregated similarly to the aggregation rules of the calibration interval. In this way, the categories of the two time intervals became comparable and were influenced by hidden changes the least as possible. Throughout the analysis, this aggregation is referred to as *behavior-based (BB)* aggregation.

Table 1. *Table edited on the basis of information provided by behavior-based category aggregation VBA macro when performing aggregation of study site Malung, zoom level L. The table shows different types of change metrics and their decrease while advising aggregation of certain pairs of categories.*

Number of Categ.	Total Change (%)	Swap (%)	Net change (%)	Order of Aggregation	Type of Aggregated Categories
13	1.36	1.25	0.10	Pastures & Inland marshes	N/Dormant & N/Dormant
12	1.36	1.25	0.10	Aggregated 13 & Broad-leaved forest	N/Dormant & N/Dormant
11	1.36	1.25	0.10	Non-irrigated arable land & Aggregated 12	N/Dormant & N/Dormant
10	1.36	1.25	0.10	Aggregated 11 & Water courses	N/Dormant & N/Dormant
9	1.36	1.25	0.10	Aggregated 10 & Complex cultivation patterns	N/Dormant & N/Dormant
8	1.36	1.25	0.10	Discontinuous urban fabric & Aggregated 9	N/Dormant & N/Dormant
7	1.36	1.25	0.10	Aggregated 8 & Water bodies	N/Dormant & N/Dormant
6	1.36	1.25	0.10	Land principally occupied by agriculture, with significant areas of natural vegetation & Peat bogs	L/Loser only & L/Net losing
5	1.36	1.25	0.10	Aggregated 6 & Coniferous forest	L/Net losing & L/Net losing
4	1.35	1.25	0.10	Mixed forest & Transitional woodland-scrub	G/Net Gaining & G/Net Gaining
3	1.32	1.21	0.10	-	-

3.3.3. Threshold-based category aggregation

This aggregation is an arbitrary manner of aggregating the categories, where the user can determine a threshold which they respect as being important. Here, a 0.1% of changes of the total actual study area was determined to be the threshold, and all categories that showed a change less than 0.1% of the actual study area, were aggregated into a

collective category. This collective category was called *Other*, referring to its character of collecting every category that did not meet the requirement. Therefore, this category can be thematically diverse. If in a specific study area there were only categories which show larger changes than 0.1% of that study area, respectively, then neither of them would be aggregated into a category called *Other*. This situation occurred in 6 areas from the 120 areas altogether, consequently there were 114 model runs at all. Throughout the analysis, this aggregation is referred to as *threshold-based (TB)* aggregation.

Figure 5 is a comprehensive summary of category aggregations, presented via the example of study site Malung, zoom level L. This area originally consisted of 13 categories in CLC Standard Level 3, and these categories were aggregated into Standard Level 2 and Standard Level 1 according to the CLC hierarchical scheme. L2 has broader definitions of categories than L3, and L1 has even general categories with basic LULC definitions that opens the door for categories with mixed characteristics. BB aggregation aggregated classes into thematically extremely diverse categories, e.g. aggregating urban areas and water in a common category, while focusing on maintaining changes in the area. TB also focused on enhancing the changes in the area, but with determining a strict threshold of changes that they cannot exceed. However, the aggregations could vary with modifying this strict threshold.

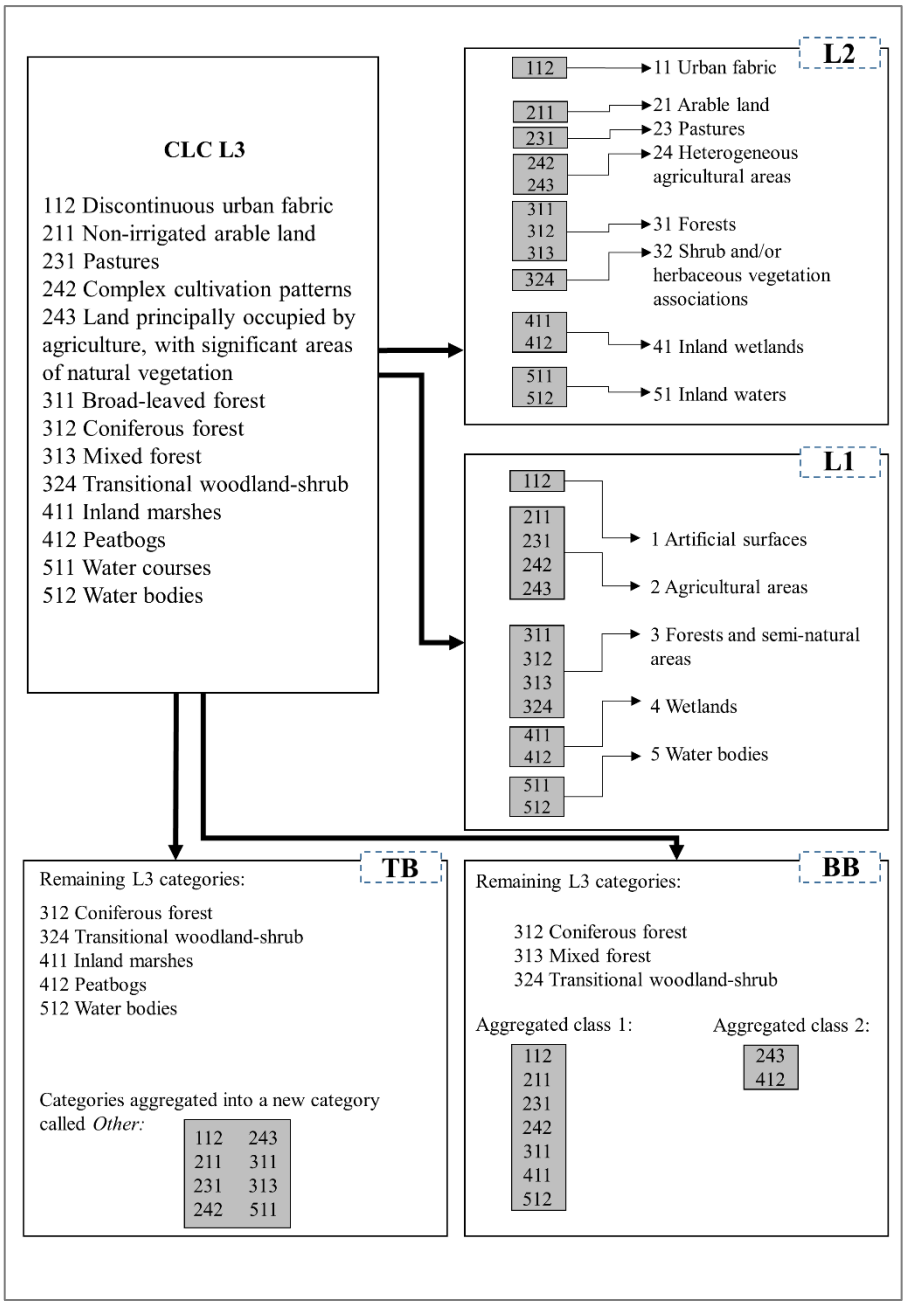

Figure 5. *Flowchart of various aggregation methods in study site Malung, Zoom level L.*

3.4. CA-Markov Model

In this research, a combined CA-Markov modelling method was applied in Idrisi Selva software environment. Markov-chain models are stochastic models, their output is distribution among states, which is based on the probability of transitions (Baker, W., 1989). They can model the future state of a system based purely on its preceding states (Eastman, 2012a). Markovian models are often used for projections of land cover in GIS workflows, based on transition probability tables, which are set up by gaining information about transitions concerning the areas of land cover categories (Mas, J. F. and Vega, 2012).

Markov analysis in land cover change prediction is based on the state of a system at time #1 and time #2. During the Markov analysis, transition probability matrix, transition area matrix and conditional probability maps are produced based on the input data, which can be considered as the training data. The transition probability matrix gives the likelihood of change of a pixel of a given class in the next time interval (between time #2 and time #3) and provides information about the probability of that a pixel characterized with a certain category transitions into a different category (Bruzzone and Serpico, 1997; Lóczy, 2010; Schweitzer, 1968; Singh et al., 2015). A collection of transition probability matrices shows the probabilities of all relevant combinations of transition.

The transition area matrix gives the number of pixels expected to change from a given class to all other classes. The conditional probability maps give the probability that each pixel will belong to the designated class in the next time interval, reporting the probability that each class type would be found in each pixel, as a projection from the time#2 map (Eastman, 2012a; Eastman, 2012b).

The cellular automaton was introduced by Neumann and Ulam in the 1940's through the problem whether self-reproduction of biological systems can be described only by mathematical formulas and logical rules in case of driving factors (Benenson and Torrens, 2004). Cellular automata consist of a regular grid of cells which sets out a dynamical system considering time and space as being discrete. The neighboring cells' previous states determine the state of the cell itself, and it is updated in discrete time steps based on identical rules (Sipper, 1997).

The CA-Markov model is a combination of the cellular automaton and the Markov chain analysis, implementing the capability of Markov analysis to project forward in time and also the cellular automaton's sensitivity to the neighborhood. Thus, the module is capable of projecting a state for time #3 based on states of time #1 and time #2,

according to the matrices of Markov analysis step, while considering the states of neighboring cells and suitability of pixels for each category of the map. Therefore, the Markov analysis helps to determine the quantity of change and CA-Markov analysis step helps to determine the spatial allocation of change (Eastman, 2012a; Mas, Jean-François et al., 2014). There are specific transition rules for the model, which can be mathematically expressed, and they govern the changes of cell characteristics during the projection, i.e. the simulation process (Mitsova et al., 2011).

It is important to declare the following references that I use while describing the study design:

- I refer to the time interval between time #1 reference map and time #2 reference map as *calibration interval*, which is used for training or calibrating the model;
- I refer to the time interval between time #2 reference map and time #3 reference map as *validation interval*, which is used for the validation of the model;
- I refer to the time interval between time #2 reference map and time #3 simulation map as *simulation interval*, where time #3 map assigns the map produced by the simulation model.

In study site groups 1 and 2, as in these cases CLC data is applied, the following statements are true:

- the time #1 map is the LULC categorical map of 2000;
- the time #2 map is the LULC categorical map of 2006;
- the time #3 map is the LULC categorical map of 2012;
- the time #3 simulation map is the 2012 LULC map simulated by the CA-Markov model.

Therefore, the following time intervals are used in these study site groups:

- calibration interval is the interval between 2000 reference map and 2006 reference map;
- validation interval is the interval between 2006 reference map and 2012 reference map;
- simulation interval is the interval between 2006 reference map and 2012 simulation map.

In study site group 3, the following statements are true for the site located in Amazonia:

- the time #1 reference map is the LULC categorical map of 1990;

- the time #2 reference map is the LULC categorical map of 2000;
- the time #3 reference map is the LULC categorical map of 2010;
- the time #3 simulation map is the 2010 LULC map simulated by the CA-Markov model.

Therefore, the following time intervals are used in the study site located in Amazonia:

- calibration interval is the interval between 1990 reference map and 2000 reference map;
- validation interval is the interval between 2000 reference map and 2010 reference map;
- simulation interval is the interval between 2000 reference map and 2010 simulation map.

In study site group 3, the following statements are true for the site located in the Atchafalaya Bay:

- the time #1 reference map is the LULC categorical map of 1990;
- the time #2 reference map is the LULC categorical map of 2003;
- the time #3 reference map is the LULC categorical map of 2016;

Therefore, the following time intervals are used in the study site located in the Atchafalaya Bay:

- calibration interval is the interval between 1990 reference map and 2003 reference map;
- validation interval is the interval between 2003 reference map and 2016 reference map;
- simulation interval is the interval between 2003 reference map and 2016 simulation map.

The models were solely based on preceding states, and no driving factor was included. Besides the input time#1 and time#2 categorical maps, an iteration number and contiguity filter are further obligatory parameters in the model. Iteration number is advised to be the number of years that the modeler wishes to project forward (Eastman, 2012a), so an iteration number of 6 in study site group 1 and 2, and iteration numbers of 10 and 13 in study site group 3. The contiguity filter is a 5x5 spatial filter as default, with a possibility to change, but there was no specific or obvious reason to change this parameter.

3.5. Change analysis

3.5.1. The error matrix

The definition of error matrix was introduced by Congalton (1991). "*An error matrix is a square array of numbers set out in rows*

and columns which express the number of sample units (i.e., pixels, clusters of pixels, or polygons) assigned to a particular category relative to the actual category as verified on the ground." (Congalton, 1991). Error matrix is also known as either crosstabulation matrix, confusion matrix or contingency table. Crosstabulation matrix is a commonly used tool as a basis for accuracy assessment in remote sensing and various metrics can be derived from the values of the matrix (Foody, 2002). Frequently derived values are commission and omission errors, also known as user's and producer's accuracy (Story and Congalton, 1986). Gopal and Woodcock (1984, 2010) advised a fuzzy approach, where the interpretation of the matrix exceeds the idea of simple agreement and disagreement, but extends to a larger set of possible responses, like acceptable or understandable situations.

In *Table 2* and *3* different interpretation options of the crosstabulation matrix are presented. *Table 2* shows the interpretation of a crosstabulation matrix used in accuracy assessment of remote sensing applications, for the purpose of the comparison of reference and classified image data. In these cases, columns assign the reference data, rows assign the comparison data.

Table 2. *The interpretation of a crosstabulation matrix with an approach of comparison of reference and classified image data, based on the published theoretical description of Congalton (1991).*

		REFERENCE DATA		
		Class A	**Class B**	**Class C**
COMPARISON DATA	**Class A**	**pixels classified correctly**	number of pixels that belong to class B in reference data and belong to class A in comparison data	number of pixels that belong to class C in reference data and belong to class A in comparison data
	Class B	number of pixels that belong to class A in reference data and belong to class B in comparison data	**pixels classified correctly**	number of pixels that belong to class C in reference data and belong to class B in comparison data
	Class C	number of pixels that belong to class A in reference data and belong to class C in comparison data	number of pixels that belong to class B in reference data and belong to class C in comparison data	**pixels classified correctly**

Table 3 shows the interpretation of a crosstabulation matrix used in change analysis, for the purpose of the comparison of categorical maps from different dates. In LULC change analysis, maps can be compared on the basis of a confusion matrix from different dates, and unlike thematic accuracy assessment, there is no reference in this case. Rows (first date) and columns (second date) have equal role and the result is not the error, but the quantified change. The raw information are the pixel quantities in the matrix diagonal which indicate the persistent areas (Pontius Jr et al., 2004) that corresponds the overall accuracy in thematic accuracy assessment (Congalton, 1991). We can also calculate the changes of the first map against the other one, or vice versa, and can reveal what class another one turned into, therefore we can also reveal what was the previous land class before the conversion.

***Table 3**. The interpretation of a crosstabulation matrix with an approach of changes in the landscape between two categorical maps of different dates (Time 1 and Time 2), based on the published theoretical description of Pontius et al. (2004)*

		TIME 2 MAP		
		Class A	**Class B**	**Class C**
TIME 1 MAP	**Class A**	**persistence**	pixels changed from class A to class B	pixels changed from class A to class C
	Class B	pixels changed from class B to class A	**persistence**	pixels changed from class B to class C
	Class C	pixels changed from class C to class A	pixels changed from class C to class B	**persistence**

3.5.2. Intensity analysis

A deeper change analysis was performed in study site group 1 based on the error matrices. Intensity analysis is a method to quantify the change intensity of a categorical variable at interval, category and transition levels across different time intervals. The method was applied to examine LULC changes with a Microsoft Excel VBA macro introduced by Aldwaik and Pontius (2012). We can use different error matrices as inputs according to the number of time intervals we intend to investigate. Aldwaik and Pontius (2012) published an equation which

gives the uniform rate of change for the entire time extent of investigation, and this uniform rate would exist if the rate of overall change would be perfectly stationary through the entire temporal extent of investigation. The relation of actual changes in a specific time extent and uniform change is a key factor in intensity analysis.

Interval level shows that during the temporal extent of examination the change of land cover was slow or fast according to the uniform intensity. This uniform intensity can be expressed by a hypothetical value that concerns a perfectly stationary change pattern during overall change. If the annual change value exceeds this uniform intensity value, then change can be regarded as fast for that time interval. If annual change value is less than uniform intensity value, then change can be regarded as slow for that time interval.

Category level shows whether a category is active or dormant within a given time interval, based on a uniform intensity value as well – for that specific interval. This level concerns the annual gain and loss of each categories and relates them to the uniform intensity value of that time interval. If the annual gain or loss intensity value is less than uniform intensity, then the category's gain or loss is dormant concerning that time interval. If the annual gain or loss intensity value is more than uniform intensity, then the category's gain or loss is active concerning that time interval. If the change was uniform across the landscape, then all the categories' annual gain and loss intensity value would equal the uniform intensity. We can regard a category as being stationary in terms of losses or gains if its intensity value is more or less than the uniform intensity across all the examined time intervals.

Transition level focuses on given transitions from one category to another. This level focuses on which category gains the loss of another one, and vice versa, and based on these observations, we can determine which categories are targeted or avoided by another category's loss or gain (Aldwaik and Pontius Jr, 2012; Aldwaik and Pontius Jr, 2013).

It is useful to highlight that intensity analysis determines the followings:

- the changes in a certain time interval are fast or slow related to uniform change;
- a category is active or dormant in terms of gains and losses;
- a category is targeted or avoided by transitions in the actual spatial extent.

For performing intensity analysis the equations of Pontius et al (2013) (Pontius Jr et al., 2013) were used, because the durations of time intervals are identical throughout the time extent in this case. Aldwaik and Pontius

(2012) equations concerning intensity analysis focus on time intervals with different durations.

3.6. Model validation

3.6.1. Approaches that does not distinguish land persistence and model performance

3.6.1.1. Kappa Index of Agreement

Kappa Index of Agreement, henceforth referred to as KIA, is often called as Kappa statistics or Kappa coefficient as well. KIA originates from Galton (1982), but its origin was frequently associated with Cohen (1960). It was later invoked for the purposes of accuracy assessment of remotely sensed data (Rosenfield and Fitzpatrick-Lins, 1986), on the grounds of Congalton et al. (1991) and Congalton and Mead (1983) articles. The exact calculation is often cited from Bishop et al. (1975), but it can be found in literature in a way easier to interpret as well (Banko, 1998). *Equation 1* gives the formula of Kappa coefficient based on Bishop (1975), published by Mather (2004):

$$K = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+} x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+} x_{+i}} \qquad \text{(Eq. 1.)}$$

where K = Kappa coefficient; x_{ii} = diagonal entries of the error matrix; x_{i+} = sum of row i of the error matrix; x_{+i} = sum of column i of the error matrix; N = total number of elements in the error matrix.

There are more Kappa variations which were introduced by Pontius (2000), but Pontius and Millones (2011) advised to use quantity and allocation disagreement indicators instead of these kappa variations. There is no uniform scale to interpret Kappa value, however, several approaches exist for assessing the results e.g. Viera and Garrett (2005) or Fleiss (1981).

I used KIA for the purpose of measuring the agreement between the reference 2012 and simulation 2012 maps. As it is described above, it is an often used method for validation of a land change model, but it is an incorrect approach to the problem. By calculating KIA, I intend to present the flaw of this index in case of using it for model validation.

3.6.1.2. Quantity and allocation disagreement as a tool for accuracy assessment in remote sensing applications

Quantity and allocation disagreements are indices introduced for accuracy assessment purposes instead of Kappa variations and give a

more reliable insight of disagreement between the maps concerning errors of quantity and allocation differences. Pontius and Millones (2011) defined quantity disagreement as „*the amount of difference between the reference map and a comparison map that is due to the less than perfect match in the proportions of the categories"*. Pontius and Millones (2011) defined allocation disagreement as *"the amount of difference between the reference map and a comparison map that is due to the less than optimal match in the spatial allocation of the categories, given the proportions of the categories in the reference and comparison maps"*. Both of them can be calculated by values derived from the error matrix of the comparison and reference maps, and their sum returns the total disagreement of the comparison and reference map, which is equal to the complement of the proportion of pixels that belong to the same class in both maps (Pontius and Millones, 2011). The following formulas of quantity disagreement (*Eq. 2.)* and allocation disagreement (*Eq. 3.)* were published by (Warrens, 2015a) based on (Pontius Jr, R. G. and Millones, 2011):

$$Q = \frac{1}{2}\sum_{i=1}^{c}|p_{+i} - p_{i+}| \qquad \text{(Eq.2.)}$$

where Q = quantity disagreement; pi+ and p+i are row and column total of the error matrix; c=number of categories; C=number of units (pixels) classified correctly.

$$A = \left[\sum_{i=1}^{c} MIN(p_{+i}, p_{i+})\right] - \sum_{i=1}^{c} p_{ii} = \left[\sum_{i=1}^{c} MIN(p_{+i}, p_{i+})\right] - C$$

(Eq.3.)

where A = allocation disagreement; pi+ and p+i are row and column total of the error matrix; c=number of categories; C=number of units (pixels) classified correctly.

The sum of quantity and allocation disagreement gives the total disagreement (Pontius Jr, R. G. and Millones, 2011). The calculations in *Equation 2* and *3* can be conducted automatically by a macro called the PontiusMatrix, which is freely available at Dr. Robert Gilmore Pontius Jr's website (http://www2.clarku.edu/~rpontius/) and was developed especially for this purpose. The spreadsheet returns the components of these calculations (quantity, exchange, shift) which gives the quantity disagreement (quantity component), allocation disagreement (sum of

exchange and shift component) and total disagreement (sum of quantity and allocation disagreement) (Pontius and Santacruz, 2014). Warrens, (2015b) also published formulas for relative quantity and allocation disagreement indices which are category-level variants of original quantity and allocation disagreement indices. The complement of total disagreement is the overall agreement (OA), also referred to as overall accuracy, which is the sum of correctly classified pixels in the crosstabulation matrix. I did not calculate quantity, allocation or total disagreement values, only their complement, the overall agreement between reference and simulation time#3 maps. Using these metrics for calculating disagreement between reference and simulation time #3 maps would be just as misleading as calculating Kappa and overall accuracy. I presented quantity and allocation disagreement to underline the differences between them and their namesake: the quantity and allocation errors used for determining quantity and allocation errors of a simulation (*Section 3.6.2.2*). I calculated overall agreement in case of study site groups 2 and 3 in order to present the flaw of this concept in model validation applications.

3.6.2. Approaches that distinguish land persistence and model performance

3.6.2.1. Figure of merit and components

The figure of merit (FOM) is a measurement which characterizes the match of observed and simulated change, latter projected by a simulation model. The FOM is calculated as dividing the intersection of observed and predicted change by the sum of observed and predicted change (Pontius et al, 2008; Klug et al. 1992; Perica and Foufoula-Georgiou 1996). If the observed and simulated change did not match at all, the FOM would return a value of 0%. If the observed and simulated change matched perfectly, the FOM would return a value of 100%. By calculating FOM components, we can get the various types of errors and agreements expressed as a ratio of the actual study area. The FOM components provide a deeper insight into the errors of changes, as follows:

- *Hits* = area of reference change simulated as change to the right category (agreement);
- *Misses* = area of reference change simulated as persistence (error)
- *Wrong Hits* = area of reference change simulated as change to a wrong category (error)

- *False Alarms* = area of reference persistence simulated as change (error)

FOM can be calculated based on FOM components (Pontius Jr, R. G. et al., 2011) as expressed in *Equation 4*:

$$FOM = \frac{Hits(100\%)}{Misses+Hits+Wrong\ Hits+False\ Alarms} \quad \text{(Eq.4.)}$$

FOM components were calculated by 'lulcc' package (Moulds et al., 2015) available in R software. A further component interpretable as agreement, called Correct Rejection, can be described as the persistence simulated as persistence. These metrics were calculated in case of all study site groups.

FOM components can be visualized for each pixel of the LULC map by a raster calculator command, in softwares where the implementation of conditions is possible when performing raster calculations. I visualized FOM maps by applying the following nested conditional expression in ArcGIS raster calculator:

```
Con((SIM2 == REF2)&(REF2 == REF1),1,Con((REF1 == REF2)&(REF2 != SIM2),2,Con((REF1 != REF2)&(REF1 != SIM2)&(REF2 != SIM2),3,Con((REF1 != REF2)&(REF2 == SIM2),4,Con((REF1 != REF2)&(REF1 == SIM2),5,0)))))
```

where REF1 = time #2 reference map; REF2 = time#3 reference map; SIM2 = time #3 simulation map, and the numbers return the FOM components according to the conditions.

3.6.2.2. Quantity and allocation disagreement as a tool for validation of a simulation model

There are two other metrics with a different purpose, but with an identical name of quantity and allocation disagreement. These two metrics aim to determine the error of simulation due to quantity of predicted change (quantity disagreement) and due to allocation of predicted change (allocation disagreement). I aim to distinguish these two metrics from quantity and allocation disagreement of Pontius and Millones (2011) unambiguously, by adding the abbreviation of the word simulation (A_S and Q_S) when referring to them. These metrics were described by Liu et al. (2014) and Chen and Pontius (2010). According to Chen and Pontius (2010), quantity disagreement in terms of observed and predicted change, measures "*how much less than perfect is the match*

between observed and predicted quantity of change". According to Chen and Pontius (2010), allocation disagreement in terms of observed and predicted change, measures "*how much less than optimal is the match in the spatial allocation of the changes, given the specification of the quantities of the changes in the observed and predicted change maps."* *Equation 5, 6* and *7* determine these metrics based on Chen and Pontius (2014) presenting the calculation of these metrics based on FOM components, as follows:

$$Q_S = |Predicted\ Change - Observed\ Change| = |(False\ Alarms + Hits) - (Misses + Hits)| = |False\ Alarms - Misses| \quad \text{(Eq 5)}$$

$$A_S = (False\ Alarms + Misses) - Q_S = 2\ x\ MIN(False\ Alarms, Misses) \quad \text{(Eq 6)}$$

$$T_S = False\ Alarms + Misses = Q_S + A_S \quad \text{(Eq. 7)}$$

where Q_S= error due to quantity of predicted change; A_S= error due to allocation of predicted change; T_S = total error in predicted change; FOM components are as defined in *section 3.6.2.1.*, and all the variables are expressed as a percent of the study area.
These metrics were calculated in case of study site groups 2 and 3.

3.7. Variables concerning Stationarity

We measured Runfola's R index in the relation of either calibration and validation intervals or calibration and simulation intervals. R index characterizes the temporal instability between time intervals by returning a proportion of change to be reallocated to the other time interval in order to achieve a uniform change during the whole time extent. If R index is 0, then change is perfectly stable. If R is increasing, the change is getting more unstable (Runfola and Pontius Jr, 2013). This measurement is influenced by three factors, one is the duration of the investigated time interval and another one is the temporal extent. Since this research includes models that use calibration and validation time intervals with the same durations, these two factors were constant throughout most of the research. One further factor, the annual change during each time interval, may influence Runfola's R value. Also based on Runfola and Pontius (2013) Runfola's R index is calculated as in *Equation 8*:

$$R = \frac{\sum_{t=1}^{T-1}\{MAXIMUM[0,(S_t - U)] * (Y_{t+1} - Y_t)\}}{U * (Y_T - Y_1)} \quad \text{(Eq. 8)}$$

where S_t = annual change; U = uniform annual, observed in case of change was perfectly stable during the whole examined time extent; Y_t = year at time point assigned with t.

