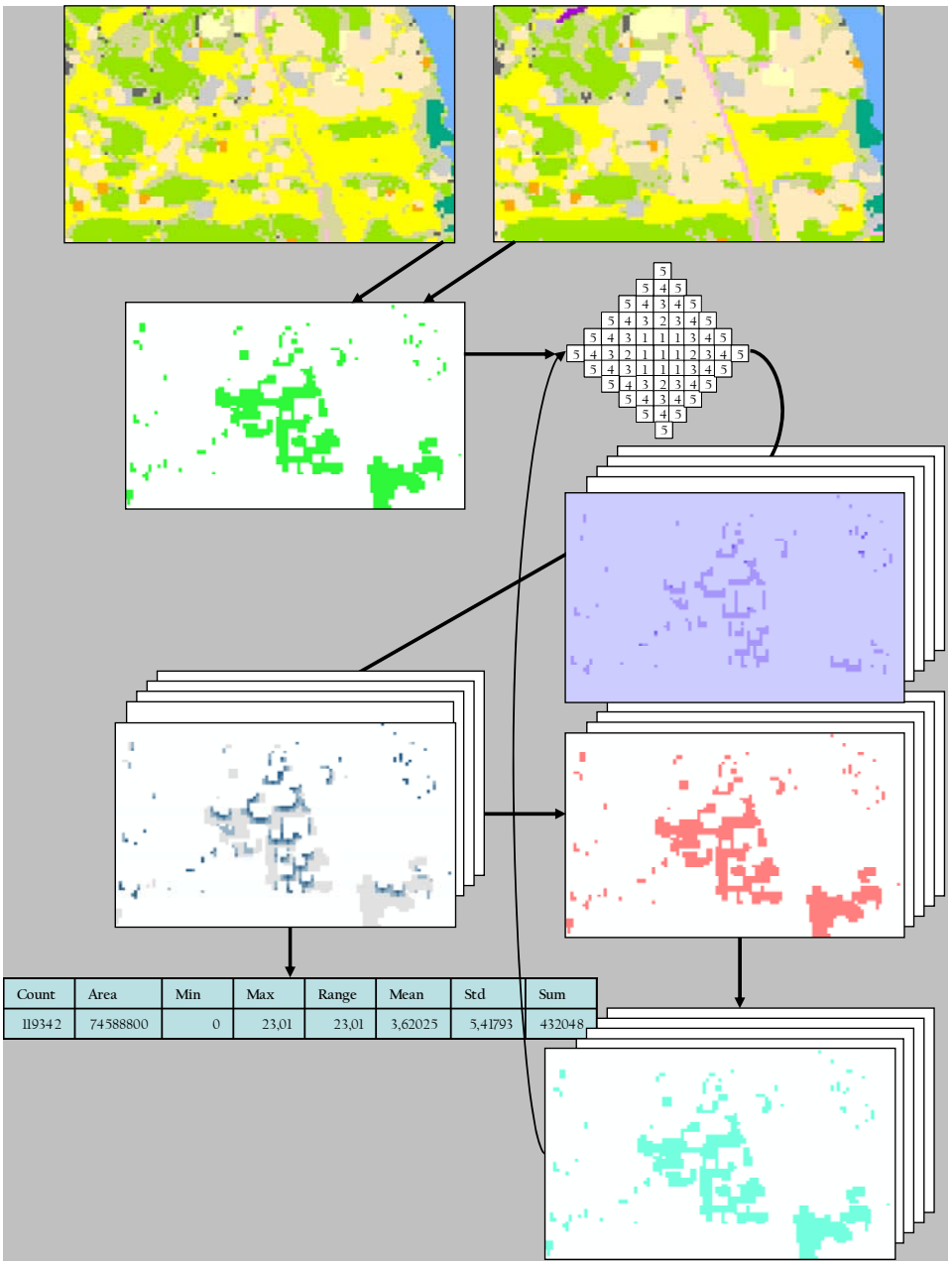
Towards a Methodology for Quantifying Neighbourhood Interaction



Eighth Semester Project

Geoinformation Technology and Management

Aalborg University in Copenhagen

Spring 2008

Aalborg University Copenhagen
Lautrupvang 15, 2750 Ballerup,
Denmark

Secretary: Dortea Koldborg Jepsen
Phone: 9940 2468
dkj@imi.aau.dk

Title: Towards a Methodology for Quantifying Neighbourhood Interaction

Semester: 8

Semester theme: Geographic Systems for Participation, Collaboration and Analysis

Supervisor: Henning Sten Hansen

Lena Hallin-Pihlatie

Abstract: The development of cellular automata-based land-use models, the increasing availability of good-quality geographical land-use data and the growing need for integrating spatial scenarios in decision-making, are surrounded with technological opportunities and challenges. Present land-use models have been criticized for not being based on empirical studies regarding land-use dynamics and actual land-use change. In this project, the role and components of neighbourhood interaction is explored from a land-use modelling perspective. The conceptual understanding of neighbourhood interaction is turned into practical knowledge by developing a programme for empirical quantification of neighbourhood interaction using existing land-use data and GIS.

Foreword

This project was completed as part of the eighth semester of the Master of Science Program in Geoinformation Technology and Management at Aalborg University. This semester's theme, Geographic Systems for Participation, Collaboration and Analysis, were addressed within the width of this project's material. While looking deeper into neighbourhood interaction - one crucial factor used in land-use models to allocate the location of land-use change - practical knowledge was built from the conceptual lessons relevant for this semester's theme.

I want to thank my colleagues at the Geoinformatics and Land Use Division at the Finnish Environment Institute for being helpful and providing me with data and valuable data documentation. I also want to thank the staff of the Department of System Analysis, at the National Environmental Research Institute of Aarhus University, for inspiring me to continue my studies. Finally, I want to thank my family, especially Mikko, for their support during this project.

Table of Contents

1. Introduction

We don't exactly know what consequences world-spanning phenomena, such as globalisation and climate change will have on land-use. To promote environmental sustainability and to minimize the risk for future environmental catastrophes, we however need to prepare ourselves and tackle the future already today. Land-use planning is a future-oriented activity, where the "building bricks" for the future are being laid and therefore it is very relevant to integrate projections and spatial scenarios the planning process.

There are a lot of political initiatives that directly work towards an environmentally more sustainable future. Future-oriented goals - strategies and legislative - on a local (e.g. Agenda 21) and an international (e.g. Kyoto, the Biodiversity convention) and European level (e.g. the Flood Directive, Water Framework Directive), are also very much being taken into account the land-use planning process, since it has been recognized that a sustainable future is also very much related to a sustainable land-use pattern. Common for many of these policies is their spatial dimension – i.e. they want to improve the state of the coast, watersheds, seas and halt the decline of biodiversity for instance in urban areas. In land-use planning these multi-scale and multi-target goals meet.

Projections and scenarios can help decision makers, planners and participating citizens to take future uncertainties, challenges, and alternative possibilities and e.g. climate related threats into account in their area of interest. In western countries, such as Finland and Denmark, it is easy to get hold of statistical projections on a national or even at the municipal level regarding population increase and economy. These are very useful for dimensioning land-use plans. However, land-use planning generally includes a more local dimension – for instance, as part of the planning process, plan maps are made for a region or a municipality. To actors in land-use planning, spatial scenarios, for visualising the development alternatives of an area on a map would be very useful. However, today spatial scenarios are not generally being used as part of the decision making process, even though the technology, such as land-use models and participatory Internet-based techniques, for integrating them is do exist.

Integration of good-quality data is regarded as the key for successful implementation of spatial policies. Data is equally important for the development of scientifically justified spatial scenarios. Problems have been found regarding data availability, quality and when combining data from different sources and scales. In this light, the international geographical community and the European Union (EU) saw the need to promote accessibility and usability of spatial data. This has resulted in initiative to create so called Spatial Data Infrastructures on different levels in society. The EU have ratified the so-called INSPIRE Directive to improve the situation regarding geographical information (INSPIRE 2008).

This project is addressed to

1. the increasing availability of good-quality, spatial data relevant to land-use modelling,
2. the growing and challenging need and technical possibilities to integrate land-use models in the decision making process and
3. the challenges that are yet to be thoroughly solved in the field of urban land-use modelling, and most specifically it is related to the methodology for justifying neighbourhood rules, which will be proven to be a relevant subject for improving present land-use models that can simulate alternatives of the future land-use pattern.

After this introductory chapter follows a *Land-use dynamics and modelling* chapter, which describes relevant concepts and the state-of-the-art technique for urban land-use modelling. In the third *Problem statement* chapter the problem statement and the research questions related to it are defined and in the *Methods* chapter the methods to answer these questions are described. Next, in the *Theoretical components of Neighbourhood Interaction* chapter the concepts and GIS-related techniques behind neighbourhood interaction will be explained. The *Software Development* chapter again presents the conceptual model developed and the tool that was programmed according to it. This is followed by the *Results and Discussion* chapter, where the findings of this project are brought forward. We

conclude with the *Conclusion and Prospects* chapter, where the findings are summed up and a few prospects of the future are pointed out.

2. Land-use dynamics and modelling

Changes in land cover and land use are among the most important human induced changes that have an impact on earth. Land cover and land use have direct and indirect affects on biodiversity, climate change and global warming. Additionally, changes in land-use can influence the vulnerability of places to climatic, economic or socio-economic perturbations. For the consequences on local and regional level the actual spatial pattern of land-use may be of high importance. It is not only a question about what land-uses are changing, but also where they are changing. For example new residential area may be vulnerable for climatic changes in one area, but not in another. You can therefore say that land-use patterns play a key role in the environmental stability of our future earth (Verburg et al. 2004a: 668).

Cities and their complexity has been the focus of many geographers and urban scientists for several decades. The first and most famous attempts to model the urban morphological structure or the land-use pattern of cities was the sector model by Burgess in 1925, the sector model made by Hoyt in 1939 and the multiple nuclei model developed by Harris and Ullman in 1945 (Wikipedia 2008). Common for these models are that they explain the spatial configuration of land-use classes within and proximate to a city with socio-economic so called spatial externalities. These spatial externalities can be defined as the radiated priced or unprized effects of one land-use on another. Hoyt's sector model positions the residential areas of the low income households adjacent to railways and the industry sector, where negative externalities like noise and pollution make them less attractive for living. In present time, the Not-In-My-Backyard (NIMBY) and Locally-Unwanted-Land-Use (LULU) can be regarded as part of the same phenomena (Hagoort et al. 2008:42). Common for the old-time models is that they are descriptive and highly static in their nature, not directly taking the time factor into account.

Today researchers from different disciplines address land-use change issues to better understand the causes and consequences of land-use change (Verburg et al. 2004a:668). The land-use pattern that form our urban landscape of today, has according to recent research emerged from a complex interaction between the human and the natural environment (Verburg et al. 2004b:125), a phenomena that is often referred to as the land-use dynamics. There is no overall theory on the so called driving forces or driving factors that are the triggers of land-use change. It is however generally recognized that these determinants of land-use change are diverse; they act on different levels and have a cross-disciplinary nature. On micro level the determinants include biophysical conditions, economic factors, social factors, spatial interaction and neighbourhood characteristics, and spatial policies; while it on macro level is the population growth, migration, and economic change among other things that affect the pattern of land-use change (Verburg et al. 2004b:126). It is the latter ones that influence the magnitude and extent of land-use change (Verburg et al. 2004b:668), while the micro level determinants affect changes in land-use pattern on a local and regional scale.

All processes affecting the land-use dynamics are intertwined and work both from a top down and bottom up perspective. A typical example of a top down derived process is a land-use plan that steer the expansion of urban land use in a particular area in a particular direction. However you can also regard global phenomena as top down steered processes, because of its effects on land-use on the local scale. For example the allocation of jobs to countries, where costs of labour is lower, have consequences on the local level. A small town that have lost its main industry will have it hard to attract new inhabitants, and subsequently the expansion of residential will decline. On the other hand, the land-use dynamics also works the other way: changes on the local scale can have a regional or even a national impact. The construction of a new motorway or a major bridge can trigger off new building activities. A good example of this case is the areas proximate to the Øresund's bridge, where huge changes in land-use have taken place after the bridge was constructed. The Øresund's bridge has strengthened the region by literally building a bridge between people, universities and innovation-based economies. Small scale patterns give in the long run give birth to larger scale patterns, such as a certain spatial configuration of residential, industrial and other urban land-use.

Land use models have been used for several decades to predict the location and extent the of land-use change. Generally the ways to use science in these land-use models has changed. While 50 years ago, the view was that one can simplify reality, to make an absolute model of the phenomena taking place within our complex cities, today urban land-use modelling is less oriented to aid in understanding and to structure debate, not to make absolute predictions. This can be seen as a shift to the use of so called "What if scenarios", which dominate present model-building (Batty and Torrens 2001:3). However, it does not make it less valuable to base these models on real-life information.

Recent research has also aimed at exploring and predicting the extent and location of future land-use change (Verburg et al. 2004a:668) using dynamic land-use models. Many methods have been developed in attempting to model land-use change. Of the models developed, including statistical and transition probability models, optimisation models and linear programming, dynamic simulation models, agent-based models, the Cellular Automata (CA) models have by many been regarded as most important for land-use modelling purposes. We have already presented the determinants of urban land-use as being complex. There are always uncertainties about the outcome of processes of change that originate from bottom up and this is often referred to as complexity (Batty 2005:preface). During recent years, focus has been changed from top down to bottom up approaches to explain and to model this complexity (Batty 2005:preface) and CA has been found a very useful tool. CA models stand out for example in their ability to simulate existing urban forms using a bottom-up based perspective of self-organisation (Hagoort et al. 2008:42-43).

Cellular Automata date back to the beginnings of digital computation (Batty 2005:74). Standard CA is based on four characteristics: cells, state, neighbourhood and transition rules. First there is a regular lattice of identical cells. Second, each cell may only have one cell state at a time. These discrete states define the outcomes of the system. Third, the state of any cell depends on the states and configurations of other cells in the neighbourhood of that cell. In a strict, traditional CA the neighbourhood cells are those that are immediately adjacent with the cell in question. Finally there are transition rules that drive changes of state in each cell. The transition rules are a function that describe

what exists or what is happening in the cell's neighbourhood (Batty 2005:68). In short, the state of a cell is determined by transition rules that according to the prevailing state of a cell's neighbourhood at a time t, returns an outcome cell state at time (t+1) (O'Sullivan & Torrens 2000:2).

CA has so far been applied to the simulation of a wide range of urban phenomena (O'Sullivan & Torrens 2000:1). One of the most useful applications of cellular automata, at least from the land-use planning point of view, is their use in simulation of urban growth at local and regional level (Barredo et al. 2003:145). Several approaches have been proposed for modifying standard CA making them more suitable for urban simulation (Barredo et al. 2003:145). Common for the urban implementations of CA is that they somewhat differ from the original CA structure (O'Sullivan & Torrens 2000:1). In applications applied in urban geography, the cells are usually represented by pixels in a two-dimensional grid-based lattice, which also is characteristic of the cellular presentation of data in raster-based GIS (Engelen et al. 2002:9). The cell states are usually defined as various categories of land-use classes, such as residential, service and industrial land use classes. These are often divided into three categories. The first category being the active classes expands as a result of external driving forces. Residential cells are an example of these. The second category is the passive classes, on which expense the active classes can expand. Examples of these are agricultural land and forest. The third category are the static classes, that stay unchanged and only affects the other classes by different push and pull effects. Infrastructure and watersheds belong to this category. (Hansen 2007).

What happens within the neighbourhood is of particular importance in the context of modelling urban growth. In comparison with the strict CA approach where all actions of interest are local i.e. taking place in the immediate vicinity of the cell, urban geographical phenomena often include actions at distance and the neighbourhood is therefore defined accordingly (Batty 2005:73). For instance the placement of a new shopping centre will influence a larger area than the closest cells and this also needs to be taken into account in urban CA applications. Therefore it is regarded important to quantify the extent and type of this land-use interaction within a neighbourhood and to integrate this

information in the transition rules of urban CA. For each land-use function, the transition rule is a weighed sum of distance functions calculated relative to other land-ust functions and features. These transition rules represent the competition of human activities within the urban area (Engelen et al. 2002: 6).

The basic idea of CA is to simulate global patterns and structures from local elements. Many urban phenomena have this kind of bottom-up structure: air pollution, neighbourhood upgrading and decline, and so on, making traditional CA applicable. However, socio-economic systems, such as cities, are also highly shaped by interaction processes that take place at various geographical scales, some of which are very local and within the reach of the extended neighbourhood, and some of which are beyond the reach of cellular automata. In order to incorporate the dynamics caused by these long range macro level processes and phenomena beyond the reach of the neighbourhood, you can link cellular automata models to macro level drivers. Socio-economic and other macro-level models can be linked to force growth in a certain way acting, as a constraint, upon the local CA-model (Engelen et al. 2002:8). For instance, you can use the municipal population projections to set the growth demand of each municipality and let it steer the allocation of new residential cells on a regional level. Land-use plans and transport infrastructure are also obvious examples macro scale constraints. These kinds of global structures need be imposed as external global constraints on local interaction in urban land-use models (O'Sullivan & Torrens 2000:2). That is why urban implementations of CA usually are constrained.

White and Engelen (Engelen et al. 2002:8) simulated the development of land use in year 1966 in a hypothetical city with the same kinds of dimensions as the medium sized US cities of Cincinnati, Houston, Milwaukee and Atlanta (figure 1). This model used trend lines as a growth constraint and broke with the standard CA notion that neighbourhoods should be local. The neighbourhood chosen was based on a circular neighbourhood of radius six cells containing some 113 cells in total and being divided in 19 distinct distance bands. However, no sensitivity testing was made regarding the choice of the neighbourhood size. Most certainly the use of another neighbourhood would have given another result. (Batty and Torrens 2001:26-27).

Later, other models such as MOLAND (Engelen et al. 2002:23-25) and LUCIA (Hansen 2007) have been applied using constrained CA methods and extended neighbourhoods. Common for these models is that they calculate the potential for each cell to change or the so-called transitional potential. The transitional potential is calculated based on weighed factors, such as the cumulative neighbourhood interaction, the accessibility of a cell and its suitability for building activities and moreover based on the binary constraints imposed by zoning. Once the transitional potential for each cell has been calculated, the transition rule is to change each cell to the state for which it has the highest potential – with the constraint that the number of cells in each cell state must be equal to the number demanded at that iteration. The type and rate of the macro scale demands that impose changes to land-use at local scale are modelled externally of the CA model using for instance socio-economic data (Engelen et al. 2002:6-7, Hansen 2007).

