USING SPATIAL LOGISTIC REGRESSION ANALYSIS TO SUPPORT CA BASED MODELLING OF INFORMAL DEVELOPMENT

DAR ES SALAAM, TANZANIA

DEMEKE ASHENAFI BITALKO

February, 2012

SUPERVISORS:

Dr., Johannes Flacke
Dr., Richard Sliuzas

USING SPATIAL LOGISTIC REGRESSION ANALYSIS TO SUPPORT CA BASED MODELLING OF INFORMAL DEVELOPMENT

DAR ES SALAAM, TANZANIA

DEMEKE ASHENAFI BITALKO

Enschede, the Netherlands, February, 2012

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Urban Planning and Management

SUPERVISORS:
Dr. Johannes Flacke
Dr. Richard Sliuzas

THESIS ASSESSMENT BOARD:
Prof. Dr. Ir. M.F.A.M. van Maarseveen (Chair)
[Dipl.-Ing., Johannes Lückenkötter, Name (External Examiner, TU Dortmund)]
Dr. Johannes Flacke (1st Supervisor)
Dr. Richard Sliuzas (2nd Supervisor)

ABSTRACT

The application of different types of land use models has been a way to better understand the mechanisms of IS expansion, which is a manifestation of rapid urbanisation in Dar es Salaam. In tacking the challenges of IS development, modelling has been used as tool to support planning and policy making processes through accentuating proactive measures. So far, empirical logistic regression (LR) and the dynamic cellular automata (CA) modelling approaches have been used for modelling IS developments in Dar es Salaam. Their capacity in making use of explanatory variables to predict the probability of cells' to be developed to new IS lands and showing the pattern of the expansion in a spatially explicit way has made (LRM) powerful tools in modelling land use developments. However, their limitation in showing the temporal dynamics and developing scenarios has been a limitation in manipulating the full potential of LR the models. On the other hand, CA models except for their complication in calibration and inclusion of global factors of land use change, are able to make a better representation in land use dynamics and are good in developing scenarios to see 'what if' conditions. Therefore, the focus of this research has been making a structured approach to integrate LR analysis with dynamic CA based modelling with a theme of providing more representative interpretable and structured information on the dynamics of IS expansion by making use of the cumulative benefits of the integrated model.

The integration of the two approaches has been done by incorporating both global and local factors of IS expansion for predicting the potentials the cells to be developed in LR model, while the CA based approach uses the predictions to simulate IS expansion dynamics by constraining the quantity based on IS expansion land demand. Two different levels of integration were used to model the IS expansion dynamics in Dar es Salaam. The first level of integration was made between three LRMs, 1982-1992 (model-A), 1982-2002 (model-B), and 1992-2002 (model-C), based on the IS expansion, in order to achieve better model performance., while the second level of integration was between LRM and dynamic CA based model in order to increase the interpretability and representativeness of the models to the phenomena of IS expansion. The evaluation of the models and comparisons of the results have provided with the fact that integrated models have a better performance in simulating IS expansion in DAR. In the evaluation of the first level model integration has been very high for integrated models than the individual LRMs while the results from the overall integrated model were able to provide a more realistic IS expansion simulations which was found to be representative of the observed IS expansion in DAR. According to the results, the most important key drivers in the rule definition for calculating probability of cells were distance to minor roads, distance to existing ISs with both negative relationship with the expansion of ISs in DAR. The probabilities of the cells to be informal is hence constrained by the IS expansion land use demand observed in order to simulate the IS expansion. At every simulation the models update the overall probabilities of the cells through the updating local probabilities based on the simulation results. The logical justification and approaches of model the integration have shown a promising results and increased the interpretability of the IS expansion dynamics in DAR. Moreover, the application of these results and approaches of the study can facilitate an informed policy and decision making processes decision making processes related to IS expansions

Key words: Informal settlements, LR modelling, CA modelling, model integration

ACKNOWLEDGEMENTS

Above all, I would like to give thanks to almighty God who given me the strength to accomplish this thesis.

I am truly indebted to The Netherlands Fellowship program (Nuffic) for giving me the grant to my whole stay in ITC and accumulate a priceless knowledge in my field of study from ITC. I am sincerely and heart fully grateful to all my lecturers and the staff at ITC without whom this dissertation would not have been possible.