3.8. Other variables concerning change

Other simple metrics were calculated concerning the number of categories and overall change between reference maps, or between reference and simulation maps used in the analysis. These simple metrics can give insight from different aspects into the variation of changes with applying various category aggregations. These variables were the followings:

- number of categories;
- calibration, validation and simulation interval changes;
- difference between calibration and validation interval annual changes in each case of study site groups 2 and 3;
- difference between calibration and simulation interval annual changes in each case of study site groups 2 and 3;
- difference between errors of simulation due to quantity (Q_s) and errors of simulation due to allocation (A_s) in each case of study site groups 2 and 3.

3.9. Statistical analysis

First of all, statistical analysis aimed to reveal the effects of category aggregation. A Tukey test was applied to investigate this issue in study site group 2. The distribution in the data did not follow normal distribution according to a preliminary Shapiro-Wilk test. An analysis with the same purpose was performed in Varga et al. (2020), but with a two-way ANOVA, with the median as an estimator and with bootstrapping (599 repetitions), where H_0 of the analysis were the followings:

- the medians of the five different aggregation approaches were equal;
- the medians of the eight study sites were equal;
- there was no interaction between aggregation approaches and study sites in a statistical sense;

Here, the difference between the aggregation methods was in focus, ignoring the possible effects of study sites, and ignoring it as a factor. Here, the H_0 of the analysis was purely that the medians of the five

different aggregation approaches were equal. Tukey analysis was performed which gave opportunity for pairwise comparisons, as full-factorial comparison between each of the aggregation methods. If medians of datasets related to different aggregations were significantly different based on the statistical test besides a 95% confidence interval, then they were assigned with different letters in the boxplots used for visualization (Piepho, 2004). Statistical analysis was performed in R software (R Core Team, 2020), with the following packages: 'multcomp'(Hothorn et al., 2008),'WRS2' (Mair and Wilcox, 2019) and 'ggplot2' (Wickham, 2016).

In order to test the correlation between model performance and all other variables calculated concerning changes and stationarity, a correlation matrix was set up in Past Statistics software. The correlation matrix was produced along with applying a 95% confidence interval and Spearman's r_s non-parametric rank-order test that does not have an assumption of normal distribution (Hammer et al., 2001). Spearman's correlation coefficient applies a Pearson's equation, after ranking the data (Field, 2013).

In *Figure 6*, there is a comprehensive visualization of the workflow used in the whole dissertation. The figure summarizes the characteristics of the study site groups, the model approach and the full analysis after running the models. Hopefully, this figure supports the overview and comprehension of the whole analysis workflow.

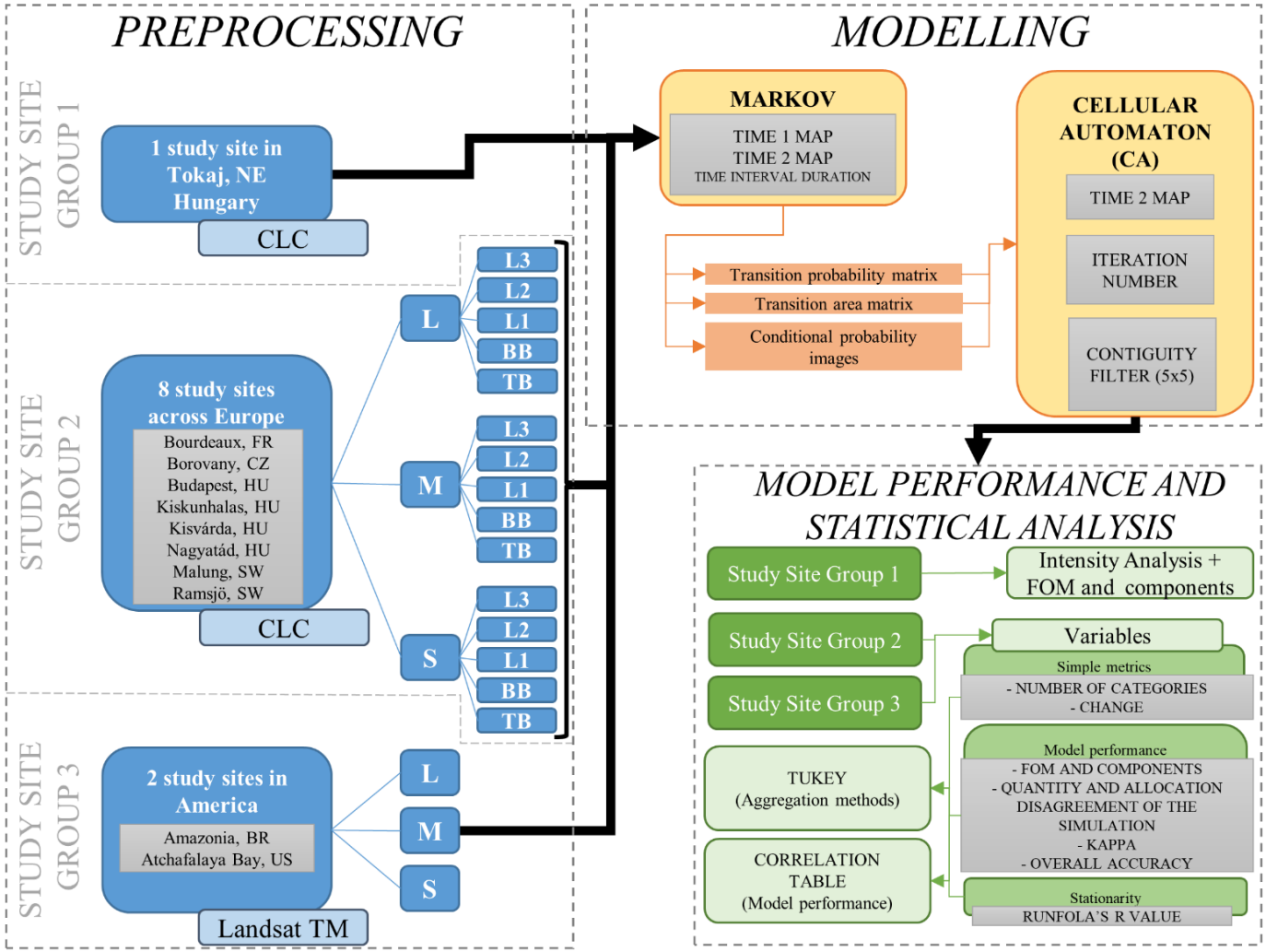

Figure 6. *Full summary of the workflow of the research as described in Methods section.*

4. RESULTS AND DISCUSSION

In this section, the results are described according to study site groups. Since the applied methodology was partly different in the study site groups, it is critically important to interpret the results separately. While reporting the results, I discuss the importance of the results as well. The main conclusions are summarized in the *Conclusions section*.

4.1. Results FOM components and Intensity analysis in Study Site Group 1

In study site group 1, FOM and components were calculated in order to characterize model performance and intensity analysis was performed in order to characterize change. It is important to highlight that the ratio of changing areas was extremely low in the study area, therefore the ratio of correctly simulated change (meaning Hits) in the area must have been low. The overall change was 1.7% in the calibration interval, 1.1% in the validation interval, and 1.5% in the simulation interval. The annual change (overall change divided by the number of years in the time interval) is presented in *Figure 7*. Figure 7 shows that the change decelerated from the calibration to the validation interval, since the annual change in the validation interval was much less than in the calibration interval. The annual change of calibration and simulation interval shows more similarity than calibration and validation interval.

FOM is calculated as Hits divided by the sum of Hits and erroneously simulated pixels in the study area. *Figure 8* shows the FOM components visualized in a map of the study site, where the colored pixels represent Hits and erroneously simulated pixels due to various reasons, such as persistence simulated as change or change simulated as persistence. Hits added up to only 0.02% of the study area. False Alarms were present mainly on the edges of the original categories. Misses were present in the form of compact patches in the landscape. Correct Rejections added up to 97.41% of the study area. Overall FOM was equal to 0.007% in the study area that refers to an extremely low model performance.

On the category level of intensity analysis, the gains and losses of each category were investigated in case of either calibration, validation or simulation intervals. This approach made it possible to observe the dynamics of changes per category in each time interval and to compare the validation and simulation changes on the basis of a more detailed collection of information. The relevant barplots (*Figure 9 A, B and C*) show annual gains and losses on the left side and show gain and loss intensities and dormant or active status on the right side, per

category. A category's gain or loss is active if its gain or loss intensity exceeds the uniform intensity that is assigned with a dashed line. A category's gain or loss is dormant if its gain or loss does not exceed the uniform intensity.

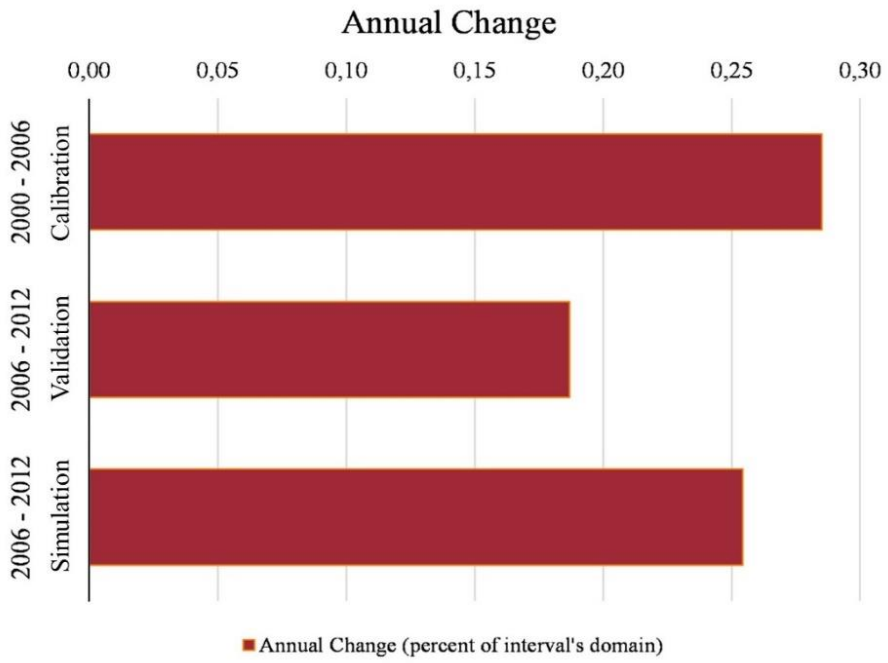

***Figure 7**. Annual change in calibration, validation and simulation intervals in study site Tokaj, NE Hungary.*

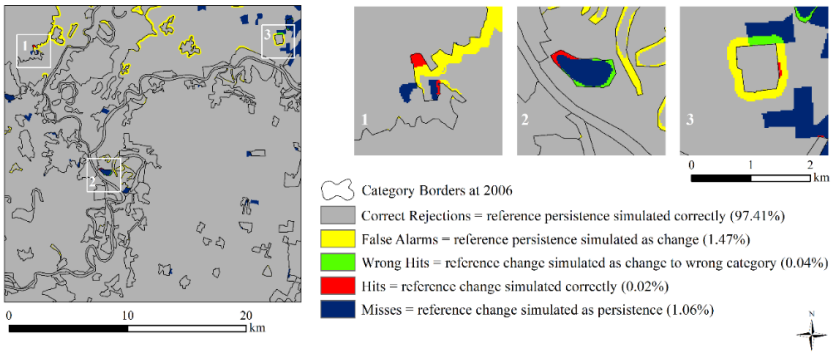

***Figure 8**. Annual change in calibration, validation and simulation intervals in study site Tokaj, NE Hungary. Certain areas are highlighted in boxes 1,2 and 3, where Hits could be observed. FOM components are assigned with different colors. Category borders of 2006 reference map are assigned with lines in the map. This figure was originally published in Varga et al. (2019).*

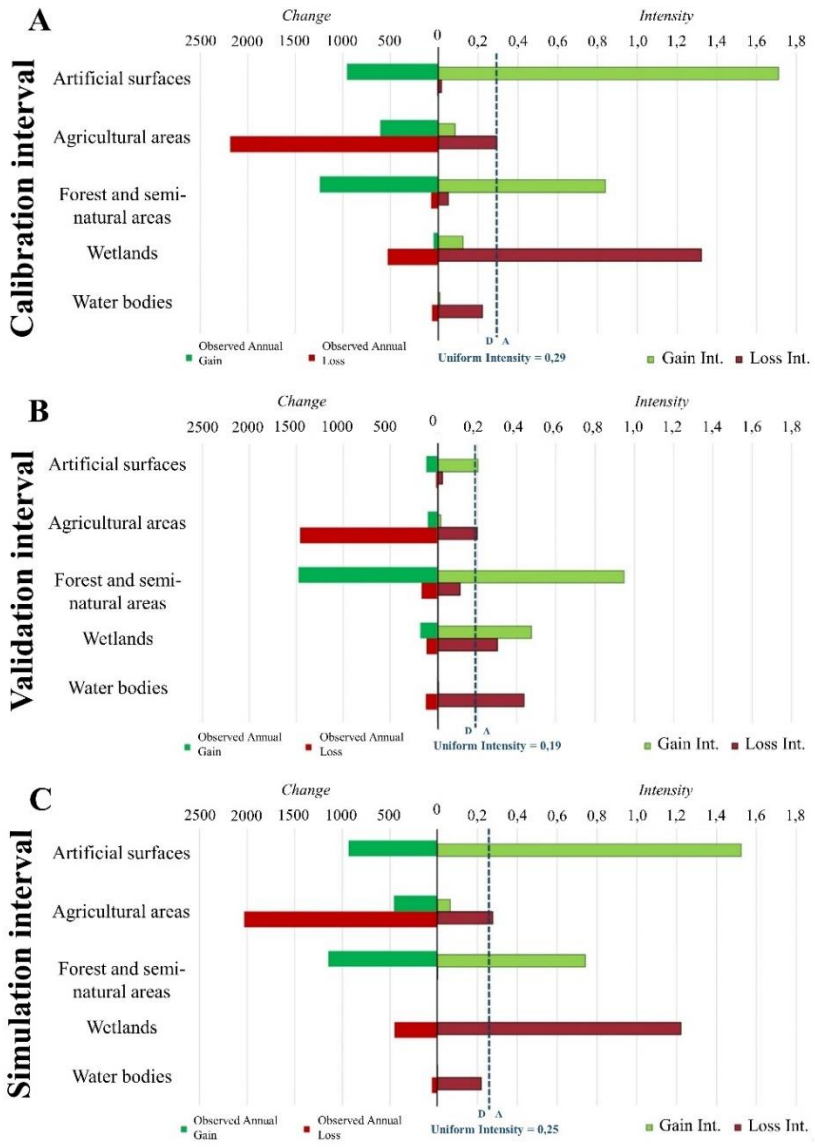

Figure 9*. Summary figure of the result of intensity analysis in the calibration (A), validation (B) and simulation (C) intervals, in terms of either annual change area expressed in pixels (left side) or gain and loss intensity values (right side) (D = dormant ; A = active). Blue dashed lines assign uniform intensity. The results presented in this figure were partly published with a different design in Varga et al. (2019)*

In *Figures 9A and 9C*, representing calibration and simulation intervals, either the annual gain/loss or gain/loss intensities were similar to each other. In both plots, the Agricultural areas showed the largest annual losses, and two categories – Forest and semi-natural areas and

Artificial surfaces – showed the largest annual gains. Category intensities and active/dormant status characteristics also showed similar patterns in calibration and simulation intervals.

In *Figures 9A and 9B* it can be clearly observed that Artificial surfaces' gain and Water bodies' loss showed a much larger value from the calibration to the validation interval. The active and dormant status of the categories were similar, except for Water bodies' loss, but with much different intensity values. These results suggest quite different change dynamics between calibration and validation interval. The common point of the results in this case was that the Agricultural areas category sustained a large loss in terms of size, but its intensity was close to uniform. In addition, Forest and semi-natural areas also showed large gain sizes and intensities, similarly to the simulation data as. In the validation interval, wetlands had a high intensity of loss, but with a high intensity of gain, meaning it was an active gainer and loser at the same time, but this dynamic did occur in neither calibration nor simulation intervals (*Figure 9*).

Transition level of intensity analysis revealed that the gain of Artificial surfaces category, which was the largest gainer in the calibration interval, targetted Agricultural areas in calibration and simulation interval. However, Artificial surfaces' gain targetted Forest category, according to validation interval analysis. It means that the simulation's dynamics on transition level matched the calibration's dynamic in terms of targeting a certain category, in case of this particular category. Further analysis could reveal further results of the dynamics of transition between every category pairs.

4.2. Discussion of FOM components and Intensity analysis in Study Site Group 1

In the study area an extremely low FOM value was calculated, meaning extremely weak model performance. A low value of Hits (0.02%) is not surprising, since the ratio of changing areas is also low, and Hits metric means the ratio of correctly simulated changes in the area. Therefore, Hits value could not exceed the ratio of the intersection of changing areas in the validation and simulation interval, expressed as a ratio of the study area. While ratio of Hits was low, *Figure 8* showed that False Alarms and Misses were relatively higher than Hits, with values over 1%. False Alarms were mostly located around the patches of the original categories which refers to the fact that the model simulated changes on the edges of the original category patches, while these areas were persistent in the reference data. Misses were located in compact patches characteristically, which refers to the fact that the model did not

match larger changes in sparsely located areas. In this example, the quantity disagreement of simulated changes were less than the allocation disagreement, according to *Equations 5 and 6*. Quantity disagreement of the simulation was 0.41%, as calculated by taking the absolute value of the difference between Misses and False Alarms. Allocation disagreement of the simulation was 2.12%, calculated by choosing the lower of Misses and False Alarms and multiplying by 2. It means that in CA-Markov model CA caused more error than Markov, since CA is responsible for the allocation control of the simulation and Markov is responsible for quantity control of the simulation. The larger allocation error can be a result of this characteristic placement of simulated changes on the edges.

In this case, ratio of Misses were lower than False Alarms, which means that more error originated from simulating persistence as change than the opposite. It is possible when reference change is less than simulated change, and this situation was also clearly visible in *Figure 9* barplots in case of this study site. Wrong Hits means that a pixel changed according to both reference and simulation data, but to a wrong category. Wrong Hits and Hits converged to zero, which means that the simulation hardly matched reference changes in the landscape, neither in a sense that the pixel exactly change to a certain category nor the poor presence of the change to any category.

Intensity Analysis revealed the pattern of real and simulated changes in the landscape, but changes were not in accordance with each other. The analysis also revealed that the change decelerated from the calibration to the validation interval. In this case, if the model followed the pattern of changes in the calibration data exactly, it would not match the validation interval changes, since there is a strong difference between calibration and validation interval changes. Intensity analysis and consideration of both calibration and validation interval changes helped to reveal this reason for weak model performance. Only by calculating the overall metric of FOM, this reason and any other information concerning quantity and allocation errors would have remained hidden. Therefore the intensity analysis provided essential information for the model validation process. A simple assessment of model performance together with applying intensity analysis is a new practical perspective, but considering calibration interval changes is an innovation. It can help to evaluate the LULC change simulations' predictive power while getting to know how much the trends of changes relate to real changes in the landscape. These results have been reported in Varga et al (2019).

4.3. Results concerning Study Site Group 2 analysis

After performing category aggregations, 114 models were run altogether in study site group 2. Due to the reasons described above, TB aggregation was not performed in six cases. Therefore, 24 models were run based on the maps aggregated according to L3, L2, L1 and BB rules, respectively, and 18 models were run based on maps aggregated according to TB rules. In all figures presented in this section, boxplots are based on these 114 models. In each one of *Figures 10-20*, boxplots present the median as a vertical line, the lower and upper quartiles as the upper and lower boundaries of the boxes and the minimum and maximum values as the ends of whiskers.

4.3.1. Results of Study Site Group 2 analysis concerning number of categories and change

After aggregation of categories according to various aggregation methods, the number of categories could vary with the applied methods.

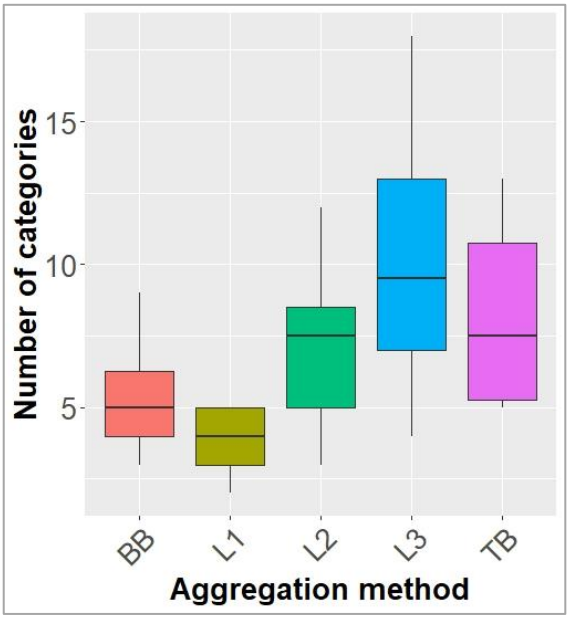

Figure 10. *Number of categories as observed in reference 2000 and reference 2006 maps, grouped by aggregation method (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation). This figure was originally published in Varga et al. (2020) with a slightly different design.*

Figure 10 shows the numbers of categories by aggregation methods in study site group 2. L3 had the largest numbers of categories,

which is trivial, since all the other aggregations are aggregations of L3 data (L3 maximum = 18). Therefore, the datasets aggregated according to other aggregation methods must not consist of more categories than L3 data. CLC L2 category scheme has a maximum of 15 categories, and the L2 dataset had a maximum of 12 categories, meaning that there was no study area where all the L2 categories were present. L1 data had the lowest numbers of categories in general, with a maximum of 5 categories, matching the maximum of CLC L1 category scheme (*Figure 10*).

BB and TB aggregations does not have a determined maximum, their theoretical maximum equals the number of categories of the data aggregated. In this case, BB dataset had a maximum of 13 categories, TB dataset had a maximum of 9 categories. TB and BB are both aggregation methods where the user has the opportunity to control the aggregation, while CLC standard levels have strict rulesets for aggregation. BB decreased the number of categories more than TB, in general. L3, L2 and L1 had less and less categories, in accordance with their decreasing determined maximum by standard levels.

The changes in the calibration, validation and simulation interval were presented in *Figure 11*, where the letters assign if the median of change values in a dataset processed according to a certain aggregation method was significantly different from the median of another dataset which was processed according to another aggregation method. This statistical difference between medians was proved by Tukey analysis ($p<0.05$). In case of all time intervals, L1 had the lowest change values, expressed as a percent of the study area, and L1 dataset was significantly different from all other datasets, while other datasets were not significantly different from each other. Since L1 had the lowest ratios of changing areas, L1 hid an enormously larger ratio of changes in the study areas related to other aggregation methods. Therefore, L1 eliminated significantly more changes in the study area than other aggregation methods. In the calibration interval BB, L2 and L3 experienced the most changes based on the similar medians. Since all aggregations are based on L3 data, it means that among all other aggregation methods, BB and L2 methods eliminated the less changes.

Already when investigating changes in the study area, it is clear that BB, L2, L3 and TB experienced more changes in the validation interval than in the calibration interval, in general. Also, simulation interval experienced much less changes according to either calibration or validation intervals with almost identical medians of BB, L2, L3 and TB datasets.

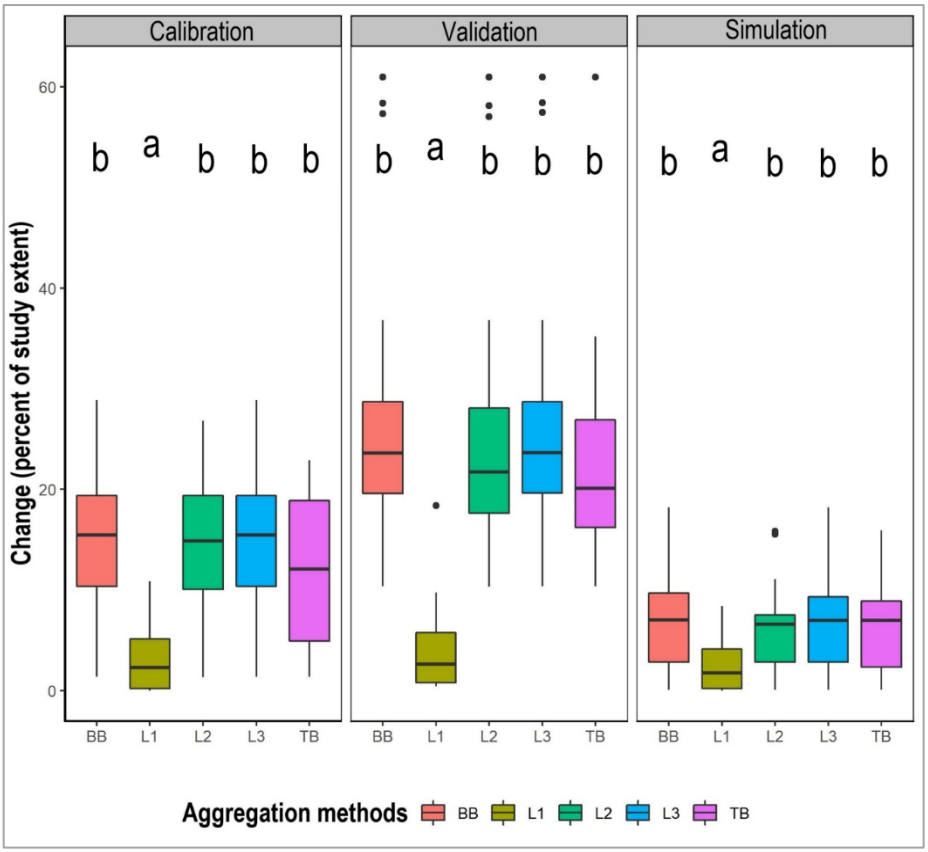

Figure 11. *Changes in the study areas expressed as percent of the study area, grouped by aggregation method (I) in the calibration interval, between reference 2000 and reference 2006 maps (II) in the validation interval, between reference 2006 and reference 2012 maps and (III) in the simulation interval, between reference 2006 and simulation 2012 maps. The groups with significantly different medians are assigned with different letters. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation). This figure was originally published in Varga et al. (2020).*

Further comparison concerning annual changes in the calibration, validation and simulation intervals gave an opportunity to have an insight to the acceleration or deceleration of changes from one time interval to another (*Figure 12*). Annual changes were calculated as dividing overall changes in the time interval by the years of duration in the same time interval. The difference of annual changes was determined by subtracting validation interval annual changes from calibration interval annual changes in each individual case, which determines if the calibration or validation interval annual changes were larger in a particular case (*Cal-Val. annual ch.*).