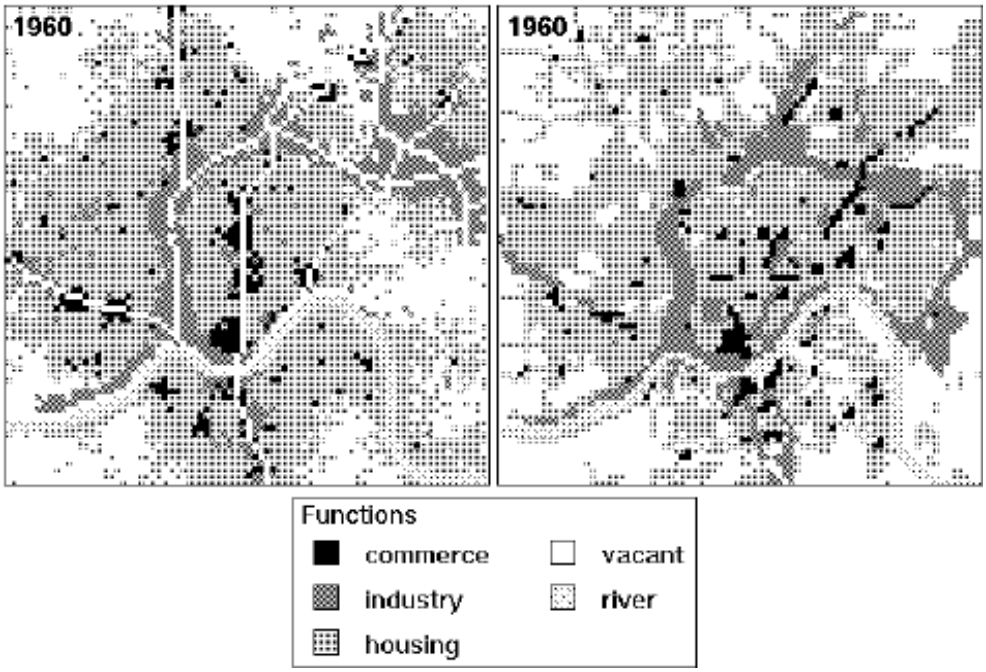


Figure 1. The actual land-use in Cincinnati in year 1960 (left) and the land-use simulated by a constrained urban CA model (right). The model was made by White and Engelen in 1993 (Engelen)

It is generally recognized that in order to make good land-use models, and to simulate alternative land-use patterns of the future, the present processes causing land-use need to be sufficiently known. Presently this is not the case. Land-use models are being

criticized for being based on too loose a scientific foundation (Malcewski 2000:21). For instant the choice of transition rules and the used neighbourhoods are usually not based on empirical studies. The field of urban CA is also been criticized for being too technology driven so far, instead of being based on actual urban dynamics (Verburg et al. 2004a:669, Geertman et al. 2007:549). Today, the technology for making simulations is out there. For instance, we already know how to make urban CA models. What is then needed? Practically, we need to quantify the drivers of land-use change, so that we can base our land-use models on empirical studies on where (location) the land-use changes are taking place and find out the extent (quantity) of the changes. In this way the "What if scenarios" can, to a certain extent be scientifically justified, which argue for their wider use in decision-making processes.

When the processes of land-use change are so diverse and complex, how shall we know what to focus on to find out what we need? A study made by Verburg et al. (2004b) in the Netherlands indicate that the historical land-use pattern can be explained by the conditions of soil and land form (e.g. height and distance to watershed), in particular regarding the location of agricultural land and of forest and nature area. Recent land-use changes were no longer determined by biophysical factors of a location, but instead mainly by accessibility, spatial policies and neighbourhood interactions (Verburg et al. 2004; 146). Based on this you can draw the conclusion that changes in land-use patterns of today follow a process of bottom-up based self-organisation, where the accessibility and proximity of existing networks and urban centres steer the direction of the growth and that this self-organising development can be altered with the aid of land-use planning implementing spatial policies. The constraints on land-use imposed by land-use plans are usually of binary nature; either a certain type of building activity is allowed or prohibited, and therefore these are easy to take into account into land-use models, if just data is available. These represent the top-down steered part of land-use change. Taking in account local interaction in the neighbourhoods of existing land-use is far more complicated. In order to do so, we need to be able to answer questions such as: Which land-use repel and attract each other? How much? And does this interaction change over distance? Where is the interaction at its maximum and when does it even out? Is the neighbourhood interaction alike or is it different in different region and what happens to

the neighbourhood interaction when we change the scale of observation? There are no straightforward answers to these questions, which of course make it interesting to look deeper into it. Many authors have also recently indicated that the validity of neighbourhood rules is an important and urgent research issue to improve the usability of CA models (Hagoort et al. 2008:40-43, Hansen 2008, Geertman et al. 2007:548-551, Verburg et al. 2004a:668-670).

3. Problem Statement

We live in a complex world, full of uncertainties, making the future hard to predict. In order to make better simulations and "What if scenarios" of the land-use of tomorrow, we need to gain deeper knowledge about the processes underlying land-use change. We can improve present land-use models by justifying the transition rules underlying them scientifically. The neighbourhood interaction has previously been proven one of the most crucial spatial factors in land-use dynamics, which inner components needs to be fully investigated. To improve our present land-use models, we need to look into the neighbourhood interactions between land-use classes and to evaluate the effect of this on land-use change. The goal of this project is to develop a tool that can assist in this process. When we know the extent and importance of neighbourhood interaction, we can use our gained knowledge to make scientifically justified land-use models and to be integrated further for instance as a participatory tool in decision support systems. Technologies for integrating neighbourhood rules in land-use models and for integrating land-use models in participatory planning already exist.

The primary goal of this project is to develop a tool with which you can describe existing neighbourhood interactions between land-use classes. Additional requirements of the tool is that it should be simple to use and easy to understand for users of desktop GIS software. It should be scale-independent so that it can be used on input data of any resolution.

With the help of the developed tool you should be able to:

- quantify the neighbourhood interaction,
- integrate different neighbourhood sizes and configurations
- compare the use of different data as input data. and
- scientifically justify the use of certain neighbourhood rules in CA based on the calculated results

During the development process, focus will be put on the following research questions:

1. What are the theoretical components of neighborhood interaction and how can they be approached?

2. Can we develop a tool that fulfills our pre-specified goals?

3. What are the challenges related to the development and use of such a tool

Due to the time limits of this project, we need to limit our focus. Our main focus will not be on data, but on the development of an applicable tool. But since, we recognize that data play a central role for the usability of the tool and for the quality of the resulting output of the tool and therefore we will allocate available time on data issues, bearing in mind that it is the tools and concepts that are out main focus.

We have earlier stated that knowing the location and extent of change is of a central importance to improve land-use models. This project will focus a certain kind of aspect of location. We will develop a tool with which you can find where changes are taking place and which kind of neighborhood interactions are behind these changes. However, we will not look at which land-use classes are likely to change and to what extent, even though you may partly get answers to these questions through the same kind of analysis. The idea is only try to capture the spatial extent of neighborhood interaction and not the extent of land-use change on a more general level.

4. Methods

This study is based on a methodologically combined approach, where a tool will be developed to quantify neighbourhood interaction, which again can be used to support the ultimate goal to derive empirically justified neighbourhood rules to be integrated in urban CA models.

In order to develop such a tool and to find answers to our research questions the following methods will be applied:

1. a literature study on the theoretical components of neighbourhood interaction;
2. input data evaluation and preparation; and
3. software development

The findings from the literature study will be presented in the chapter on *Theoretical Components of Neighbourhood Interaction*. The chapter also includes a brief description of methods in basic spatial statistics and raster-based GIS functions that are relevant for the estimation of spatial interaction and neighborhood characteristics within the field of land-use dynamics.

In the following stage, data from the study area will be evaluated regarding its suitability and quality and its contents will be prepared for data processing. These processes will be described in the report. Aspects regarding the contents of the data will be put forward and challenges related to data will be described. The major outcome of this method can be found in the chapter *Data*.

The final stage is to design a programme that can be used for quantifying neighbourhood interaction between land-use classes. A conceptual model of the programme components will be made and a programme will be developed according to it. The conceptual model and the resulting programme will be described in the chapter *Software Development*.

5. Theoretical components of Neighbourhood Interaction

5.1. Characteristics of spatial data

Statistical methods for measuring relationships are well established in traditional non-spatial statistics. These methods are based on assumptions that are not valid for spatial data. Typical of spatial data is that they may include a regional or a directional trend that vary in your data. This is for example the case with wind data and other spatially continuous environmental data. Additionally geographical features that are near each other are more likely have similar values more similar than distant feature. This phenomenon, which is referred to as spatial autocorrelation, violates the assumption that observations are independent on which traditional statistics is based (Mitchell 2005:200-201). If nearby features are more like each other than distant features, there is said to be a positive spatial autocorrelation. An example of a data set with spatial autocorrelation is urban land-use data – since for instance in the vicinity of industrial feature there tend to be other industrial features. If neighbouring features tend to be unlike each other, this is termed negative spatial autocorrelation (Mitchell 2005:105).

Spatial autocorrelation makes spatial data redundant in a statistical sense and therefore it may be of relevance to find out whether your data is spatially autocorrelated or not. Spatial autocorrelation is a typical characteristic of very precise data, holding small units. When using large geographical units the problem with spatial autocorrelation is not as present, but then again you risk loosing the local variation within the data. For example if you use a coarse resolution for your land-use data, will the land-use classes in real-life be a mixture of many kinds of land-uses. Researchers have proposed several techniques for dealing with this problematic issue. Broadly speaking, these are different resampling techniques to exclude the spatial influence or techniques to incorporate the spatial influence in the analysis for gaining a more accurate picture of real-world spatial relationships (Mitchell 2005:200-201). Of these, the latter technique is in line with this study.

Spatial interaction is closely related to the phenomenon of spatial autocorrelation. Spatial interaction can in its simplest be defined as the influence a geographic feature has on another geographic feature or in other words the spatial relationship that exist between them. Basic tools to support the quantification of spatial interaction do exist in desktop GIS. However, for studying the dynamics of land-use in a broader sense these tools have to be combined and developed further. Only then they may be able to measure real-life influence of land-use classes on each other within an area of interest. In the following we will describe GIS-based statistical methods a GIS functions, which may be useful for estimating neighbourhood interaction.

5.2. GIS and measuring spatial interaction

In desktop GIS, there are several methods, which enable you to identify and characterise patterns and clusters based on the feature or cell values of your data. So-called global statistics methods, like join count statistics for categorical data (Mitchell 2005:109); indicate if there is a spatial autocorrelation. To capture local variations, and so-called hot or cold spots in the data you can use local statistical methods, such as Local Geary's c. While global statistics calculate a single statistic that summarizes the pattern for the study area, the local methods calculate a statistics for each feature, based on its similarity to its predefined neighbours. Local statistics can therefore help pinpointing which features contribute to the spatial autocorrelation, so that you can account for it in your model (Mithcell 2005:165).

Global and local statistical methods use both the values of features and the spatial relationship between the features. In these processes, GIS compares the value of a target feature with the values of its neighbouring or all features, looping through all target features, calculating measures of the pairs of features we are interested in. We might for example want to know if the location of a certain land-use class correlated with the location of another land-use class within a dataset. A precondition for making this kind of local statistical analysis is that we need to define the area surrounding the target feature, which we are interested in. This area is what we call the neighbourhood. If you use a neighbourhood, you should define the shape and extent of it based on the spatial interaction of the particular phenomena of interest. However, the spatial interaction may

not be known or at least not empirically proved. Additionally the nature of the spatial relationship between the features needs to be defined (Mitchell 2005:135). The spatial relationship of two features could be known to decline with the distance to the target feature. However, in many cases the spatial relationship may not be known either. To carry out these statistical methods describing whether a pattern is dispersed, clustered or random, actually may require research on spatial autocorrelation and spatial relationships within your data, which is the focus of this study.

Next we will go into GIS functions that apply for raster data. Raster-based functions are relevant, since land-use data is often represented and CA models incorporated with data in raster form. Several raster-based GIS techniques are available that may help in founding out the spatial interaction within raster data. In ArcGIS raster analysis can be carried out using the Spatial Analyst extension. Spatial Analyst allows you to perform a whole range of functions, including neighbourhood functions, using the so called Map Algebra language. When working with Spatial Analyst to find out about the neighbourhood interaction, you are likely to use many of these functions, which make it relevant to describe several of them in more detail:

- local functions that work on single cell location
- focal functions that work on cell locations within a user-defined neighbourhood
- zonal functions that work on cell locations within zones, which can be a land-use class
- global functions that work on all cells within the raster dataset
- application functions that perform a specific application or task, as for instance the altering of resolution (ESRI 2007)

Local functions, or per-cell functions, compute a raster output dataset, where the output value at each location (cell) is a function of the value associated with that location on one or more raster datasets. An example of a per-cell function is the base 10 logarithm of the cells in a raster.

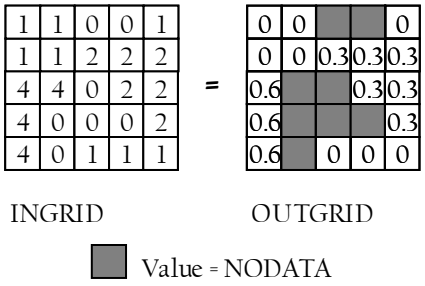


Figure 2. An example of a local function, in this case the base 10 logarithm, the expression being LOG10(INGRID)

Focal functions create output values for each cell location based on the value of a certain location and the values identified in the defined neighbourhood around that location. Characteristic for these focal functions are that they move from cell to cell with an overlap and therefore they are also called overlapping neighbourhood functions. These generally calculate a specified statistics within the neighbourhood. For example, you can find the sum of the values within a rectangular Moore neighbourhood of 3 x 3 cells (figure 3 and 6). This is useful if we are interested in how many cells of a particular land-use class is in the proximity of another land-use class. When carrying out these kinds of analysis, the neighbourhood size and configuration used play a central and will affect the calculated result. The cells on the edge of the data set will be affected by the lack of neighbours.

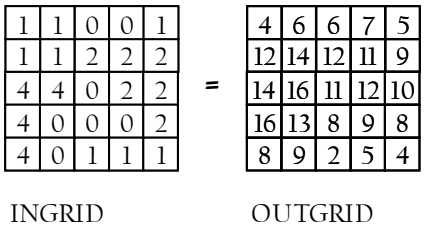


Figure 3. An example of a focal function in ArcGIS, in this case using the expression FOCALSUM(INGRID, Rectangle, 3, 3).

The zonal functions provide a set of tools for zonal analysis and computing zonal statistics. A zone is all the cells in a raster that have the same value. An example of a zone is a land-use class of interest. Zonal statistical functions perform operations on a zone-to-zone basis, so that a single output is computed for every zone defined by the input

zone dataset. Based on the zone dataset, you calculate the statistics from a value raster, containing the input values to be used in the calculation. You can for instance calculate the mean land value in different land-use zones with the ZonalMean funtion. In that case the output is a raster dataset. You can also choose to calculate the zonal statistics an get a table as an output using the ZonalStatisticsAsTable function. In this case you do not need to choose what statistics you are interested in, since these fields will be automatically created: Value, count, area, min, max, range, sum, mean, and std fields will be created regardless of the input values. Majority, minority, median, and variety fields will only be created when the input value raster is integer.

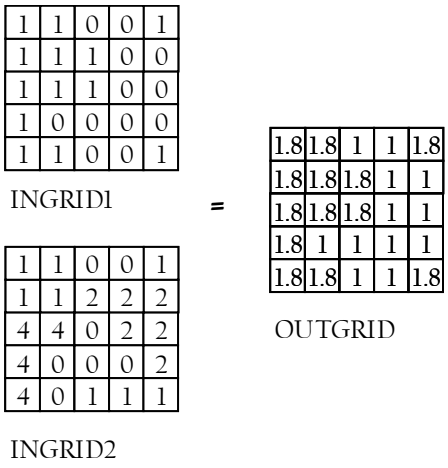


Figure 4. An example of a zonal function, with the expression being ZONALMEAN(INGRID1, INGRID2)

So-called conditions (Con) can be used to restrict the scope of another function in Map Algebra. Conceptually, the Con function visits each cell location and, based on the cell's value and the conditional statement, determines if the cell evaluates to true or false. If the cell evaluates to true, the output value for that location is identified in the true input raster or constant. If the cell evaluates to false, the output value for that location is identified in the false input raster or constant. It is by setting these kinds of conditions that you can implement for instance Cellular Automata using ArcGIS Spatial Analyst.

1	1	0	0	1
1	1	2	2	2
4	4	0	2	2
4	0	0	0	2
4	0	1	1	1

INGRID

=

1	1	0	0	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	1	1	1

OUTGRID

Figure 5. An example the use of a condition, in this case (Con(INGRID == 1), 1, 0)

5.3. Neighbourhood configurations

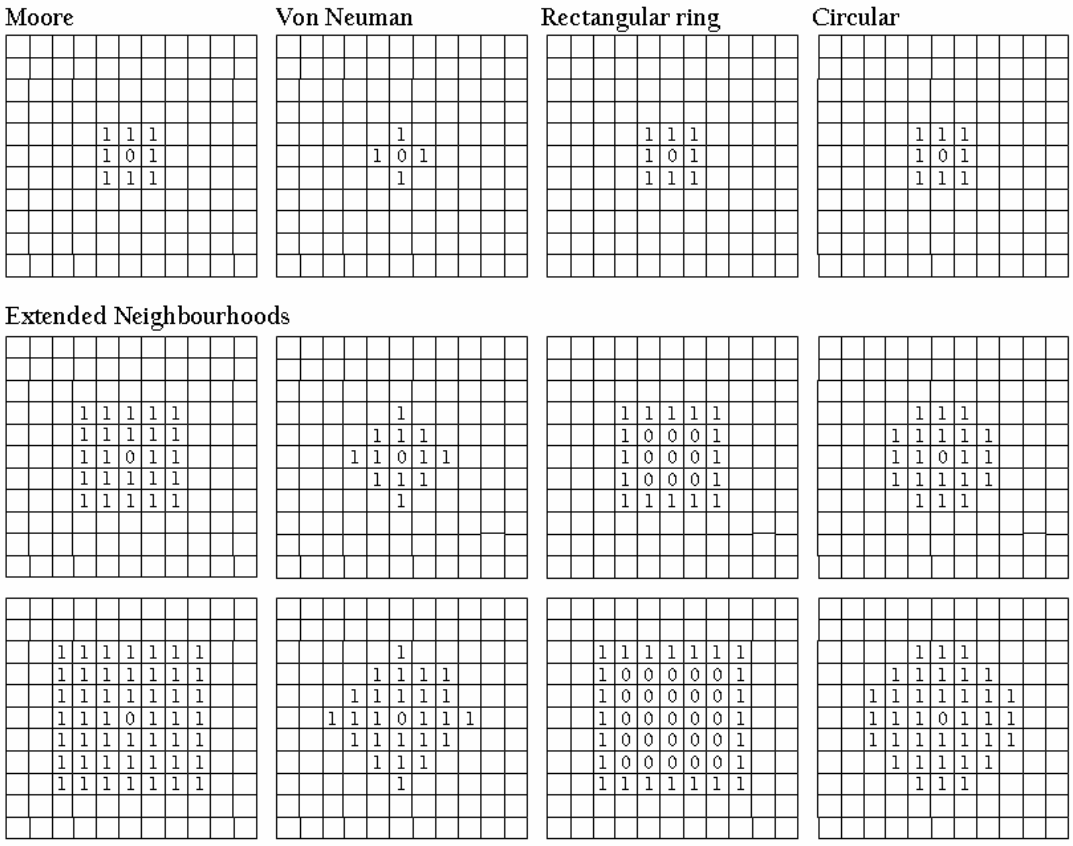


Figure 6. Examples of neighbourhoods, where 0 represents the target feature (cell) and 1 represents the neighbourhood. The central cell is not included in the neighbourhoods of these examples - usually it is.