My exceptionally heartily gratitude goes to my supervisors Dr. Johannes Flacke and Dr. Richard Sliuzas for the greatest opportunity you gave me to share your rich experience and learn from your constructive advices and support throughout the thesis work. I am sure it would have been unthinkable without your support to see the work accomplished.

I would like acknowledge the great contribution of Prof. Dr. Ir. M.F.A.M. van Maarseveen for his critical comments in erecting the directions in of the research while assessing the proposals work.

I would also like express my gratitude my friend HaileMchael Mitiku Worku (Dr.) to his support in enriching the thesis while in dealing with logistic regression and generalized linear mixed models.
To my dear friend Abebe F. Kassahun, I am very grateful to the whole support you have made and for the discussion we had for the developments of the thesis.

To my friend Feven S. Desta thank you so much for your help and good advices.
I am also highly grateful to UPM students and all individuals who have contributed to the improvement of the thesis.
Finally I would like to express the greatest thanks and love to my wife Workie Tegen who boosted me morally and has paid a great sacrifice to the achievements of the thesis.
To all my family to my mother Almaz and my Sisters and brothers thank you to the encouragement and love you have showed while my stay away from home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ABM	Agent based model
ANN	Artificial Neutral Network
CA	Cellular automata
CBD	Central Business District
CI	Confidence interval
CL	Confidence level
DAR	Dar es Salaam
GIS	Geographic Information System
GLMM	Generalised Linear Mixed Model
ILWIS	Integrated Land and Water Information system
IS	Informal Settlement
ITC	International Training Centre, (currently, Faculty of Geo-Information Science and Earth Observation, University of Twente)
LDC	Least Developed Counties
LR	Logistic Regression
LRM	Logistic Regression Model
PSS	Planning Support System
SE	Standard Error
SFAP	Small Format Aerial Photography
SPOT	Satellite Probatoire pour l'Observation de la Terre
SPSS	Statistical Package for the Social Sciences
SSA	Sub Saharan African
TNBS	Tanzanian National Bureau of Statistics
TR	Transition Rule
TP	Transition Potential
VIF	Variance Inflation Factor

1. INTRODUCTION

Informal settlement expansion which is the manifestation of rapid urbanization in developing countires has been a prime concern in the sustainable development of their cities. In facilitating better understanding of the IS dynamics, recently, it has been a practice to make use of the GIS based modelling techniques such as, LR and CA modelling as a tool for policy and decision making processes. The different modelling types, however, have been used independently in a way which could not exploite the integrated application of the techniques.

The focus of this study would be maximizing the benefits of the models through possible logical integrations in order to better understand the mechanisms of IS expansion in Dar es Salaam and hence provide a structured way in integration LR and CA based modelling approaches.

1.1. Back ground and justification

Since the beginning of the 21st century many countries of the developing world have been facing socio-economic as well as environmental problems due to rapid urbanization experienced in their cities (Hill & Lindner, 2010). Currently the majority of the world's population is living in urban centers (UN-HABITAT, 2008). Especially, in the cities of the developing Sub- Saharan Africa (SSA) this has increased the demand for infrastructure, public services, jobs, and the need for residential land. The nature of urbanization in the developing world is creating a situation where urban development being accompanied by poverty and limited economic growth. This is because of the existence of wide mismatch between high demand of land for residential use and limited capability in resources, financial and personnel power of the local planning authorities to respond to the need (Hill & Lindner, 2010). In Dar es Salaam, the development of informal settlements (IS) which are often built on land without having legal tenure and not following established building and planning regulations, is a result of this imbalance. It brings in the emergence of settlements of low standard shelters with a pattern disregarding the planning regulations and lacking the basic amenities and infrastructure provisions (Abbott, 2002; Sliuzas, Ottens, & Kreibich, 2004).

Following this, several measures have been practiced by local authorities in order to address the issue of IS development(Abbott, 2001). Despite the fact that a number of demolishing and resettling measures were taken, IS growth is still a concern of many developing countries. Thus, the issue needs measures to be practiced to prevent future development of ISs. The development of different modelling techniques and Geographic Information Systems (GIS) is found to have a very importance in the analysis of ISs. They are used as a tool to identify the key driving factors of ISs, analyse dynamics of urban growth, develop possible future scenarios and prediction of future urban growth patterns to support planning and policy-making processes(Dubovyk, Sliuzas, & Flacke, 2011). To this end, different modelling techniques have been applied to model land use changes. However, the techniques have been applied independently according to the purpose to which they are intended for. This is also because of the limitation that a single modelling technique is not effective to perform all the intended tasks in the analysis of urban growth (Huang, Zhang, & Wu, 2009).