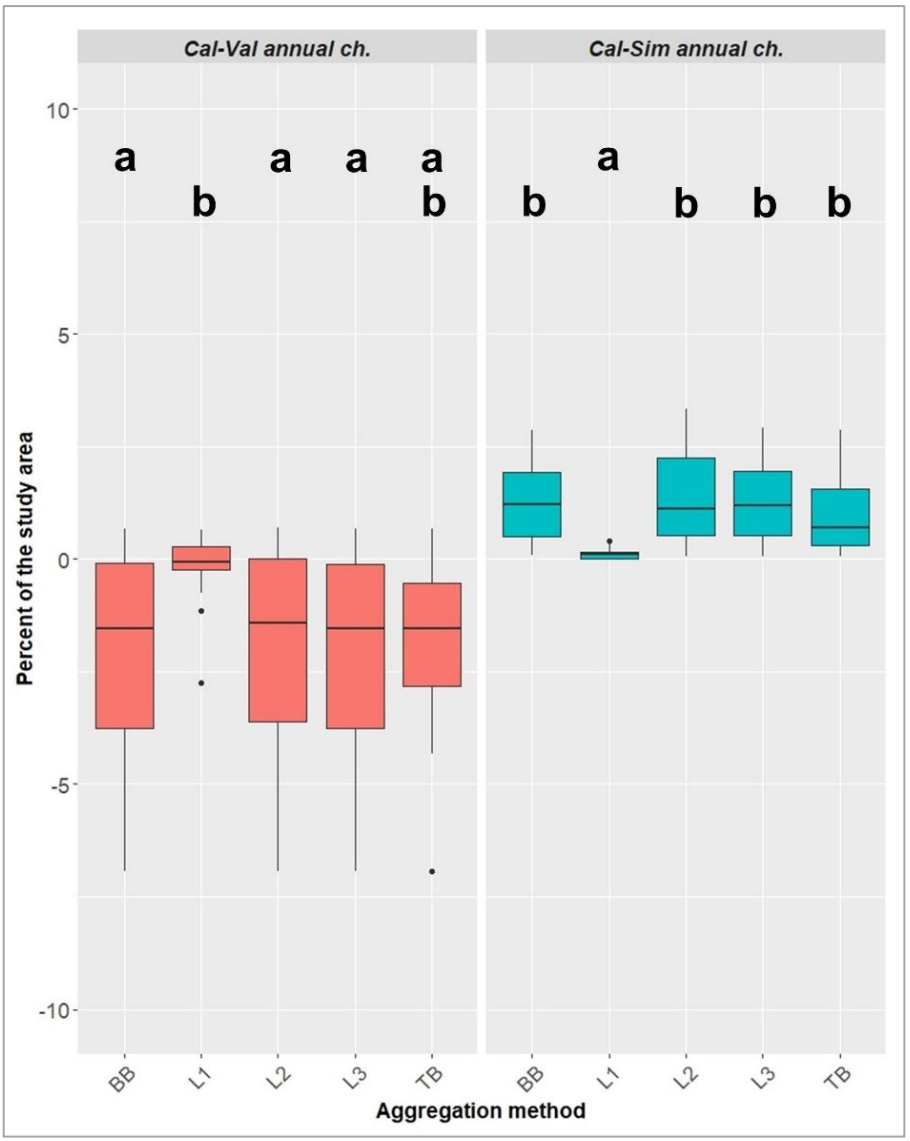

Figure 12*. Difference of annual changes in the study areas expressed as percent of the study area, based on the comparison of calibration and validation interval changes (Cal-Val. annual ch., left side) and based on the comparison of calibration and simulation interval changes (Cal-Sim annual ch., right side), grouped by aggregation method. The groups with significantly different medians are assigned with different letters (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

The same procedure was performed in the relation of calibration and simulation interval changes, by subtracting simulation interval annual changes from calibration interval annual changes in each individual case (*Cal-Sim. annual ch.*). In case of calibration and validation interval, the values were mainly negative, meaning that the rate of validation interval changes was higher than the rate of calibration interval changes, so the changes accelerated from the calibration to the validation interval. On the contrary, in case of calibration and simulation interval, the values were always positive, meaning that the rate of simulation interval changes was lower than the rate of calibration interval changes, so the changes decelerated from the calibration to the simulation interval in each particular case. In some cases of L1 annual changes, the change did not accelerate but decelerated from the calibration to the validation interval, since there were many positive values in the dataset, and the median was close to zero. In relation of calibration and simulation interval changes, the median of L1 dataset was significantly different from all other datasets in terms of aggregation methods and the median was close to zero again. It means that the change decelerated in L1 dataset, similar to all other datasets, but had a really slight difference related to the calibration interval, meaning a rate of change quite similar to the calibration interval. In *Tables 4-8*, examples for crosstabulation matrices of the calibration and validation intervals are presented, together with the transition area and transition probability matrices concerning the simulation model of the same study site. *Table 8* provides information about the probabilities of each possible inter-category transition in the model based on the calibration interval changes.

Table 4. *Crosstabulation matrix of the time interval between 2000 and 2006, in study site Borovany, zoom level S. Row and column headings assign the L3 categories in accordance with Appendix 1.*

		2006					
		112	**211**	**231**	**243**	**312**	**313**
2000	**112**	46	0	0	0	0	0
	211	0	868	26	4	0	0
	231	0	7	310	0	0	0
	243	0	0	9	154	0	0
	312	0	0	0	0	997	0
	313	0	0	0	0	0	79
Sum of persistent pixels (matrix diagonal pixels)							**2454**

Table 5. *Crosstabulation matrix of the time validation interval (2006-2012), in study site Borovany, zoom level S. Row and column headings assign the L3 categories in accordance with Appendix 1.*

		2012					
		112	**211**	**231**	**243**	**312**	**313**
2006	**112**	46	0	0	0	0	0
	211	0	541	325	0	9	0
	231	0	0	345	0	0	0
	243	0	0	0	158	0	0
	312	0	0	1	0	996	0
	313	0	0	0	0	0	79
Sum of persistent pixels (matrix diagonal pixels)							**2165**

Table 6. *Crosstabulation matrix of the simulation interval (2006-2012), in study site Borovany, zoom level S. Row and column headings assign the L3 categories in accordance with Appendix 1.*

		2012					
		112	**211**	**231**	**243**	**312**	**313**
2006	**112**	46	0	0	0	0	0
	211	0	851	18	6	0	0
	231	0	0	345	0	0	0
	243	0	0	12	146	0	0
	312	0	0	0	0	997	0
	313	0	0	0	0	0	79
Sum of persistent pixels (matrix diagonal pixels)							**2464**

Table 7. *Transition area matrix generated by the Markov component of CA-Markov simulation model in study site Borovany, zoom level S. Row and column headings assign the L3 categories in accordance with Appendix 1.*

		2012					
		112	**211**	**231**	**243**	**312**	**313**
2006	**112**	46	0	0	0	0	0
	211	0	846	25	4	0	0
	231	0	8	337	0	0	0
	243	0	0	9	149	0	0
	312	0	0	0	0	997	0
	313	0	0	0	0	0	79
Sum of persistent pixels (matrix diagonal pixels)							**2454**

Table 8*. Transition probability matrix generated by the Markov component of CA-Markov simulation model in study site Borovany, zoom level S. Row and column headings assign the L3 categories in accordance with Appendix 1.*

		2012					
		112	**211**	**231**	**243**	**312**	**313**
2006	**112**	1.0000	0.0000	0	0	0	0
	211	0	0.9666	0.0290	0.0045	0	0
	231	0	0.0221	0.9779	0	0	0
	243	0	0	0.0552	0.9448	0	0
	312	0	0	0	0	1.0000	0
	313	0	0	0	0	0	1.0000

The similarity of the sum of persistent pixels in *Tables 4 and 7* shows that the transition area matrix determined the exact same quantity of persistence as the calibration data did, therefore the matrix dictated the exact same quantity of overall change as well. However, the quantity of persistent pixels in categories 211 and 231 were substantially different while the changes of each categories were similar, when comparing the calibration data and the transition area matrix. The sum of persistent pixels in the simulation interval was larger, meaning less changes in the simulation interval, than in the calibration interval. The sum of persistent pixels was much less in the validation interval, than in the calibration interval. These dynamics also refer to decelerating changes from the calibration to the simulation interval and accelerating changes from the calibration to the validation interval.

4.3.2. Results of Study Site Group 2 analysis concerning FOM, FOM components, and quantity and allocation disagreements of the models

In terms of Figure of merit (FOM) values, there was no significant difference between datasets of various aggregation methods (*Figure 13*). While Figure 13 shows insignificant difference, effect sizes indicated a larger effect in case of L1 (L1-BB: 0.30; L1-L2: 0.29; L1-L3: 0.31; L1-TB: 0.30; where the numbers are effect sizes which determine the magnitude of differences between each pair of datasets). In some cases, L1 dataset showed larger FOM values related to other aggregation methods, but its median was close to zero, meaning many cases with extremely low model performances. FOM provides an overall characterization of model performance, and all other datasets showed FOM values similar to each other, meaning quite similar performance.

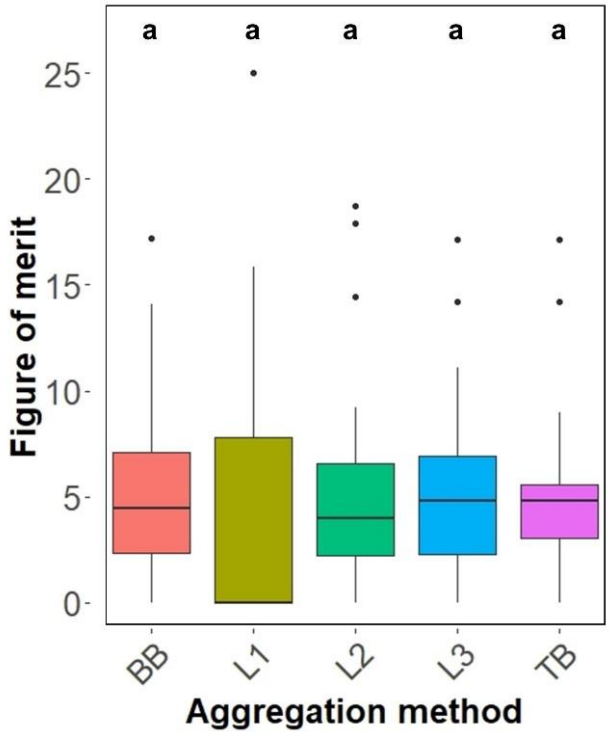

Figure 13. *Figure of merit values in Study site group 2, grouped by aggregation method. The groups with indistinguishable medians are assigned with similar letters (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation). This figure was originally published in Varga et al. (2020), with a slightly different design.*

FOM components (False Alarms, Misses, Wrong Hits, Hits) had similar characteristics concerning L1 dataset, since L1 showed the lowest values in case of all FOM components (*Figure 14*). Since L1 had the lowest values in case of each component, it means that it had the lowest ratio of correctly simulated pixels (Hits), but it had the lowest ratio of erroneously simulated pixels (False Alarms, Misses, Wrong Hits) as well. In case of False Alarms, L1 was significantly different from BB and L3 datasets. In case of Misses, L1 was significantly different from all other datasets. In case of Wrong Hits, L1 was significantly different from BB, L2 and L3 datasets. Finally, in case of Hits, L1 was significantly different from BB, L2 and L3 datasets again.

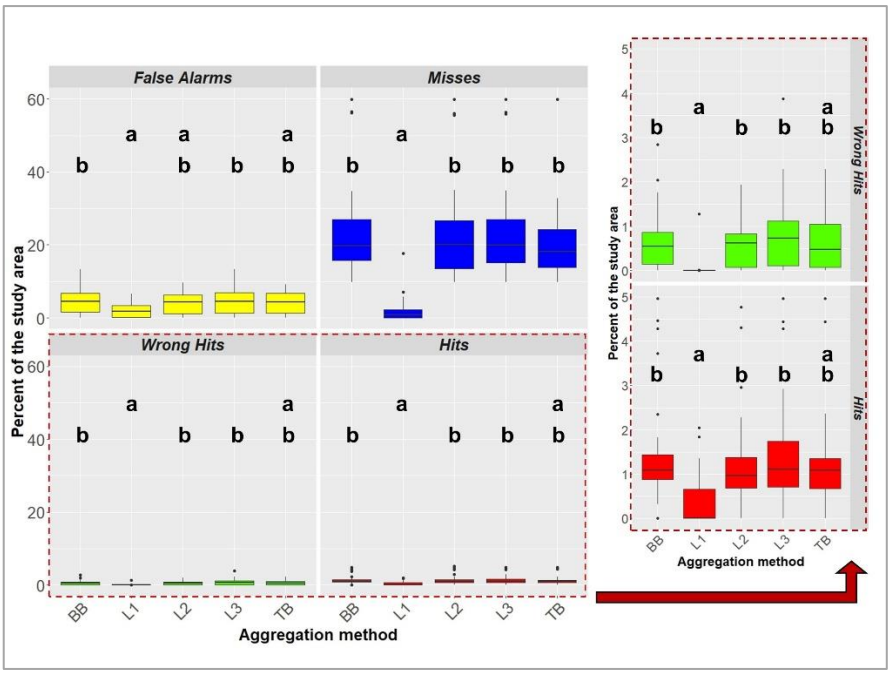

Figure 14. *Figure of merit (FOM) components values in Study site group 2, grouped by aggregation method. The groups with significantly different medians are assigned with different letters. The Wrong Hits and Hits values are highlighted on the right side of the figure with a different scale, in order to make the differences between datasets visible. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation). This figure was originally published in Varga et al. (2020), with a different design.*

Quantity (Q_s) and allocation (A_s) disagreement of the simulation, derived from False Alarms and Misses, were calculated for each particular model. Quantity disagreement, Allocation disagreement and Wrong Hits were reported together (*Figure 15*), since the sum of these three values is equal to the Total disagreement (T_s) of the model. By this way of visualization, it may be clearer how different types of disagreement contribute to the Total disagreement of the model. In general, L1 showed the lowest values of disagreement in terms of either quantity or allocation. Since these values are calculated from Misses and False Alarms, it was somewhat presumed that L1 would show the lowest values, because L1 showed the lowest values in case of False Alarms and Misses as well.

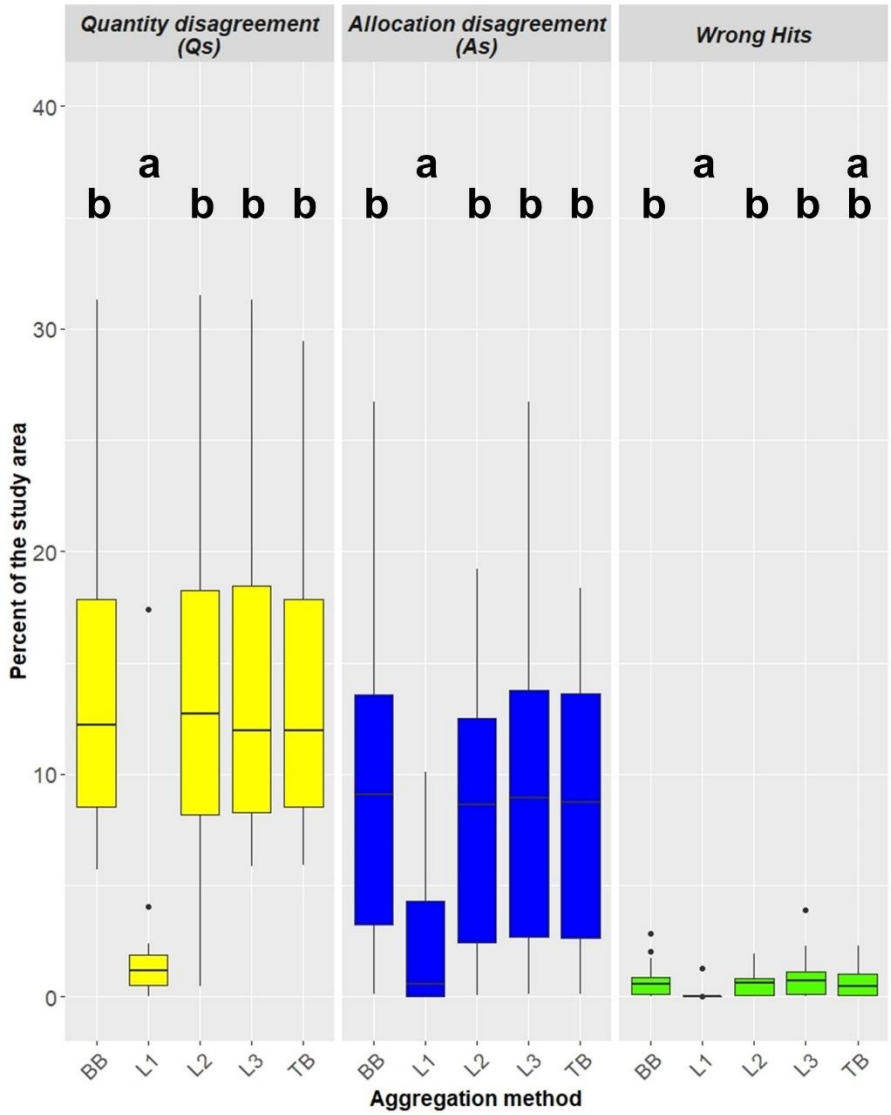

Figure 15. *Different types of simulation disagreement (Quantity disagreement, Allocation disagreement and Wrong Hits) in Study site group 2, grouped by aggregation method. The groups with significantly different medians are assigned with different letters. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

In case of BB, L2, L3 and TB aggregation methods, allocation disagreement values were generally lower than quantity disagreement values. The allocation disagreement median of L1 was slightly lower than quantity disagreement median of L1. Since Wrong Hits had

extremely low values related to quantity and allocation disagreement, Wrong Hits contributes to the Total disagreement the less from all disagreement components. In *Figure 16*, the difference between quantity and allocation disagreement was visualized for each cases. By calculating this metric for each case, it is possible to investigate if the quantity or allocation disagreement is larger in each case.

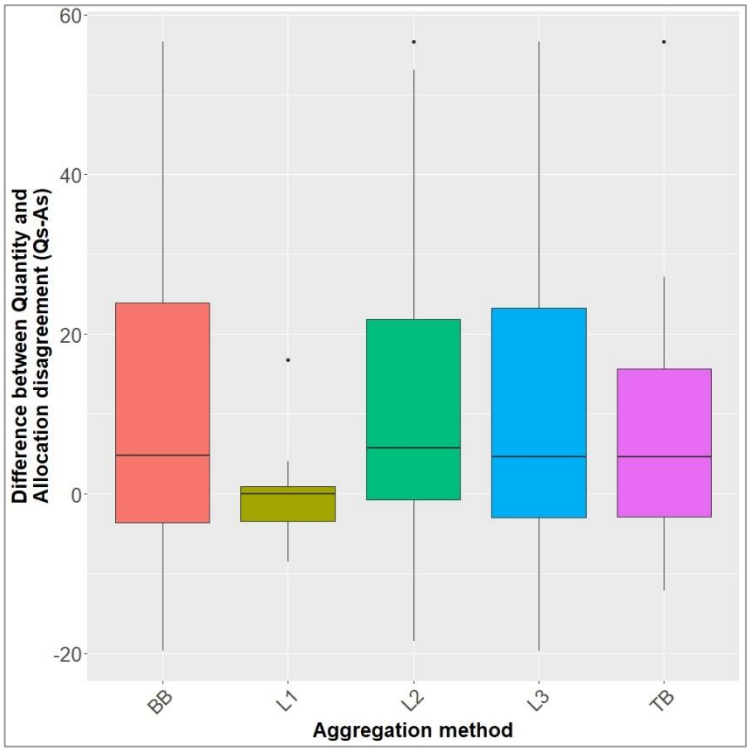

Figure 16. *Difference between quantity (Q_s) and allocation (A_s) disagreement of the simulation in each case in Study site group 2, grouped by aggregation method. If the value is positive, then Q_s is larger. If the value is negative, then A_s is larger. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

It is clear that in most cases of BB, L2, L3 and TB datasets, quantity disagreement (Q_s) was larger than Allocation disagreement (A_s) which means that the model had more error originating from quantity than from allocation issues. However, in case of L1, many cases were negative, meaning that allocation disagreement was larger than quantity disagreement. Median is close to zero, suggesting that these two alternatives occurred evenly in L1 cases.

4.3.3. Results of Study Site Group 2 analysis concerning stationarity

Temporal instability, also known as stationarity, was measured by Runfola R values, calculated between calibration and validation interval (Runfola R REF), and between calibration and simulation interval (Runfola R SIM). *Figure 17* presents the temporal instability in these terms. The more the value is close to 1, the more change should be reallocated between the two intervals in order to achieve a uniform change throughout the whole time interval – in this study design it means between the whole time interval from 2000 to 2012.

The temporal instability concerning calibration and validation interval was the highest in case of L1 dataset (Runfola R REF). The median of all the other datasets were close to each other, meaning a similar temporal instability. The statistical analysis did not prove significant difference of L1 dataset. According to the boxplots, there are cases of almost all values of temporal instability in the datasets.

On the contrary, the temporal instability concerning calibration and simulation interval was the lowest in case of L1 dataset (Runfola R SIM), also significantly different from BB, L2 and L3 datasets. It means that L1 dataset cases showed high stability of changes concerning calibration and simulation interval changes.

Finally, the difference between *Runfola R* values concerning reference and simulation data (*Runfola R DIFF*) for each individual case was the largest in case of L1 dataset, significantly. The instability of L1 literally dropped when comparing *Runfola R REF* and *Runfola R SIM* values. In case of the other aggregation methods, the medians were not significantly different from each other. It means that the difference of stability throughout calibration and validation, and throughout calibration and simulation intervals were extremely large in most L1 cases and it points to the fact that the model simulated much more stable changes in the landscape than the real situation. In case of all other aggregation methods, the Runfola R DIFF medians are negative, meaning that in many cases, the model simulated less stable changes related to the real situation.

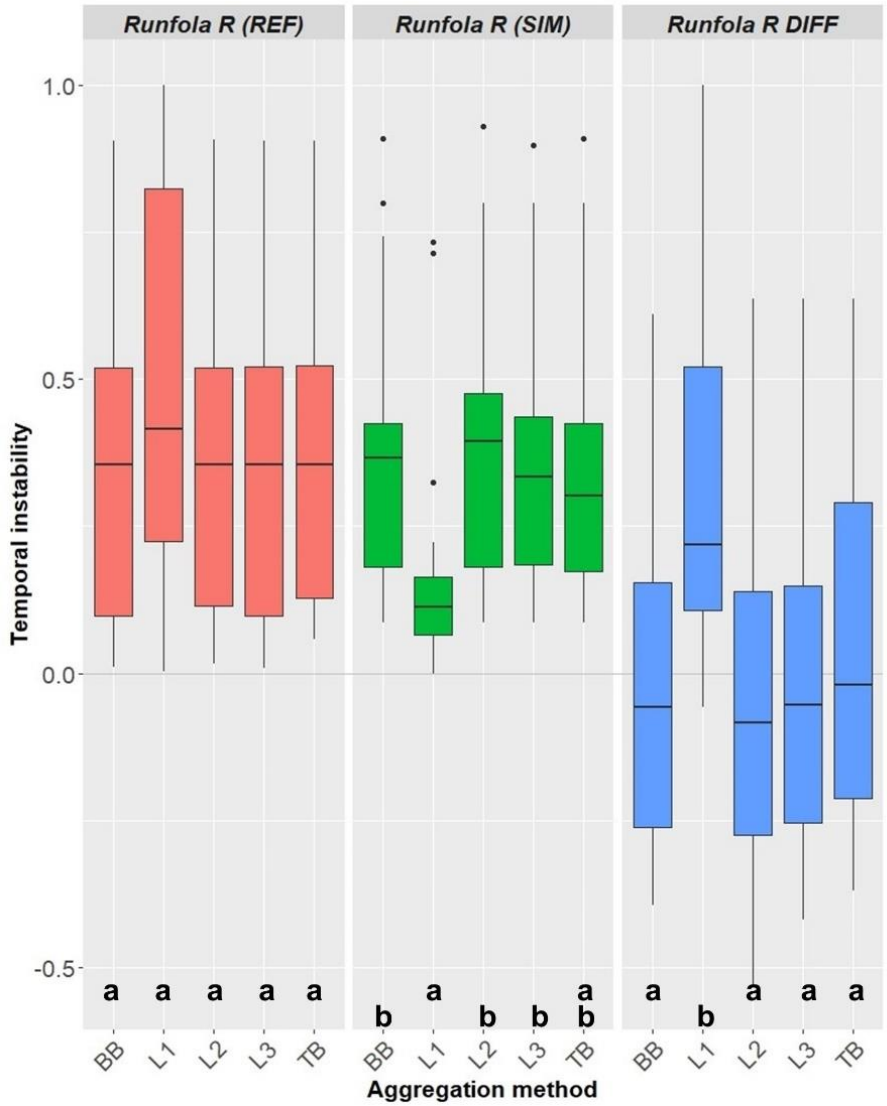

Figure 17. *Runfola's R values measuring temporal instability between time intervals. R values concerning temporal instability between calibration and validation intervals [Runfola R (REF)], concerning temporal instability between calibration and simulation intervals [Runfola R (SIM)], and the difference between the two Runfola R values calculated for each case [Runfola R DIFF = Runfola R (REF) - Runfola R (SIM)]. If Runfola R DIFF value is positive, then Runfola R (REF) is larger. If Runfola R DIFF value is negative, then Runfola R (SIM) is larger. Values of all three variables are grouped by aggregation method. The groups with significantly different medians are assigned with different letters. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

4.3.4. Results of Study Site Group 2 analysis concerning Kappa Index of Agreement and Overall Agreement

As discussed in literature review, there are validation approaches frequently reported in contemporary literature, where the modeler compares time #3 reference map to time #3 simulation maps by calculating metrics used in accuracy assessment of remotely sensed images. Two popular metrics for this purpose are Kappa coefficient and overall agreement.

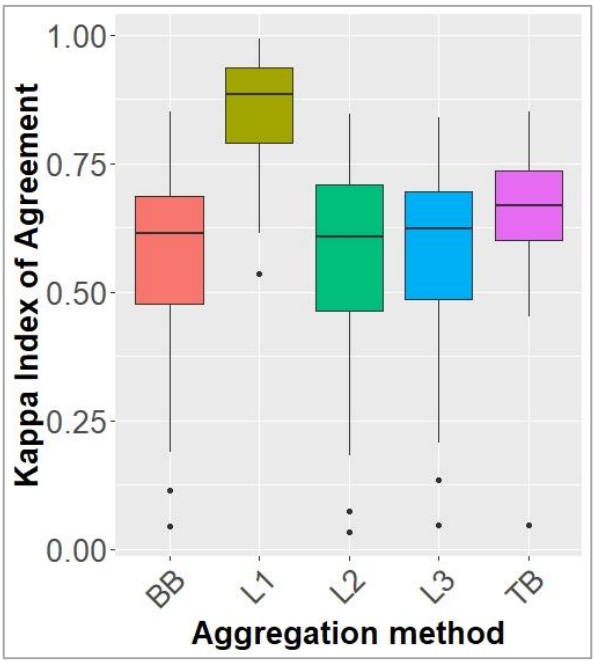

Figure 18. *Kappa Index of Agreement values in each case in Study site group 2, grouped by aggregation method (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

In *Figure 18*, Kappa coefficient values are presented for each model, based on the comparison of 2012 reference maps and 2012 simulation maps in each case. Kappa returned high values (around 0.85) in L1 dataset, while all other aggregation methods returned values around 0.6–0.7, indicating a lower agreement between the two maps. Overall agreement also characterized the agreement between 2012 reference and 2012 simulation maps in each case. Overall agreement values are presented in *Figure 19* along with the ratio of persistent areas in the validation interval and with correctly simulated persistent areas, also known as Correct Rejections. The differences of the medians are

seemingly similar in all four metrics, which is in accordance with the hypothesis that Kappa and overall agreement return large values if the ratio of persistent areas is large, because large ratio of persistent areas result in high ratio of pixels that belong to the same category in both maps.