There are many kinds of neighbourhoods in use. Figure 6 presents a few examples of these. There is no general agreement on which neighbourhood to use – it depends on the purpose and application you are using it for. Generally very little research has been

addressed to evaluating the use of different neighbourhood configurations (Geertman et al. 2007:551). In most cases the size and configuration of neighbourhood is chosen arbitrarily and only the direct neighbourhood of a location if taken into account – either 4 cells according to the Von Neuman or 8 adjacent cells according to the Moore neighbourhood. In many cases no sensitivity testing is made regarding the choice of the neighbourhood size that is being used in for instance in Cellular Automata-based urban application. However, most certainly the choice of neighbourhood affects the result. (Batty and Torrens 2001:26-27).

it is recognized that the neighbourhood interaction between land-use classes in cities may extend beyond the most adjacent cells (Verburg et al. 2004a:671). This is why extended neighbourhoods have been used in recent research on neighbourhood interaction and in newly developed urban land-use models (Verburg et al.,2004a; 671, Geertman et al., 2008:554, Hansen 2008). For example, Geertman et al. (2008:554) and (Verburg et al. 2004a:688) used eight specific neigbourhoods on their 100 x 100 square meter data. Each zone was 100 meter wide, leading to a maximum distance of 800 meter around a specific location. Why that particular neighbourhood size was chosen was not reflected on in these articles and according to one paper this square shape was only chosen due to computational benefits. It was mentioned that the use of circular neighbourhoods would be better, since the distance between the neighbourhood and the central cell stay more or less the same in these (Verburg et la. 2004a:688).

5.4. Weights, distance decay functions and neighbourhood rules

In land-use models that aggregate the land-use development over many years, such as constrained urban CA models, it may be relevant to use weights representing the spatial interaction between the features or cells or. In this way the neighbourhood interaction influencing a cell can be aggregated. The simplest and most often used method in GIS for doing this is by specifying a matrix of weights for example between land-use classes either ad hoc or through a trial and error based method. In desktop GIS, you can also specify the rate at which the influence or spatial interaction decreases as the distance decreases. This is termed the distance decay. If the influence decreases at a constant rate, the inverse of the distance is a suitable weight measure. It can be derived by dividing one

(1) with the distance the feature is from the target feature. This means that the weight of a feature decreases with its distance from the observed feature. Other straightforward and integrated methods to include the spatial influence as weights are exponential distance decay, proportional weights and row-standardized weighting (Michell 2005: 135-145). The drawback of these weights is that they are not in most cases based on real relations, from empirical studies. The use of these kinds of non-empirical weights as a basis for transition rules in land-use models, have been highly questioned and criticized (Hagoort et al. 2008:40, Malczewski 2000:16-19).

Using weights between two separate land-use classes also address another problem. The spatial interaction doesn't only differ between land-use classes for example within a defined neighbourhood, it also differs according to the distance from the target feature/cell. According to Tobler's so called first law of geography "Everything is related to everything else, but near things are more related than distant things". The meaning of it plays a significant role for understanding importance of spatial interaction in urban land-use dynamics (Barredo et al., 2003:146). Tobler's law stresses that the neighbourhood of a feature, also beyond the most adjacent space, can influence the feature as a function of distance. The influence of different land-use types on the land-use type under observation differs of course. A certain land-use can have a attracting, repulsive effect or no effect at all on another land-use at a certain distance. Often the effect changes over distance. For example a motorway in the direct proximity is not desirable for residential areas due to noise, pollution and other negative effects. However, having good access to a motorway from your home is a benefit of a residential area, when it is for instance one to two kilometres away.

Empirically based distance decay functions can be used to describe the influence a factor have on another factor over distance. These functions can be integrated in land-use models as so called neighbourhood rules. They can be defined as transition rules that are based on neighbourhood interaction and its aggregated effect. Seven generalised types of neighbourhood rule shapes have been identified (figure 7). The seventh, not shown in the figure, represents those where no or an indifferent spatial relationship has been identifie

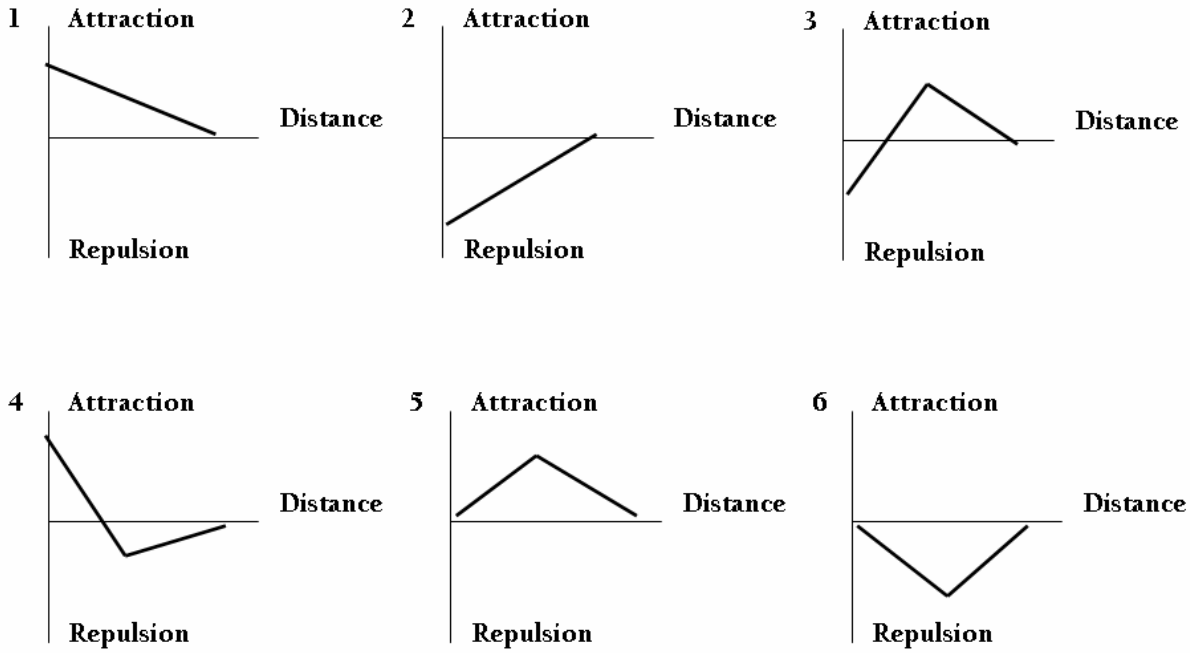


Figure 7. A neighborhood rule shape typology modified after Hagoort et al., 2008:45

But how can we actually derive these distance decay functions? First of all, you need to have access to good-quality temporal data, from which you can produce land-use classes, which are convenient to represent the land-use dynamics. The land-use classes need to be suitable for the scale of investigation and since they will have an effect on the observed distance decay function, they may not directly be applicable for another scale. Secondly, you need to analyse and quantify the amount of neighbourhood interaction that has been taken place in your study area since it has been found that the factors influencing the land-use dynamics may differ from place to place. Ideally you would have access to data the temporal resolution of one year, so that you could follow the development from year to year (Hansen 2008). Unless you have access to this you need to be aware of the consequences on the results. If you compare land-use with a temporal resolution of 10 years, the urban cells of the later land-use data will according to the

previous data have neighbours such as arable land. This would probably not be the case if the temporal resolution was one year, then the urban evolution would proceed so that the earlier datasets would give a more realistic picture of the neighbours of a new urban cell. When analysing the spatial interaction of land-use classes, you also have to take into account that the neighbourhood shape (form) and size influence the results of your analysis and you should therefore evaluate which neighbourhoods capture the spatial interaction the best. After this process you should be able to identify which neighbourhood rules are relevant for your case study area.

A formal theory of shaping neighbourhood rules is only slowly developing. A difficulty is to recognize how many neighbourhood rules that need to be defined (Hagoort et al. 2008:44). If your land-use data have 10 land-use classes, there are 100 combinations of land-use classes, whose neighbourhood interaction to study. If you additionally want to test the influence of different kinds of neighbourhoods the amount of analysis increases by 100 for each new neighbourhood type. It is unnecessary to carry out such an extensive process. The number of land-use relations to be analysed, can be cut down without compromising with the result, by not analysing static land-use classes, which do not change and by not analysing land-uses that are known not to interact, such as the passive classes (Hagoort et al. 2008:44). Instead main focus can be put on the active expanding urban classes. You can also use methods such as join count statistics to find out which land-use classes that show a spatial interaction of interest. Alternatively you can ask experts on land-use issues (Hagoort et al 2008:45).

5.5. Spatial metrics and the enrichment factor

One way of identifying neighbourhood rules and the processes they describe is to analyse spatial processes with the help of so-called spatial metrics. Spatial or landscape metrics have been used for a couple of decades in landscape ecology and recently also outside that field (Geertman et al. 2007:552). They were developed to describe the landscape structure on a patch, class and a landscape level. While in landscape ecology a class is represented by a biotope or a habitat type, it is represented by a land-use class when studying urban land-use dynamics. Although many spatial metrics have been developed and applied for the characterization of various, also urban, landscape patterns, their use

in urban studies has not been fully explored (Geertman et al. 2007:552). With spatial metrics you can describe spatial characteristics, such as the size, shape, number, kind, and configuration of the urban morphological structure, in an effective, innovative way. Spatial metrics cannot of course explain the causes of observed land-use patterns and processes, but they can give scientifically justified indications of causal relationships taking place (Geertman et al. 2007:552) and this is what is needed to justify the use of neighbourhood rules in urban CA.

Several authors have found a landscape metric – the so-called mean enrichment factor – particularly appropriate for quantifying and analysing neighbourhood characteristics (Verburg et al. 2004a:685, Geertman et al. 2007:552-554, Hansen 2008). The enrichment factor characterises the over- or under-representation of different land-use types in a neighbourhood of a specific raster cell. To measure this over- or under-representation, it compares the amount of cells of a particular land-use type in the vicinity of a specific location as relative to the volume of cells of that land-use type in the study area in total. When the proportion of a land-use type in a neighbourhood equals the national average, the neighbourhood possesses an enrichment factor of 1 for that land-use type. If the neighbourhood of a specific location (cell) consists of 20% residential areas, whereas the proportion of residential areas in the study area as a whole in total is 5%, we can characterise the neighbourhood by an enrichment factor of 4 for residential areas. Contrary an under representation of a certain land-use type in the neighbourhood will result in an enrichment factor between 0 and 1.

The equations for the enrichment factor are specified in Verburg et al. (2004a:671-672) as follows:

$$F_{i,k,d} = \frac{n_{k,d,i} / n_{d,i}}{N_k / N} \qquad (1)$$

Where $F_{i,k,d}$ characterises the enrichment of neighbourhood d of location i with land-use type k. The shape of the neighbourhood and distance of the neighbourhood from the central grid cell i are identified by d. This neighbourhood characteristic results from each

grid cell *i* in a series of enrichment factors for the different land use types (*k*). The procedure is repeated for different neighbourhoods located at different distances (*d*) from the grid cell to study the influence of distance on the relation between land-use types.

The average neighbourhood characteristic for a particular land-use type *l* ($\overline{F}_{i,k,d}$) is calculated by taking the average of the enrichment factors for all raster cells belonging to a certain land-use type *l*, following:

$$\overline{F}_{i,k,d} = \frac{1}{N}\sum_{i \in L} F_{i,k,d} \qquad (2$$

where *L* is the set of all locations with land-use type *l*, and N_l is the total number of cells belonging to this set.

To conclude, it seems that if we can make a programme that derives, the enrichment factor, we can arrive at theoretically and empirically more justifiable neighbourhood rules, which in their turn can help to improve urban CA models. To make and to use such a programme we need to have access to spatio-temporal land-use data, which is the focus of the next chapter.

6. Data

6.1. Description of data

Data is important, because without data this kind of programme development could only be carried out on a theoretical level. Additionally it is of course relevant that there are data available that you can use in your programme. However, the main focus of this project is to understand how to analyse neighbourhood interaction and to actually learn how to develop a program characterising neighbourhood effects. It is not a data driven project, but a method-driven. Anyway, it is good to take a critical look at the data you work with, to be aware of its quality and possible disadvantages so that these do not cause unexpected problems in the software development process. In the long run, the availability of most suitable data is the key to get the best possible results, also regarding the quantification of neighbourhood interaction.

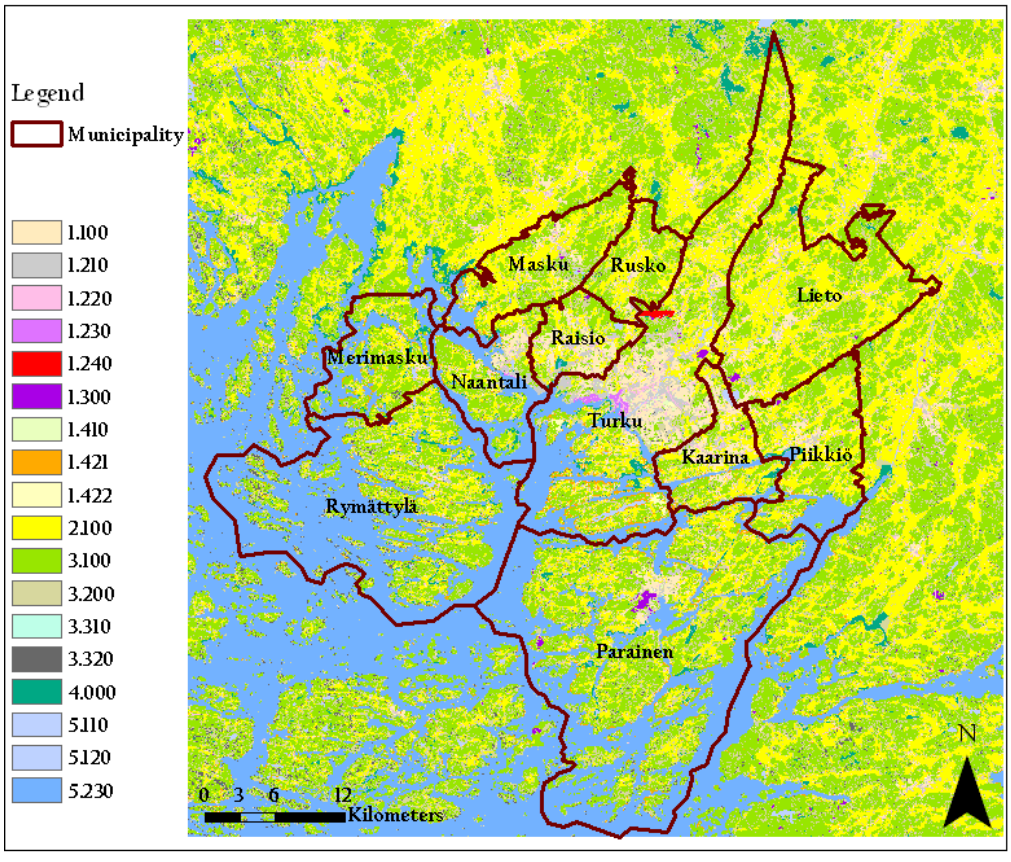


Figure 8. The study area in South-Western Finland. The land-use codes are explained in table 1.

The data, which is used in the development process, is from South-Western Finland. It covers the area of eleven municipalities and the areas surrounding them; Turku, Kaarina, Lieto, Masku, Merimasku, Naantali, Parainen, Piikkiö, Raisio, Rusko, Rymättylä (figure 8). In all the study area is 4900 square kilometres. The area within the municipalities is 1913 square kilometres in all, including inland, coastal municipalities and municipalities situated in the archipelago. Both delineations are used.

To study neighbourhood characteristics, temporal land-use data is required. Three land-use datasets exist from the study area: two representing the situation in year 1990 and one representing the situation in year 2000. The land-use data has been produced according to the principles and classification of the European CORINE Land Cover project. The data therefore represents more the land-cover on the surface and not so much directly the purpose or use of land. It will however be referred to as land-use data in this report.

CORINE LC is being produced differently for Finland than in most other European countries. The production is based on a combined method consisting of automated interpretation of satellite images and the integration existing digital map data. The quality of the source data and how these are benefited in the production process therefore affect the quality of CORINE. In addition to the European CORINE data set a national, less generalized data set have been made and it is the national version of CORINE 2000 that is used in this project. The national data set have a spatial resolution of 25 x 25 m pixels in comparison with the European vector data, where the minimum mapping unit for most features is 25 hectares (CLC 2000:3).

The use of satellite images and map data for producing CORINE LC 2000 has been evaluated. According to the validation of the national CORINE LC 2000 dataset, the accuracy is 90 % for the most aggregated class at the co-called level 1 (for instance mire, class 4000), 80 % for level 2 classes (for instance forest, class 3100) and 70 % for level 3 classes (for instance lake, class 5110) when compared to the National Forest Inventory (NFI) information (CLC 2000:41). The NFI does not cover agricultural and the artificial surfaces classes and therefore these classes were compared with another Finnish land-

use data, SLICES, which is one of the primary sources for the artificial surfaces and for agricultural areas of CORINE LC 2000. Overall accuracies for these classes are high; 85% on level 1 and 85 on level 3 (CLC 2000:43).