In recent studies the application of integrated modelling approach has been given attention as a way for better understanding of urban dynamics and support practical problems of planning and decision making (White & Engelen, 2000). A broad number of land use modelling techniques, such as, cellular automata, logistic regression, multiple agent based have been applied for urban land use change and urban dynamics analysis (Huang, et al., 2009; White & Engelen, 2000). However, since each method has its own limitations seeking ways to have better applicability and interpretability of the modelling realm, for instance integrating different models has been the issue of today's planners and modellers.

Previously, few studies have been done on the analysis of IS modelling in Dar es Salaam. For instance, LR modelling by Abebe (2011) and CA modelling by Hill and Linder (2010) are the two recently done studies in Dar es Salaam. LRMs have been found more effective in establishing probability of land use change and the driving factors of the change but weak in modelling urban dynamics, future predictions and developing scenarios. On the contrast, CA models are efficient in in bottom-up simulation of urban dynamics and have strong capacity of scenario simulation but weak in interpretation of spatiotemporal processes of land use changes and computationally difficult for their calibration(Huang, et al., 2009).Hence, the issue of understanding IS development needs a comprehensive view and structured modelling technique for better understanding and to fill the knowledge gap created due to the shortcomings and the complexities of the independent application of the models.

Therefore, in this particular study the integration of two modelling techniques, Cellular Automata (CA) and Logistic regression (LR), is going to be introduced in order to deal with the issue of IS expansion in a more structured and comprehensive way. The technique will provide a framework to make use of the cumulative benefits of the models. Some of the benefits gained from the integrated model are: (1) identifying key drivers of IS expansion in Dar es Salaam (DAR), (2) estimating the probability of a land use to be developed, (3) facilitating the calibration computational task in CA models, (4) simulating urban dynamics and developing scenarios to support planning and policy making processes. The common advantage of LR and CA models, their application in data scarce environments of developing cities, such as Dar es Salaam, makes the study more relevant for studying the development of ISs.

1.2. Research problem

Urbanization in Dar es Salaam is characterized by rapid population growth, and poverty resulting in the development of ISs. High population growth creating high demand for residential land and housing has been a critical issue in the urban centres of least developed countries (UN-HABITAT, 2008). The emergency of ISs has become inevitable as the formal public sector is unable to satisfy the high residential land and housing demand (Kironde, 2006). In the cities of SSA, such as Dar es Salaam, the problem of IS development is manifested in complex stages of growth, expansion and consolidation.

The application of GIS based modelling techniques has become a widely used approach to tackle the problem of ISs in a proactive manner (Dubovyk, et al., 2011; Sliuzas, et al., 2004). Several urban growth modelling techniques are applied independently for modelling in a way of identifying key drivers of future informal developments, finding probable areas of future settlement, and developing scenarios (Abebe, 2011b; Dubovyk, et al., 2011; Hill & Lindner, 2010).The driving factors, which are fundamentally the inputs of the local knowledge, are the core components of the IS development modelling. In the analysis of urban growth different models, such as LRMs, have been using these factors in order to explain the development of ISs in a spatially explicit way (Wu, 2002).In fact, the land use models are also required to

explain how the dynamics of physical expansion and consolidation of ISs is experienced in a way to assist planning and decision making to project future urbanization under various scenarios.

However, since different land use models have their own limitations to incorporate all the intended needs, still there is a need for modelling informal settlements in a way of integrating different modelling techniques to better understand the influence of IS drivers as well as the dynamics of informal developments. Hence, in this study, an integrated modelling technique will be developed by using Logistic regression (LR) for supporting Cellular Automata (CA) modelling.

1.3. Research objectives and questions

1.3.1. Main objective

To develop an approach to integrate Logistic regression and cellular automata modelling in order to better understand the development of informal settlements in Dar es Salaam.