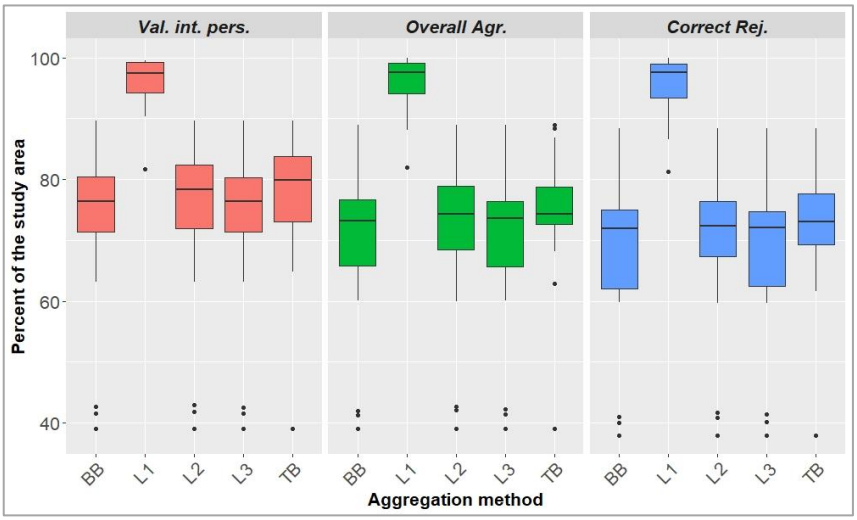

Figure 19. *Ratio of persistent areas in the validation interval (left), overall agreement (middle) and Correct Rejections (right), grouped by aggregation method. (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

4.3.5. Results of statistical analysis in Study Site Group 2

In order to reveal the correlations between variables used in the analysis, a comprehensive statistical analysis was conducted. In this analysis, the correlation between variables was checked for all pairs of variables and the results are presented in *Figure 20*. The correlations concerning variables characterizing model performance are highlighted. The thinner the ellipses, the larger the correlation is between the variables of which the ellipse is intersected.

During the interpretation, the trivial correlations were ignored. For instance, a strong correlation between allocation disagreement of the simulation (A_s) and False Alarms was presumed, since allocation disagreement is equal to the double of the minimum of False Alarms and Misses, thus in many cases, the double of False Alarms is equal to allocation disagreement. This relation establishes a strong correlation between these two variables.

FOM did not show a strong correlation, in a statistically significant manner, with any other variables that are not used for calculating FOM. Among FOM components, Misses returned strong correlation with validation interval changes (R^2=0.95) and validation interval persistence (with the same correlation, since persistence and change complement each other in the study area). False Alarms returned moderate correlation with calibration changes (R^2=0.56), calibration persistence, and temporal instability between calibration and validation interval (R^2=0.72). Quantity disagreement values of the simulation showed strong correlation with validation interval changes (R^2=0.82), and mild correlation with the temporal instability between calibration and simulation interval (R^2=0.38) and with the difference between calibration and validation annual changes (R^2=0.65) (which latter means the acceleration or deceleration of changes). Allocation disagreement values of the simulation showed moderate correlation with the temporal instability between calibration and validation interval (R^2=0.67) and with the calibration interval persistence (R^2=0.61) and change. Overall agreement and Kappa showed a high correlation with validation interval persistence and change (OA R^2=0.92; Kappa R^2=0,85), and with Correct Rejections (OA R^2=0,96; Kappa R^2=0,87) as well.

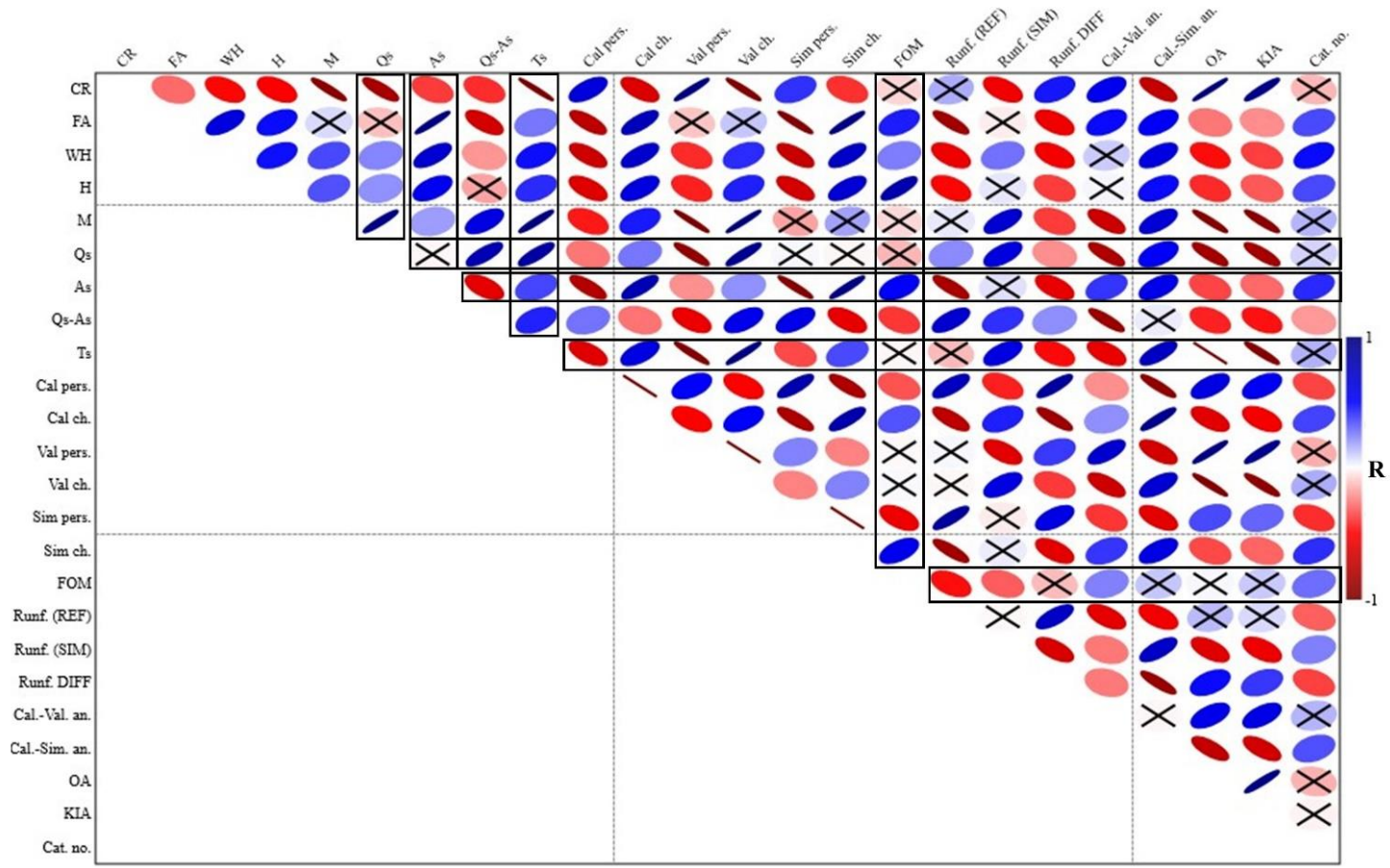

Figure 20. *Correlation table of variables based on study site 2 data, created in Past statistics software. Correlations are assigned to colors and ellipse width, according to Spearman's r_s correlation statistic ($p<0.05$). Correlations concerning FOM and quantity, allocation and total disagreement of the simulations are highlighted with black boxes. Correlations that are not significant are assigned with a cross. Variables are assigned with abbreviations referring to the complete variable name. Abbreviations are described in detail in Appendix 2.*

The difference between the rate of change of the calibration and validation intervals (Cal-Val annual ch.) also showed strong correlation with the difference between quantity and allocation disagreement of the simulation (Q_s-A_s, R^2=0.76).

4.4. Discussion of Study Site Group 2 analysis results

After a comprehensive analysis of number of categories, changes, model performance metrics and stationarity metrics, this section discusses the scientific importance of the results presented. The statistical analysis results are discussed together with the detailed discussion of the relevant variables.

4.4.1. Discussion concerning number of categories and changes in the landscape

It was clear throughout the study that L1 dataset dissevered from other aggregation methods in almost all manners, whether proved by statistical analysis or by simple visual interpretation of the boxplots. It is important to interpret the number of categories together with the changes in the landscape, according to various aggregation methods. In this way, the effect of aggregation methods on category numbers becomes clear, while the extent of the elimination of changes also becomes clear. When aggregating categories, the user should pay attention for not eliminating important changes in the area, because the model cannot simulate the changes which are not present in the study area anymore. However, by decreasing the number of categories in the study area, the interpretation of land changes gets much easier (Aldwaik et al., 2015). Furthermore, running a model which needs to handle fewer categories, meaning fewer combinations of category interactions, demands much less computing capacity. As *Figure 10* showed, the number of categories decreased the most in the case of L1, related to all other aggregation methods, however, L1 had the lowest possible maximum of the number of categories, with a maximum of 5. BB had the second less categories with a maximum of 9, and then L2, TB and L3 in increasing order. While BB and TB aggregation methods are change-focused in a manner that the aggregation is performed with respect to the presence of change in the study area, the L1, L2 and L3 category schemes are dominated by a thematic ruleset, not comprehending changes at all. All the other aggregation methods dissevered from L1 in terms of changes in all three time intervals: calibration, validation and simulation, based on the statistical analysis. It means that, in a statistical sense, BB, L2, L3 and TB did not differ from each other, but all of them differed from L1. While all BB, L1, L2 and TB was an aggregation of L3 data, L1 showed the

less changes, meaning it definitely eliminates a large ratio of changes in the study area. BB showed the least maximum number of categories after L1, meanwhile not eliminating the changes from the study area, due to the characteristics of the applied behavior-based aggregation process. When aggregating the categories according to BB method, it was a critical condition to maintain total change in the area, thus the aggregation process stopped before an upcoming aggregation step would have decreased the total change. This process scheme resulted in the phenomenon that BB and L3 had the exact same ratio of changing areas in the calibration and the validation interval in each individual case, while BB had much less categories in each individual case. Therefore, BB eliminated zero change from the study areas, while decreasing the number of categories, thus making an opportunity for easier interpretation and less computing demand of a possible simulation. TB is also a change-focused aggregation method, but applies an arbitrary threshold when assigning the categories subject for aggregation. TB was less effective in either reducing the number of categories or maintaining changes, because the changes not meeting the requirement were not taken into account.

The analysis of annual changes makes it possible to see the deceleration or acceleration of changes in the landscape. From the calibration to the validation interval, the change accelerated in most cases, while from the calibration to the simulation interval the change decelerated in all cases. It means that the model always simulated less changes than the change observed in both the calibration and the validation interval. The simulation quantity of change matched neither calibration nor validation interval quantities of change, which leads to the following conclusions:

- the simulation models underestimated the changes related to the calibration interval data, based on which the model is trained;
- the rate of change simulated by the models did not match the rate of change observed in the validation time interval, because the dynamics of changes were reverse.

Olmedo et al. (2015) found in three examples of CA-Markov models trained with CLC data that the change accelerated from the calibration to the validation interval, while the CA-Markov output simulated less change than the Markov matrix would extrapolate. In the example presented in *Table 4-8*, the exact same phenomenon could be observed, where the simulation returned less change than the Markov's transition area matrix would have dictated. Study site group 1 also showed a deceleration of changes in the simulation interval, meaning the same as

the findings of Olmedo et al. (2015), and the cases of Study site group 2 proves a similar pattern in a large dataset of models. However, in case of L1, the aggregation affected the changes in the study area so much that in some L1 cases the rate of changes even turned into a deceleration from the calibration to the validation interval, while other aggregation methods showed acceleration. L1 showed values closest to zero, meaning really slight differences in general, between calibration and validation interval rates of changes, and also between calibration and simulation interval rates of changes. The changes in the landscape correlated with various metrics concerning model performance or temporal stability, discussed in the following sections.

4.4.2. Discussion concerning FOM and FOM components

Model performance metrics aimed to characterize whether the model could simulate changes in accordance with real landscape changes. In *Figure 13 and 14*, the FOM and FOM components were presented along with the statistically significant difference between the medians of the datasets of each aggregation method. In terms of FOM, there was no significant difference between the aggregation methods, although the median of L1 was zero, meaning a complete error of the model in half of the L1 cases. Hypothesis testing succeeded in a limited way in this case, however with an effect size larger in case of L1, but contemporary results also supported the idea that a clearly significance-focused interpretation can turn out to be misleading (Baker, M., 2016; Kim, J. and Bang, 2016; Szabó et al., 2016; Szucs and Ioannidis, 2017).

In case of FOM components L1 also showed the lowest values, but again, L1 was significantly different from all other aggregation methods in terms of Misses only. In a statistical sense, L1 was not different from TB in all other cases. False Alarms and Misses showed enormously larger values related to Wrong Hits and Hits (*Figure 14*), and Wrong Hits and Hits values did not exceed 3% of the study areas in any of the cases. It means that the errors originating from simulating persistence instead of change, and the opposite, were much more characteristic than the error of matching the change but to wrong category or than the ratio of correctly simulated changes.

It is important to consider changes when interpreting FOM components, because these components are calculated based on changes in the study area, as described in *Section 3.6.2.1.* Therefore, if a study area shows less changes, then it will probably show a lower ratio of FOM components in the study area, like less erroneously or less correctly simulated pixels, as it could be observed in L1 cases. While L1 aggregation eliminated changes, it lost the ability to extrapolate the

eliminated changes, and the elimination of changes also led to less ratio of FOM components in the study areas. In Varga et al (2019) it was stated that FOM is not enough to qualify model performance, and in Varga et al (2020) it was stated that all four FOM components were lower in L1 dataset. In this research, a further statistical analysis supports the correlation between False Alarms and simulation interval changes (R^2=0.91) and between Misses and validation interval changes (R^2=0.95). In Study site group 1, False Alarms concentrated around the patches of existing categories, while Misses could be observed in sparsely located patches. In Study site group 1, mainly the spatial filter of the model caused False Alarms – persistence simulated as change – around the existing patches. If the simulated change is large, and the simulated changes are influenced by the spatial filter to locate around the existing patches while real changes are not located around the existing patches, that phenomenon can result in a large ratio of False Alarms. In Study site group 1, the Misses –change simulated as persistence – were located in compact patches in sparsely located areas, meaning that the real changes were also not located around the existing patches but in sparsely located areas. If the validation interval change is large, and the simulated changes are influenced by the spatial filter to locate around the existing patches while real changes are not located around the existing patches, that situation can result in a large ratio of Misses. The correlations concerning False Alarms and Misses suggests that the same pattern may influence the models in Study site group 2. According to a set of models, Pontius et al. (2018) suspected that smaller amounts of change is associated with lower predictive accuracy, but significant correlation between calibration or validation interval changes and FOM was not found in this research (*Figure 20*).

4.4.3. Discussion concerning quantity and allocation disagreement of the simulation (Q_s and A_s)

Quantity and Allocation disagreement (Q_s and A_s) of the simulation models were derived from False Alarms and Misses components, as described in *Section 3.6.2.2.* According to *Figure 16,* Quantity disagreements were larger than Allocation disagreement, in general, and Wrong Hits were the lowest in all aggregation methods, related to A_s and Q_s. L1 dissevered unambiguously from the other aggregation methods again, in a statistically significant manner, however, L1 median was not significantly different from TB in case of Wrong Hits. L1 dissevered in a way that either Q_s, A_s and Wrong Hits values were lower than in case of other aggregation methods. Since these metrics were derived from FOM components, the effect of changes on

these metrics is obvious again. *Figure 16* showed if Q_s or A_s was larger in each individual cases, and in most cases, Q_s was larger, which means more error originated from the Markov than from the CA part of the model, since Markov controls quantity and CA controls allocation of the simulation. In case of L1, CA caused more error than Markov, because mostly the allocation error was larger and the median was around zero, referring to an almost equal relation between the two types of errors. In case of Study site group 1, the allocation error was larger as well, where an L1 dataset was the subject of research, too. The difference between quantity and allocation error (Q_s-A_s) showed a strong inverse correlation with the difference between annual changes of calibration and validation interval, meaning that it is sensitive of the rate of changes. Moreover, it is sensitive in a way that whether the changes accelerate from the calibration interval to the validation interval, than the quantity of changes will be larger than allocation error. This result is in accordance with the systematic deceleration pattern of simulated changes, since the quantity error would possibly be larger, if the rate of changes move in the opposite way in the validation and in the simulation interval.

4.3.3. Discussion concerning temporal stability in the landscape

Temporal instability was measured by Runfola R values, concerning either stability between calibration and validation intervals or between calibration and simulation intervals. Temporal instability is also related to the quantity of change in the landscape by definition. Olmedo et al (2015) claimed that non-stationarity of the changes was the most obvious reason for a lower model performance, since the changes in CLC data accelerated from the calibration to the validation interval. Temporal instability, stationarity and the difference of the annual changes between the time intervals all characterize the acceleration or the deceleration of the data, from a slightly different aspect. Temporal instability was the largest in case of L1 according to *Figure 17*, concerning the calibration and validation intervals (Runfola R REF), while it was the lowest concerning the calibration and simulation intervals (Runfola R SIM). Therefore, it is not surprising that the difference between these two variables (Runfola R DIFF) was the largest in case of L1, in a statistically significant manner in this case. The temporal stability of BB, L2, L3 and TB datasets were similar in case of Runfola R REF. The temporal stability of BB, L2, L3 and TB datasets were similar in case of Runfola R SIM as well, with a slightly more stable character in L2, L3 and TB in increasing order. These results refer to the fact that the model simulated more stable changes from the calibration to the simulation interval, which is in accordance with the fact that the

model simulates decelerating changes. Runfola R REF also showed a strong inverse correlation with False Alarms (R^2=0.72), and Allocation disagreement of the simulation (A_s, R^2=0.67), suggesting that the more instable the changes are throughout the reference time intervals, the less False Alarms and Allocation disagreement are present in the simulation. According to some examples in literature (Mertens and Lambin, 2000; Runfola and Pontius Jr, 2013) the land change in reality does not match the idea of stationarity, that is a reason for predictive inability of Markov models.

4.3.3. Discussion concerning Overall Agreement and Kappa Index of Agremeent

In *Figures 18 and 19*, Kappa Index of Agreement and Overall Agreement were presented in a comparison with Correct Rejections and validation interval changes. As already presented when analyzing changes in different time intervals, L1 had the lowest ratios of changes in the study area, consequently, it is trivial that L1 had the largest ratios of persistence in the study areas. Here, L1 had also the largest Correct Rejections, as known as correctly simulated persistence in the area. Due to the fact that in case of a simulation, overall agreement and Kappa Index of Agreement measures the agreement between a pair of time #3 maps, they does not distinguish correctly simulated changes and simple persistence, because they are incapable of comprehending this information. Traditional agreement index results of this research between simulation 2012 and reference 2012 maps, were in accordance with previous researches' modelling results concerning CA-Markov method (Memarian et al., 2012; Singh et al., 2015). However, high ratio of persistent area between the two dates could be a considerable reason for a seemingly successful model performance, as previous researches delineated (Kityuttachai et al., 2013; Subedi et al., 2013). Scientists warned to take into account that the high agreement in the models can be a consequence of high persistence and/or meaning small changes in landscape over time (Pontius Jr, R. G. et al., 2011; van Vliet, 2009). Correlation study also showed that OA and KIA had an outstandingly strong correlation with validation interval persistence (OA R^2 = 0.92; KIA R^2=0.83). L1 cases showed large OA and KIA values, while Hits, also known as correctly simulated changes, were under 3% of the study areas. OA and KIA does not have a strong inverse correlation with Hits, but the results mean that these metrics show large agreement even if the ratio of correctly simulated pixels is extremely low. These examples reveal that the usage of these metrics for model validation is systematically misleading and their usage can seriously make the

modeler believe that the model performance is acceptable even when a model is totally incapable of simulating the real changes.

4.5. Results concerning Study Site Group 3

The study design of study site group 3 is different from study site groups 1 and 2 in the following relations:

- the input data is different, since the models are not based on CLC datasets, but Landsat images;
- the input data had only 2 LULC categories in all models in this study site group, since the classification was performed with the goal of creating categories that focus on the phenomenon that is the specific subject of modelling;
- the overall number of cases in study site group 3 (6 cases) is much lower than in any datasets grouped by aggregation method in study site group 2.
- the model parameter, which is determined by the duration of the time intervals, was different, because the time intervals did not matched the duration between CLC datasets (6 years), since they were mainly determined by the accessibility of cloud-free images.

All these differences resulted in the situation that the models performed in study site 3 are not comparable to study site 1 and 2 in a statistical sense. Although, they are not comparable on a ground where statistical correlation information can be derived under appropriate circumstances, but a comparison on the ground of empirical observations was conducted. The same variables were calculated in this study site group as well, but I will present selectively those variables which demonstrate substantial differences related to study site group 1 and 2.

Figure 21 shows the results of FOM components presented in maps of the study areas where the pixels represent the erroneously simulated, correctly simulated and persistent areas in accordance with the legend of *Figure 8*. It can be observed that the Hits were concentrated on the edges of original patches in either the Amazonian or the Atchafalaya Bay cases. Misses were located as larger, more concentrated patches again in case of the Amazonian site. Misses in the Atchafalaya Bay example are more distant from the location of the initial changes in the area. False Alarms are sparsely located areas in the Amazonian example, and they follow the leads of the rivers in the Atchafalaya example. In 2012 simulation map in the Atchafalaya example, the rivers were literally closed as a result of the sprawl of changes. In both areas a relatively larger ratio of changes were characteristic, while the changes

were presumed to sprawl around the edges of the original category, due to the nature of the causes of these phenomena.

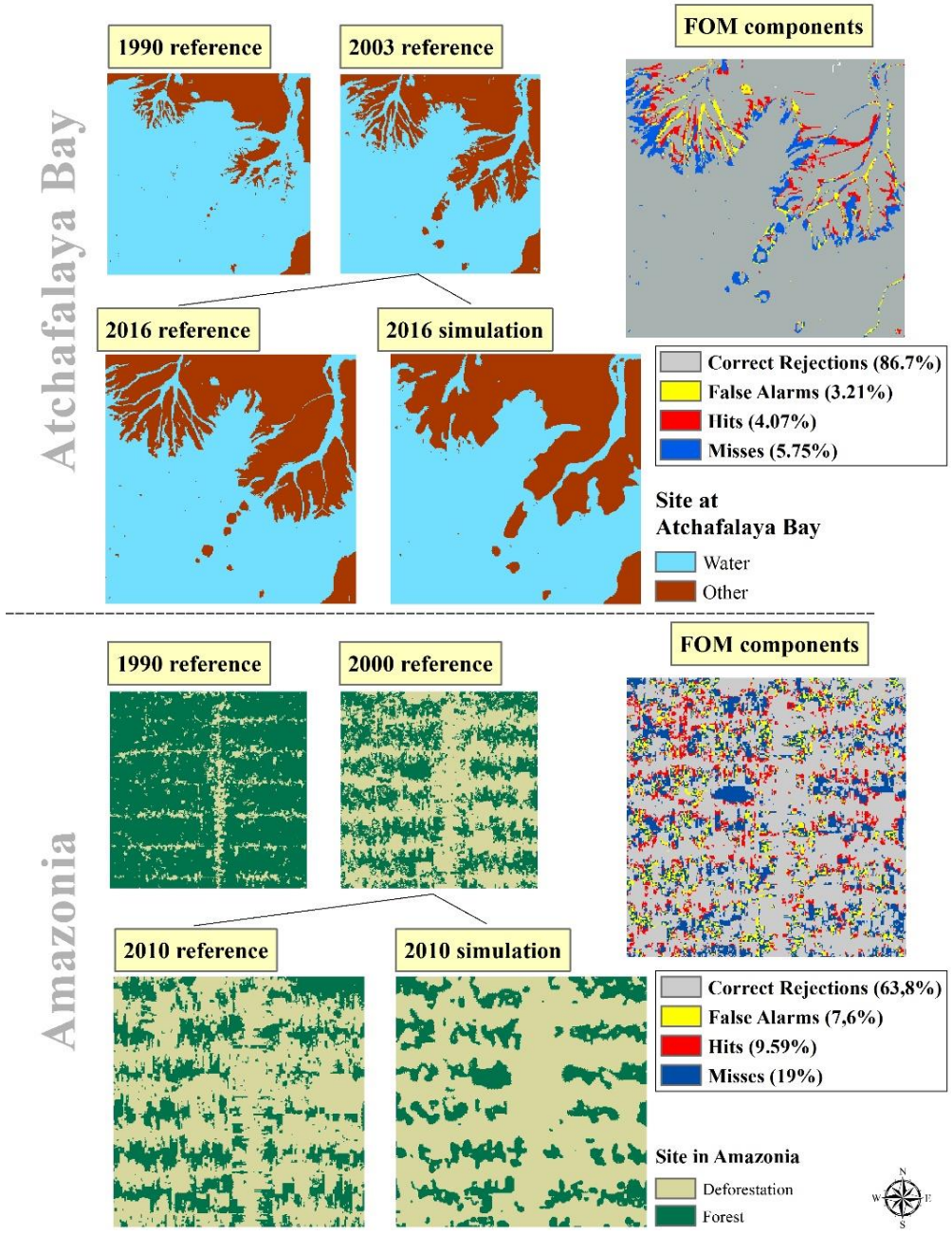

Figure 21. *Figure of merit component values of study site group 3, zoom level L areas. The figure presents the time #1, time #2 and time #3 maps in both areas.*

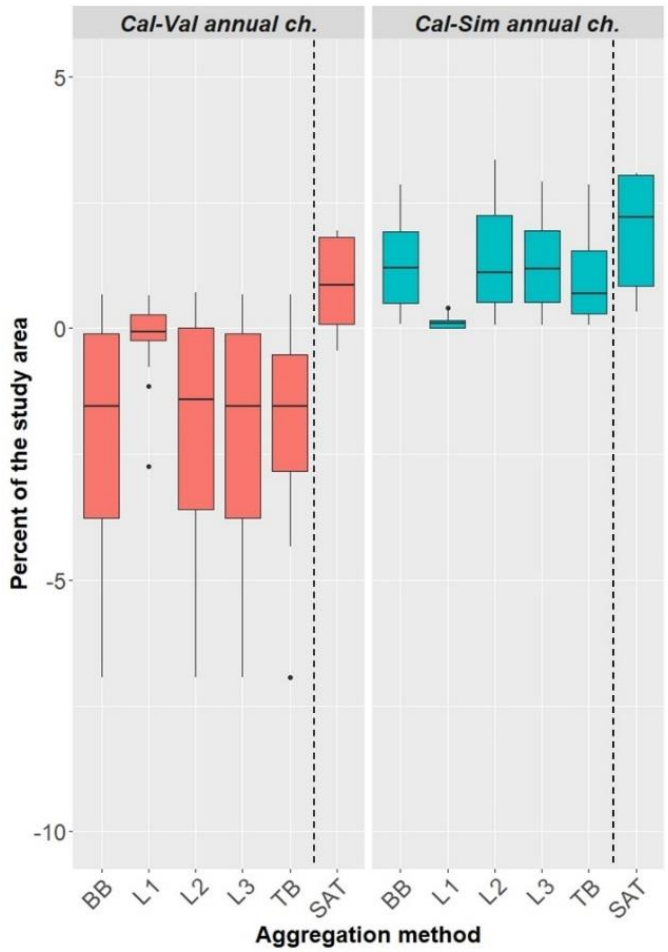

***Figure 22**. Presentation of difference of annual changes in study site group 2 and study site group 3, expressed as percent of the study area, based on the comparison of calibration and validation interval changes (Cal-Val. annual ch., left side) and based on the comparison of calibration and simulation interval changes (Cal-Sim annual ch., right side). The cases based on satellite image analysis are separated with a dashed line and assigned with label "SAT". (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation, SAT=satellite image-based analysis).*

An important difference related to study site 2 is presented in Figure 22, where the annual changes between calibration and validation interval and the annual changes between calibration and simulation interval were presented. The calculation of annual changes provides a good basis for comparison of the results, since the study designs in study

site 2 and 3 applied different durations of time intervals, and durations are even different within Study site group 3 (13 years in Atchalaya Bay, 10 years in Amazonia), as labelled in *Figure 21*. If we compared the overall changes, then the result would not be weighted by the duration of the time intervals and this issue would be an appropriate basis for a misleading interpretation. *Figure 22* shows that the changes mostly decelerated from the calibration to the validation interval, and the changes decelerated in all cases from the calibration to the simulation interval in study site group 3. The deceleration from calibration to the validation interval could have been observed in only some cases of the L1 dataset. According to the decelerating pattern of both reference and simulation data, the patterns are matching in terms of the rate of changes. Furthermore, the rate of deceleration was the most considerable in case of the satellite-based dataset, related to other datasets, all derived from CLC.