An official CORINE LC 1990 does not exist for Finland. The land-use data for year 1990 has been produced after CORINE LC 2000 as an attempt to cover this data gap and to develop a method to make a data set with which land-use changes between these the years 1990 and 2000 can be derived. The 1990 data has only been produced for a pilot area in South-Western Finland. It has the same spatial resolution as the national CORINE 2000, i.e. 25 m cells. To ensure comparability, the production chain for making CORINE 2000 was repeated as much as possible, integrating data sets from approximately year 1990. Separate themes describing artificial surfaces, agricultural areas, forests, wetlands and water were produced using old digital map datasets, registers and satellite data received close to year 1990 (CLC 1990). Since a complete set of input map databases representing year 1990 was not available, like for year 2000, a fully compatible land-use cover classification was not possible to produce. So far there is little information on the quality of this data. What we know is, that data availability differed from the situation in year 2000 and that the lack of data sources representing the land-use situation in 1990 could partly be replaced by more extensive image interpretation and segmentation (CLC 1990). At the time of writing, a master's thesis is being made at the Finnish Environment Institute to evaluate the quality in more detail.

A common challenge in the production of CORINE LC 2000 and the test version of 1990, is that the delineation of urban land-use classes or build-up areas, is partly based on the national Building and Dwelling register, containing information on the location, purpose and size of buildings. The problem with using register data that only are fixed to one geographical point, is that neither the point nor its attribute data can sufficiently explain the spatial extent and configuration of the buildings and their surrounding artificial land. This is especially essential for industrial and service buildings which often include extensive parking lots and other supporting facilities, all of which should be classified as urban artificial surfaces. This problem was reduced in the production of CORINE LC 2000 by integrating interpretation of satellite images with available GIS data sets. For

CORINE LC 1990 an additional segmentation method that was based on multidate satellite data was used to capture these artificial areas . As part of data integration made for CORINE LC 1990 separate GIS and remote sensing-based (RS) datasets were merged according to several rules. Not only map data from the BDR was integrated, but also data from, but also data from other sources, such as SLAM, Unofficial Finnish Corine of year 1990 and Corine LC 2000 (CLC 1990).

The most relevant data quality issue of this project is related to how well the two land-use data sets that are being used in the study match together or in other words how well they can be compared to describe the actual changes of urban land-use. For this analysis, we will use the Building and Dwelling Register (BDR) from year 2006. The register contains extensive attribute data, such as the coordinates of building, the construction year and the use of the building etc. The register is being updated by the municipalities, and the quality can therefore vary (BDR 2006).

Both the land-use data and the building and dwelling register are concerned by INSPIRE. INSPIRE requires geographic data to be well-described, easily evaluated and easily accessible, of a good quality, available for broad use and harmonised across Europe (INSPIRE 2008). In relation to the INSPIRE Directive, several requirements are filled, but there is still work to be done.

Against one of one major so-called INSPIRE principles it may be hard to find and to actually get access the data. A juridical agreement was made in order to obtain the data to this project. The situation of the data accessibility and availability of at least the land-use data will be improved as part of the implementation of INSPIRE in the environmental administration of Finland. Within the same project the coordinate system will be changed from the national grid (Gauss Krûger projection) to the European standard ETRS89.

For the official CORINE LC 2000 data, descriptions are available on the Internet, both regarding productions procedures and data. Also the BDR is well described. Although these descriptions are not made based on standards, they can with not too much an effort be converted to the metadata imposed by INSPIRE. The unofficial test data from

year 1990 is only described in a few articles that are of a more technical nature. A drawback of this data was that its quality is not yet known, but more relevant for this project is to know how well the land-use data sets match thematically and spatially. A few analyses are carried as part of the data preprocessing to get a picture of this.

6.2. Data preprocessing

Table 1. The land use classes used in this project.

Land-Use class in Corine	Used land-use class	Description
1110 and 1120	1100	Residential
1210	1210	Industry and service
1220	1220	Traffic areas
1230	1230	Port areas
1240	1240	Airports
1310 - 1330	1300	Extraction and dump sites
1410	1410	Green urban areas
1421	1421	Summer cottages
1422	1422	Sports facilities
2110 - 2430	2100	Arable and semi-natural areas
3111-3133	3100	Forest
3241 -3246	3200	Grasslands and shrubs
3310	3310	Beaches, dunes and sand plains
3320	3320	Bare rock
4111-4212	4000	Mire
5110	5110	Lake
5120	5120	River
5230	5230	Sea

We took a closer look at the data in order to check its compatibility, and to improve its suitability for this project, taking into account the time limits. The national classification of the land-use classes is unnecessarily narrow for this project and for urban land-use modelling purposes in general. The land-use data for 1990 and 2000 used in this study contained 38 classes and 40 classes respectively. CORINE is a land-cover data, not an urban land-use data set, and data that represent the urban land-use should preferably

include only relevant land-use classes to explain the processes and spatial interaction of interest. For instance, it is not relevant in respect to urban land-use modelling purposes and for this project, to know the exact type of forest. Rather it is relevant to know if there is forest or not, to find out if the proximity of forests have an attractive or repulsive effect on urban land-use. A need for reclassification was also supported by the indications that imply that the data quality improves by using a broader thematic classification (CLC2000; 41-44). A reclassification of the original land-use data was made, with ArcGIS *Reclass by ASCII file* function, to respond to these issues. The used classification is presented in table 1. In the rest of the report the classes 1100 to 1422 will be referred to as urban classes.

Next, the land-use classes were compared. Inconsistent changes within a land-use class would indicate that the datasets used in the study are not comparable. There was also another purpose of this comparison. Two versions of the land-use data of 1990 had been made and we needed to decide which one of these to use. Since a sufficient evaluation of the quality of these data sets did not exist, we hoped this comparative analysis would help us to decide on that.

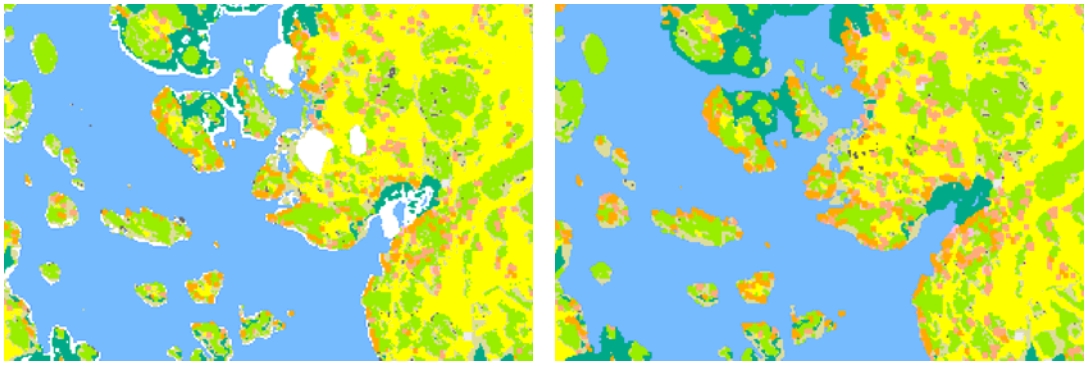


Figure 9. The original 1990 land-use data with white holes and the 1990 land-use data, where the holes have been filled with Corine 2000 data.

Before we look at the comparison, we need also to reveal that there were 172 745 cells that lacked a class in the land-use data of year 1990 (figure 9). These data gaps were filled with cell from the land-use data of year 2000. The majority of the cells were replaced

with agricultural, forest and mire cells, but about 20 000 urban cells were added as well. The replaced cells were the same for both the 1990 data versions.

Table 2. A comparison on the amount of cells of a certain land-cover type in the original data sets from year 1990 and year 2000, having a resolution of 25 times 25 meter.

Used land-use class	Description	Land-use 1990 All arable	Land-use 1990 Not all arable	Land-use 2000
1100	Residential	302560	302560	421902
1210	Industry and service	60675	60675	75712
1220	Traffic areas	60611	60611	91518
1230	Port areas	3529	3529	3666
1240	Airports	2417	2417	2512
1300	Extraction and dump sites	12256	12256	20098
1410	Green urban areas	2271	2271	3067
1421	Summer cottages, leisure homes	90408	90408	140837
1422	Sports facilities	7988	7988	11022
2000	Arable and semi-natural land	1870787	1719873	1715662
3100	Forest	2519714	2547196	2439737
3200	Grasslands and shrubs	769878	889997	769271
3310	Beaches, dunes and sand plains	18	20	65
3320	Bare rock	78394	81590	44474
4000	Mire	115951	115966	123542
5110	Lake	12963	12963	12963
5120	River	24290	24290	26458
5230	Sea	1905390	1905390	1937494
	Total	**7840000**	**7840000**	**7840000**

No major inconsistencies were found (table 2) on this level of classification; all urban classes, those from 1100 to 1422, have increased on the expense of mainly the forested and arable land areas. However, how well this increase corresponds to actual changes cannot be evaluated here. It is certain that the use of slightly different methods and different source data also have a share in it. An indication of the effect of the use of different

methods and source data can be observed in the changes of the values of the most stable classes, where changes are not likely to have taken place in reality; such as for sea and mires.

The differences between the 1990 data sets were obvious. In the first version (All arable), all possible arable lands from image interpretation and other data sources have been classified as arable on the expense mainly of the forest class, grasslands and shrubs, and the bare rock class. For the other version, the situation is the opposite. However, the so-called urban classes are identical. In this research we are not interested in which cells that are being changed into urban cells, but instead we are interested in what type of cells those cells that have changed where in the proximity of before changing. Therefore we can conclude that it is indifferent which of the 1990 data sets we use for this study.

Next it was checked how well the land-use data correspond with the buildings of the building and dwelling register. Generally the buildings from the BDR correspond very well with the location of the urban land-use classes that is the classes 1100 to 1422. A visual comparison show a very good spatial match between the location of buildings and the residential areas (1100) and summer cottage areas (1421). The same cannot be said for the aggregated business and service class (1210) and the port class (1230). In figure 10 all buildings are presented by the same symbol type, but in real-life service and business-related buildings are generally bigger. This has been taken into account in the production of these land-use datasets by expanding or increasing the size of the 25 x 25 building pixels, by 25, 50 and 75 meters according by their type of use (CLC2000:20). Part of the difference can also be explained by the extensive parking lots and other pavements that are related both to this class and the port areas that have been captured by the help of image interpretation and other map data as already stated before.

The visual comparison also revealed that the road network represented by the traffic class (1220) was not continuous. Due to this, the traffic class is not representative of road accessibility. Therefore, if you include the traffic areas in the neighbourhood analysis, it will affect the result in an inconsistent way. To improve the land-use data to examine neighbourhood interaction the inconsistent traffic areas should preferably be eliminated.

This also is supported by the fact that the Road and rail network and associated land, that we have referred to as the traffic class is the worst class accuracy wise, according to a data quality study made on the national CORINE 2000 (CLC 2000:44). As part of the traffic elimination, you could also try to minimize the influence of the data gaps in the dataset of 1990 data that were replaced with land-use data from 2000.

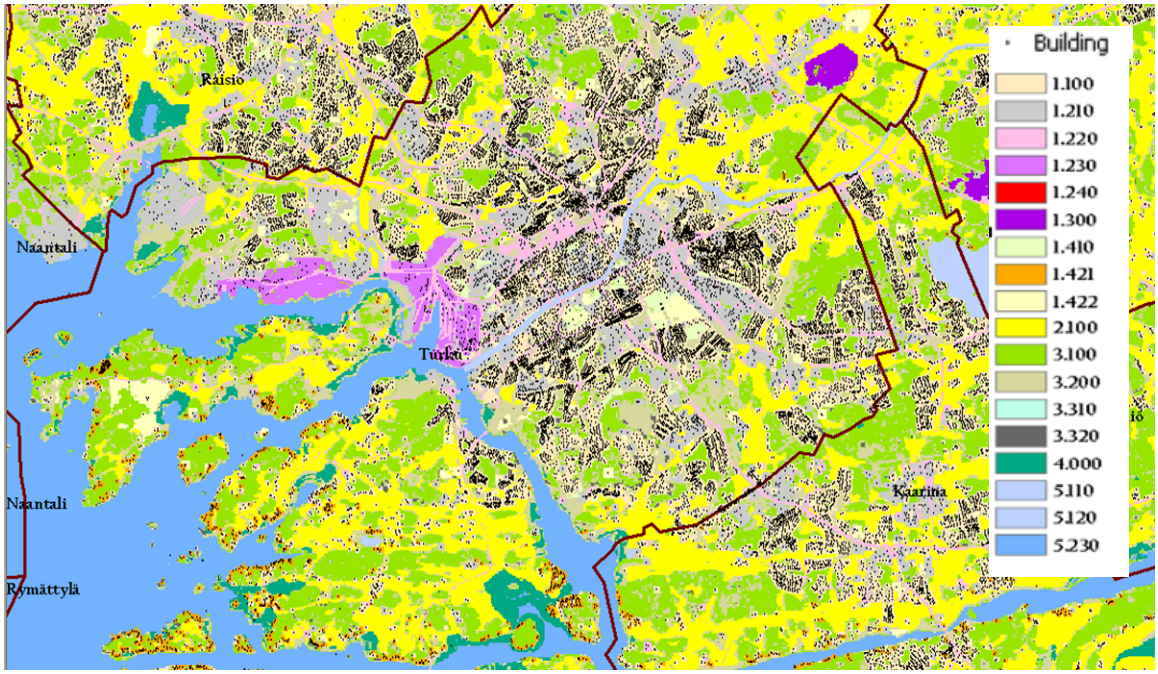


Figure 10. Visual comparison of the correspondence between land-use classes and buildings, showing land-use from year 1990 (including all arable lands) and buildings from the end of 1990.

One possible way to eliminate linear traffic features and to diminish the effect of the data gaps, would be to change the data into a coarser resolution, which actually could be quite a good idea. The use of a coarser resolution may also be better to capture urban land-use dynamics - it would at least be interesting to compare the 25 meter resolution with a coarser one to find this out. A drawback of the 25 meter cell resolution is that you need to analyse a four times more cells to reach the distance where the neighbourhood interaction ebbs away, than if you use data with the resolution of 100 meter. The use of a coarser resolution therefore reduces processing time.

A simple way to change the resolution in ArcGIS is to resample your data. To completely eliminate traffic areas you need to reclassify the traffic cells into NODATA before doing

the resampling. An example of the result of resampling the land-use data based on the cells that have the majority is presented in figure 11. According to a visual comparison of the result with the original data set, it appears to be an effective tool to eliminate linear traffic feature. However, the disadvantages of this resampling method to the data quality are also obvious; it results in the elimination of other linear elements, such as a river s (light blue in the figure), the elimination of small features, such as stand-alone summer cottages (orange in the figure), distorting the composition of land-use classes and their location among themselves, which are crucial when studying neighbourhood interaction. To fix for instance the absence of for instance summer-cottages and rivers, we would need to separately rasterise these features into 100 meter cells and update the resampled data with them.

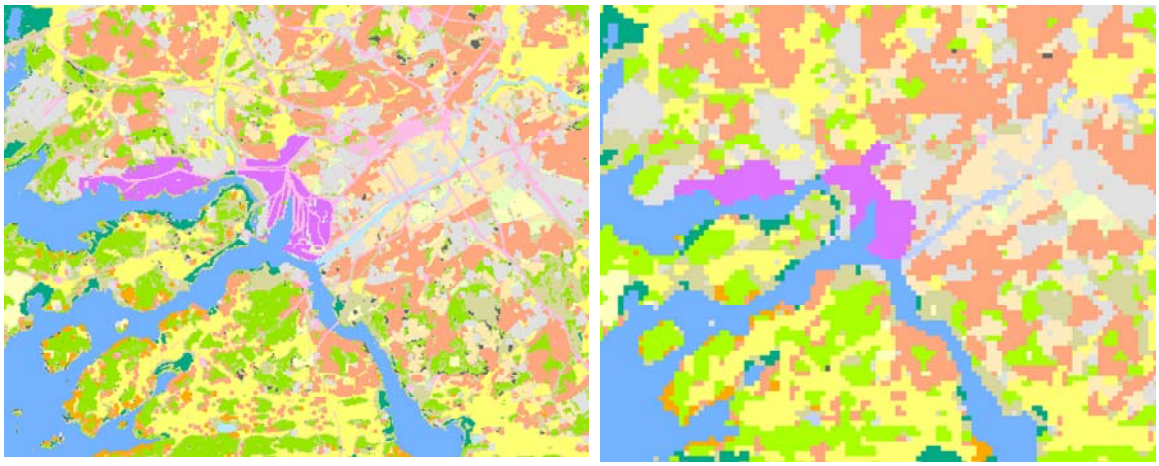


Figure 11. The original data set of 25 meter cells (left) and the resampled data set of 100 meter cells (right) showing the city centre of Turku and the surroundings.

A few other drawbacks of using resampling were also realised. If we resample by majority, the spatially most connected and extensive land-use classes will cluster and expand. In urban areas for instance residential areas will expand on the expense of parks and other less dominant land-use types. In more scarcely populated locations, for instance single residential cells and summer cottages that traditionally are not situated too close to each other, will be resampled into the surrounding dominant forest or arable class. You can regard this as a good method for eliminating non-urban stand-alone

houses in the countryside. However, in this way the good match with the BDR would be lost (figure 12).

You can also argue that single stand-alone buildings are relevant to take into account in the neighbourhood calculations. Excluding these from the neighbourhood analysis would give the wrong result; since we found that it is in the vicinity to these stand-alone residential cells that further building development take place at least in the vicinity of attractive urban centres, such as the city centre of Turku. Also new summer cottages are likely to be built on free spots on the coast, despite the existence of neighbours, when more private locations no longer exist. These single residential cells and summer cottages cannot simply be added the resampled dataset as 100 meter residential a summer cottage cells. This way they might be expanded too much, so that new buildings from the subsequent year will coincide with already built-up cell, which mean that they will not be observed as new, changed cells. These things can be observed in figures 12 and 13.