1.3.2. Sub-objectives and questions

1. To develop a conceptual method to integrate LR and CA modelling for analysis of IS expansion.
   - What outputs of LRM can be used to support CA modelling?
   - What are the major challenges in CA modelling to be supported by LR model outputs?
   - What is the relevance of using LR modelling outputs in CA modelling?
2. To revise Logistic regression model of IS Expansion in Dar es Salaam to support CA modelling of informal development
   - What logical reorganisations should be made for integration of LR and CA modelling?
   - What is appropriate in the development of IS expansion CA modelling in terms of the potential drivers of IS expansion?
   - How robust the LRMs would be?
   - What measures can be applied to increase the accuracy of the LRM of IS expansion modelling in DAR?
3. To model IS expansion dynamics in Dar es Salaam using LR integrated CA based simulation.
   - How to make use of LRM output, cells' probability, as input to the CA rule definition for IS expansion dynamics in DAR
   - How would the exogenous IS expansion demand be applied for simulating of IS expansion in DAR?
   - How stochastic perturbation effects be incorporated in the CA based simulation model?
   - How the dynamics of IS expansion in DAR would look like?
   - How robust the LR integrated CA based model would be?

1.4. Conceptual framework

1.4.1. Stages of IS growth

Informal settlements as manifestations of urban growth in the cities of the developing world are dynamic processes. ISs do not show a linear development pattern, rather at certain stage of their development they explode. Hence, they manifest different growth patterns with three development stages are seen in the growth of informal settlements: Infancy, Consolidation and Saturation (Abebe, 2011b). The ***infancy*** stage of IS development which is otherwise called expansion the stage at which agricultural land is squatted by

informal low income households. At infancy stage ISs manifest a characteristic of scattered layout in the built areas and comprises the largest proportion compared to other stages of IS growth in DAR. The stage of ***consolidation*** which is often known as densification a booming sate in the IS development. Middle income residents would become the integral parts in the IS at this stage and each piece of open space would tend to be covered by IS. At the pick of densification almost 80 % of the land will be covered by ISs. The last stage of IS development which is not yet a case in Dar es Salaam is ***Saturation*** stage. This is a case where vertical growth of ISs is observed. Saturation occurs mostly on previously densified areas of ISs.

Dar es Salaam (DAR) IS development is expressed in the first two stages of development: Infancy (expansion) and Consolidation (densification) (Sliuzas, et al., 2004). However, as the expansion of ISs is a continuing manifestation of IS development in DAR the study would be made on the IS expansion modelling of DAR.

1.5. Self-organizing Systems

Recent studies of nonlinear and open systems have shown that the process of urban development is based on open system theory. Liu (2009) argues that 'a city can be viewed as an open and complex self-organizing system that is far from being in equilibrium, and it exists in a constant exchange of goods and energy with other cities and its inter-land' (Liu, 2009). According to this theory, the overall pattern of the system, for instance urban space, emerges from local actions where uncontrolled local interactions (decision making) give rise to the coordinated global patterns. Hence, any urban development is considered to be a spatially dynamic process, manifesting the fundamental features of a self-organization system. Urban land conversion is by no means random. The behaviour of land use conversion can be seen accordance with the fact in urban economics. Urban economics establishes the development probability through regression methods. For example, 'the characteristics of land use conversions can be revealed through the plotting of the quality of land use conversion against distance from the city Centre' (Liu, 2009).

Logistic regression methods can be used to examine the relationship between land use changes and their locational characteristics. They can provide probability maps of informal settlement growth of expansion and densification as well as the key drivers of the growths. That is, LRMs reflect the global distribution of land conversion it the urban areas as a function of the factors (Wu, 2002).However, LRMs do not show self-organization nature of informal growth, that is the clustering process of informal expansion and densification at local level.