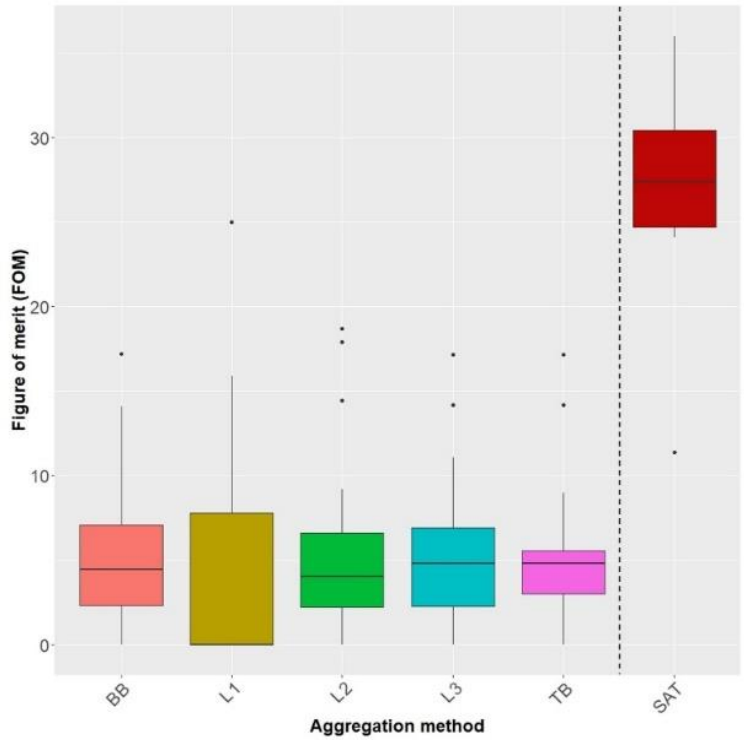

Figure 23. *Comparison of Figure of merit (FOM) values in study site groups 2 and 3, grouped by aggregation method. The cases based on satellite image analysis are separated with a dashed line and assigned with label "SAT". (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

Figure 23 reports the values of FOM as compared to study site group 2 datasets again. In this context, FOM was outstandingly high, since FOM median converged to 28%, which value could be achieved by only outliers of the study site group 2. It means that model performance was much better in these sites as compared to study site group 2.

Figure 24 reports FOM components, as compared to Study site group 2 results. There was no substantial difference in terms of False Alarms and Misses, since the medians had similar values as compared to BB, L2, L3 and TB datasets of study site group 2. However, Wrong Hits

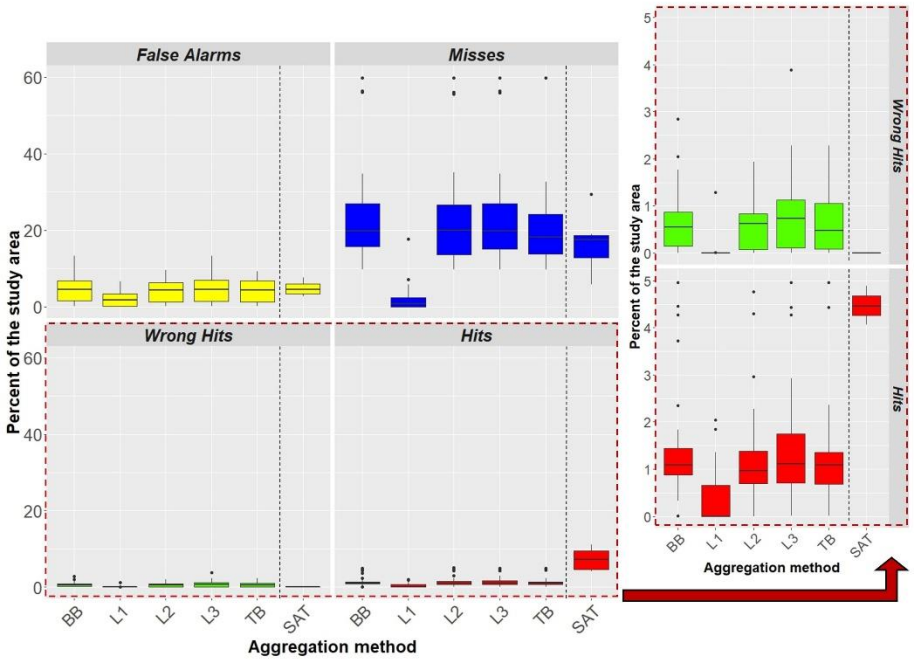

Figure 24. *Different types of FOM components in study site groups 2 and 3, grouped by aggregation method. The cases based on satellite image analysis are separated with a dashed line and assigned with label "SAT". (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

and Hits were substantially different, as Wrong Hits were all zero and Hits were all over 4%. It means that Hits values were all larger than in all study site group 2 cases, but Wrong Hits were always zero, meaning no error originating from the simulation of changes to wrong category. In this sense, the study site group 3 cases are similar to L1 datasets.

However, here the chance of simulating to wrong category was zero, since there were only 2 categories in the landscape with a single opportunity for changing into one another, meaning no opportunity for changing to a wrong category.

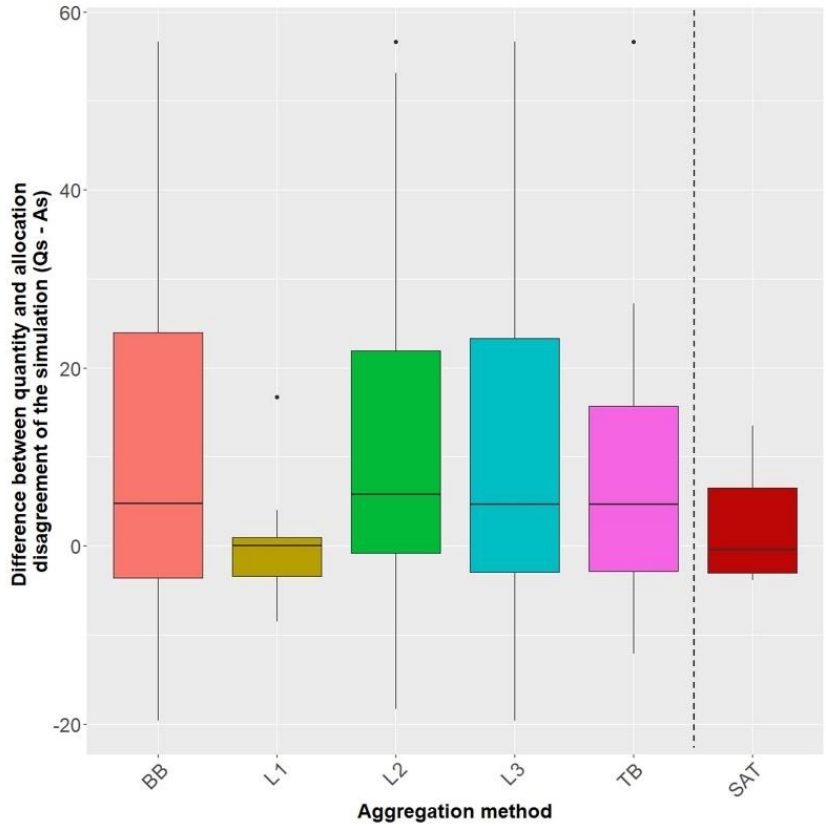

Figure 25. *Difference between quantity (Q_s) and allocation (A_s) disagreement of the simulation in each case in study site groups 2 and 3, grouped by aggregation method. If the value is positive, then Q_s is larger. If the value is negative, then A_s is larger. The cases based on satellite image analysis are separated with a dashed line and assigned with label "SAT" (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

In *Figure 25*, the difference between quantity and allocation disagreement was visualized for each cases. Since some cases are over zero, it suggests that quantity disagreement (Q_s) was larger than Allocation disagreement (A_s) which means that the model had slightly more error originating from quantity than from allocation issues. However, the median was close to zero, suggesting that Q_s and A_s were

approximately equal which means that almost equal error originated from quantity and allocation issues.

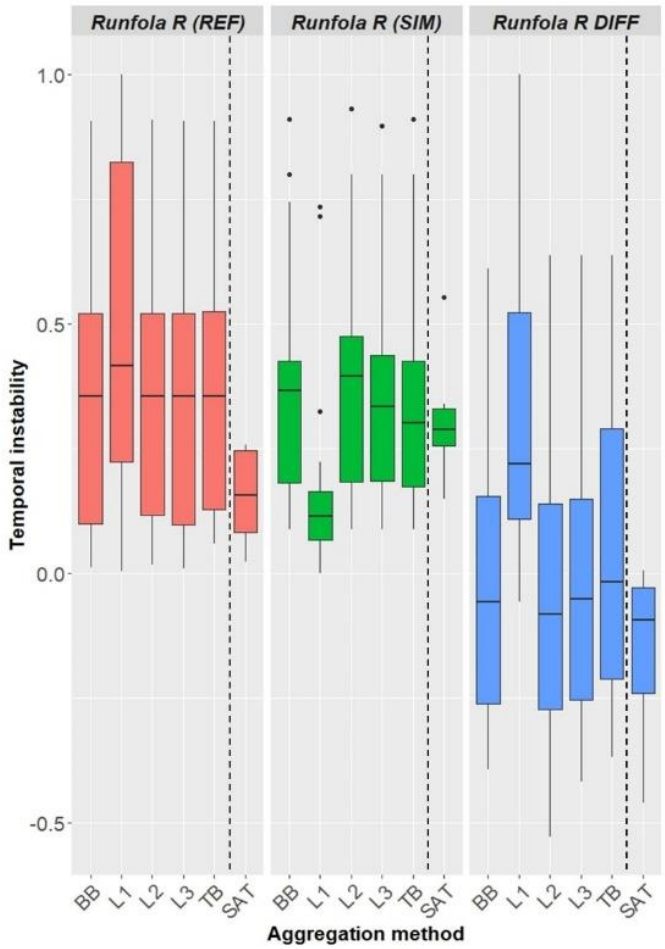

Figure 26. *Runfola's R values measuring temporal instability between time intervals. R values concerning temporal instability between calibration and validation intervals [Runfola R (REF)], concerning temporal instability between calibration and simulation intervals [Runfola R (SIM)], and the difference between the two Runfola R values calculated for each case [Runfola R DIFF = Runfola R (REF) - Runfola R (SIM)]. If Runfola R DIFF value is positive, then Runfola R (REF) is larger. If Runfola R DIFF value is negative, then Runfola R (SIM) is larger. Values of all three variables are grouped by aggregation method. The cases based on satellite image analysis are separated with a dashed line and assigned with label "SAT" (BB=Behavior-based aggregation; L1 = CLC Level 1 aggregation; L2 = CLC Level 2 aggregation; L3 = CLC Level 3 aggregation; TB = Threshold-based aggregation).*

Temporal stability values are presented in *Figure 26*, where variations of Runfola's R values are compared to study site group 2

values. According to Runfola R REF, the temporal instability was the lowest in case of satellite-based dataset, meaning it had the most stable changes from the calibration to the validation interval. On the contrary, satellite-based datasets had the most instable changes from the calibration to the simulation interval. Therefore, the instability increased as compared the simulation to the reference data, presented by Runfola R DIFF values. This metric also highlights that instability increased in all cases of study site group 3 from the reference to the simulation data, because Runfola R DIFF maximum value was zero, meaning that Runfola R REF was lower than Runfola R SIM in all cases.

4.6. Discussion concerning Study Site Group 3

It is important to discuss study site group 3 results in the context of study site group 2 results, because these cases highlight even more specific relations between model performance and changes in the study area. Study site group 3 presented sprawl-like changes, which are likely to sprawl around the original categories, this type of change is in accordance with the operation of the model. If the model simulates changes around the edges of the original categories, it simulates a sprawl-like change, even if the real change is not a sprawl-like phenomenon.

Figure 21 presented that the Hits concentrated on the edges of the original categories, which were persistent pixels, while the Misses concentrated into compact patches (Amazonian example) or near to Hits (Atchafalaya example). This latter example clearly shows that the Hits were located around the original category where the changes had been expected to occur, and Misses were located besides the Hits as a result of the model's underestimation of changes. If the model did not underestimate changes and simulated more changes, probably the Misses would have become Hits as well. The model were likely to consider only the spatial filter when allocating the changes, since it did not consider the lead of the rivers, just simulated changes into all direction from the original category *Other*, even closing up the rivers in simulation map 2012. Therefore, False Alarms followed the leads of the closed rivers, where the model simulated changes instead of persistence.

The Amazonian example shows a slight salt-and-pepper effect in case of Hits, Misses and False Alarms. Hits seemed to be located at the edges of the original categories, since the logging in the forest is also a sprawl-like phenomenon. However, there was a salt-and-pepper effect in the real changes, since the deforestation advanced in a manner that some individual forest patches sparsely remained in the area and sometimes forest patches witnessed deforestation around them. Meanwhile the model simulated a situation where the deforestation sprawls perfectly

around the existing patches. Consequently, False Alarms were located where these forest patches were still present, Misses were located where the advance of deforestation concentrated in an area and Hits were present where the advance of deforestation matched a perfectly sprawling dynamic. Pontius Jr et al. (2007) used a simulation model for projecting deforestation in the Amazon basin and found that the phenomenon is likely to occur near the local main and secondary roads. In this model there was no input information or driver in the model, only the land changes in the calibration interval. It is logic to assume that the deforestation occurs near the main roads in a harsh environment like Amazonia, since the proximity of roads helps the accessibility of the forests. However, in the model presented in *Figure 21*, the 1990 reference maps helps to identify the possible location of roads, but the sprawl did not follow a uniform spreading pattern from the possible location of roads towards the forests, since the deforestation left some remaining forest patches while spreading. These patches caused the salt-and-pepper effect in the FOM component map.

The FOM values of this study site group were especially high, meaning especially high model performances related to study site group 2. However, it is really important to see how these FOM values evolved on the ground of various FOM components. FOM is calculated based on FOM components, where the Hits values are divided by the sum of Hits and erroneously simulated pixels. These erroneously simulated pixels are the bases for various errors, expressed as Misses, False Alarms and Wrong Hits. If the Misses, False Alarms and Wrong Hits are high and Hits are low, then the nominator of FOM calculation will be low and the denominator will be high, consequently, FOM will be low. If Hits increases, besides the same values of errors, the FOM will be higher. If Hits increases, but the errors also increase, the FOM will not necessarily increase. If Hits are the same, but errors increase, then FOM will decrease. These combinations show why it is not enough to consider only the correctly simulated changes, as known as Hits. In study site group 3, the Hits increased substantially, but the False Alarms and Misses did not increase, while Wrong Hits even decreased, all related to study site group 2. This combination resulted in higher FOM values related to study site group 2.

Concerning the tendency of annual changes, the rate of annual changes from the calibration to the validation interval matched the rate of annual changes from the calibration to the validation interval, regarding the fact that the changes decelerated from one interval to another. However, the rate of deceleration was different, since the changes decelerated more from the calibration to the simulation interval.

It means that the model matched the tendency of deceleration, but underestimated the quantity by which the change decelerates. It may be the reason for the Runfola's R values, where the Runfola's R was larger in case of the simulation indicating more instable changes. Runfola's R characterizes the changes that should be reallocated in order to achieve a uniform change throughout the whole time interval (Runfola and Pontius Jr, 2013). In order to achieve a uniform change throughout the calibration and simulation intervals, more quantity of changes should be reallocated, because the model simulated a dynamic deceleration. Throughout the calibration and validation intervals, the real changes decelerated more slowly, so less quantity of changes should be reallocated to achieve a uniform change. It is a good example of that the temporal stability does not necessarily reveal the tendency of changes that the comparison of annual changes reveals, namely the acceleration or deceleration of changes, because it does not inform about the tendency, only about the necessary quantity of changes to reallocate in order to achieve the uniformity.

5. DISCUSSION OF OVERALL RESULTS IN THE CONTEXT OF CONTEMPORARY LITERATURE AND FUTURE PERSPECTIVES

In literature there are few examples of comprehensive analysis of LULC change model performance. There were researches in contemporary literature about the different modelling approaches, like Dinamica EGO, CLUE, Land Change Modeler, CA-Markov (Mas, et al., 2014; Olmedo, et al., 2018; Olmedo et al., 2015; Paegelow et al., 2014; Paegelow and Olmedo, 2005) or specifically the Markovian matrix (Takada et al., 2010). Mas et al. (2014) provided a comprehensive overview of possible errors of the following simulation models: CA-Markov, Land Change Modeler (LCM) and Dinamica. In that case, the examined models all applied Markov chains in order predict the quantity of changes, and used different approaches for spatial allocation of changes. They found that although all models used Markov matrices throughout the simulation, but the CA-Markov predicted substantially less change than the other two models and that model was closer to null hypothesis that showed only persistence in the relevant time interval.

In this research, all the models predicted a deceleration of changes, meaning a decrease in annual change from the calibration to the validation interval. In our article (Varga et al., 2019) that described the analysis in study site group 1, we also discussed that the model *"simulated fewer and smaller transitions than an extrapolation of a Markov chain would dictate"*. However, in that case, the model

simulated correctly the deceleration of the changes in the study area. The allocation error was larger than the quantity error which could be observed in some L1 cases in study site group 2, while study site group 1 can be considered as an L1 case, since it was a five-category aggregation of L3 data, however, with a 25 m spatial resolution. Our article (Varga et al., 2019) was the first example of using intensity analysis and FOM components together for model validation purposes. Intensity analysis helped to reveal the deceleration of changes and important category-level and transition-level changes in the study area, as compared to simulation. It would be an exciting future perspective of the research to apply intensity analysis on all datasets of study site group 2 and 3 as well. There is an intensity analysis framework where the scientists can perform the analysis on their own datasets, but in 2019 an R software package was published with the name of 'intensity.analysis' package, by which intensity analysis could be performed *en masse*. It creates a good basis for deeper insights into the category-level and transition-level dynamics of the datasets.

Based on all the study site groups, the results enlightened that the models always simulated a decelerating tendency. It refers to the fact that study site group 1 simulation would have not matched the tendency in the validation data, if the validation data had showed accelerating changes. The model matches this kind of tendency in the changes only if the real change decelerates. The cases of study site group 2 support this finding, since the cases in study site group 2 showed that quantity disagreement was mostly larger than allocation disagreement. Furthermore, study site group 3 cases demonstrated the relation of the contiguity filter and the mechanism of changes, while the map of FOM components revealed the errors originating from this relation. In study site 3, quantity and allocation errors were mostly even, but the change decelerated more in the simulation than in the validation interval.

Mas et al. (2014) claimed that CA model was suitable for only applications where there is a specific rule of neighborhood in changes, since the model was designed for urban growth simulation. In the Atchafalaya example, the Hits concentrated on the edges of the original categories, in a great unison with the logic of contiguity filter. However, the very presence of the contiguity filter caused land changes that would be implausible in reality, like the close up of rivers.

For the best of my knowledge, there is not any study of running a large set of CA-Markov models and examine the model performance under similar circumstances. Our article (Varga et al., 2020) was the first example of investigating the effect of category aggregation on model performance, based on FOM and FOM components, in a large set of

models. Our article (Varga et al 2019), based on study site group 1 results, revealed that FOM is not enough to qualify the model since it does not distinguish quantity and allocation errors, and the interpretation of a combination metrics can help to evaluate the reasons for errors. There are quite few examples of using Figure of merit for the validation of a simulation model in literature (Cao, M. et al., 2019; Memarian et al., 2012; Tajbakhsh et al., 2018). The analysis of FOM components in details or in context of allocation and quantity disagreements were performed even more rarely, and it seems to be a really special and focused subproblem of simulation modelling (Chen and Pontius Jr, 2010; Feng et al., 2019; Mejean et al., 2019; Wang et al., 2019).

There is a model parameter in Idrisi's CA-Markov model, where the user can set a proportional error for the model that originates from the input map error. In this research, the proportional error was set to zero, because in a couple of previous empirical observations, the model returned simulations with weaker model performance when setting this parameter to a 85% accuracy of the input maps (set to a 0.15 proportional error), compared to models with zero proportional error. Mas et al (2014) reported that when this proportional error option was applied in CA-Markov model approach, this action resulted in affecting area estimations significantly. The effect of input map error could hardly be checked in study site group 2 dataset, because all the input maps are derived from CLC data, where the reported thematic accuracy is over 85%, and the exact same parameter should be set for all the models. Consequently, the study design did not create an opportunity for testing the systematic effects of various input map accuracies on model performance.

Another model parameter is the iteration number of the model which is advised to be the number of years of the duration of the analyzed time interval (Eastman, 2012a). The spatial filter is also modifiable, and the user can modify the default 5x5 spatial filter to a user-defined filter in the analysis. There are various classic types of spatial filters that are used for edge detection in image processing (Birchfield, 2016), even in a combination of cellular automaton and spatial filter for a purpose independent from LULC modelling (Sharma et al., 2013). Although Deep and Saklani (2014) used Kappa coefficient for model validation in their paper, which is a wrong approach for validation of the model, but they tested the effects of iteration number and neighborhood on the simulation and found that 12 iterations and a 17x17 neighborhood returned models with the highest kappa values. In Varga et al. (2019) we also tested spatial filter variations and we found that a larger filter would allow changes to extend farther from the original patches, but still

concentrate changes near the original patches. It was obvious throughout this research design that the spatial filters caused a concentration of changes on the edges of the original categories, and partly due to this phenomenon, cases in study site group 3 resulted in better model performance. If the modeler checked patch patterns of the calibration and validation interval changes before running the model, it would possibly provide a trace whether the simulation design match the changes that the modeler wants to simulate. The whole dataset of study site group 2 could be checked in FRAGSTAT (McGarigal, K. et al., 2012) that is capable of calculating landscape metrics based on raster datasets, and shape-focused indices (Haines-Young and Chopping, 1996) of changing patches could lead to a useful pre-check in the analysis. There were researches concerning the variability of land changes using landscape metrics to characterize this feature of land change (Szilassi, 2017).

The issue of spatial resolution was not tested in this research, since most of the models were derived from CLC dataset, thus the spatial resolution applied in the study design was determined by the resolution of the input dataset. There were a few investigations on the effects of resolution on model performance in literature (Olmedo et al., 2018; Pontius Jr, et al., 2011) and even on the scale-dependency of the driving factors of change (Verburg et al., 1999), but this study design did not create an opportunity for testing the systematic effects of spatial resolution on model performance. If the models received input data derived from satellite-images with various resolutions, then an analysis on the effects of spatial resolution would make sense, which is an interesting subject for future research.

The practical utilization of the results is not limited to scientists who intend to run land change models. Due to the fact that the models in this research did not integrate land change drivers, it is rather a research focused on practical issues and provides insights in order to identify or avoid possible errors in the CA-Markov model. It is really important to know how the model works before integrating any drivers which could complicate the model and then the exact origin of errors could not be distinguished. According to Verburg et al. (2004a) we still do not have enough information or understanding of land change processes to decide which land change modelling approach suits our purposes the most. Land change modelling is a good basis for setting up future scenarios in a world where many recent changes concerning climate and habitat may be irreversible. Land change policy and decision-makers need complex information on the changing processes, and landscape ecology is in a good position to investigate the causes and interactions of these processes (Mayer et al., 2016). Schuwirth et al (2019) declared certain

conditions that helps increasing usefulness of models for ecological and land management, like sufficient predictive performance, among others. They also claimed it is important to suit these requirements when the policy-makers need models to support decisions presented to the public. It is a substantial demand towards land change modelling science to investigate possible errors in modelling issues in order to being capable of recognizing, correcting or avoiding these errors. This dissertation intended to take a step in this direction.

6. CONCLUSIONS

The analysis focused on CA-Markov models and their abilities of predicting LULC changes in the study areas. A variety of metrics were calculated in order to measure the numbers of categories, the ratio of changes, the model performance and temporal stability under specific circumstances. The following points summarize the most important findings:

(1) Intensity analysis substantially contributed to the validation of the simulation model, since it revealed the real and simulated changes in detail, thus helped to reveal the reasons for the unsuccess of the model.

(2) The combined usage of comparison of the calibration to the validation interval and comparison of the calibration to the simulation interval revealed patterns that the FOM and components could not reveal, therefore it is recommended to consider either calibration interval changes or usage of a combination of metrics when validating a model.

(3) Category aggregation decreased changes in the study sites. In case of Corine Land Cover (CLC) Standard Level 1 change decreased the most and in a statistically significant manner, related to other aggregation methods. Behavior-based category aggregation maintained changes the most and absolutely, while model performance did not decrease significantly and number of categories decreased substantially. Therefore, CLC Level 1 aggregation is not recommended, and behavior-based aggregation is recommended to use when aggregating categories.