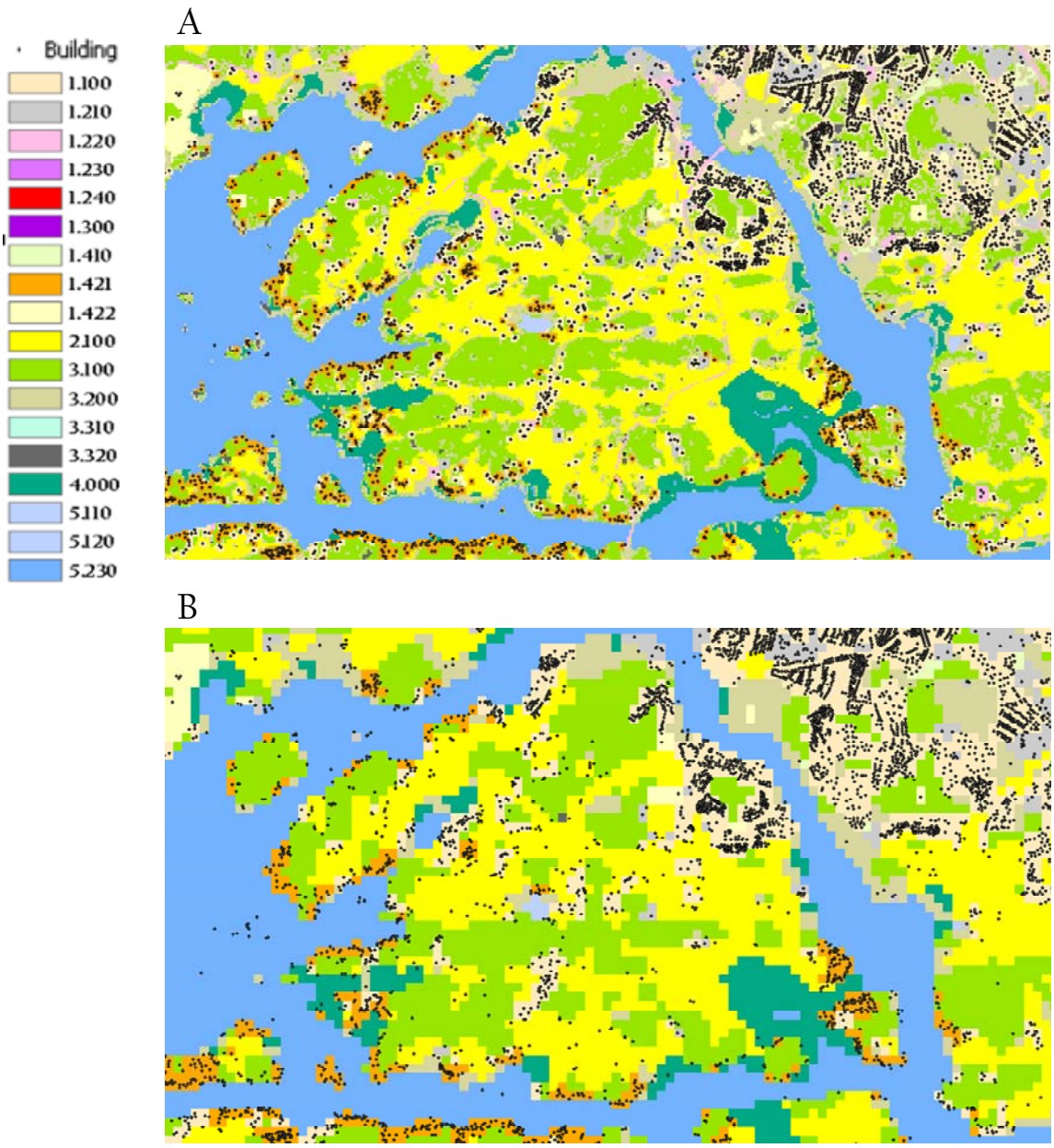


Figure 12. **A.** Land use in year 1990 (All arable land) and existing buildings (black). Cell size is 25 m. The area is 10 kilometres wide. **B.** Land use in year 1990 (All arable land) and existing buildings (black) after resampling. Cell size is 100 m. The area is 10 kilometres wide.

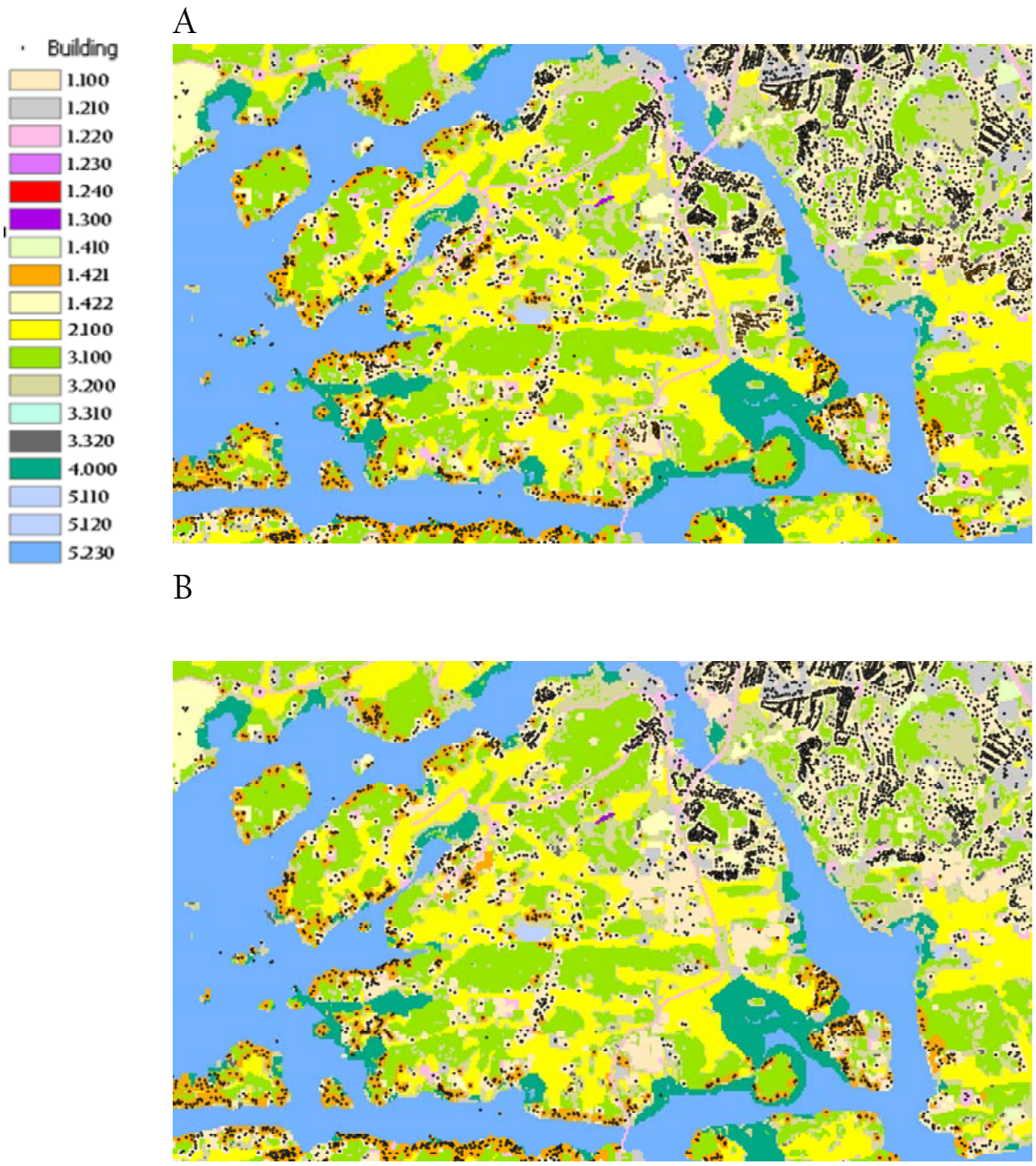


Figure 13. **A**. Land use in year 2000 and existing buildings (black) in the end of year 2001. Cell size is 25 m. The area is 10 kilometres wide. **B.** Land-use in year 2000 and existing buildings (black) in the end of year 1990. Cellsize is 25 m. The area is 10 kilometres wide.

Based on these observations, we can conclude that the resampling method studied is not good enough in our case and we discarded the option to use resampling in this project.

We are fully aware of that if the goal of our project would be to derive actual neighbourhood rules, the resolution of 25 meter cells may not be the most convenient for studying neighbourhood interaction between land-use cells. In other neighbourhood studies a spatial resolution of 50-500 meter has been used, and of this range the 100 m resolution have been used most often (Engelen 2002:5, Hansen 2008, Verburg et al. 2004:674, Geerman et al. 2007:554). Using small the 25 resolution grids, require the use of larger neighbourhoods, which increases the execution time considerably (Engelen et al. 2002:5-6). However, since we focus on developing a tool with which neighbourhood characteristics can be quantified, the scale of the test data is not of major importance. We also recognize the probable need to improve the land-use data thematically and temporally, if actual neighbourhood rules were to be derived from it. The use of land-use representing time series of subsequent years has been highly recommended by several authors (Verburg et al. 2004a:687, Hansen 2008). If you would have access to data representing the land-use of subsequent years, you could make the neighbourhood analysis based on actual yearly changes and on the neighbourhood interactions of each year. This would of course give much more reliable results, than if you are using data of a temporal resolution of ten years and where the neighbours have had time to change several times.

A way to improve the data quality and thematic and temporal resolution to better fit the purposes of neighbourhood analysis, you could develop a method to update of the original land-use dataset(s) using the building and dwelling register. In the process, you could use the information about the annual increase of buildings, their building type and size. In this way you could for instance develop a good consistent way to extract service buildings from the aggregated service and industry class. This would be very desirable, since the repulsive and attractive effects service and industry have on itself and on other land-use classes may greatly vary. However, since the extent urban land-use classes do not directly correspond to the building and dwelling register, as we earlier demonstrated, but also to image interpretation, this may not be a straightforward task. Unofficial documents of the construction methods of the Finnish CORINE LC contains valuable information on how much each building type of BDR have been expanded, or in other words how much a the single buildings have been exaggerated in order to

represent it's actual area in a two-dimensional space. Perhaps by using this information we could develop a method to extract the service class and to make an urban data set for subsequent years. The suggested method was briefly tested, but we realized that the suggested method development need more focus than there was time for in this project.

7. Software development

7.1. The structure of the programme

In the first phase of the software development, a conceptual model of the programme was developed (figure 14). The conceptual model explains which parts and processes the model should consist of. How to process input data and how to derive neighborhood rules from the output is not included.

The programme should be able to find changed cells between land-use in "year 0" and land-use in "year 1" (Calculate Changes). The changes should then be classified according to the type of change we are interested to look deeper into (Reclassify). In the case of urban sprawl and expansion, it is most interesting to analyse the neighbourhoods of where new residential, industry or summer cottage cells have emerged. In the following the amount of a particular cell-type (e.g. residential, industry) within a specified neighborhood will be calculated (Calculate Focal Sum). Focal Sum is a function in ArcGIS Spatial Analyst, which enable you to do this. Next, the neighborhood enrichment factor for all the neighborhood pairs (e.g. residential-residential, residential-industry) should be calculated, on cell-to-cell basis (Calculate Enrichment). Based on the detailed neighborhood enrichment data a general statistics table (Calculate Statistics) and a mean enrichment factor raster can be produced (Calculate Mean Enrichment). The base 10 logarithm of the mean enrichment factor will be calculated (Calculate log10). The programme should loop trough all user-specified neighbourhoods and make the described output tables and rasters for all neighbourhoods. Next you should run the model on data from the following years, "year 1" and "year 2", instead of "year 0" and "year 1". The data the model produces should be usable to evaluate the effects of using different neighbourhood configurations and to make scientifically founded neighborhood rules.

Due to lack of time, focus was limited down to implement this conceptual model for the neighbourhood effect of existing residential areas/cells on new residential areas/cells. The same principles can be adopted for the rest of the land-use types of interest. Of course the Reclassify and Focal Sum functions have to be altered according to the land-use pairs of interest. It is a question of taste if you would like to keep the analysis

separate for separate urban classes or if you would like to combine them to a separate script.

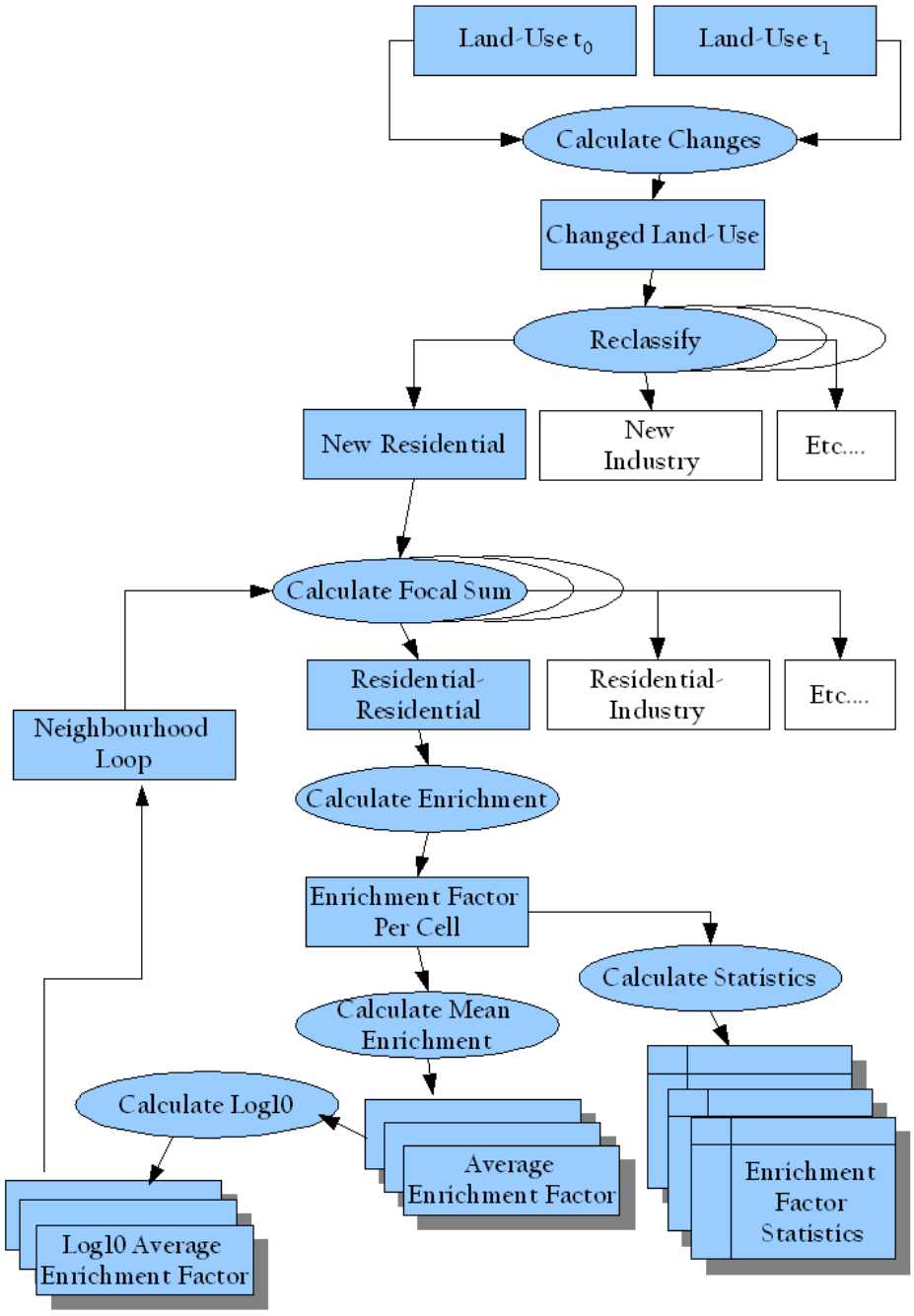


Figure 14. A conceptual model of the programme structure

7.2. Choosing software

In the beginning we thought of two options as programming environment; Octave and Python. The first alternative was to develop the program in Octave, the open-source equivalent to Matlab. Ethically, it would have been more convenient to make a program using Octave. Octave is a programme under GNU licence and the use of it is therefore free of charge and it is easy to share with any one. In this option the raster data would have been converted into text-based matrices and the calculation would have been based on simple mathematics and the programme would therefore have been very fast and effective.

The other alternative was to use Python, which is also a GNU licensed programming environment. However, if ArcGIS functionalities, such as Spatial Analyst functions, are used in your Python script, you of course need to have the corresponding ArcGIS functionalities installed on your computer. This requires an expensive license. Due to the time limits of the project and due to our deeper knowledge of the implementation of ArcGIS functionalities in Python, than of the use of Octave, Python together with ArcGIS were chosen as the programme development environment. This choice went hand in hand with pre-specified requirements. The benefits of the Python and ArcGIS constellation is that it enables an easy integration of raster and vector GIS-data and that it makes visualising your results a straightforward task. It also means that the tool can be developed on using Map Algebra functions, which are easy to understand for users of GIS. Earlier efforts to quantify the neighbourhood interaction have been developed using C ++ (Verburg et al., 2004a:672) and Delphi (Hansen 2008), so in that sense use of Python together with ArcGIS will bring forward a new kind of an approach.

7.3. The Python script

The goal is to make a programme to support the estimation of neighbourhood influence between land-use classes. For this an ArcGIS-based Python script was created.

The enrichment factor is regarded as a useful measure for calculating neighbourhood interaction. In order to calculate the enrichment factor we need to know:

1. the number of cells that that have the value 1100 within a certain neighbourhood
2. the total amount of the cells in the neighbourhood
3. the total amount of 1100 cells within the area to be analysed
4. the number of land-use cells that are being analysed

The total amount of land-use cells needs to be inserted in the Python script by the user of the script. The rest of the numbers, the script calculates. Based on these values, the script generates the cell-specific enrichment factor, the mean enrichment factor, the log10 mean enrichment and statistics regarding the enrichment factor. This information is calculated for each neighbourhood size that has been pre-specified by the user.

The script requires the following input data:

- the original land-use
- the land-use of a later year
- neighbourhood text files, defining the extent of the focal analysis

and additionally the following input parameters:

- the number of land-use cells to be analysed
- the cell size
- the input, temporary and output folders

All input, temporary and output map data are in ERDAS IMAGINE (.img) raster format.

In the following we'll go through the developed script step by step, bringing forward which things need to be changed if the script was to be applied for another area using other input data or for testing other neighbourhood relationships. On the way will also point out different kinds of findings and try to visualize what processes the script is actually carrying out.

```
# ---------------------------------------------------------------------------
# CalcNeigbMetrics.py
# created on: on may 1 2008
# Lena Hallin-Pihlatie
#
# A programme for estimating neighbourhood characteristics
# ---------------------------------------------------------------------------
```

This first part of the script serves as an introduction. Everything that is after the sign # will not be interpreted by Python as code, but as additional comments. First the name of the script is specified (CalcNeigbMetrics.py). Secondly the script contains a date when this script was created. In this case the date expresses when the script was elaborated the last time (May 1st 2008). The author of the script is Lena Hallin-Pihlatie.

```
# Import system modules
import sys, string, os, arcgisscripting
```

The actual script starts by importing system modules. Arcgisscripting makes it possible to use ArcGIS 9.2's functionalities.

```
# Create the Geoprocessor object
gp = arcgisscripting.create()
```

To make Python able to actually use the tools of the ArcGIS 9.2 geoprocessor, several additional steps have to be carried out. First the Geoprocessor object needs to be created.

```
# Check out any necessary licenses
gp.CheckOutExtension("spatial")

# Load required toolboxes...
gp.AddToolbox("C:/Program Files/ArcGIS/ArcToolbox/Toolboxes/Spatial Analyst Tools.tbx")
gp.AddToolbox("C:/Program Files/ArcGIS/ArcToolbox/Toolboxes/Data Management Tools.tbx")
gp.AddToolbox("C:/Program Files/ArcGIS/ArcToolbox/Toolboxes/Conversion Tools.tbx")
```

Next, since the Spatial Analyst extension is required to carry out most of the raster-based analysis, the availability of the extension Spatial Analyst is checked. Finally the toolboxes the script is using need to be loaded. It should be mentioned that, the path to these Toolboxes may be different on another computer.

```
# Allow output to overwrite
gp.OverwriteOutput = 1
```

The script will be run several times. There is a need to be able to automatically overwrite the output data, both in the iterative construction phase but also when running the finalized script itself. The loop is constructed so that it creates temporal data sets that will be overwritten during the next loop. This will be pointed out in more detail later.