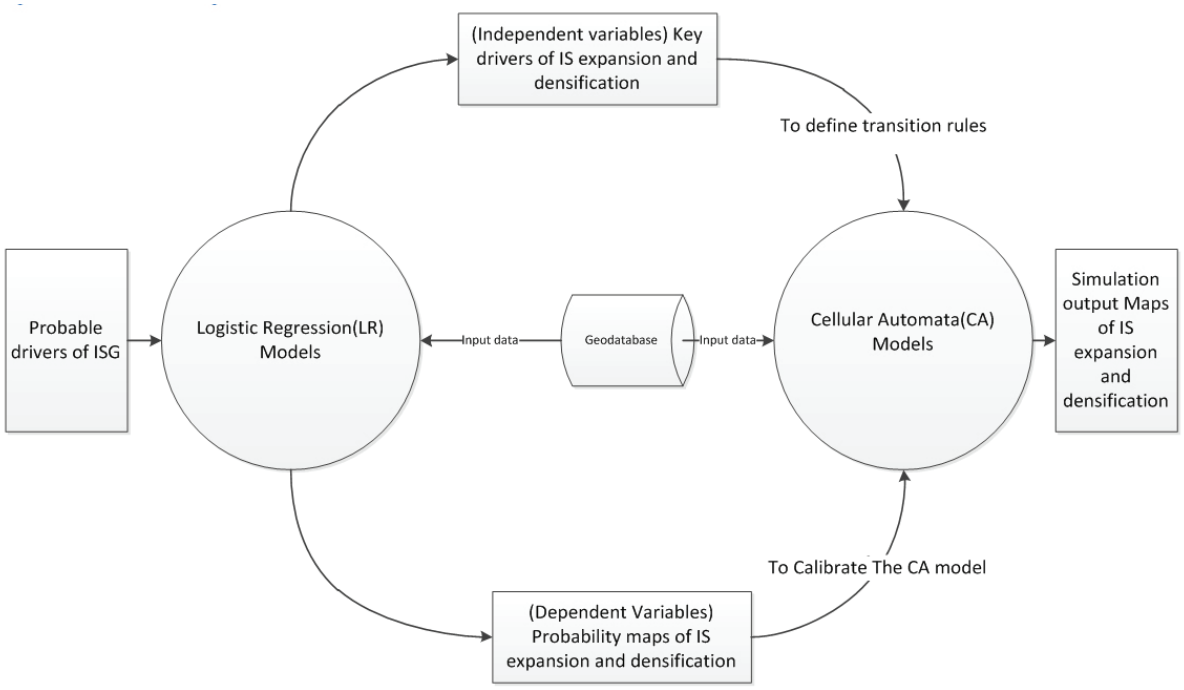


Figure 1: Conceptual Model for simulation of IS expansion with integrated of Logistic regression and Cellular automata models.

1.6. Research design and content outline

The research overview on (figure 2) has shown the major steps applied in the overview configuration of the research. The conceptual framework has been started first by first putting the research problem which is made on the application of LR integrated CA modelling for the application of IS expansion dynamics in DAR. The basic contents of the six chapters which make up the conceptual framework will be discussed in this chapter.

Chapter One-Introduction

Chapter one starts with background information provided as a justification to develop the research problem, design the objectives and questions to achieve the objectives. This all together make up the conceptual design on model integration.

Chapter Two-Modelling and informal development

Chapter two provides detail ground information which is relevant to scientifically justify further the problem, objective and questions of the research. The issue of urbanisation in developing counties, the phenomenon of IS expansion and modelling IS expansion by using LR and CA modelling techniques shall be discussed in the chapter.

Chapter Three-Data and Research methodology

Chapter three provides general information on the data and makes major focus on the methodological steps undertaken to achieve the research objectives. The questions rose to achieve the objectives under the

umbrella of integrating LR and CA modelling for modelling IS expansion in DAR will be addressed by the methodology chapter.

Chapter Four-Results

Chapter four is the part which presents the results of the applied methodology to answer the research questions. The results from the LR analysis which are basically the parameter values of the drivers and probabilities of the cells for modelling IS expansion dynamics as well as the simulation results shall be presented and interpreted in LRMs.

Chapter Five-Discussions

Chapter Five makes discussions of the obtained results of the research. The discussion is basically on the possible ways of integrations obtained and comparison of the evaluation the performance of the models.

Chapter Six-Conclusion and Recommendation

Chapter Six is the concluding chapter of the research. It basically explains how the three objectives were addressed in the study and provide points for further research.