(4) Quantity disagreement of the Cellular Automaton (CA)-Markov models was mostly larger than allocation disagreement, meaning that the quantity control of Markov caused more errors than the allocation control of cellular automaton.

(5) The model simulated decelerating changes systematically, meaning a systematic underestimation of changes, which resulted

in large quantity errors of the models, because the changes were mostly accelerating in the study sites.

(6) FOM values, characterizing model performance, were substantially larger in case of modelling sprawl-like mechanisms. However, the model had quantity errors due to underestimating the changes and allocation errors due to the uneven sprawl mechanism.

(7) Kappa Index of Agreement and Overall Agreement showed a strong correlation with validation interval changes, moreover showed high agreements also in the cases where correctly simulated changes were extremely low. This demonstration in a large set of datasets clearly shows why the usage of these metrics is not recommended for the validation of simulation models in the context of comparing reference and simulation time #3 maps.

These results may help scientists see behind the scenes of CA-Markov model, its logic and operation, when it is free of any drivers or influencing factors of change. I still hope that my dissertation helps a better understanding of category aggregation consequences and model validation approaches, and contributes to the dissemination and propagation of good practices and possible errors in land change modelling science.

S U M M A R Y

The main purpose of my research was to analyze land change models that were also capable of demonstrating the capability or incapability of certain model performance metrics. Furthermore, another important purpose was to analyze the changes in the landscape in order to reveal the detailed background of model performance. The analysis was based on a large set of CA-Markov models, by which I drew conclusions concerning the following issues:

- how the detailed change analysis help the analysis of model performance;
- how certain category aggregation methods influence the model performance;
- which methods are not suitable for a correct model performance analysis;
- how the operation of the model influence model performance.

The innovation of the research was that the research questions have not been analyzed in a large set of models before. It is an innovation from a methodological point of view that in the context of model performance analysis, intensity analysis have not been used and the effect of category aggregations have not been investigated.

The analysis was performed in 3 study site groups. Study site group 1 contained only one study site located around Tokaj city with an extent of 25 x 25 km and a quite heterogeneous land character. The facts that it is located at the joint of five microregions and all categories of the level 1 of Corine Land Cover standard nomenclature are present in the area, also show its heterogeneity. The analysis was based on Corine Land Cover (CLC) data concerning the years 2000, 2006 and 2012, using subsets of the vector database resampled to 25 m raster datasets. 5 categories were used according to CLC standard nomenclature Level 1, which were the followings: artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands and water bodies. Partly due to the presence of protected areas, the changes in the area did not exceed 2% of the study area in both time intervals examined (2000-2006 and 2006-2012). CA-Markov models were run based on the 2000 and 2006 maps as training data and the model simulated a map for the year 2012. The model was validated based on 2006 and 2012 reference maps. Intensity analysis was used for investigating the changes in the study area in detail concerning either reference or simulated changes, then model performance metrics (Figure of merit [FOM] and its components) were calculated.

Study site group 2 consisted of 8 study sites that were chosen on the basis of Corine Land Cover change layers concerning 2000-2006 and 2006-2012 time interval changes. The main aim was to find study sites with as large ratio of changing areas as possible, so as to produce a dataset with various quantities of changing areas in the study sites. In this case, the CLC datasets of 2000, 2006 and 2012 were used, however, using 100 m spatial resolution raster version instead of resampled vector version. In each study site, two further subareas were assigned, so each study sites consisted of three subareas according to their zoom level (large = L, medium = M, small = S). The assignment of the study areas was complicated, because the CA-Markov had specific requirements. For instance, the study areas must not have had more than 20 categories and the maps representing the first two dates (2000 and 2006 in this case) must have had the exact same categories. The categories of the study areas were aggregated according to the following schemes:

- the basis of the maps were the categories of level 3 of CLC standard nomenclature (L3);
- the categories of level 3 of CLC standard nomenclature were aggregated based on the level 2 of CLC standard nomenclature (L2);
- the categories of level 3 of CLC standard nomenclature were aggregated based on the level 1 of CLC standard nomenclature (L1);
- the categories of level 3 of CLC standard nomenclature were aggregated based on the behavior-based category aggregation method (BB), where the user may decide the degree of aggregation based on a stepwise aggregation procedure and the user can monitor the status of changes in every step of aggregation;
- the categories of level 3 of CLC standard nomenclature were aggregated based on the threshold-based category aggregation (TB), where the user may decide which categories should be aggregated into a new category based on an arbitrary threshold of ratio of changes in the area. In 6 cases the changes did not meet the applied threshold, therefore the aggregation was not performed.

As a result of various study sites, zoom levels and category aggregations, 114 models were run altogether in Study site group 2. After running the models, metrics concerning model performance, changes and other variables were calculated, they were analyzed by statistical methods and then comprehensive conclusions were drawn.

The characteristics of Study site group 3 were substantially different from the other two study site groups. I aimed to investigate phenomena with sprawl-like dynamics where the changes affect the neighboring areas of the original categories. Study sites with these characteristics were selected in North and South America. In North America, the study site was located in Atchafalaya Bay where a delta accumulation could be observed. In South America, the study site was located in Amazonia, where massive deforestation could be observed. Since the study sites were located outside Europe, CLC could not be used in this analysis, so time-series Landsat image datasets were processed. By segmenting and classifying the images, 2-category maps were created that enhanced the target phenomena and after performing resample procedure they matched the 100 m spatial resolution of study site group 2 maps. In each study site, two further subareas were assigned, so each

study sites consisted of three subareas according to their zoom level (large = L, medium = M, small = S). In this case, CA-Markov models were run again, and after running the models, metrics concerning model performance, changes and other variables were calculated, similarly to study site group 2. Due to the different input data and parameters of study site group 3, a statistical comparison to study site group 2 cases could not have been well-grounded, hence the conclusions of the comparison were rather empirical.

The most important element of methodology was the CA-Markov model, which is a land change model, and it is capable of simulating a categorical land use/land cover (LULC) map based on input LULC maps representing two different dates. I ran the models in Idrisi software environment. The model consists of cellular automaton (CA) and Markov components, where the latter is responsible for the quantity of simulated changes and the cellular automation is responsible for the allocation of changes. Throughout the dissertation, I referred to the time interval used for training or calibrating the model as the calibration interval. I referred to the time interval between the reference dates used for validation as the validation interval, which was the time interval between 2006 and 2012 in case of study site group 2. Finally, I referred to the time interval between the reference and simulation dates, where the latter date is the date to which the simulation model projects forward, as the simulation interval. The model produces conditional probability maps, and transition area and probability matrices based on the calibration interval changes and then a contiguity filter determines the allocation of changes. I did not include any specific drivers of changes in the model.

Model performance was measured by the Figure of merit index and its components (Hits = correctly simulated changes; Wrong Hits = changes simulated as changes to wrong category; False Alarms = reference persistence simulated as change; Misses = reference change simulated as persistence), and by Quantity (Q_s) and Allocation disagreement (A_s) of the simulation that can be derived from Figure of merit components. The FOM components reveal the agreement and disagreement of reference and simulated changes. Moreover, I calculated Kappa coefficient and overall accuracy metrics based on the comparison of reference and simulated maps of time #3 maps (2012 in case of study site group 2 and 2010 and 2013 in case of study site group 3). Further variables were concerned, as follows:

- number of categories

- quantity of changes in all examined time intervals
- quantity of annual changes in all examined time intervals
- difference of changes between time intervals
- temporal stability between the calibration and validation interval changes, and between the calibration and simulation interval changes

Statistical analysis was conducted in study site group 2 exclusively, by ANOVA and Tukey pairwise comparison tests. Statistical analysis aimed to reveal whether there is significant difference between the medians of aggregation groups concerning the variables measured throughout the analysis. In order to measure correlation between the variables and model performance, a correlation matrix was set up where Spearman's rs coefficient was applied ($p<0.05$).

In case of study site group 1, the ratio of changes remained under 2% in either the calibration, validation or simulation intervals, while the annual changes decelerated in both validation and simulation intervals, related to the calibration interval. The ratio of correctly simulated changes (0.02%, expressed as a ratio of the study area) and model performance (FOM=0.007%) were extremely low. However, the allocation disagreement of the simulation (2.12%, expressed as a ratio of the study area) was larger than quantity disagreement (0.41%, expressed as a ratio of the study area). This refers to the fact that CA component of the model caused more errors, than the Markov component. Based on the location of False Alarms and Misses values, the model placed the changing areas to the neighboring areas of the original categories, which was probably an effect of the contiguity filter. Intensity analysis results showed that more similarity could be observed between the calibration and simulation interval change dynamics, than between the calibration and validation interval change dynamics and that could partly lead to the unsuccess of the model. Furthermore, intensity analysis revealed the dynamics of changes in all three intervals that a comprehensive metric, like FOM, could have not revealed.

The results of study site group 2 showed that the changes in L1 group were significantly lower than the other aggregation groups in either calibration, validation or simulation intervals. The number of categories drastically decreased in L1 group, but a substantial decrease could be observed in BB group as well. However, in BB group, the changes did not decrease at all, since an important aspect was to maintain changes when performing BB aggregation. By analyzing annual

changes, it became clear that the model always simulated decelerating changes, although the validation interval changes showed mostly accelerating tendencies.

There was no significant difference between the aggregation groups concerning FOM, but L1 group FOM, Hits and Wrong Hits medians converged to zero. In case of L1 group, all FOM components were lower than other groups, but the ratio of these components is even more important when interpreting model performance and FOM. Misses and False Alarms were higher than Hits and Wrong Hits, which leads to the assumption that the contiguity filter affected the results, such as in case of study site group 1. The statistical analysis returned strong correlation between Misses and validation interval changes (R2=0.95) and between False Alarms and simulation interval changes (R2=0.91). In study site group 2, the quantity disagreement of the simulation was characteristically larger than the allocation disagreement, and concerning both metrics, L1 group values were significantly lower than other aggregation groups' values. Regarding individual cases, quantity disagreement was mostly larger than allocation disagreement again, however, in L1 group, both cases were characteristically present. It refers to the fact that Markov component of the model caused more errors, than the CA component. The statistical analysis returned strong correlation between Quantity disagreement and validation interval changes (R2=0.82) and mild correlation between Quantity disagreement and the difference between calibration and validation interval annual changes (R2=0.65).

Based on the measurement of the difference comparison of the instability between the calibration and validation intervals and the instability between the calibration and simulation intervals, L1 group showed substantial decrease from the reference to the simulation instability. It means that the changes between the calibration and the simulation intervals were much more stable than the changes between the calibration and the validation intervals.

The map of the last reference date (2012, 2010 and 2016 in study site group 2, Amazonian case and Atchafalaya Bay case, respectively) and the relevant simulated map were compared by calculating Kappa coefficient and Overall agreement metrics. Both metrics returned significantly higher values in L1 group, where the less changes could be observed. Statistical results supported the strong correlation between either Kappa index of agreement or Overall agreement and validation interval persistence and Correct Rejections, latter meaning the correctly simulated persistent areas. Therefore the usage of these indices can be seriously misleading when using for the purpose of model performance

assessment, since they return large agreement values, even if the model hardly matched reference changes. This idea has already been published before, but a systematic relationship has not been proved in a large set of models.

In case of study site group 3, Hits were present mainly around the borders of the original patches, which is in accordance with the mechanism of sprawl-like changes and the effect of contiguity filter. In the case of Atchafalaya Bay, Misses were located near Hits frequently, meaning the model did not simulate as much changes as the reference data showed. This phenomenon is in accordance with the results of study site group 2, where the model systematically underestimated the changes. In the Amazonian study site, a salt-and-pepper effect could be observed, due to the allocation of reference changes. Here, Hits were located near the original patches again. The sites of study site group 3 returned drastically higher FOM values than sites of study site group 2. However, among FOM components, only Hits were drastically higher than study site group 2 cases. Wrong Hits always returned zero, because the maps consisted of 2 categories only, therefore it was impossible to simulate changes to a wrong category. The stability of changes in the reference time intervals was substantially larger than in study site group 2 cases. The sprawl-like change mechanism was much more in accordance with the logic of the model than sparsely located changes.

As summarizing the conclusions of the whole study, intensity analysis and the investigation of calibration interval changes substantially helped to reveal the reasons for the unsuccess of the model. The CLC L1 category aggregation hid important changes in the landscape that is a disadvantageous circumstance when performing land change simulation model. Quantity disagreements were mostly larger than Allocation disagreements of the simulation which means that Markov component of the model caused more errors, than the cellular automaton (CA) component. All the models in the study simulated decelerating changes, even if the reference changes were mostly accelerating changes, therefore the model is able to match the tendency of reference changes only if it is decelerating as well. Contiguity filter caused a concentration of changes to the neighboring areas, which is advantageous when simulating sprawl-like changes. The research presented systematic relations and errors based on a large set of simulation models and these conclusions can help the work of the modelers directly, and the workflows that support decision making concerning land change issues indirectly.

Ö S S Z E F O G L A L Á S

Kutatásom fő célja olyan tájváltozás modellek részletes vizsgálata volt, amelyek képesek demonstrálni egyes teljesítmény mérési módszerek alkalmasságát, illetve alkalmatlanságát. Célom volt továbbá a tájban lejátszódó változások részletes vizsgálata annak érdekében, hogy felfedjem a modell teljesítményének részletes okait. Vizsgálatom a CA-Markov típusú modell nagy esetszámon történő futtatására épül, melynek alapján következtetéseket vontam le a következőkre nézve:

- hogyan segíti a részletes változásvizsgálat a modell teljesítményének vizsgálatát;
- hogyan befolyásolják az egyes kategória aggregációs módszerek a modell teljesítményét;
- mely módszerek nem alkalmasak a modell teljesítményének érdemi vizsgálatára;
- a modell sajátos működése hogyan befolyásolja a modell teljesítményét.

Kutatásom újszerű megközelítését az adja, hogy a vizsgált összefüggéseket korábban nagyszámú modellen még nem bizonyították. Továbbá módszertani értelemben új megközelítés, hogy a vizsgálatban alkalmazott intenzitás-vizsgálat nevű módszert korábban modell teljesítmény mérésének kontextusában nem alkalmazták, és hogy a földhasználati-felszínborítási kategória összevonások modell teljesítményre vonatkozó hatásait nem vizsgálták.

A vizsgálatot 3 mintaterület-csoport példáján végeztem el. Az 1. mintaterület-csoport konkrétan egy mintaterületet tartalmaz, amely egy Tokaj-környéki, 25 x 25 km kiterjedésű, igen heterogén táji adottságokkal rendelkező terület. Heterogenitását mutatja, hogy öt kistáj találkozásánál helyezkedik el, illetve, hogy a Corine sztenderd nómenklatúra 1. szintje szerinti összes kategória (mesterséges felületek, mezőgazdasági területek, erdők és természetközeli területek, vizenyős területek, vízfelületek) megtalálható a területén. A vizsgálatot ebben az esetben Corine Land Cover (CLC) adatbázis segítségével végeztem, és a 2000., 2006. és 2012. évi vektoros adatbázisok kivágatának 25 méteres térbeli felbontású raszterizált verzióját használtam. A vizsgálat során 5 kategóriát alkalmaztam a CLC sztenderd nómenklatúra 1. szintje szerint, amely szintén maximum 5 kategóriát engedélyez. A mintaterületen a védett területek jelenléte miatt kismértékű változás volt megfigyelhető: a mintaterület 2 százalékánál kisebb arányú változás mindkét vizsgált időszakban. CA-Markov modellt futtattam a 2000. és 2006. évi adatok segítségével, melynek alapján a modell 2012. évre egy becsült kategória

térképet hozott létre. A tájban bekövetkező változásokat egy intenzitás-vizsgálat nevű módszer segítségével azonosítottam mind a referencia, mind a szimulált változásokat tekintve, valamint különböző mutatókat számoltam a modell teljesítményének mérésére (Figure of merit [FOM] mutató és komponensei).

A 2. mintaterület-csoport 8 mintaterületből állt. A 8 mintaterületet a Corine Land Cover változásrétege alapján választottam, amely tartalmazza a változásokat többek között 2000-2006 és 2006-2012 közötti időszakokra. A fő cél az volt, hogy a választott mintaterületeken minél nagyobb mértékű változás menjen végbe, ezáltal támogatva az esetekben előforduló változások mértékének sokszínűségét. Ebben az esetben szintén a CLC 2000., 2006. és 2012. évi rétegeit használtam, azonban az adatbázisok 100 méteres térbeli felbontású raszter verzióját. Minden mintaterületen két további, egyre kisebb alterületet jelöltem ki, így minden mintaterületen összesen három nagyítási szintnek megfelelő terület jött létre (nagy = large [L]; közepes = medium [M]; kicsi = small [S]). A mintaterületek kijelölését nehezítette, hogy az alkalmazott CA-Markov modell sajátosságai miatt egyik területen sem lehetett több, mint 20 kategória, illetve hogy minimum az első két időpontban egyforma számú kategóriának kellett jelen lennie. A mintaterületeken jelenlévő kategóriákat különböző megközelítések alapján aggregáltam, melyek a következők:

- A CLC sztenderd nómenklatúra 3. szintje (L3) volt a kategória térképek alapja, a további összevonások e beosztás kategóriáit vették alapul;
- a CLC sztenderd nómenklatúra 2. szintje (L2);
- a CLC sztenderd nómenklatúra 1. szintje (L1);
- viselkedésalapú kategória összevonás (BB), amely a felhasználó döntése alapján lépésenként vonja össze a kategóriákat, az egyes összevonások következményeként fellépő változások figyelembe vételével;
- határértékalapú kategória összevonás (TB), amely a felhasználó által meghatározott határérték alapján vonja össze azokat a kategóriákat, amelyek változásai a határérték szerint meghatározott minimális változási szintet nem haladják meg. Ezt az összevonási módszert 6 esetben nem alkalmaztam, mert minden kategória meghaladta a minimális változási szintet, így nem volt szükség a kevés változást mutató kategóriákat tömörítő új kategória létrehozására.

A különböző mintaterületek, a mintaterületek nagyítási szintjeinek és az összevonási módszerek alkalmazásának eredményeképp 114 esetet

vizsgáltam. Ezekre az esetekre CA-Markov modellt futtattam, a modell teljesítményét, a változásokat és egyéb ismérveket mérő mutatókat számítottam, majd az eredményeket statisztikai módszerekkel értékeltem, és az eredményekből átfogó következtetéseket vontam le.

A 3. mintaterület-csoport merőben eltér az első két mintaterület-csoport sajátosságaitól. Ebben az esetben olyan mintaterületek vizsgálatára törekedtem, amelyek terjedésszerű változást mutatnak, tehát a változás jellemzően a meglévő kategóriákkal szomszédos területeket érinti. Ilyen jellemzőkkel bíró mintaterületet választottam Észak- és Dél-Amerika területén, az Atchafalaya-öbölben és az Amazonas-vidéken. A vizsgált időszakokban az Atchafalaya-öbölben található területen delta akkumuláció ment végbe (érintett időpontok: 1990, 2003, 2016), míg az Amazonas-vidéken található területen nagyfokú erdőirtás volt tapasztalható (érintett időpontok: 1990, 2000, 2010). A mintaterületek nem Európában találhatóak, tehát CLC adatbázist nem alkalmazhattam a vizsgálat során, ezért Landsat-felvételek idősoros elemzésével hidaltam át az adathiányt. A felvételek szegmentálása és osztályozása révén 2 célkategóriából álló térképeket hoztam létre, amelyek a vizsgált jelenségek változásait hangsúlyozták, és újramintavételezés után 100 méteres térbeli felbontással rendelkeztek. Minden mintaterületen két további, egyre kisebb alterületet jelöltem ki, így minden mintaterületen összesen három nagyítási szintnek megfelelő terület jött létre (nagy = large [L]; közepes = medium [M]; kicsi = small [S]). Ebben az esetben szintén CA-Markov modelleket futtattam, majd a 2. mintaterület-csoporthoz hasonlóan a modell teljesítményét, a változásokat és egyéb ismérveket mérő mutatókat számítottam. A statisztikai összevetés lehetősége a 2. mintaterület-csoporttal az eltérő alapadatok és paraméterek miatt nem volt szakmailag megalapozott, ezért az összehasonlítás a két mintaterület-csoport tapasztalataiból levezetett következtetéseket eredményezett.

A módszerek központi eleme a CA-Markov modell, amely egy tájváltozás szimulálására alkalmas modell, és két bemeneti időpont kategória térképe alapján létrehoz egy kategória térképet egy harmadik időpontra. A modellt Idrisi szoftverkörnyezetben futtattam. A modell a sejtautomata (cellular automaton = CA) és a Markov komponensekből áll, melyek közül a Markov a szimulált változás mértékét határozza meg, míg a sejtautomata a változások térbeli elhelyezkedéséért felel. A dolgozatban a betanításra használt időszakot következetesen kalibrációs időszaknak, a szimulált változásokat jelző időszakot szimulációs időszaknak, míg a validációra használt változásokat jelző időszakot validációs időszaknak neveztem. A modell a kalibrációs időszak

változásaihoz igazítva átalakulási mátrixokat, valamint feltételes valószínűségeket jelző térképeket hoz létre, melyek alapján a sejtautomata egy szomszédossági szűrő segítségével lokalizálja a változó területeket. A modellekben tájváltozást befolyásoló tényezőket nem határoztam meg.

A modell teljesítményét a Figure of merit (FOM) mutatóval és annak komponenseivel (találatok=helyesen szimulált változás; helytelen találatok = helyesen szimulált változás, de nem megfelelő kategóriába; téves riasztások = referencia szerint nem változó területek, változó területként szimulálva; mulasztások = referencia szerint változó területek, nem változó területként szimulálva) mértem, illetve az ezekből levezetett mennyiségi és helyzeti eltérés mutatókkal. A FOM komponensek betekintést engednek a referencia és a szimulált változások közti egyezések és eltérések részleteibe. Továbbá az utolsó referencia időpont (1. és 2. mintaterület-csoport esetében 2012, a 3. mintaterület-csoport esetében 2010 és 2016), illetve az utolsó szimulált időpont közti egyezés mérését végeztem el a teljes egyezés, illetve a Kappa egyezési index mutatókkal. Mindemellett a következő egyéb változókat vizsgáltam:

- kategóriák száma;
- változás mennyisége a vizsgált időintervallumban;
- évenkénti változás mennyisége a vizsgált időintervallumban;
- változások különbségei a vizsgált időintervallumok között;
- időbeli stabilitás a kalibrációs és validációs időszak között, illetve a kalibrációs és szimulációs időszak között.

A statisztikai vizsgálatokat kizárólag a 2. mintaterület-csoport esetében végeztem el ANOVA-teszt és Tukey-féle páros összehasonlítás segítségével. A statisztikai teszt annak feltárására irányult, hogy az egyes kategória aggregációs módszerek mediánjai között van-e szignifikáns különbség a vizsgált változók tekintetében. A vizsgált változók és a modell teljesítmény összefüggéseinek vizsgálatára korrelációs mátrixot állítottam fel, ahol Spearman-féle korrelációs koefficienset használtam ($p<0.05$).

Az 1. mintaterület-csoport esetében a változások a kalibrációs, validációs és szimulációs időszakban is 2% alatt maradtak, az évenkénti változás a validációs és a szimulációs időszakban is lassult a kalibrációs időszakhoz képest – bár eltérő mértékben. A helyesen szimulált változások aránya (0,02%, a mintaterület viszonylatában) és a modell teljesítménye (FOM=0,007%) is extrém alacsony volt. Ugyanakkor a helyzeti eltérés (2,12%, a mintaterület viszonylatában) magasabb volt, mint a mennyiségi eltérés (0,41%, a mintaterület viszonylatában), ami

arra utal, hogy több eltérés származott a modell sejtautomata összetevőjéből, mint a Markov összetevőből. A téves riasztás és mulasztás értékek és elhelyezkedésük alapján a modell a szomszédos területekre koncentrálta a változó területeket, ami vélhetően a szomszédossági szűrő hatása. Az intenzitás-vizsgálat kimutatta, hogy sokkal több hasonlóság mutatkozott a kalibrációs és a szimulációs időszak változásainak dinamikája között, mint a kalibrációs időszak és a validációs időszak változásainak dinamikája között. Ez azt jelenti, hogy nem feltétlenül a modell által szimulált változások térnek el a kalibrációtól, hanem a valós változások dinamikája és részben ez vezet a modell alacsony teljesítményéhez. Az intenzitás-vizsgálat emellett részletesen feltárta a három időszak változásait, amelyet egy egyszerű mérőszám (FOM) nem tárhatott volna fel.

A 2. mintaterület-csoport eredményei rávilágítottak, hogy az L1 csoport változásainak mennyisége szignifikánsan alacsonyabb volt, mint a többi aggregációs módszer esetében, mind a kalibrációs, mind a validációs és mind a szimulációs időszakban. Emellett az L1 csoportban drasztikusan csökkent a kategóriák száma, bár a kategóriák számának jelentős csökkenése a BB csoport esetében is megjelent. Ugyanakkor utóbbi esetében a változások mértéke egyáltalán nem csökkent, hiszen a kategóriák e módszerrel történő összevonásánál a változások megőrzése mérvadó szempont volt. Az évenkénti változások vizsgálata alapján kiderült, hogy a modell minden esetben csökkenő változást szimulált a kalibrációs időszakhoz képest, bár a validációs időszak változásai sok esetben gyorsuló tendenciát mutattak.

A FOM tekintetében nem volt szignifikáns különbség az aggregációs módszerek között, de az L1 esetében a FOM medián a többi csoporttól eltérően nullához közelített, valamint szintén nulla értékhez közelített a találatok és helytelen találatok értéke. Az L1 esetében minden FOM komponens értéke alacsonyabb értéket mutatott a többi csoport értékeinél, de a FOM esetében e komponensek aránya a mérvadó. A téves riasztás és a mulasztás értékek minden esetben jellemzően magasabbak voltak, mint a találat és helytelen találat értékek, ami ebben az esetben is – az 1. mintaterület-csoporthoz hasonlóan – a szomszédossági szűrő hatását feltételezi. A statisztikai eredmények alapján a mulasztások és a validációs időszak változásai között (R^2=0,95), valamint a téves riasztások és a szimulációs időszak változásai között (R^2=0,91) szoros korreláció állt fenn. A 2. mintaterület-csoport esetében a modell mennyiségi eltérései jellemzően magasabbak voltak, mint a helyzeti eltérései, és mindkét mutató esetében az L1 csoport értékei szignifikánsan alacsonyabbak voltak a többi csoport értékeinél. Az egyedi esetek többségében a mennyiségi eltérés magasabb

volt, mint a helyzeti eltérés (az L1 csoport esetében mindkét lehetőség jellemző), ami arra utal, hogy általában több hiba származott a modell Markov összetevőjéből, mint a sejtautomata összetevőből. A statisztikai eredmények alapján a mennyiségi eltérés és a validációs időszak változásai között szoros (R^2=0,82), valamint a mennyiségi eltérés és a változás referencia időszakokban mutatott lassulása/gyorsulása között számottevő (R^2=0,65) korrelációs kapcsolat állt fenn.