The local variables and the geoprocessing environment specified below need to be changed in order to use this script on another data of perhaps another resolution. The lines that need to be synchronized according to user needs are marked in red.

```
# Set the Geoprocessing environment...
WS = "C:/Data/"
gp.Extent = WS + "clc_1990_25m_eilaaj_classes.img"
gp.CellSize = 25

# Set the number of cells in the input
CountTotal = 7840000

```

Next the environment, where the script is to be run, is chosen. All input data and results will be fetched and put in folder "C:/Data", unless another folder is specifically pointed out. For flexibility the Workspace is given the variable WS. The WS variable is used when setting the extent, the maximum and minimum x and y-coordinates of lower left and upper right corner. In this case it is the extent of the data set "clc_1990_25m_eilaaj_classes.img", which is being used. All analysis in this Python script will be carried out within the limits of the extent, and all output data sets will be of the same extent. The cell size should go hand in hand with the input data, which in this case have a resolution of 25 times 25 meters. A precondition for the script to work is of course also that the data sets to be analyzed have to be in the same coordinate system, covering the same extent. If the study area does not have a rectangular form, which can be specified by the extent, you might want to specify which cells to be included in the calculations with gp.Mask = WS + dataset_name. This gives you flexibility to test different kinds of mask for the same data. A mask can be useful for instance, if you want to study if the enrichment factor is different in a region or a municipality than in another.

You also need to give the amount of cells that are included in your study area. In this case there are 7 840 000 cells. This figure is used in the calculation of the enrichment factor later on in the script.

```
# Local variables...
# Input data ..
LandUse1990 = WS + “clc_1990_25m_eilaaj_classes.img”
LandUse2000 = WS + “clc_2000_25m_classes.img”

# Temporary data
Changes = WS + "changes.img"

# Temporary data, overwritten in each loop
ResidentialSum = WS + "Temp/residential_sum.img"
ResidentialResidential = WS + "Temp/residential_residential.img"
CountNeighbourConstant = WS + "Temp/CountNeighbourConstant.img"
test1 = WS + "Temp/test1.img"
test2 = WS + "Temp/test2.img"
test3 = WS + "Temp/test3.img"
test4 = WS + "Temp/test4.img”
Constant = WS + "Temp/constant.img"
```

In the following, the main part of script the local variables are set. Variables are very useful, since it is much easier to write a Python script using variables than whole file names. Only two input data sets are needed, representing the land-use at two times and these are marked in red. The temporary data sets are overwritten by each loop. As it is now, you need to go into this script and change the name of the land-use data to be compared each time you want to run it for new data sets. This could of course be improved, for instance by using Tkinker.

```
# Make a comparison of the changes between the input land-use images
gp.Diff_sa(LandUse2000, LandUse1990, Changes)
gp.Reclassify_sa(Changes, "Value", "1100 1100 1; 1112 5550 0" , ChangeResidential, "NODATA")
```

Next, the changes between the land-use data will be extracted. A comparison is made between the data sets LandUse2000 and LandUse1990, using the Statistical Analyst (Diff_sa) tool. This local tool compares the cells of the first data set (LandUse2000) with the cells of the second data set (LandUse1990) and returns the cells that are different.

In this particular version of the script we are only interested in which cells that are new residential cells i.e. having the Value 1100. Therefore the residential cells (1100) of the Changes data set is being reclassified to one (1), and the rest is classified to zero (0). Of course if you are interested in the changes of summer cottages, the corresponding classification has to be done, by exchanging "1100 1100 1; 1112 5550 0" with “1421 1421 1; 1100 1420; 0 1422 5550; 0”

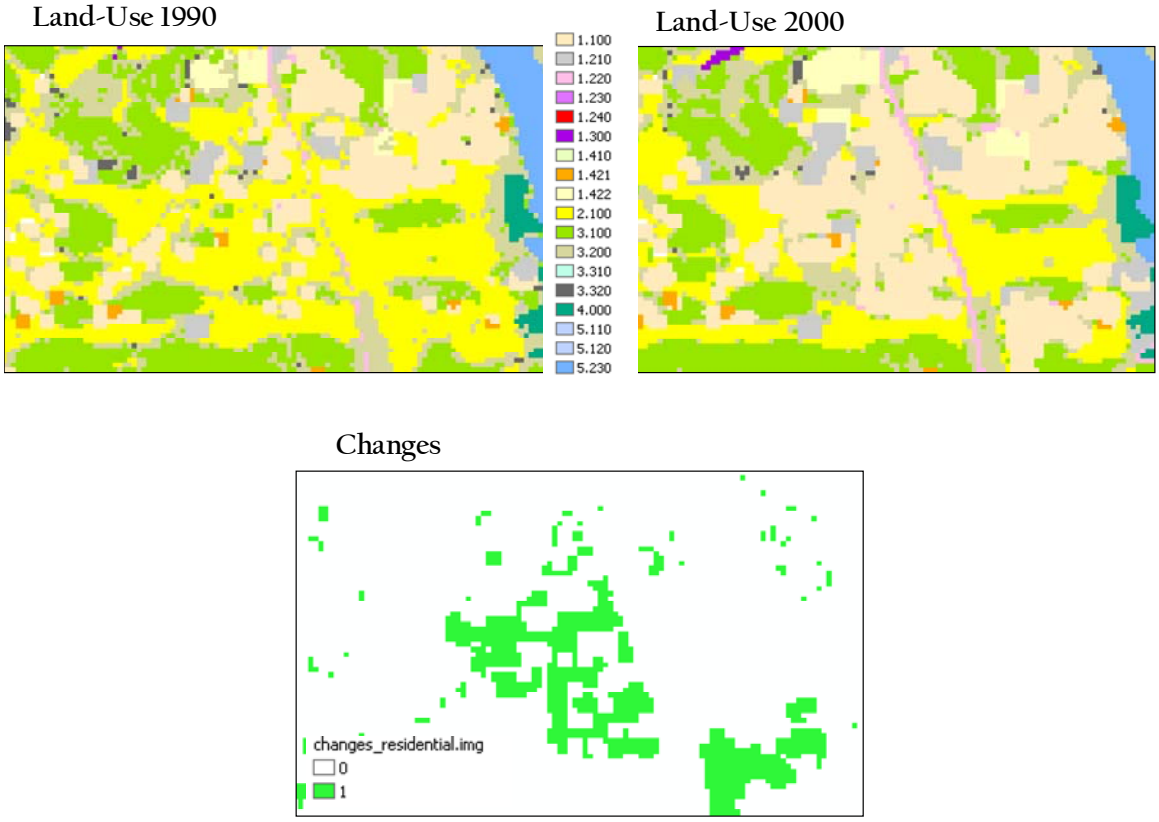


Figure 15. The new residential areas that have emerged between year 1990 and year 2000.

To check that the Diff_sa function captures the right changes, and that the data sets can be compared, the data set where the new residential cells have value 1 and the rest value 0 (ChangeResidential) was multiplied with the original dataset from year 1990 (LandUse1990) in the Raster Calculator of Spatial Analyst: *changes_residential.img* * *clc_1990_25m_eilaaj_classes*. We found that 7 720 659 cells were unchanged, 34 256 have changed into residential cells from arable and semi-natural lands (2100), 52 727 have changed from forest (3100), 30 385 from grasslands and shrubs (3200), 1 966 from bare rock (3320) and 8 mire cells (4000). It seems that the Diff_sa function captures the changes of this data. Also figure 15 supports this statement. However, when we tested this same comparison with resampled 100 cell data, inconsistencies were found: other urban cells had changed class into residential, which is at least to some extent is unlikely to have happened. If you are interested in changes in the urban fringes and for instance

not in changes in proximity of stand-alone cells in the countryside, this Diff_sa function is not enough and should be changed.

```
# Set initial conditions for the neighbourhood loop
NumNeighbourhood = 5                                    # You need to know the number of neighbourhoods

# Start neighbourhood loop
Neighbourhood = 1
while (Neighbourhood <= NumNeighbourhood):
  gp.AddMessage("Neighbourhood:")
  gp.SingleOutputMapAlgebra_sa("FocalSum((" + LandUse1990 + " == 1100), IRREGULAR, " + WS +
"Neighbourhood_Circular_" +str(Neighbourhood) + ".txt, DATA)", ResidentialSum)
```

You need to specify how many neighborhoods you want to use. The programme will iterate the rest of the script for all pre-specified neighbourhoods, where you expect that some neighbourhood interaction can take place. Iteration means to repeat a process and is sometimes referred to as looping. Iteration is a key concept in most programming languages enabling you to execute a process over and over using different data in each iteration. In this case the different data is the neighbourhood text files referred to as " + WS + "Neighbourhood_Circular_" +str(Neighbourhood) + ".txt

The number of neighbourhood specifies how many times this script should loop through, before being finished. NumNeighbourhood = 5 means that we are using five neighbourhoods. This value has to be changed according to the amount of different neighborhoods used.

The initial value of the Neighbourhood = 1. The script will run as long as the condition for the While loop "while (Neighbourhood <= NumNeighbourhood): " is true or in this case five times. The commands that are to be carried out within the loop are all intended. In the Python script, the processes to be carried out fit on a line, while in this document the line continues unintended on the next line.

Next, we calculate the amount of cells that is residential cells (i.e. has the value of 1100). The calculation is made for each cell. We use the data set that represents the initial state or in this year the land-use of year 1990. The function FocalSum, sums up (1 + 1 + 1 and so on) the amount of 1100 that exist within a specified neighborhood. This procedure can be demonstrated with a simple picture (figure 15).

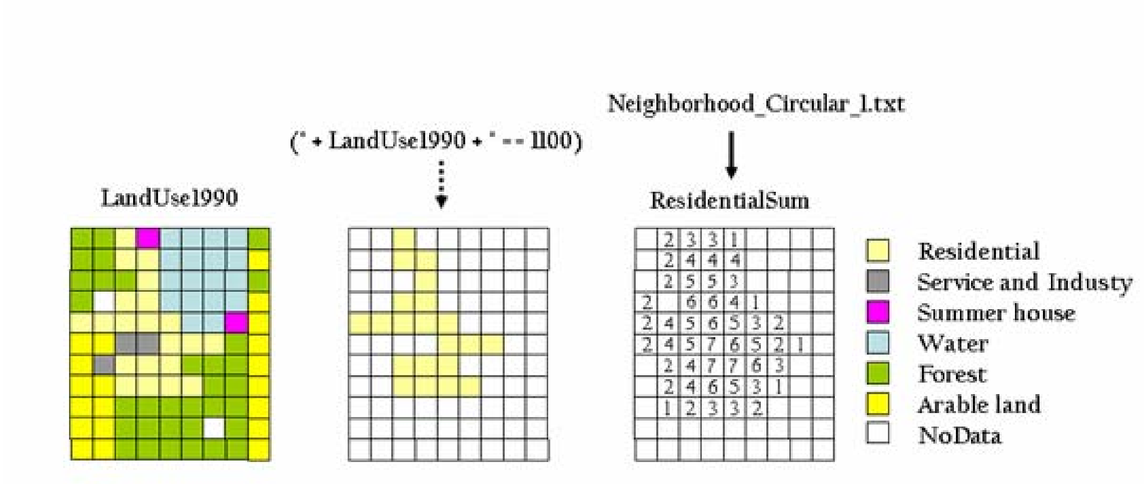
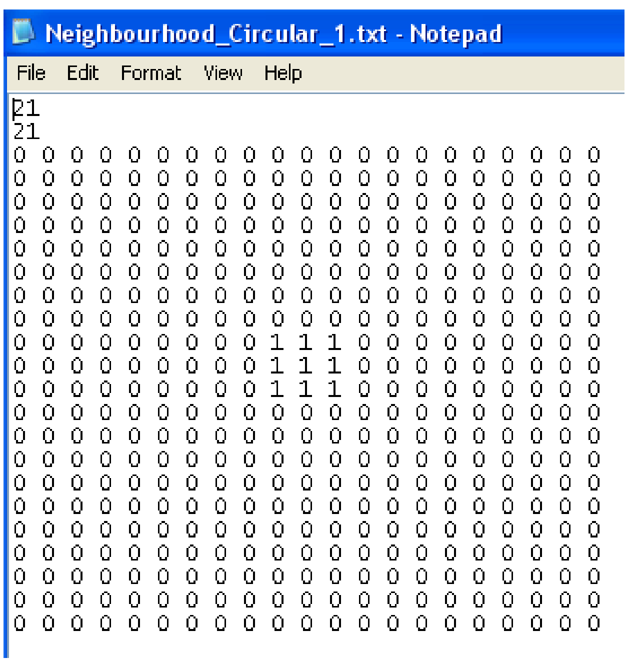


Figure 16. An example of: gp.SingleOutputMapAlgebra_sa("FocalSum((" + LandUse1990 + " == 1100), IRREGULAR, " + WS + "Neighbourhood_Circular_" +str(Neighbourhood) + ".txt, DATA)", ResidentialSum)




```
21
21
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 17. The neighbourhood text file used in figure 16.

The neighborhood shape is defined by a text-file, where the neighborhood has value 1 and the rest value 0. The neighborhood can have any shape as long as its name is in line with the loop number. str(Neighbourhood) stands in this case for 1 and the whole pathname of the first neighborhood file is: C:/Data/Neighbourhood_Circular_1.txt. In the

second loop number the text file ending with “2.txt” is used to specify the extent of the neighborhood and so on.

In the example, figure 16, the impact of the edges on the analysed result is big. However, in a real situation the share of the edge is much less. To eliminate this problem, you could use two masks. In this case a mask where the area of the study area has been buffered to include cells for example on the other side of a municipality border in the FocalSum calculation. For the next steps of the script another mask can be used, covering only the actual study area, not taking into account the cells outside the municipality border anymore.

```
# Get the FocalSum values for new residential areas only
   gp.SingleOutputMapAlgebra_sa(ResidentialSum + " * " + ChangeResidential, ResidentialResidential)
```

We are not interested in the FocalSum values of all cells, but only in the values of those cells that are new residential cells in year 2000. This is why the ResidentialSum is being multiplied with ChangeResidential, where the cells that have changed into a residential cell (Value =1100) have the value 1 and the rest have the value 0. The affect of this multiplication can be seen by comparing figure 18a and 18b.

In figure 19 you can compare the location of the new residential cells with the urban cells of a land-use raster from 1990. Many new residential cells have emerged in the vicinity of existing urban, and especially residential, cells. The FocalSum function, that has been carried out, captures this phenomenon very well.

Figure 18a. An example of the FocalSum values of ResidentialSum raster. A neighbourhood of 61 cells have been used.

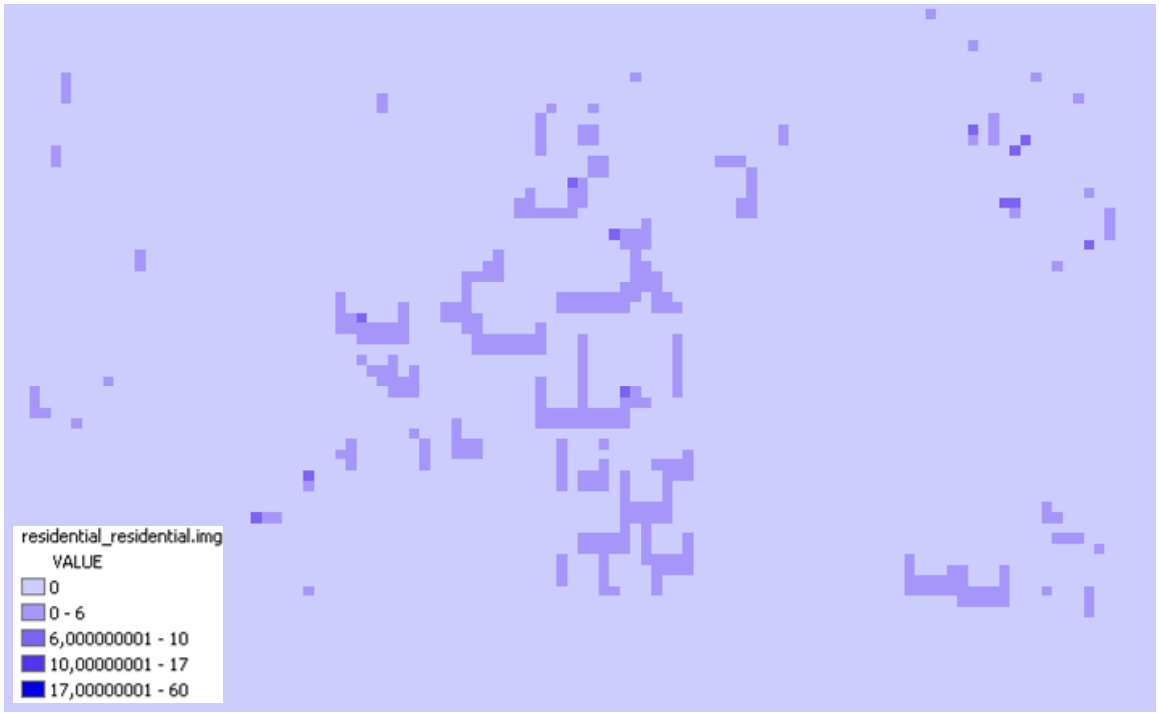


Figure 18b. An example of the ResidentialResidential raster, which includes only the previous FocalSum values of new residential cells . A neighbourhood of 61 cells have been used.

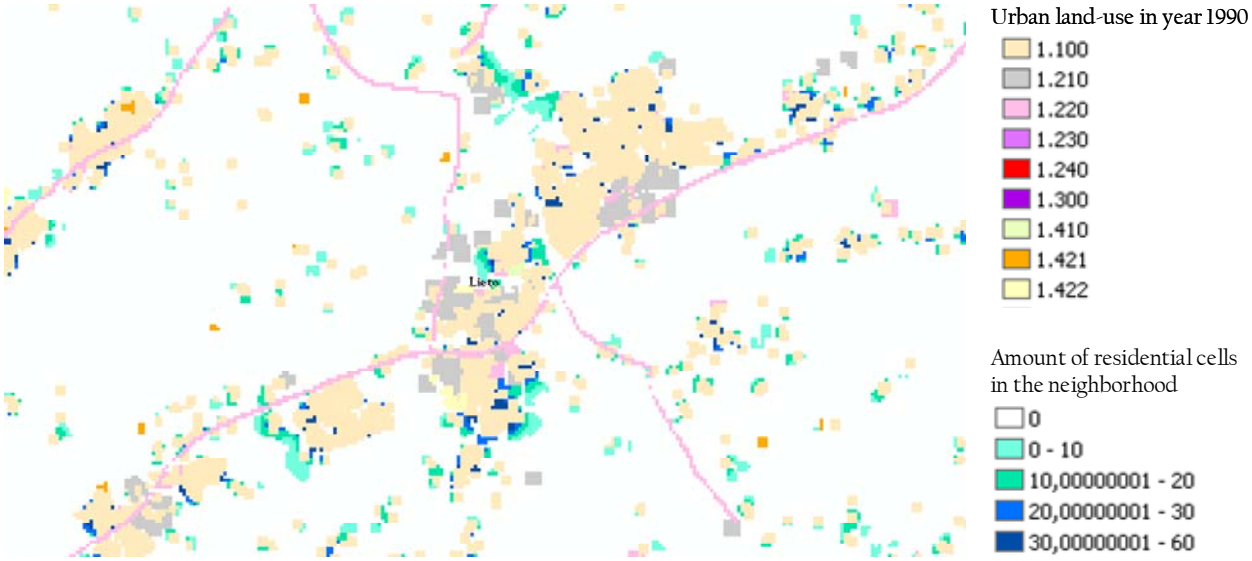


Figure 19. The figure shows the amount of residential cells in the vicinity of new residential cells and their location in relation to the urban land-use of year 1990. The cell number of the neighbourhood is 61. The area is 7 kilometres wide.

The next steps show how the values needed for the enrichment factor are being derived and how the enrichment factor output data is generated.