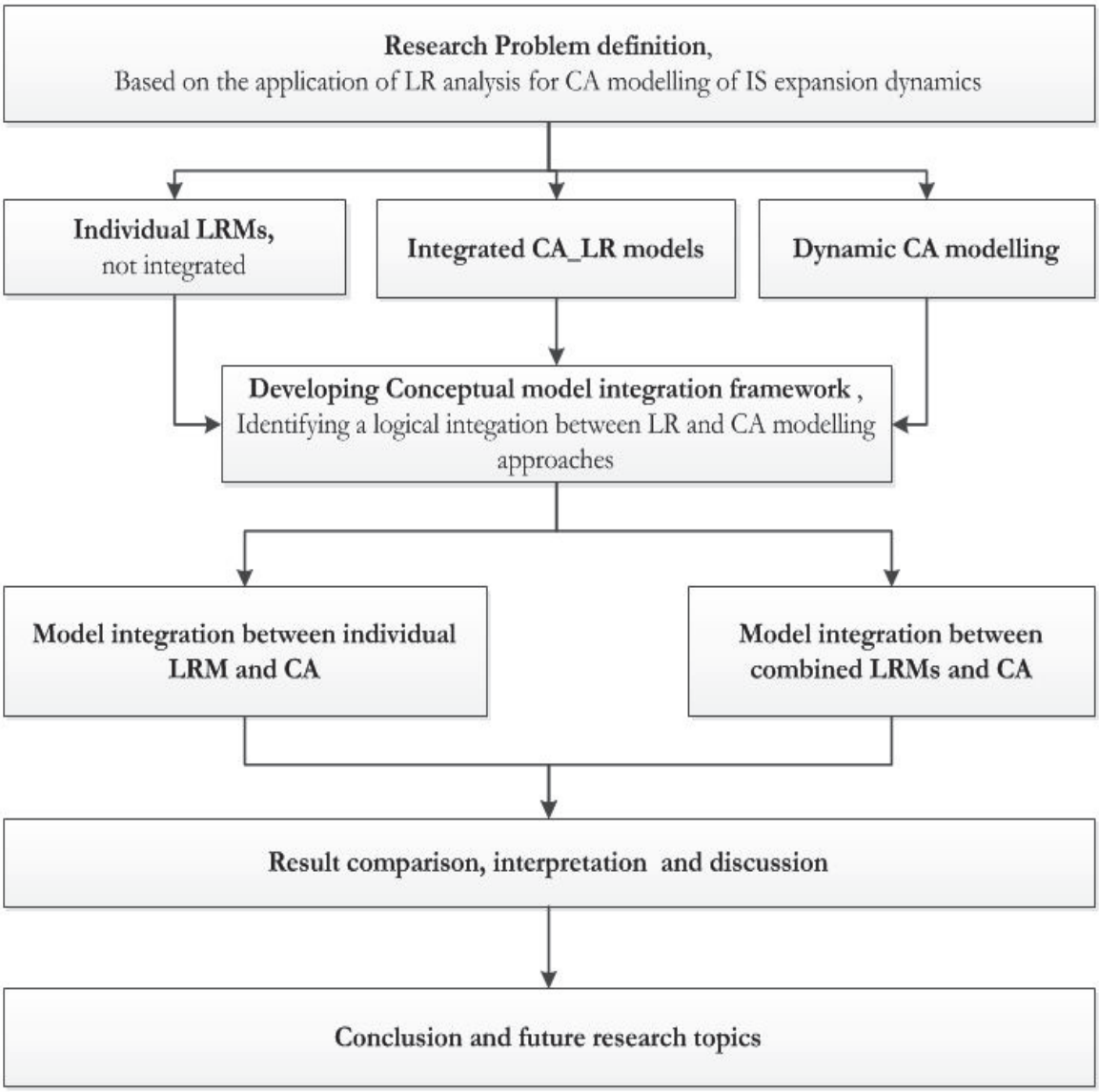


Figure 2: Research design overview

2. MODELLING AND INFORMAL DEVELOPMENT

The chapter provides detail ground information which is relevant to scientifically support and further justify the problem, objective and questions of the research. The main points which were given attention in the chapter are the phenomena of informal development as a concern of urbanization developing counties, the drivers of rapid urbanization and more elaborated discussion on the modelling techniques which can be applied in IS development modelling. The concept of model integration as well as the nature of LR and CA modelling by referring to the works of other researchers has also been given attention in the chapter.