A változások időbeli stabilitásának mérése alapján a kalibrációs-validációs időszakok között fennálló stabilitás és a kalibrációs-szimulációs időszakok között fennálló stabilitás között az L1 számottevő különbséget mutatott, amelyből kiderül, hogy a szimulációs időszak változásai sokkal stabilabbak voltak, mint a validációs időszak változásai, mindkét esetben a kalibrációs időszakhoz viszonyítva.

Az utolsó felhasznált időpont (2012) referencia és szimulált térképének összehasonlítása során a Kappa egyezési index és a teljes egyezés mutató is szignifikánsan és kiugróan magasabb értékeket adott vissza az L1 csoportban, ahol a legkevesebb változás volt megfigyelhető. A statisztikai eredmények alátámasztották a Kappa egyezési index (R^2=0,85) és a teljes egyezés (R^2=0,92) mutatók szoros korrelációját a validációs időszak perzisztens területeinek arányával, valamint a helyesen perzisztens területként szimulált területek arányával (R^2=0,96). Tehát ezen indexek alkalmazása félrevezető a szimuláció értékelésekor, mert akkor is magas egyezést adnak, ha a változások mértéke alacsony és a helyesen szimulált változások találati aránya is alacsony. Ez a megállapítás a szakirodalomban leírtak alapján ismert, de nagyszámú modellen a szisztematikus összefüggést nem bizonyították.

A 3. mintaterület-csoport esetében a találatok javarészt az eredeti kategóriahatárok mentén jelentek meg, ami összhangban van a terjedő jellegű változás mechanizmusával és a szomszédossági szűrő hatásával. Az Atchafalaya-öbölben található mintaterület esetében a mulasztások sok esetben közvetlenül a találatok szomszédságában helyezkedtek el, ami azt jelzi, hogy a modell nem szimulált annyi változást, mint amennyi a referencia adat szerint történt. Ez a jelenség összhangban van a 2. mintaterület-csoport eredményeivel, miszerint a modell szisztematikusan alábecsülte a változások mennyiségét. Az Amazonas-vidéken található mintaterületen egyfajta só-bors hatás volt megfigyelhető, ami a referencia változások hasonló elrendeződéséből adódott. A találatok ebben az esetben is jellemzően az eredeti kategóriahatárok mentén voltak láthatóak. A 3. mintaterület-csoport mintaterületei drasztikusan magasabb FOM értékeket produkáltak, mint a 2. mintaterület-csoport mintaterületei. A FOM-komponensek közül ugyanakkor csak a találat értékek különböztek nagymértékben, amelyek

sokkal nagyobb arányú helyesen szimulált változást mutattak, mint a 2. mintaterület-csoport modelljei esetében. A helytelen találatok minden esetben nulla értéket adtak vissza, mert helytelen kategóriába történő változás nem volt lehetséges, hiszen összesen két kategória szerepelt a térképeken. A változások stabilitása a referencia időszakban jellemzően magasabb volt, mint a 2. mintaterület-csoport mintaterületei esetében. A terjedő jellegű változás modellezése a vizsgálat alapján sokkal inkább összhangban volt a modell működési mechanizmusával, mint az elszórt elhelyezkedésű változások.

A vizsgálatok tapasztalatai alapján összefoglalva megállapítható, hogy az intenzitás-vizsgálat és a kalibrációs időszak vizsgálatának bevonása nagyban hozzájárult a változások megismeréséhez, és képes volt felfedni a modell alacsony teljesítménye mögött húzódó okokat. Továbbá a kategóriák CLC sztenderd 1. szint szerinti aggregációja elrejtheti a tájban lejátszódó fontos változásokat, ami tájváltozás modellezés esetén hátrányos körülmény. A vizsgált modellek esetében a mennyiségi eltérések általában magasabbak voltak, mint a helyzeti eltérések, ami azt jelzi, hogy a Markov komponens több hibát okozott, mint a sejtautomata komponens. A vizsgálatban szereplő minden modell lassuló változásokat szimulált, függetlenül a valós változások lassuló vagy gyorsuló tendenciájától, ezért a modell csak akkor képes eltalálni a valós tendenciát, ha az szintén lassuló. A szomszédossági szűrő a változások koncentrálódását okozza, ami a terjedő jellegű változásoknál kifejezetten előnyös. A kutatás nagyszámú modell segítségével mutatott be szisztematikus összefüggéseket és hibákat, amelyek nagyban segíthetik a modellező szakemberek munkáját, és ezen keresztül a releváns döntéshozást támogató munkafolyamatokat.

7. STATEMENT OF RESEARCH CREDITS

It is important to declare which parts of the research are results of external cooperation. The results concerning study site group 1 are results of a cooperation with Professor Robert Gilmore Pontius Jr Ph.D. and my counselor, Professor Szilárd Szabó DSc. These results were published in Varga et al. (2019). The results concerning study site group 2, in connection with the effects of aggregation methods on changes, FOM and FOM components are results of cooperation with Professor Robert Gilmore Pontius Jr Ph.D. and my counselor, Professor Szilárd Szabó DSc. These results were published in Varga et al. (2020).

8. ACKNOWLEDGEMENT

Special thanks to my counsellor, Professor Szilárd Szabó DSc, who always had a really good sense to see when and what would be necessary – help, support or chase me with my research. The perfect ratio of these components resulted in completing my dissertation.

Special thanks to Prof. Robert Gilmore Pontius Jr Ph.D., who supported my work with many pieces of advice and helped to find the way through the mistery of land change metrics.

Special thanks to my previous counsellor, Zoltán Túri Ph.D., that he has supported my research activity since my undergraduate years and set the idea in my mind that I would have a chance for being a candidate for a Ph.D. sometime.

9. APPENDICES

Appendix 1

Nomenclature of Corine Land Cover Category based on the guidelines of Copernicus Land Monitoring Service and Kosztra et al. (2019). A nomenclature with detailed definitions of each category is available at the website cited in the footnote. [1]

Standard levels	CLC Standard Level 1	CLC Standard Level 2	CLC Standard Level 3
Classification and category labels in different standard levels	1 Artificial surfaces	11 Urban fabric	111. Continuous urban fabric
			112. Discontinuous urban fabric
		12 Industrial, commercial and transport units	121 Industrial or commercial units
			122 Road and rail networks and associated land
			123 Port areas
			124 Airports
		13 Mine, dump and construction sites	131 Mineral extraction sites
			132 Dump sites
			133 Construction sites
		14 Artificial, non-agricultural vegetated areas	141 Green urban areas
			142 Sport and leisure facilities
	2 Agricultural areas	21 Arable land	211 Non-irrigated arable land
			212 Permanently irrigated land
			213 Rice fields
		22 Permanent crops	221 Vineyards
			222 Fruit trees and berry plantations
			223 Olive groves
		23 Pastures	231 Pastures
		24 Heterogeneous agricultural areas	241 Annual crops associated with permanent crops
			242 Complex cultivation patterns
			243 Land principally occupied by agriculture, with significant areas of natural vegetation

[1] *URL: https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/html*

			244 Agro-forestry areas
	3 Forests and semi-natural areas	31 Forest	311 Broad-leaved forest
			312 Coniferous forest
			313 Mixed forest
		32 Shrub and/or herbaceous vegetation associations	321. Natural grasslands
			322 Moors and heathland
			323 Sclerophyllous vegetation
			324. Transitional woodland-shrub
		33 Open spaces with little or no vegetation	331 Beaches, dunes, sands
			332 Bare rocks
			333 Sparsely vegetated areas
			334 Burnt areas
			335 Glaciers and perpetual snow
	4 Wetlands	41 Inland wetlands	411. Inland marshes
			412. Peatbogs
		42 Coastal wetlands	421 Salt marshes
			422 Salines
			423 Intertidal flats
	5 Water bodies	51 Inland waters	511. Water courses
			512. Water bodies
		52 Marine waters	521 Coastal lagoons
			522 Estuaries
			523 Sea and ocean

Appendix 2

Abbreviations for variables used throughout the analysis, in an order matching the variable order in *Figure 20.*

Abbreviation	Description
CR	Correct Rejections
FA	False Alarms (Figure of merit component)
WH	Wrong Hits (Figure of merit component)
H	Hits (Figure of merit component)
M	Misses (Figure of merit component)
QS	Quantity disagreement of the simulation
AS	Allocation disagreement of the simulation
TS	Total disagreement of the simulation
QS-AS	Difference between quantity and allocation disagreement of the simulation
Cal pers.	Ratio of persistent area in the calibration interval
Cal ch.	Ratio of changing area in the calibration interval
Val pers.	Ratio of persistent area in the validation interval
Val ch.	Ratio of changing area in the validation interval
Sim pers.	Ratio of persistent area in the simulation interval
Sim ch.	Ratio of changing area in the simulation interval
FOM	Figure of merit
Runf. (Ref)	Runfola's R value calculated for the stationarity of calibration and validation interval
Runf. (Sim)	Runfola's R value calculated for the stationarity of calibration and simulation interval
Runf. DIFF	Difference between Runf. (Ref) and Runf. (Sim)
Cal-Val an.	Difference between calibration and validation interval annual changes
Cal-Sim an.	Difference between calibration and simulation interval annual changes
OA	Overall Agreement
KIA	Kappa Index of Agreement
Cat no.	Number of categories in the actual study area

10. REFERENCES

Aabeyir, R. – Agyare, W.A. – Weir, M.J.C. – Adu-Bredu, S. (2017): Multi-Level Land Cover Change Analysis in the Forest-Savannah Transition Zone of the Kintampo Municipality, Ghana. Journal of Natural Resources and Development, 7:1-11.

Abd El-Kawy,O.R. – Rød,J.K. – Ismail,H.A. – Suliman,A.S. (2011): Land use and land cover change detection in the western Nile delta of Egypt using remote sensing data. Applied Geography 31:483-494, DOI: 10.1016/j.apgeog.2010.10.012.

Abriha, D. – Kovács, Z. – Ninsawat, S. – Bertalan, L. – Boglárka, B. – Szabó, S. (2018): Identification of roofing materials with Discriminant Function Analysis and Random Forest classifiers on pan-sharpened WorldView-2 imagery – a comparison. Hungarian Geographical Bulletin, 67(4):375-392, DOI: 10.15201/hungeobull.67.4.6.

Aldwaik, S.Z. – Onsted, J.A. – Pontius Jr, R.G. (2015): Behavior-based aggregation of land categories for temporal change analysis. International Journal of Applied Earth Observation and Geoinformation, 35, Part B:229-238, DOI: 10.1016/j.jag.2014.09.007.

Aldwaik, S.Z. – Pontius Jr, R.G. (2013): Map errors that could account for deviations from a uniform intensity of land change. International Journal of Geographical Information Science, 27:1717-1739, DOI: 10.1080/13658816.2013.787618.

Aldwaik, S.Z. – Pontius Jr, R.G. (2012): Intensity analysis to unify measurements of size and stationarity of land changes by interval, category, and transition. Landscape and Urban Planning, 106:103-114, DOI: 10.1016/j.landurbplan.2012.02.010.

Almeida, C.A. – Coutinho, A.C. – Esquerdo, J.C.D.M. – Adami, M. – Venturieri, A. – Diniz, C.G. – Dessay, N. – Durieux, L.G. – Gomes, A.R. (2016): High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. Acta Amazonica, 46(3), DOI: 10.1590/1809-4392201505504.

Alo, C.A. – Pontius Jr, R.G. (2008): Identifying Systematic Land-Cover Transitions Using Remote Sensing and GIS: The Fate of Forests inside and outside Protected Areas of Southwestern Ghana. Environ Plann B Plann Des, 35:280–295, DOI: 10.1068/b32091.

Alphan, H. – Doygun, H. – Unlukaplan, Y.I. (2009): Post-classification comparison of land cover using multitemporal Landsat and ASTER imagery: the case of Kahramanmaraş, Turkey. Environmental Monitoring and Assessment, 151:327–336.

Anderson, J.R. – Hardy, E.E. – Roach, J.T. – Witmer, E. (1976): A Land use and Land Cover Classification System for use with Remote Sensor Data, Geological Survey Professional Paper 964, Washington, DC, US Government Printing Office.

Baker, M. (2016): Statisticians issue warning over misuse of P values. Nature, 531(7593):151, DOI: 10.1038/nature.2016.19503.

Baker, W. (1989): A review of models of landscape change. Landscape Ecology, 2:111–133, DOI: 10.1007/BF00137155.

Banko, G. (1998): A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data and of Methods Including Remote Sensing Data in Forest Inventory. IIASA Interim Report. IR-98-081.

Baret, F. – Guyot, G. – Major, D.J. (1989): Crop Biomass Evaluation using Radiometric Measurements. Photogrammetria, 43(5):241–256, DOI: 10.1016/0031-8663(89)90001-X

Benenson, I. – Torrens, P.M. (2004): Geosimulation: Automata-Based Modeling of Urban Phenomena, John Wiley & Sons, Ltd., ISBN: 978-0-470-84349-9

Bielecka, E. – Jenerowicz, A. (2019): Intellectual Structure of CORINE Land Cover Research Applications in Web of Science: A Europe-Wide Review. Remote Sensing, 11(17), DOI: 10.3390/rs11172017.

Birchfield, S. (2016): Image Processing and Analysis, Cengage Learning, Mason, OH, United States, ISSN/ISBN: 9781337515627.

Bishop, Y. – Fienberg, S. – Holland, P. (1975): Discrete Multivariate Analysis — Theory and Practice, MIT Press, Cambridge, MA.

Brown, D.G. – Verburg, P.H. – Pontius Jr, R.G. – Lange, M.D. (2013): Opportunities to improve impact, integration, and evaluation of land change models. Current Opinion in Environmental Sustainability 5:452–457, DOI: 10.1016/j.cosust.2013.07.012.

Bruzzone, L. – Serpico, S.B. (1997): An Iterative Technique for the Detection of Land-Cover Transitions in Multitemporal Remote-Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 35(4):858–867.

Burai, P. – Deák, B. – Valkó, O. – Tomor, T. (2015): Classification of Herbaceous Vegetation Using Airborne Hyperspectral Imagery. Remote Sensing, 7(2):2046–2066, DOI: 10.3390/rs70202046.

Büttner, G. (2014): CORINE Land Cover and Land Cover Change Products, In: Manakos, I. – Braun, M. (Eds.): Land use and Land Cover Mapping in Europe, Springer, Dordrecht, pp. 55–74.

Büttner,G. – Feranec, J. – Jaffrain, G. – Mari, L. – Maucha, G. – Soukup, T. (2004): The CORINE Land Cover 2000 Project, In: The CORINE Land Cover 2000 Project EARSeL eProceedings , pp. 331–346.

Büttner, G. – Kosztra, B. (2017): CLC2018 Technical Guidelines. 25 October 2017, Service Contract No 3436/R0-Copernicus/EEA.56665

Cao, M. – Zhu, Y. – Quan, J. – Zhou, S. – Lü, G. – Chen, M. – Huang, M. (2019): Spatial sequential modeling and predication of global land use and land cover changes by integrating a global change assessment model and cellular automata. Earth's Future, 7:1102–1116, DOI: 10.1029/2019EF001228.

Cao, X.R. – Wan, Y.W. (1998): Algorithms for sensitivity analysis of Markov systems through potentials and perturbation realization. IEEE Transactions on Control Systems Technology, 6:482–494, DOI: 10.1109/87.701341.

Carvalho, W.D. – Mustin, K. – Hilário, R.R. – Vasconcelos, I.M. – Eilers, V. – Fearnside, P.M. (2019): Deforestation control in the Brazilian Amazon: A conservation struggle being lost as agreements and regulations are subverted and bypassed. Perspectives in Ecology and Conservation, 17:122–130, DOI: 10.1016/j.pecon.2019.06.002.

Castro, G.G.H. – Rocha, W.P. (2015): Change Analysis of Land Use and Urban Growth in the Municipalities of Culiacan and Navolato, Sinaloa, Mexico Using Statistical Techniques and GIS. Journal of Geographic Information System, 7(6):620–630.

Chan, H. – Darwiche, A. (2005): Sensitivity analysis in Markov networks, In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers, pp. 1300–1305.

Charitos, T. – van der Gaag, L.C. (2006): Sensitivity Analysis of Markovian Models, In: FLAIRS Conference.

Chen, H. – Pontius Jr, R.G. (2010): Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. Landscape Ecology, 25:1319–1331.

Clarke, K.C. – Hoppen, S. – Gaydos, L. (1997): A self-modifying cellular automata model of historical urbanization in the San Francisco bay area. Environment and Planning, 24:247–261.

Cochran, W. G. (1977). Sampling techniques, 3rd Edition, New York: John Wiley & Sons

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

Congalton, R. (1991): A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of the Environment, 37:35–46.

Congalton, R. – Green, K. (1999): Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, Third Edition, CRC Press, Boca Raton, ISSN/ISBN: 978-1-4987-7666-0.

Congalton, R. – Mead, R. (1983): A quantitative method to test for consistency and correctness of photointerpretation. Photogrammetric Engineering and Remote Sensing, 49:69–74.

Congalton, R. – Oderwald, R.G. – Mead, R. (1983): Assessing Landsat classification accuracy using discrete multivariate statistical techniques. Photogrammetric Engineering and Remote Sensing, 49:1671.

Convertino, M. – Valverde Jr, J. (2013): Portfolio Decision Analysis Framework for Value-Focused Ecosystem Management. PLoS One, 9(3):e92951.

Council of Europe (2000): European Landscape Convention. Florence, 20.X.2000.

Csorba, P. – Szabó, S. (2009): Degree of human transformation of landscapes: a case study from Hungary. Hungarian Geographical Bulletin, 58(2):91–99.

Dadashpoor, H. – Azizi, P. – Moghadasi, M. (2019): Land use change, urbanization, and change in landscape pattern in a metropolitan area. Science of The Total Environment, 655:707-719, DOI: 10.1016/j.scitotenv.2018.11.267.

de Area Leão Pereira,Eder Johnson – Silveira Ferreira,P.J. – de Santana Ribeiro,L.C. – Sabadini Carvalho,T. – de Barros Pereira,H.B. (2019): Policy in Brazil (2016–2019) threaten conservation of the Amazon rainforest. Environmental Science & Policy 100:8–12, DOI: 10.1016/j.envsci.2019.06.001.

De Rosa,M. – Knudsen,M.T. – Hermansen,J.E. (2016): A comparison of Land Use Change models: challenges and future developments. Journal of Cleaner Production 113:183–193, DOI: 10.1016/j.jclepro.2015.11.097.

Deák, B. – Valkó, O. – Török, P. – Tóthmérész, B. (2016): Factors threatening grassland specialist plants - A multi-proxy study on the vegetation of isolated grasslands. Biological Conservation, 204:255–262, DOI: 10.1016/j.biocon.2016.10.023.

Deák, M. – Telbisz, T. – Árvai, M. – Mari, L. – Horváth, F. – Kohán, B. (2017): Heterogeneous forest classification by creating mixed vegetation classes using EO-1 Hyperion. International Journal of Remote Sensing, 38(18):5215–5231, DOI: 10.1080/01431161.2017.1325529.

Deep, S. – Saklani, A. (2014): Urban sprawl modeling using cellular automata. The Egyptian Journal of Remote Sensing and Space Science, 17(2):179–187, DOI: 10.1016/j.ejrs.2014.07.001.

DeLaune,R.D. – Smith,C.J. – Patrick,W.H. – Roberts,H.H. (1987): Rejuvenated marsh and bay-bottom accretion on the rapidly subsiding coastal plain of U.S. Gulf coast: a second-order effect of the emerging Atchafalaya delta. Estuarine, Coastal and Shelf Science 25:381–389, DOI: 10.1016/0272-7714(87)90032-1.

Di Gregorio, A. – Jansen, L.J.M. (2000): Land Cover Classification System (LCCS): Classification Concepts and User Manual, FAO

Diaz-Pacheco, J. – Gutiérrez, J. (2014): Exploring the limitations of CORINE Land Cover for monitoring urban landuse dynamics in metropolitan areas. Journal of Land Use Science, 9(3):243–259, DOI: 10.1080/1747423X.2012.761736.

Dövényi, Z. (2010): Inventory of Microregions in Hungary, MTA Földrajztudományi Kutatóintézet, Budapest.

Eastman, J.R. (2012a): Idrisi Selva Manual.

Eastman, J. R. (2012b): Idrisi Selva Tutorial.

El-Hattab, M.M. (2016): Applying post classification change detection technique to monitor an Egyptian coastal zone (Abu Qir Bay). The Egyptian Journal of Remote Sensing and Space Science 19: 23–36, DOI: 10.1016/j.ejrs.2016.02.002.

Farina, A. (2013): Principles and Methods in Landscape Ecology, Springer Science & Business Media, ISSN/ISBN: 9401589844.

Feng, Y. – Wang, R. – Tong, X. – Shafizadeh-Moghadam, H. (2019): How much can temporally stationary factors explain cellular automata-based simulations of past and future urban growth? Computers, Environment and Urban Systems, 76:150–162, DOI: 10.1016/j.compenvurbsys.2019.04.010.

Feranec, J. – Soukup, T. – Hazeu, G. – Jaffrain, G. (2016): European Landscape Dynamics: CORINE Land Cover Data, CRC Press.

Field, A. (2013): Discovering Statistics using IBM SPSS Statistics, 4th Edition, SAGE Publications Ltd.

Filho, B.S.S. – Cerqueira, G.C. – Pennachin, C.L. (2002): DINAMICA - A stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. Ecological Modelling, 154(3):217–235, DOI: 10.1016/S0304-3800(02)00059-5.

Fleiss, J.L. (1981): Statistical Methods for Rates and Proportions, 2nd Edition, pp. 38–46., Wiley, New York.

Foody, G.M. (2002): Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80(1):185–201, DOI: 10.1016/S0034-4257(01)00295-4.

Forman, R.T.T. (1995): Land Mosaics: The Ecology of Landscapes and Regions, Cambridge University Press, ISSN/ISBN: 0521479800.

Forman, R.T.T. (1983): An ecology of the landscape. Bioscience, 33(535).

Fosberg, F.R. (1961): A classification of vegetation for general purposes. Tropical Ecology, 2:1–28.

Galton, F. (1892): Fingerprints, McMillan and Co., London & New York.

Gehlke, C.E. – Biehl, K. (1934): Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. Journal of the American Statistical Association, Supplement 29:169–170.

Gopal, S. – Woodcock, C. (1994): Theory and Methods for Accuracy Assessment of Thematic Maps Using Fuzzy Sets. Photogrammetric Engineering & Remote Sensing, 60:181–188.

Grigorescu, I. – Kucsicsa, G. – Popovici, E. – Mitrică, B. – Mocanu, I. – Dumitraşcu, M. (2011): Modelling land use/cover change to assess future urban sprawl in Romania. Geocarto International, DOI: 10.1080/10106049.2019.1624981.

Gulácsi, A. – Kovács, F. (2018): Drought monitoring of forest vegetation using MODIS-based normalized difference drought index in Hungary. Hungarian Geographical Bulletin, 67(1):29–42, DOI: 10.15201/hungeobull.67.1.3.

Guttenberg, A. (2002): Multidimensional Land Use Classification and How it Evolved: Reflections on a Methodological Innovation in Urban Planning. Journal of Planning History, 1(4):311–324, DOI: 10.1177/1538513202238308.

Hagen-Zanker, A. (2006): Map comparison methods that simultaneously address overlap and structure. Journal of Geographical Systems, 8:165–185.

Haines-Young, R. – Chopping, M. (1996): Quantifying landscape structure: a review of landscape indices and their application to forested landscapes. Progress in Physical Geography, 20(4):418–445, DOI: 10.1177/030913339602000403.

Halls, J.N. – Kraatz, L. (2006): Estimating error and uncertainty in change detection analyses of historical aerial photographs, In: Caetano, M. – Painho, M. (Eds.): 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, pp. 429–438.

Halmy,M.W.A. – Gessler,P.E. – Hicke,J.A. – Salem,B.B. (2015): Land use/land cover change detection and prediction in the north-western coastal desert of Egypt using Markov-CA. Applied Geography 63:101–112, DOI: 10.1016/j.apgeog.2015.06.015.

Hammer,Ø – Harper,D.A.T. – Ryan,P.D. (2001): PAST - PAlaeontological STatistics.

Heistermann, M. – Müller, C. – Ronneberger, K. (2006): Land in sight? Achievements, deficits and potentials of continental to global scale land-use modeling. Agriculture Ecosystems & Environment, 114(2–4):141–158, DOI: 10.1016/j.agee.2005.11.015.

Herold,M. – Hubald,R. – Di Gregorio,A. (2009): Translating and evaluating land cover legends using the UN Land Cover Classification System. GOFC-GOLD Report No. 43.

Hothorn, T. – Bretz, F. – Westfall, P. (2008): Simultaneous Inference in General Parametric Models. Biometrical Journal, 50(3):346–363.

Jamal, J.A. (2012): Dynamic Land use/Cover Change Modelling - Geosimulation and Multiagent-Based Modelling, Springer-Verlag Berlin Heidelberg.

Jelinski, D.E. – Wu, J. (1996): The Modifiable Areal Unit Problem and Implications for Landscape Ecology. Landscape Ecology, 11(3):129–140, DOI: 10.1007/BF02447512.

Jensen, J.R. (1996): Introductory Digital Image Processing: A Remote Sensing Perspective, Prentice Hall Series in Geographic Information Science, Upper Saddle River, New Jersey.

Johnston,R. (2002): Manipulating maps and winning elections: measuring the impact of malapportionment and gerrymandering. Political Geography 21:1–31, DOI: 10.1016/S0962-6298(01)00070-1.

Jollife, I.T. – Stephenson, D.B. (2003): A Practitioner's Guide in Atmospheric Science, Wiley, Hoboken, New Jersey.

Kerényi, A. (2015): Loess Features on Tokaj Hill, In: Lóczy, D. (Ed.):Landscapes and Landforms of Hungary, First ed., Springer International Publishing, pp. 219–225.

Kerényi, A. (2007): Tájvédelem, Pedellus Tankönyvkiadó Kft., Debrecen, Hungary, ISSN/ISBN: 963-9612-54-5.

Keshtkar, H. – Voigt, W. (2016): A spatiotemporal analysis of landscape change using an integrated Markov chain and cellular automata models. Modeling Earth Systems and Environment, 2:10.

Kim, C. (2016): Land use classification and land use change analysis using satellite images in Lombok Island, Indonesia. Journal Forest Science and Technology, 12(4):183–191, DOI: 10.1080/21580103.2016.1147498.

Kim, J. – Bang, H. (2016): Three common misuses of P values. Dent Hypotheses, 7(3):73–80., DOI: 10.4103/2155-8213.190481.

Kityuttachai, K. – Tripathi, N.K. – Tipdecho, T. – Shrestha, R. (2013): CA-Markov Analysis of Constrained Coastal Urban Growth Modeling: Hua Hin Seaside City, Thailand. Sustainability, 5:1480–1500., DOI: doi:10.3390/su5041480.

Klug,W. – Graziani,G. – Grippa,G. – Pierce,D. – Tassone,C. (1992): Evaluation of long range atmospheric transport models using environmental radioactivity data from the Chernobyl accident: The ATMES Report.