```
# Derive the enrichment factor for each cell for the neighbourhood
# Find out the amount of 1100 cells in the LandUse1990 table
# Create search cursor
   rows = gp.SearchCursor(LandUse1990, "Value = 1100")
   row = rows.Next()
   while row:
      Count1100 = row.GetValue("Count")
      row = rows.Next()
   del row
   del rows
   print(Count1110)
```

The enrichment factor is calculated according to formula (1), described in detail in chapter 5.5. In order to calculate the enrichment factor we need to know in this case:

1. the number of cells that that have the value 1100 within a certain neighbourhood size

2. the total amount of the cells in the neighbourhood

3. the total amount of 1100 cells within the area to be analysed

4. the amount of land-use cells that are being analysed

We already know the amount of cells that have the value 1100 within a certain neighbourhood, the "ResidentialResidental" data contains that information. We also know that the dataset to be analysed contains 7840000 pixels. We now need to find out the amount of residential cells (Value = 1100) in the land-use data that represents the initial state. First we search the Value 1100. The GetValue method returns a field's value, with the field name being the only input parameter. When found we can get the number of residential cells from the Count-field and use it later as the variable Count1100. The figure is printed to the Python Shell with print (Count1100) so that the user can follow and double check the calculations. For table 3 the Count1100 is 302 560. This number stays the same during the whole looping process, so you could save processing time by putting it outside the loop, for example after specifying the amount of cells to be analysed. If we were interested to study the effect of another class, the same procedure could be carried out to obtain e.g Count1421 by finding the value of the summer cottage class, 1421.

Table 3. Example of LandUse1990.

Value	Count
1100	302560
1210	60675
1220	60611
1230	3529
1240	2417
1300	12256
1410	2271
1421	90408
1422	7988
2000	3287
2100	1716586
3100	2547196
3200	889997
3310	20
3320	81590
4000	115966
5110	12963
5120	24290
5230	1905390

```
  # Calculate the number of cells in the neighbourhood
  gp.ASCIIToRaster_conversion(WS + "Neighbourhood_Circular_" +str(Neighbourhood) +
"_to_image.txt", WS + "/Temp/Neighbourhood_" +str(Neighbourhood) + ".img")
  rows = gp.SearchCursor(WS + "/Temp/Neighbourhood_" +str(Neighbourhood) + ".img", "Value = 1")
  row = rows.Next()
  while row:
      CountNeighbour = row.GetValue("Count")
      row = rows.Next()
 del row
 del rows
 gp.CreateConstantRaster_sa(CountNeighbourConstant, str(CountNeighbour), "Float")
 print(CountNeighbour)
```

In order to calculate the enrichment factor, we also need to know the size of the neighborhood used. This number is different for each loop. We start by converting a text file representing the neighbourhood into an image. The text file used for making a raster is somewhat different from the text file used in the previous calculation. In addition to the row and column numbers, the x and y-coordinates of the lower left corner need to be specified and of course within the extent set earlier on in the script. The cellsize and the value of NoData also have to be included. The "cells" representing the neighbourhood have the value zero, while the rest have value -9999, the value of NoData (figure 20). The value of the rest could also be 0 – it wouldn't make any difference. When we know the "cells" of the neighborhood has the value 1, we can search for them and get their amount (CountNeighbour) the same way as in the previous case.

Next, we can create a raster (CountNeighbourConstant) from the CountNeghbour value, by changing it to a string object (str(CountNeighbour)). It automatically generates a raster, in the cell size specified in the beginning of the script. We choose to make it into a floating point value (Float), but unfortunately the constant raster, but is still made into integer. There seems to be a bug here.

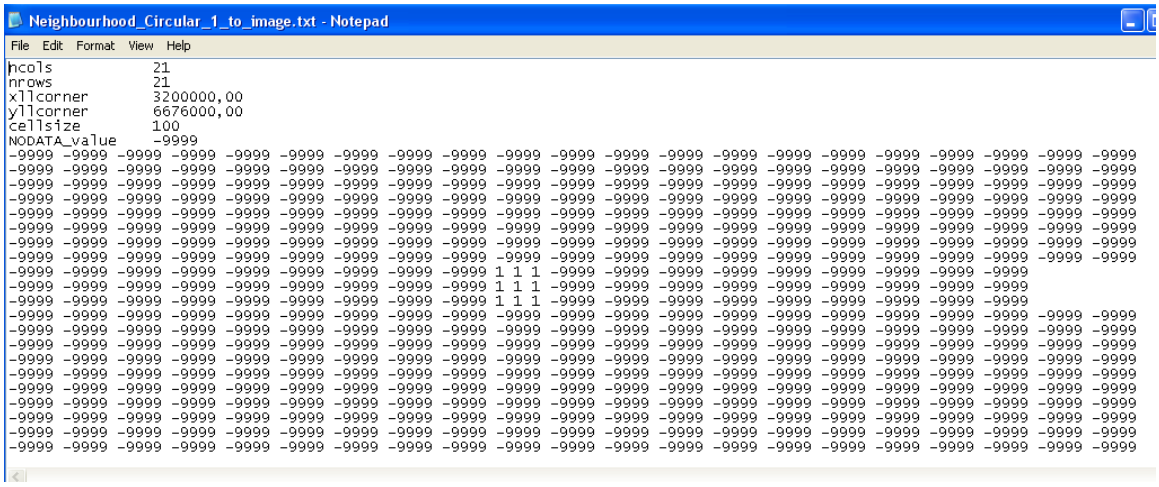


Figure 20. An example of a neighbourhood text file that can be used for making a raster.

As referred to earlier, the numbers to be used for the enrichment factor calculations printed by for example the command print(CountNeighbour), so that you can follow the looping process in the Python shell (figure 21). This might also be useful to check that the input values to the following enrichment factor calculation are correct.

```
>>> ===========
>>>
7840000
90408.0
9
90408.0
13
90408.0
25
90408.0
41
90408.0
61
>>> ===========
```

Figure 21. The values of each loop are printed in the Python shell.

```
gp.Times_sa(ResidentialResidential, 1000, ResidentialDivide)

# Calculate the Neighbourhood Enrichment Factor
gp.Divide_sa(ResidentialDivide, CountNeighbourConstant, test1)
gp.SingleOutputMapAlgebra_sa(str(CountTotal) + " / " + str(Count1100), test2)
gp.SingleOutputMapAlgebra_sa(test1 + " * " + test2, test3)
gp.Divide_sa(test3, 1000.00, test4)
```

Divide_sa is a local function, where the values of the first raster is divided with the values of the second raster on a cell-by-cell basis According to ArcGIS help, you get a floating point value if either value you are using in your division is floating point. Unfortunately, neither of the rasters contained floating values (due to the bug), so when we divided ResidentialResidential (e.g. containing values up to 9) with CountNeighbourConstant (e.g 9), the result was returned as integer and in this case all cells in the output raster gets the value 0. In order to solve this problem, we needed to multiply the ResidentialResidential data first, by another local function, Times_sa. In this example the ResidentialResidential is multiplied by 1000 to create the new raster ResidentialDivide.

According to the equation (1) presented in chapter 5.5, we can obtain the enrichment factor by the following divisions and multiplication, creating the temporal datasets test1, test2 and test3. One additional step is added to the calculation, since we need to divide it with 1000 to compensate for the earlier multiplication.

```
# Specifies what dataset the calculation should be put into
OutputEnrichment = WS + "Temp/Enrichment_ResRes_N" + str(Neighbourhood) + ".img"

# We are interested only in the situation of changed 1100
gp.SingleOutputMapAlgebra_sa("Con((" + Changes + " == 1100)," + test4 + ")", OutputEnrichment)
```

Now we know the enrichment factor on a cell-to-cell basis and can assign the name of the output files for it to be saved in. The number of the loop is used, so that it will not be overwritten by the following loop. The condition (Con) ensures that only the value of the cells from test4 that have been changed into residential cells, will be put into the output raster (OutputEnrichment). In figure 22 you can see that actually many new residential cells have the value zero.

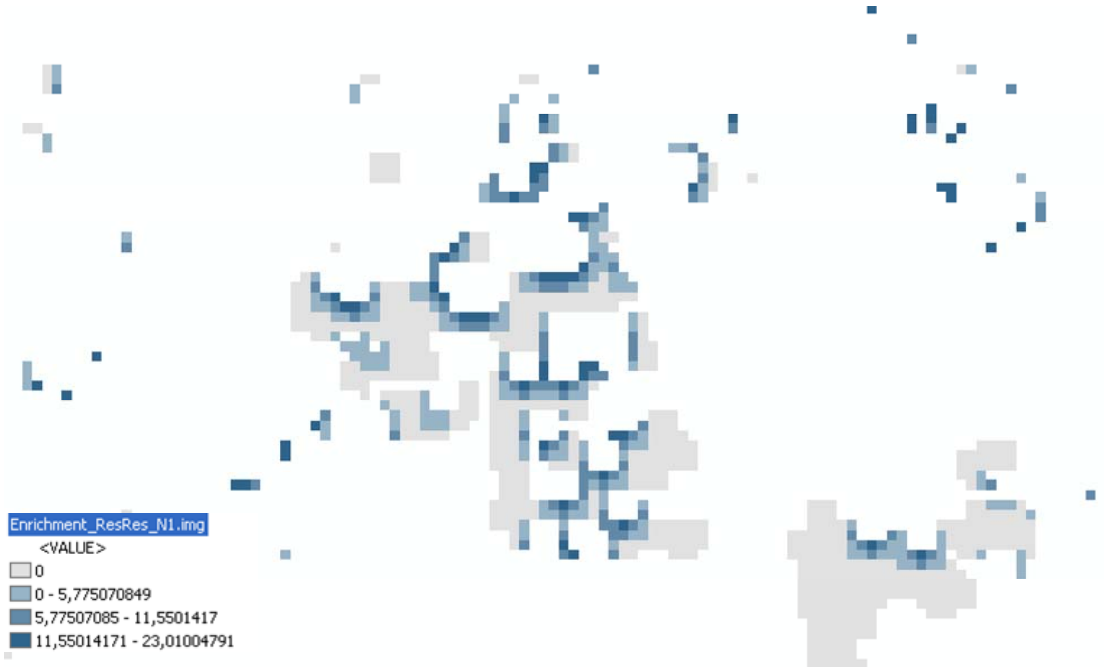


Figure 22. The enrichment factor for new residential cells using the first neighbourhood (Moore) presented in figure 17.

```
# Calculate the average enrichment factor and put the result into a table

gp.Reclassify_sa(Changes, "Value", "1100 1100 1" , Constant, "NODATA")
OutputTable = WS + "AveEnrichment_ResRes_N" + str(Neighbourhood) + ".dbf"
gp.ZonalStatisticsAsTable_sa(Constant, "Value", OutputEnrichment, OutputTable)
OutputImage = WS + "AveEnrichment_ResRes_N" + str(Neighbourhood) + ".img"
gp.ZonalStatistics_sa(Constant, "Value", OutputEnrichment, OutputImage, "Mean", "Data")
```

After we have the enrichment factor for each cell, we can now calculate the average enrichment factor according to equation (2) in chapter 5.5. We want to make a zonal function, where the zone consists only of the new residential cells. A separate raster, Constant, is made by a reclassification function to represent this zone. In this way the cells that have the enrichment value zero is also included in the calculation. When the output table name has been specified, we can calculate the zonal statistics as a table (ZonalStatisticsAsTable_sa), using the Constant raster as the zone and the OutputEnrichment file as the input data. The output table contains the minimum and maximum values, the value range, the mean value, the standard deviation value and an aggregated sum of the enrichment factor values, within the specified zone (table 4).

Table 4. An example of the output of the ZonalStatisticsAsTable function.

Attributes of AveEnrichment_ResRes_N1

OID	VALUE	COUNT	AREA	MIN	MAX	RANGE	MEAN	STD	SUM
0	1	119342	74588800	0	23,01	23,01	3,62025	5,41793	432048

We also want to make a raster representing the average enrichment factor. We can do it by first specifying an output file name, OutputImage, and by then using the ZonalStatistics_sa function, where it is specified that we want to calculate the mean value and that the calculation will be based only on data, not on possible NoData values.

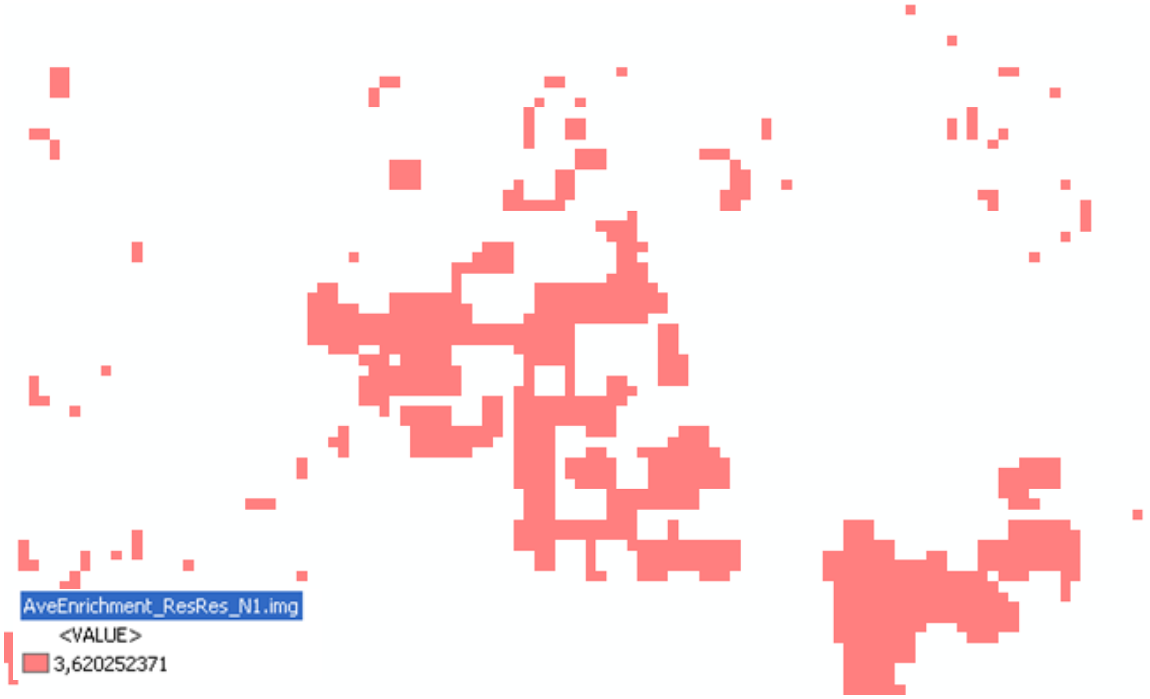


Figure 23. An example of the average enrichment factor.