2.1. Urbanisation and Population growth

Although the development pattern varies among different countries, rapid urbanization is expected to happen as a consequence of high population growth in many countries. Between 2000 and 2005 the population of the world was estimated to have grown showing an increase of 380 million. Hence, in 2005 the total population of the world was estimated to be 6.5 billion. This number is projected to reach 9.1 billion in 2050.The increase in population between these years is equivalent to the current population of China and India (United Nations, 2006). However, this growth figure is found to be irregular, specially, between the developing and the developed world. While the urban population in many of the developed countries is expected to be in a process of stagnation or even decline, many developing cities would be in a rapid population growth. 51 countries, the majority of which belong to the developed countries, are estimated to have less population in 2050 than their current population. On the other hand, the report by the United Nations (UN) has shown that virtually all population growth of the world will be happening in the least developed countries (LDC). Hence, the population number of the least developed countries in 2040 is estimated to be twice of the figure in 2005. In a similar report UN has put 50 developing countries including Tanzania as LDCs in which the population growth is expected to be at a high rate (United Nations, 2006).

2.1.1. Urban growth and informal settlements in Sub-Saharan Africa (SSA)

Historical urbanization trends show that Africa is in a fast process of urban growth. In the beginning of the main decolonization time, the total population of the continent living in urban areas was only 13%. According to the estimation this rate has increased to 18% in 1960. However, the distribution of the growth in the continent was not even. While the rate in the Southern and northern parts of Africa was estimated to be 42 and 30% respectively, the urbanization rate in the eastern Africa was only 7.3%. This increasing rate of urbanization has raised up the total number of urban inhabitants to 43% between 1990 and 2000 (Hill & Lindner, 2010). The UN report also shows that by 2007 two-thirds of the world's population has started living in the urban centres of the developing countries. In this list Africa had an estimated urbanization rate of 3.4% between 2000 and 2005. 'For the first time, in 2009, Africa's total population exceeded one billion, of which 395 million, almost 40 per cent, lived in urban areas. This urban population will grow to one billion in 2040, and to 1.23 billion in 2050, by which time 60 per cent of all Africans will be living in cities' (UN-HABITAT, 2010).

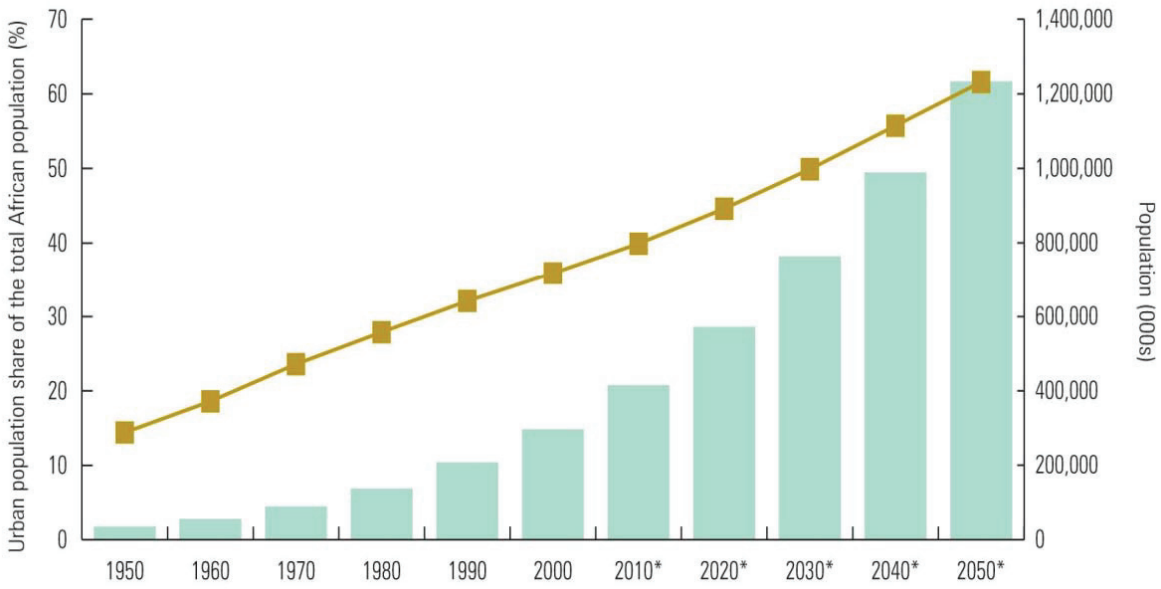


Figure 3: African urban population trend 1950-2050 (source UN-HABITAT 2010)

Recent studies have shown that the highest urbanization has been taking place in the SSA. The Eastern part of Africa including Tanzania, as part of the SSA region, has been in a high rate of urbanization. Hence, the highest annual growth rate of the region between 1950 and 1995 was 5.6 to 6.5.This figure is by far the highest compared to the rate of other regions of the continent which was between 4.5 and 4.9%. Among the SSA countries, between 1990 and 2003 Tanzania and Mozambique were the only two countries with the yearly growth rate higher than 6% (UN-HABITAT, 2008). Thus, having a focus on this part of Africa would have valuable contributions to tackle the multidimensional problems of uncontrolled urban growth in the developing world.