Kosztra,B. – Büttner,G. – Hazeu,G. – Arnold,S. (2019): Updated CLC illustrated nomenclature guidelines. https://land.copernicus.eu/user-corner/technical-library/corine-land-cover-nomenclature-guidelines/docs/pdf/CLC2018_Nomenclature_illustrated_guide_20190510.pdf

Kristóf, D. – Csató, É – Ritter, D. (2002): Application of High Resolution Satellite Images in Forestry and Habitat Mapping - Evaluation of Ikonos Images Through a Hungarian case study. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 34(4):602–607.

Küchler, A.W. – Zonneveld, I.S. (1988): Vegetation Mapping. Handbook of Vegetation Science, 10 ed., Kluwer Academic, Dordecht, ISSN/ISBN: 978-9401078856.

Ladányi,Z. – Blanka,V. – Deák,ÁJ. – Rakonczai,J. – Mezősi,G. (2016): Assessment of soil and vegetation changes due to hydrologically driven desalinization process in an alkaline wetland, Hungary. Ecological Complexity 25:1–10, DOI: 10.1016/j.ecocom.2015.11.002.

Lambin, E.F. – Rounsevell, M.D.A. – Geist, H.J. (2000): Are agricultural land-use models able to predict changes in land-use intensity? Agriculture, Ecosystems and Environment, 82:321–331, DOI: 10.1016/s0167-8809(00)00235-8 .

Lambin, E.F. – Geist, H.J. – Lepers, E. (2003): Dynamics of Land-use and land-cover change in tropical regions. Annual Review of Environment and Resources, 20(28): 205–241.

Lerman, D.C. – Tetreault, A. – Hovanetz, A. – Bellaci, E. – Miller, J. – Karp, H. – Mahmood, A. – Strobel, M. – Mullen, S. – Keyl, A. – Toupard, A. (2010): Applying signal-detection theory to the study of observer accuracy and bias in behavioral assessment. Journal of Applied Behavior Analysis, 43(2):195–213, DOI: 10.1901/jaba.2010.43-195.

Li, X. – Tong, Z. – Guo, E. – Luo, X. (2017): Quantifying Spatiotemporal Dynamics of Solar Radiation over the Northeast China Based on ACO-BPNN Model and Intensity Analysis. Advances in Meteorology, 9:1–15.

Lippe, M. – Hilger, T. – Sudchalee, S. – Wechpibal, N. – Jintrawet, A. – Cadisch, G. (2017): Simulating Stakeholder-Based Land-Use Change Scenarios and Their Implication on Above-Ground Carbon and Environmental Management in Northern Thailand. Land, 6(4):85, DOI: 10.3390/land6040085

Liu, Y. – Feng, Y. (2016): Simulating the Impact of Economic and Environmental Strategies on Future Urban Growth Scenarios in Ningbo, China. Sustainability, 8(10):1045.

Liu, Y. – Feng, Y. – Pontius Jr, R.G. (2014): Spatially-Explicit Simulation of Urban Growth through Self-Adaptive Genetic Algorithm and Cellular Automata Modelling. Land, 3(3):719–738.

Lóczy, D. (2010): Tájdinamika – módszertani fejlemények. In: Földrajzi tanulmányok 5, pp. 11–30., ISBN: 978 963 315 021 4

Lóczy, D. (2015): Landscapes and Landforms of Hungary. Springer International Publishing, DOI: 10.1007/978-3-319-08997-3

Madrigal-Martínez, S. – García, J.L.M. (2019): Land-change dynamics and ecosystem service trends across the central high-Andean Puna. Scientific Reports, 9, DOI: 10.1038/s41598-019-46205-9.

Mair, P. – Wilcox, R. (2019): Robust statistical methods in R using the WRS2 package. Behavior Research Methods, DOI: 10.3758/s13428-019-01246-w.

Mallinis,G. – Koutsias,N. – Arianoutsou,M. (2014): Monitoring land use/land cover transformations from 1945 to 2007 in two peri-urban mountainous areas of Athens metropolitan area, Greece. Science of The Total Environment 490: 262–278, DOI: 10.1016/j.scitotenv.2014.04.129.

Mallupattu, P.K. – Reddy, J.R.S. (2013): Analysis of Land Use/Land Cover Changes Using Remote Sensing Data and GIS at an Urban Area, Tirupati, India. The Scientific World Journal, DOI: 10.1155/2013/268623.

Martínez-Fernández, J. – Ruiz-Benito, P. – Bonet, A. – Gómez, C. (2019): Methodological variations in the production of CORINE land cover and consequences for long-term land cover change studies. The case of Spain. International Journal of Remote Sensing, 40(23):8914–8932, DOI: 10.1080/01431161.2019.1624864.

Mas, J.F. – Perez Vega, A. – Andablo Reyes, A. – Castillo Santiago, M.A. – Flamenco Sandovala, A. (2015): Assessing Modifiable Areal Unit Problem in the Analysis of Deforestation Drivers Using Remote Sensing and Census Data, In: Assessing Modifiable Areal Unit Problem in the Analysis of Deforestation Drivers Using Remote Sensing and Census DataThe International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-3/W3, 2015, ISPRS Geospatial Week 2015 La Grande Motte, France.

Mas, J.F. – Vega, E. (2012): Assessing yearly transition probability matrix for land use / land cover dynamics, In: Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Florianopolis-SC, Brazil, 10-13 July, 2012.

Mas, J. – Kolb, M. – Paegelow, M. – Olmedo, M.T.C. – Houet, T. (2014): Inductive pattern-based land use/cover change models: A comparison of four software packages. Environmental Modelling & Software 51, 94–111, DOI: 10.1016/j.envsoft.2013.09.010.

Mather, P. M. (2004): Computer Processing of Remotely-Sensed Images: An Introduction, John Wiley & Sons, ISSN/ISBN: 9780470849194.

Mayer, A.L. – Buma, B. – Davis, A. – Gagné, S.A. – Loudermilk, E.L. – Scheller, R.M. – Schmiegelow, F.K.A. – Wiersma, Y.F. – Franklin, J. (2016): How Landscape Ecology Informs Global Land-Change Science and Policy. BioScience, 66(6):458–469, DOI: 10.1093/biosci/biw035.

McGarigal, K. – Cushman, S.A. – Ene, E. (2012): FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. v4.

McGarigal, K. (2002): Landscape Pattern Metrics, DOI: 10.1002/9780470057339.val006

Mejean, R. – Paegelow, M. – Saqalli, M. – Kaced, D. (2019): Improving Business-as-Usual Scenarios in Land Change Modelling by Extending the Calibration Period and Integrating Demographic Data , In: Kyriakidis, P. – Hadjimitsis, D. – Skarlatos, D. – Mansourian, A. (Eds.): Geospatial Technologies for Local and Regional Development - Proceedings of the 22nd AGILE Conference on Geographic Information Science. 17-20 June, Limassol, Cyprus, Springer, pp. 243–260.

Memarian, H. – Balasundram, S.K. – Talib, J.B. – Teh Boon Sung, C. – Sood, A.M. – Abbaspour, K. (2012): Validation of CA-Markov for Simulation of

Land Use and Cover Change in the Langat Basin, Malaysia. Journal of Geographic Information System, 4:542–554.

Mertens, B. – Lambin, E.F. (2000): Land-cover-change trajectories in southern Cameroon. Annals of the Association of American Geographers, 90(3):467–494.

Mishra, V.N. – Rai, P.K. (2016): A remote sensing aided multi-layer perceptron-Markov chain analysis for land use and land cover change prediction in Patna district (Bihar), India. Arabian Journal of Geosciences, 9(4):1–18.

Mitsova, D. – Shuster, W. – Wang, X. (2011): A cellular automata model of land cover change to integrate urban growth with open space conservation. Landscape and Urban Planning, 99:141–153, DOI: 10.1016/j.landurbplan.2010.10.001.

Moulds, S. – Buytaert, W. – Mijic, A. (2015): An open and extensible framework for spatially explicit land use change modelling: the lulcc R package. Geoscientific Model Development, 8:3215–3229.

Näschen, K. – Diekkrüger, B. – Evers, M. – Höllermann, B. – Steinbach, S. – Thonfeld, F. (2019): The Impact of Land Use/Land Cover Change (LULCC) on Water Resources in a Tropical Catchment in Tanzania under Different Climate Change Scenarios. Sustainability, 11(24):7083, DOI: 10.3390/su11247083.

Olmedo, M.T.C. – Paegelow, M. – Mas, J. – Escobar, F. (2018): Geomatic Approaches for Modeling Land Change Scenarios, In: Cartwright, W. – Gartner, G. – Meng, L. – Peterson, M.P. (Eds.): Lecture Notes in Geoinformation and Cartography, Springer International Publishing AG, Cham, Switzerland.

Olmedo, M.T.C. – Pontius Jr, R.G. – Paegelow, M. – Mas, J. (2015): Comparison of simulation models in terms of quantity and allocation of land change. Environmental Modelling and Software, 69:214–221, DOI: 10.1016/j.envsoft.2015.03.003.

Olofsson, P. – Foody, G.M. – Herold, M. – Stehman, S.V. – Woodcock, C.E. – Wulder, M. A. (2014): Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148:42–57.

Openshaw, S. – Taylor, P.J. (1979): A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem, In: Wrigley, N. (Ed.): Statistical Applications in the Spatial Sciences, London, pp. 127–144.

Ozsahin, E. – Duru, U. – Eroglu, I. (2018): Land Use and Land Cover Changes (LULCC), a Key to Understand Soil Erosion Intensities in the Maritsa Basin. Water, 10(3):335, DOI: 10.3390/w10030335.

Paegelow, M. – Olmedo, M.T.C. – Mas, J. – Houet, T. (2014): Benchmarking of LUCC modelling tools by various validation techniques and error analysis. Cybergeo - European Journal of Geography. Systèmes, Modélisation, Géostatistiques, DOI: 10.4000/cybergeo.26610.

Paegelow, M. – Olmedo, M.T.C. (2005): Possibilities and limits of prospective GIS land cover modelling—a compared case study: Garrotxes (France) and Alta Alpujarra Granadina (Spain). International Journal of Geographical Information Science, 19(6):697–722.

Perica, S. – Foufoula-Georgiou, E. (1996): Model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. Journal of Geophysical Research: Atmospheres, 101:26347–26361.

Pham, M. (2005): Land-use Change in the Northwestern Uplands of Vietnam: Empirical Evidence from Spatial Econometric Models and Geo-Referenced Analyses and Policy Implications for Sustainable Rural Development, Cuvillier Verlag.

Piepho, H. (2004): An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons. Journal of Computational and Graphical Statistics, 13(2):456–466.

Pontius Jr, R.G. (2000): Quantification Error versus Location Error in Comparison of Categorical Maps, Photogrammetric Engineering and Remote Sensing, 66:1011–1016.

Pontius Jr, R.G. – Batchu, K. (2003): Using the Relative Operating Characteristic to Quantify Certainty in Prediction of Location of Land Cover Change in India. Transactions in GIS, 7:467–484, DOI: 10.1111/1467-9671.00159.

Pontius Jr, R.G. – Boersma, W. – Castella, J. – Clarke, K. – de Nijs, T. – Dietzel, C. – Duan, Z. – Fotsing, E. – Goldstein, N. – Kok, K. – Koomen, E. – Lippitt, C.D. – McConnell, W. – Sood, A.M. – Pijanowski, B. – Pithadia, S. – Sweeney, S. – Trung, T.N. – Veldkamp, A.T. – Verburg, P.H. (2008): Comparing the Input, Output, and Validation Maps for several Models of Land Change. The Annals of Regional Science 42:11–37.

Pontius Jr, R.G. – Malizia, N.R. (2004): Effect of category aggregation on map comparison, In: Egenhofer, M.J. – Freksa, C. – Miller, H.J. (Eds.): Geographic

Information Science. GIScience 2004. Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 251–268.

Pontius Jr, R.G. – Millones, M. (2011): Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32:4407–4429, DOI: 10.1080/01431161.2011.552923.

Pontius Jr, R.G. – Parmentier, B. (2014): Recommendations for using the relative operating characteristic (ROC). Landscape Ecology, 29:367–382.

Pontius Jr, R.G. – Peethambaram, S. – Castella, J. (2011): Comparison of Three Maps at Multiple Resolutions: A Case Study of Land Change Simulation in Cho Don District, Vietnam. Annals of the Association of American Geographers, 101:45–62, DOI: 10.1080/00045608.2010.517742.

Pontius Jr, R.G. – Si, K. (2014): The total operating characteristic to measure diagnostic ability for multiple thresholds. International Journal of Geographical Information Science, 28(3):570–583.

Pontius Jr, R.G. – Gao, Y. – Giner, N.M. – Kohyama, T. – Osaki, M. – Hirose, K. (2013): Design and Interpretation of Intensity Analysis Illustrated by Land Change in Central Kalimantan, Indonesia. Land, 2:351–369, DOI: 10.3390/land2030351.

Pontius Jr, R.G. – Neeti, N. (2010): Uncertainty in the difference between maps of future land change scenarios. Sustainability Science, 5:39–50, DOI: 10.1007/s11625-009-0095-z.

Pontius Jr, R.G. – Shusas, E. – McEachern, M. (2004): Detecting important categorical land changes while accounting for persistence. Agriculture, Ecosystems and Environment, 101:251–268, DOI: 10.1016/j.agee.2003.09.008.

Pontius Jr, R.G. – Walker, R. – Yao-Kumah, R. – Arima, E. – Aldrich, S. – Caldas, M. – Vergara, D. (2007): Accuracy Assessment for a Simulation Model of Amazonian Deforestation. Annals of the Association of American Geographers, 97(4):677–695, DOI: 10.1111/j.1467-8306.2007.00577.x.

Popovici, E.A. – Kucsicsa, G. – Bălteanu, D. – Grigorescu, I. – Mitrica, B. – Dumitrașcu, M. – Damian, N. (2018): Past and future land use/land cover flows related to agricultural lands in Romania. An assessment using CLUE-S model and Corine Land Cover database. Carpathian Journal of Earth and Environmental Sciences, 13, DOI: 10.26471/cjees/2018/013/052.

Quan,B. – Tao, G. – Liu, P.L. – Ren, H.G. (2017): Spatiotemporal patterns and its instability of land use change in five Chinese node cities of the belt and road, In: Spatiotemporal patterns and its instability of land use change in five Chinese node cities of the belt and road ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information .

R Core Team (2020): R: A language and environment for statistical computing . R Foundation for Statistical Computing,Vienna, Austria .

Raphael John, L. – Hambati, H. – Ato Armah, F. (2014): An Intensity Analysis of land-use and land-cover change in Karatu District, Tanzania: community perceptions and coping strategies. African Geographical Review, 33:150–173, DOI: 10.1080/19376812.2013.838660.

Renooij, S. (2010): Efficient sensitivity analysis in hidden Markov models, In: Myllymaki, P. – Roos, T. – Jaakkola, T. (Eds.): Proceedings of the Fifth European Workshop on Probabilistic Graphical Models, HIIT Publications 2010-2, Helsinki, Finland, pp. 241–248.

Risser, P.G. – Karr, J.R. – Forman, R.T.T. (1984): Landscape Ecology: Directions and Approaches, Illinois Natural History Survey, Champaign.

Rocha, W.P. – Barraza, G.C. – Castro, G.G.H. – Armenta, S.A.M. – González, J.C.B. – Lozoya, H.E.R. (2017): Spatial-Temporal Analysis of Territorial Transformations in the State of Sinaloa Mexico Using Geographic Information Systems. Agricultural Sciences, 8:171–182.

Rosenfield, G.H. – Fitzpatrick-Lins, K. (1986): Coefficient of agreement as a measure of Thematic Classification accuracy. Photogrammetric Engineering and Remote Sensing, 48:131–137.

Ruelland, D. – Dezetter, A. – Puech, C. – Ardoin-Bardin, S. (2008): Long-term monitoring of land cover changes based on Landsat imagery to improve hydrological modelling in West Africa. International Journal of Remote Sensing, 29(12):3533–3551, DOI: 10.1080/01431160701758699.

Runfola, D. – Pontius Jr, R.G. (2013): Measuring the temporal instability of land change using the Flow matrix. International Journal of Geographical Information Science, 27(9):1696–1716, DOI: 10.1080/13658816.2013.792344.

Saltelli, A. – Tarantola, S. – Campolongo, F. – Ratto, M. (2004): Sensitivity Analysis in Practice : A Guide to Assessing Scientific Models, John Wiley & Sons, Hoboken, NJ.

Sang, X. – Guo, Q. – Wu, X. – Fu, Y. – Xie, T. – He, C. – Zang, J. (2019): Intensity and Stationarity Analysis of Land Use Change Based on CART Algorithm. Scientific Reports, 9, DOI: 10.1038/s41598-019-48586-3.

Schuwirth, N. – Borgwardt, F. – Domisch, S. – Friedrichs, M. – Kattwinkel, M. – Kneis, D. – Kuemmerlen, M. – Langhans, S.D. – Martínez-López, J. – Vermeiren,P. (2019): How to make ecological models useful for environmental management. Ecological Modelling 411, 108784, DOI: 10.1016/j.ecolmodel.2019.108784.

Schweitzer, P.J. (1968): Perturbation theory and finite Markov chains. Journal of Applied Probability, 5(2):401–413., DOI: 10.2307/3212261.

Sharma, P. – Diwakar, M. – Lal, N. (2013): Edge Detection using Moore Neighborhood. International Journal of Computer Applications, 61(3):26–30.

Silva, E.A. – Clarke, K.C. (2002): Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. Computers, Environment and Urban Systems, 26:525–552, DOI: 10.1016/S0198-9715(01)00014-X.

Singh, S.K. – Mustak, S. – Srivastava, P.K. – Szabó, S. – Islam, T. (2015): Predicting Spatial and Decadal LULC Changes Through Cellular Automata Markov Chain Models Using Earth Observation Datasets and Geo-information. Environmental Processes, 2:61–78, DOI: 10.1007/s40710-015-0062-x.

Sipper, M. (1997): Evolving uniform and non-uniform cellular automata networks, In: Stauffer, D. (Ed.): Annual Reviews of Computational Physics, 5 ed., World Scientific, Singapore, pp. 243–285.

Stathopoulou, M.I. – Cartalis, C. – Petrakis, M. (2007): Integrating Corine Land Cover data and Landsat TM for surface emissivity definition: Application to the urban area of Athens, Greece. International Journal of Remote Sensing 28(15):3291–3304, 28(15),:3291-3304, DOI: 10.1080/01431160600993421.

Subasinghe, S. – Estoque, R.C. – Murayama, Y. (2016): Spatiotemporal Analysis of Urban Growth Using GIS and Remote Sensing: A Case Study of the Colombo Metropolitan Area, Sri Lanka. International Journal of Geo-Information, 5(12):197.

Subedi, P. – Subedi, K. – Thapa, B. (2013): Application of a Hybrid Cellular Automaton – Markov (CA-Markov) Model in Land-Use Change Prediction: A Case Study of Saddle Creek Drainage Basin, Florida. Applied Ecology and Environmental Sciences, 1:126–132.

Szabó, S. – Bertalan, L. – Kerekes, Á – Novák, T.J. (2016): Possibilities of land use change analysis in a mountainous rural area: a methodological approach. International Journal of Geographical Information Science, 30(4):708–726, DOI: 10.1080/13658816.2015.1092546.

Szilassi, P. (2017): Land cover variability and the changes of land cover pattern in landscape units of Hungary. Tájökológiai Lapok, 15(2):131–138.

Szucs, D. – Ioannidis, J.P.A. (2017): When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. Frontiers in Human Neuroscience, 11:390, DOI: 10.3389/fnhum.2017.00390.

Tajbakhsh, S.M. – Memarian, H. – Moradi, K. – Aghakhani Afshar, A.H. (2018): Performance comparison of land change modeling techniques for land use projection of arid watersheds. Global Journal of Environmental Science and Management, 4(3):263–280, DOI: 10.22034/gjesm.2018.03.002.

Takada, T. – Miyamoto, A. – Hasegawa, S.F. (2010): Derivation of a yearly transition probability matrix for land-use dynamics and its applications. Landscape Ecology, 25:561–572, DOI: 10.1007/s10980-009-9433-x.

Teixeira, Z. – Marques, J.C. – Pontius Jr, R.G. (2016): Evidence for deviations from uniform changes in a Portuguese watershed illustrated by CORINE maps: An Intensity Analysis approach. Ecological Indicators, 66:382–390, DOI: 10.1016/j.ecolind.2016.01.018.

Túri, Z. (2010): Studying landscape pattern in Great Hungarian Plain model areas, In: Rahmonov, O. (Ed.):Anthropogenic Aspects of Landscape Transformations 6, University of Silesia, Sosnowiec–Będzin, pp. 109–115.

Turner, M.G. – Gardner, R.H. (2015): Landscape Ecology in Theory and Practice, Springer-Verlag, New York, DOI: 10.1007/978-1-4939-2794-4 Turner, M.G. – Constanza, R. – Sklar, F. (1989): Methods to evaluate the performance of spatial simulation models. Ecological Modelling, 48(1–2):1–18, DOI: 10.1016/0304-3800(89)90057-4 .

Tye, R.S. – Coleman, J.M. (1989): Evolution of Atchafalaya lacustrine deltas, south-central Louisiana. Sedimentary Geology 65(1–2):95–112 , DOI: 10.1016/0037-0738(89)90008-0.

U.S. Geological Survey. (2016): Landsat—Earth observation satellites (ver. 1.2, April 2020): U.S. Geological Survey Fact Sheet.

UNESCO (1973): International Classification and Mapping of Vegetation.

Urban, D.L. – O'Neill, R.V. – Shugart Jr., H.H. (1987): Landscape Ecology. A Hierarquical Perspective Can Help Scientists Understand Spatial Patterns. Bioscience, 37:119–127, DOI: 10.2307/1310366.

Van Dessel, W. – Szilassi, P. – Van Rompaey, A. (2011): Sensitivity analysis of logistic regression parameterization for land use and land cover probability estimation. International Journal of Geographical Information Science, 25:489–508, DOI: 10.1080/13658810903194256.

van Schrojenstein Lantman, J. – Verburg, P.H. – Bregt, A. – Geertman, S. (2011): Core Principles and Concepts in Land-Use Modelling: A Literature Review , In: Koomen, E. – Borsboom van Beurden, J. (Eds.): Land-use Modelling in Planning Practice, Springer Science+Business Media, pp. 35–57.

van Soesbergen, A. (2016): A Review of Land use Change Models, United Nations Environment Programme, ISSN/ISBN: 978-92-807-3575-8.

van Vliet, J. (2009): Assessing the Accuracy of Changes in Spatial Explicit Land Use Change Models, In: Assessing the Accuracy of Changes in Spatial Explicit Land Use Change Models12th AGILE International Conference on Geographic Information ScienceLeibniz Universität, Hannover, Germany, pp. 1–9.

Varga, O.G. – Pontius Jr, R.G. – Singh, S.K. – Szabó, S. (2019): Intensity Analysis and the Figure of Merit's components for assessment of a Cellular Automata – Markov simulation model. Ecological Indicators 101:933-942, DOI: 10.1016/j.ecolind.2019.01.057.

Varga, O.G. – Pontius Jr, R.G. – Szabó, Z. – Szabó, S. (2020): Effects of Category Aggregation on Land Change Simulation Based on Corine Land Cover Data. Remote Sensing, 12(8):1314, DOI: 10.3390/rs12081314.

Varga, O.G. – Szabó, S. – Túri, Z. (2014): Efficiency assessments of GEOBIA in land cover analysis, NE Hungary. Bulletin of Environmental and Scientific Research, 3(4):1–9.

Verburg, P.H. – de Koning, G. H. J. – Kok, K. – Veldkamp, A. – Bouma, J. (1999): A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. Ecological Modelling, 116(1):45–61, DOI: 10.1016/s0304-3800(98)00156-2.

Verburg, P.H. – Schot, P. – Dijs, M. – Veldkamp, A. (2004a): Land-use change modeling: Current practice and research priorities. GeoJournal, 61(4):309–324, DOI: 10.1007/s10708-004-4946-y.

Verburg, P.H. – van Eck, J.R. – de Nijs, T. – Dijst, M. – Schot, P. (2004b): Determinants of Land-use Change Patterns in the Netherlands. Environment and Planning B: Planning and Design, 31(1):125–150, DOI: 10.1068/b307.

Verőné Wojtaszek, M. (2010): Földhasználati tervezés és monitoring 4., Földhasználati monitoring távérzékeléssel, Nyugat-magyarországi Egyetem.

Viana, C.M. – Girão, I. – Rocha, J. (2019): Long-Term Satellite Image Time-Series for Land Use/Land Cover Change Detection Using Refined Open Source Data in a Rural Region. Remote Sensing, 11:1104, DOI: 10.3390/rs11091104.

Viera, A.J. – Garrett, J.M. (2005): Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine, 37:350–363.

Visser, H. – de Nijs, T. (2006): The Map Comparison Kit. Environmental Modeling & Software, 21:346–358, DOI: 10.1016/j.envsoft.2004.11.013.

Wang, M. – Sun, X. – Fan, Z. – Yue, T. (2019): Investigation of Future Land Use Change and Implications for Cropland Quality: The Case of China. Sustainability, 11:3327, DOI: 10.3390/su11123327.

Warrens, M.J. (2015a): Properties of the quantity disagreement and the allocation disagreement. International Journal of Remote Sensing, 36:1439–1446, DOI: 10.1080/01431161.2015.1011794.

Warrens, M.J. (2015b): Relative quantity and allocation disagreement measures for category-level accuracy assessment. International Journal of Remote Sensing, 36:5959–5969, DOI: 10.1080/01431161.2015.1110265.

Wickham, H. (2016): Ggplot2: Elegant Graphics for Data Analysis. , Springer-Verlag New York.

Wong, D.W.S. (2004): The Modifiable Areal Unit Problem (MAUP), In: Janelle, D.G. – Warf, B. – Hansen, K. (Eds.): WorldMinds: Geographical Perspectives on 100 Problems, Springer, Dordrecht, ISSN/ISBN: 978-1-4020-2352-1.

Woodcock, C.E. – Gopal, S. (2010): Fuzzy set theory and thematic maps: Accuracy assessment and area estimation. International Journal of Geographical Information Science, 14(2):153–172, DOI: 10.1080/136588100240895.

Yague, J. – Garcia, P. (2004): Approaching Corine Land Cover over Castilla and Leon (central Spain) with a multitemporal NOAA-AVHRR NDVI MVC series, In: Smits, P.C. – Lorenzo, B. (Eds.): Approaching Corine Land Cover Over Castilla and Leon (Central Spain) with a Multitemporal NOAA-AVHRR NDVI MVC Series, pp. 314–321.

Yang, Y. – Liu, Y. – Xu, D. – Zhang, S. (2017): Use of intensity analysis to measure land use changes from 1932 to 2005 in Zhenlai County, Northeast China. Chinese Geographical Science, 27(3):441–455.

Yang, X. – Zheng, X. – Chen, R. (2014): A land use change model: Integrating landscape pattern indexes and Markov-CA. Ecological Modelling 283:1–7, DOI: 10.1016/j.ecolmodel.2014.03.011.

Zhu, Z. – Woodcock, C.E. (2014): Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment 144:152–171, DOI: 10.1016/j.rse.2014.01.011.