```
# Calculate base 10 logarithm for the Enrichment factor and the Average Enrichment factor

    gp.Log10_sa(WS + "Output/AveEnrichment_ResRes_N" + str(Neighbourhood) + ".img", WS +
"Output/AveEnrichment_ResRes_Log_N" + str(Neighbourhood) + ".img")
```

In most cases the mean enrichment factor is presented as the base 10 logarithm. When the base 10 logarithm is used, small values, where the residential class is underrepresented and have an average value under 1 will be assigned negative values.

This is convenient, because it shows that instead of being a positive autocorrelation, there is actually a negative one. We therefore decided to include the calculation of the base 10 logarithm in this script.

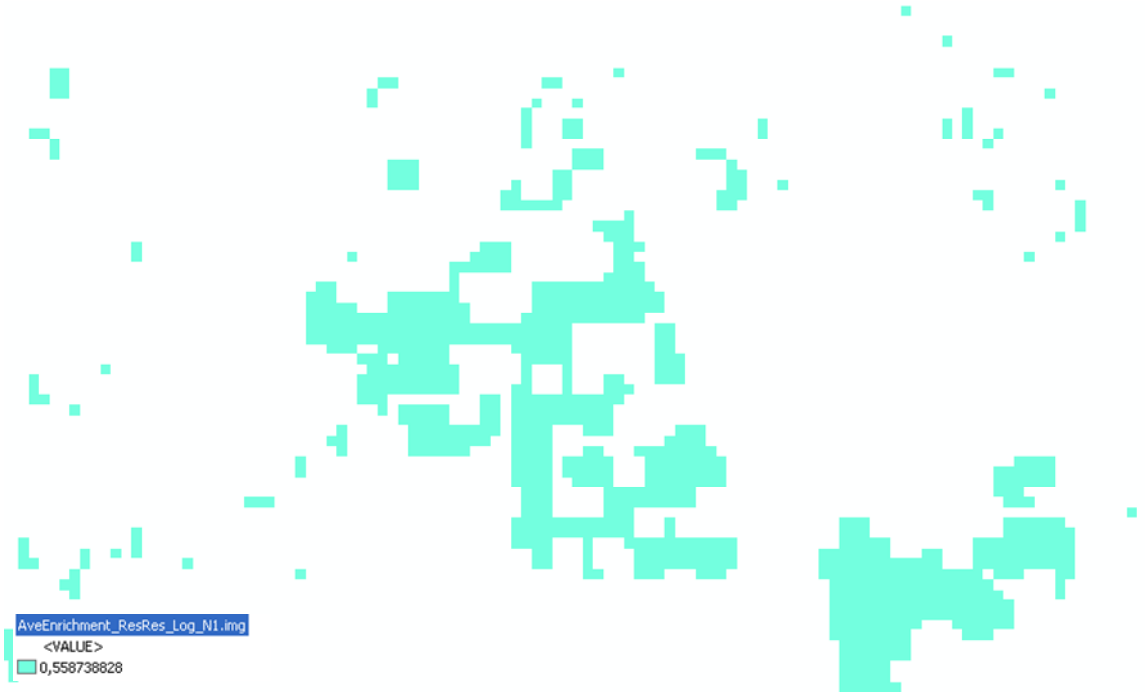


Figure 24. An example of the base 10 logarithm average enrichment factor, using the input data of figure 23.

```
Neighbourhood = Neighbourhood + 1
```

This marks the end of each loop and also the end of the programme after the final loop.

By this stage you have three separate rasters as a result of each neighbourhood loop. In the following you can combine the rasters with the *combine* function of Spatial Analyst, for combining the neighbourhood wise data sets to three combined raster datasets, containing the attribute values of all neighbourhoods. Unfortunately the function did not completely meet our needs – it turns all values into integer, eliminating all valuable variations. However, by first multiplying the values by for instance 1000 in the similar manner as earlier, this problem can be solved.

7.4. Test run

In the following we will present some test results, calculated by the script. We calculated the enrichment factor for new residential areas, both regarding existing residential cells, as in the script above, and for existing summer cottages within the eleven municipalities pointed out in the *Data* chapter. Five neighbourhoods were used (figure 25). The neighbourhoods are based on aggregation, so that the first neighbourhood contains the cells with number 1, the second neighbourhood contains the cells of number 1 and number 2, and so on. Since the cell size is only 25 meter, we can expect no radical changes in the enrichment factor within this the maximum distance, used here.

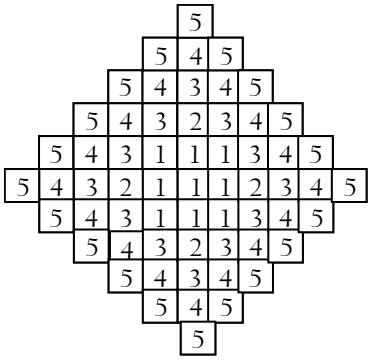


Figure 25. The extent of the neighbourhood used in the analysis.

All results earlier presented focus on the average enrichment factor and its base 10 logarithm. However, since a cell-to-cell based enrichment factor raster is generated using this programme, we take a look at these rasters first.

Table 5. The variation of the enrichment factor on a cell-to-cell basis

Neighbourhood	N1	N2	N3	N4	N5
Enrichment Factor: Residential-Residential	0-14.59	0-15.16	0-15.77	0-16.02	0-16.15
Enrichment Factor: Residential-Summer Cottage	0-51.98	0-49.52	0-46.83	0-42.79	0-43.14

It does not seem logical that the enrichment factor of residential-summer cottages is higher than the enrichment factor of residential-residential (table 5). However, we found the explanation of this. Of the study area, 6.09 % is covered by residential areas and 1.71 % by summer cottage areas. The relative low share of the summer cottages influences the calculations, so that s neighbourhoods that included summer cottages therefore get a higher value of over-representation, than if the same neighbourhood would contain the same amount of residential cells. This is problematic and has not been revealed in studies where only the average enrichment factor has been addressed.

The average enrichment factor is lower for summer cottages than for residential areas for the zone of new residential cells (table 6). There are relatively few high enrichment factor values in the cell-to-cell based data set. The broader variation of enrichment factor values can also be seen as higher standard deviation (std) values for summer cottage areas. Despite the higher range of values and higher variation of enrichment values, the average enrichment factor seems to describe the overall situation well: the neighbours of new residential cells are overrepresented by existing residential cells, while new residential cells are either under-represented or almost as present as in the study area on average in the neighbourhoods of the new residential cells. The increasing trend in the figures also seems logical, since it is more likely that there is a residential cell over 50 meters away, that really close to existing residential or summer cottage cells. It is not needed to build buildings as close as 25 meters to each other, in a country like Finland, where there is enough space for every one.

We would have expected the mean enrichment and the base 10 log values to have been higher for the residential cells. Verburg et al. (2004:676) obtained the mean enrichment value of 7.9 in the Moore neighbourhood of residential cells, when using land-use data from year 1989 and 1996. Compared with the Netherlands, the building activities in Finland are in general more dispersed and less controlled, explaining some of the difference. The use of 500 meter cells in comparison with 25 meter cells in this study, also explain part of the difference. In our study no single standing houses were eliminated

through aggregation or resampling methods. In comparison, the data used in the Dutch study, was aggregated from 25 m to 500 m cells according to the majority rule, diminishing the role of dispersedly situated settlements.

Table 6. The average enrichment factor for new residential cells in relation to residential cells and summer cottage cells in the 1990 data. Information of the observed standard division is also included.

Neighbourhood	N1	N2	N3	N4	N5
Average Enrichment Factor(std): Residential-Residential	2,37879 (3.55)	2,50902 (3.37)	2,69285 (3.15)	2,73605 (3.00)	2,7045 (2.87)
Average Enrichment Factor(std): Residential-Summer cottage	0,867761 (4.36)	0,979208 (4.31)	1,16746 (4.24)	1,28651 (4.10)	1,34008 (3.91)

In a visual comparisons, made earlier in this study, we could observe that new residential buildings are generally built in the vicinity of existing ones, in urban areas. However, if there are only a few residential cells in the neighbourhood, as is the situation in many cases outside highly urbanized areas, the result will not indicate over-representation, which evens out the neighbourhood effect of residential areas. The neighbourhood interaction is also partly evened out due to the broad temporal resolution of the data. Our data has a time gap of 10 years, it may in may cases seem that the neighbours are arable land or forest, even though in reality the process of urbanisation have continued in a certain direction cell by cell. To capture these yearly subsequent changes, requires data of a better temporal resolution (Hansen 2008), so to some extent the low mean enrichment values of residential cells were expected as a result of the land-use data used.

The neighbourhood configuration also plays a role. In the study carried out by Verburg et al. (2004:671), Moore neighbourhoods and extended Moore neighbourhoods were used. In this way, a lot more cells where included than in our configuration, for example for the fifth neighbourhood 121 cells in comparison with our 61 cells.

In some studies, the base 10 logarithm values have been found to be as high as 0.9 to 1.3 (Verburg et al. 2004:677-680, Hansen 2008). In a study carried out by Geertman et al. (2007:559) the over-representation near new residential areas for the period 1986-1993 were found to have values between 0.7 and 0.3, depending on the characteristics of the area observed. The latter value is similar to the values obtained in this test run (table 7). There have been no results published on the mean enrichment factor for summer cottage areas, so we cannot make any comparison with results of others regarding them.

Table 7. The base 10 log average enrichment factor of new residential cells in relation to residential cells and summer cottage cells in the 1990 data

Neighbourhood	N1	N2	N3	N4	N5
LogAverage Enrichment Factor: Residential-Residential	0,376356	0,399505	0,430212	0,437124	0,43087
LogAverage Enrichment Factor: Residential-Summer cottage	-0,0616	-0,00913	0,067241	0,190941	0,12713

The overall results of the test run seem logical in relation to the study area and available data. We can conclude that the role of data and neighbourhoods are big, for observing neighbourhood interaction. There also seems to be a lot of regional and scale-related variations. We also found that the enrichment factor is very sensible to the proportion of the observed land-use in the study area as a whole. This sensibility is good to be aware of when deriving neighbourhood rules from the resulting values.

8. Results and discussion

During the scope of this project we have described the components of neighbourhood interaction in land-use dynamics and modelling as it is understood and approached today. We have recognized, that neighbourhood interaction plays a central in urban land-use dynamics and that we can improve the transition rules of land-use models of today, and particularly CA-based ones, by knowing and quantifying the rules of the neighbourhood interaction. Literature indicated that a spatial metrics, the enrichment factor, is an appropriate measure of neighbourhood interaction, what can help in defining the needed neighbourhood rules and so we decided to make a tool with which the enrichment factor can be quantified.

Earlier attempts have been carried out to quantify neighbourhood interaction in the form of the enrichment factor, but no one seems to have used a raster-based Map Algebra approach for doing it. In the process of developing the tool, we ran into challenges and obstacles related to the pre-processing of raster data and to the use of the ArcGIS and Python constellation. The challenges that appeared as Python or ArcGIS bugs could be solved by adding more procedures to the script. This of course makes the programme more time-consuming to run, but hopefully in the long run the advantages of working closely with visualisation tools and easily understandable Map Algebra will win. The speed can also be cut down by using input data of a coarser resolution.

Based on our data evaluation, the land-use datasets were found to have a good spatial and thematic match with each other and with the building and dwelling register. Nevertheless, it was not ideal for analysing neighbourhood interaction. In the report, changes are suggested to the land-use data, regarding its temporal resolution, the thematic classification and the resolution. According to our experience, it is not a simple task to change the resolution of a good-quality temporal land-use data in raster form. We made an attempt in the data-processing phase, since we recognized that the original resolution of the land-use data, 25 meter cells, was not ideal for studying urban land-use dynamics. Despite the fact that we also had access to the building and dwelling register (BDR) and good documentation of the production methods of the land-use datasets, we

could not successfully carry out all the data processing we found relevant; including the deriving of a service class and the production of land-use data for subsequent years. Even though the production of our land-use data was well documented, it was not detailed enough in order to extract a land-use class (e.g. service) from within an existing land-use class (e.g. industry and service). Due to the combination of automatic image interpretation and the use of diverse map data in the production process of the land-use data, it is challenging to produce land-use data representing the situation of subsequent years. The challenges related to data could not be solved within the limits of this project. However, this is not of major importance, since the main focus of the project was to develop a tool that can be applied on more suitable data later on.

We succeeded in making a simple tool in the form of a Python script for quantifying neighbourhood interaction, in the form of the enrichment factor, between land-use classes. By using the tool, the influence of different input data and different neighbourhoods can be analysed in several ways, using the output rasters and tables of the programme. The data you use for deriving neighbourhood characteristics may vary regarding spatial, thematic and temporal resolution and this will affect the resulting enrichment factor. With the tool, you test which kind of data best captures the neighbourhood interaction that you are interested in quantifying.

More specifically, the data set representing changes in a certain land-use type and the enrichment factor raster data set, can be used in for example ArcGIS to visually locate areas where changes has been taken place and possible areas with a big deviations in the enrichment factor. In this way it you can evaluate if your calculations are representative for your whole study area or if you possibly should analyze your data in smaller parts. The DBASE table that contains zonal statistics again can be useful for evaluating the quality of the mean enrichment factor, using information on the standard deviation. You can combine the tables for all neighbourhood sizes to make summarizing graphical presentations. The average mean enrichment factor raster can be used for visualizing purposes. You can compare the mean enrichment factor with the 10 base logarithm mean enrichment factor data and consider which numbers you want to elaborate further into actual neighbourhood rules. According to several authors (Verburg et al. 2004:685), it is

possible to use the calculated enrichment factors to assist in the definition of neighbourhood rules. Unfortunately, we did not have time to test this in practice.

According to literature, surprisingly little focus have been put on the neighbourhood configuration, even though it may have a big effect on the result. This tool may also be used to indicate which neighbourhood configuration is most suitable for capturing urban land-use interaction, since the user of the tool can choose the size and the configuration of the neighbourhood delineation he wants to use, by making simple text files as demonstrated in the previous chapter.

9. Conclusion and prospects

We met most of the goals that were set up in the beginning of our project. We have explained what neighbourhood interaction is, what central components it has, how it can be measured and why it is so important in the field of land-use dynamics and modelling. We have demonstrated in practice that the neighbourhood interaction can be studies using Map Algebra-based coding in Python. We have developed a tool that met most of our expectations. However, how the enrichment factor can be transformed into actual neighbourhood rules for use in for instance a Cellular Automata-based model remain to be discovered. This is anyway better to carry out when data better meet our needs.

The challenges regarding data were surprisingly big and time consuming. A major effort should be put on this in subsequent work. In order to be able to properly evaluate the developed Python script, we need to test it on another data. Preferably this would be land-use data with a classification suitable for capturing relevant interaction between its land-use classes, land-use data of a bit coarser resolution and most importantly land-use data of subsequent years, produced in a consistent way, retaining data compability.

Data is important because only with suitable datasets can we gain true knowledge of the neighbourhood rules of a certain area. Unfortunately, your results are never better than the data you rely on in your analysis, however good a method or tool you have developed. Fortunately people are working on improving the overall data situation continuously and new data sets are being made. For the time being a European Urban Atlas and CORINE LC 2006 are under production. With the implementation of the INSPIRE, other kinds of improvements will also take place. If they are sufficient for the needs of land-use modellers remain to see.

Today, we live in a world where we are surrounded by technology. There are all kinds of tools out there. Even by using desktop GIS you can develop land-use models for making land-use simulations. There are also tools available, which could make it possible to integrate these as a part of participatory decision support systems. Imagine the

participants of a planning process being able to simulate the future alternatives using their own criteria and constraints, and giving the results in the hand of the planner. Technically this is already possible. However, there are still challenges on the way. We need to analyse the phenomena behind land-use change for making empirically justified transition rules to base our land-use models simulating "What if scenarios" on. This project is a small step towards this goal.

List of sources

Barredo, J.L., Kasanko, M., McCormick, N. and Lavalle, C. (2003). Modelling dynamic spatial processes: Simulation of urban future scenarios through cellular automata. *Landscape and Urban Planning*, 64:145-160. Page 145.

Batty, Michael (2005). *Cities and Complexity. Understanding cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press, Cambridge, Massachussets, London, England. Pages preface, 68, 73.

Batty and Torrens (2001). Modeling Complexity: The Limits to Prediction. *Working Paper Series.* Paper 36. Centre for Advanced Spatial Analysis, University College London. 36 pages. Pages 3, 26-27.

BDR (2006). *Building and dwelling register* (in Finnish). A metadata description dated 18.5.2007. The Finnish Environment Institute. Unpublished material.

CLC (2000). *CLC2000-Finland*. Final Report. Finnish Environment Institute (SYKE), Geoinformatics and Land Use Division (GEO), May 2005. Pages 3, 41, 43. Retrieved from http://www.ymparisto.fi/download.asp?contentid=38725&lan=en on May 20th 2008.

CLC (1990). *Production of Land Cover and Change Information for ENVIFACILITATE-project.* Geoinformatics and Land Use Division, Finnish Environment Institute.

Engelen, G., White, R. and Uljee, I. (2002). *The MURBANDY aqnd MOLAND models for Dublin.* Final Report. Research Institute for Knowledge Systems BV. Pages 5-9, 23-25. Retrieved from http://www.geo.ucl.ac.be/LUCC/MODLUC_Course/PDF/G.%20Engelen.pdf on May 20th 2008.

Engelen, G. (no year). *Cellular Automata for Modelling Geographical Systems.* Research Institute for Knowledge Systems BV. LUCMOD Course / Louvain-La-Neuve 12.10 – 2.11. Page 23. Retrieved on May 22 2008. Retrieved from www.geo.ucl.ac.be/LUCC/MODLUC_**Course**/Presentations/Guy_engelen/**Cellular_automata**_regional.ppt on May 20th 2008.

ESRI 2007. *Operators and functions of Spatial Analyst*. ArcGIS Desktop Help. ArcGIS 9.2.

Geertman, S., Hagoort, M. and Ottens, H. (2007). Spatial-temporal specific neighbourhood rules for cellular automata land-use modelling. International *Journal of Geographical Information Science* 21(5):547–568. Taylor & Francis. Pages 549-552, 554, 559.

Hagoort, M., Geetman, S.,Ottens, H. (2008). Spatial externlities, neighbourhood rules and CA land-use modelling. *Ann Reg Sci* 42:39-56. Springer. Pages 40-45.

Hansen, H. S. (2007). An Adaptive Land-use Simulation Model for Integrated Coastal Zone Planning. In: *Lecture Notes in Geoinformation and Cartography*. The European Information Society. Springer. Pages 35-53.

Hansen, H. S. (2008). Quantifying and Analysising Neighbourhood Characteristics Supporting Urban Land-Use Modelling. In: *Lecture Notes in Geoinformation and Cartography.* The European Information Society. Springer. Pages 293-298.

INSPIRE (2008). Retrieved from http://www.ec-gis.org/inspire/index.cfm on May 20th 2008.

Malczewski, J. (2000). On the Use of Weighted Linear Combination Method in GIS: Common and Best Practice Approaches. *Transactions in GIS*, 4(1): 5-22. Blackwell Publishers. Pages 16-19, 21.

Mitchell (2005). The ESRI Guide to GIS Analysis. Volume 2: Spatial Measurements & Statistics. 238 pages. ESRI Press. 105, 109, 135, 165, 200-201.

O'Sullivan, David & Torrens, Paul M. (2000). Cellular Models of Urban Systems. *Working Paper Series*. Paper 22. Centre for Advanced Spatial Analysis, University College London. Pages 1-2.

Verburg, P. H., de Nijs T. C. M., van Ritsema, E. J., Visser, H. and de Jong K. (2004a). A method to analyse neighbourhood characteristics of land use patterns. Computers, *Environment and Urban Systems* 28:667-690. Pages 668-672, 674, 676, 677-680, 685, 687.

Verburg, P.H., van Ritsema, E. J., de Nijs, T. C. M., Dijst, M. J. and Schot, P.(2004b). Determinants of land-use change patterns in the Netherlands. *Environment and Planning B: Planning and Design* 31:125-150. Great Britain. Pages 126-126, 146.

Wikipedia (2008): "Urban structure".
Retrieved from http://en.wikipedia.org/wiki/Urban_structure on May 20th 2008.