2.1.2. Informal settlements in Tanzania

Tanzania is one of the SSA countries with the fastest urban agglomerations in which the informal development takes a major portion of the urban landscape of its cities. The poor performance of the public sector in regulating access to land and provision of housing to the residents has been a major pressure to the growth of ISs (Kironde, 2006). The total housing share of the informal sector in Tanzania was estimated to be more than 50% of the total housing stock (Kombe, 2005). Currently, a significant portion of the Tanzanian urban centres is covered with a vast mass of ISs. ISs come up with a structure different from the normative urban land pattern and yield urban challenges such as inefficient land use distribution, and become threats to proper development patterns and health conditions because of the over increasing settlement densities over the urban landscape (Kombe, 2005; Sliuzas, et al., 2004).

2.1.3. Informal settlements in Dar es Salaam

Urban growth in Dar es Salaam, as a Capital of Tanzania, is a result of natural population growth and rural-urban migration. The national census of 2002 in Tanzania shows that between 1988 and 2002 the annual average population growth was 4.3 %. The rapid population growth together with substantial poverty problems has drastically increased the number of informal settlers. Currently Tanzania is one of the countries with the highest informal residents in the Sub-Saharan Africa. 50 to 80 % of the population is living in informal residential areas (UN-HABITAT, 2008). In 2007 the population of Dar es Salaam is estimated to be 3.3 million. It shares 29% Tanzania's total population. This figure is expected to be 5.7 million in 2010 and double of this in 2025 (Hill & Lindner, 2010).

As is a case in many developing countries, land seekers use the informal sector as their prime option to have access to residential land. Several researches have also shown that the problem of informal development is the result of the inefficiency of the formal market to accommodate the high residential land and housing demand, and the incapability of the poor to participate in the formal land and housing market(Abbott, 2001; Abebe, 2011b; Dubovyk, et al., 2011; Hill & Lindner, 2010).

2.1.4. Urban growth factors

Urban development is affected by factors of both global and local scale(Liu, 2009; Wu, 2002). Local interaction factors of land use development can be defined as those drivers to a land use conversion by its immediate neighbourhoods. On the other hand, global factors of land use change, such as, road networks which have an attraction effect to a certain land use type in their surroundings. Hence, in the study of informal expansion land use development needs to be seen as a result of the probability potentials from the local and global factors. Global probability, a probability of a certain land use cell due to global factors, will have a problem if applied at micro-level as it does not include local neighbourhood interaction factors and ignores the path dependent and self-organization nature of land development(Wu, 2002). That is to say local interaction factors should be incorporated independently so as to show that land use development is explained in terms of spontaneous and self-organized growth factors. Spontaneous growth implies land use change based on the demand and supply relationship through development tendency, which is independent of individual land use conversion/changes. On the other hand, the probability of land use change at local scale is also dependent on the interaction of the individual land uses with their neighbouring cells' uses. Wu argues that the probability of these factors can be estimated through logistic regression if the data available is at small time frequencies, for instance every year as individual land use changes are mostly in discrete yearly time basis. With this respect, both LR and CA models fail to incorporate both categories of factors at the same time in their modelling process.

The self-organization nature of CA makes the urban development a result of only local scale neighbourhood interactions, that is to say CA model builds upon the fundamental unit of behaviour. In real situations, however, factors, for instance, transportation networks affect land use development at regional scale(Liu, 2009). On the other hand, in land use conversion process forward and reverse processes will be revealed only as a result of self-organization nature CA models. The incorporation of both local and regional factors in the simulation of land use development will be of a high importance to represent land development in a more realistic way. Hence, the core of this research is exceptionally unique as it tries to define a comprehensive modelling approach through the integration of local and global factors of land use development.

2.2. Informal development and Modelling

Numerous strategies have been made and implemented to deal with the issue of informal development. Early 1970s, ISs were seen as problems of urban development; hence, the solutions made were demolishing the houses and resettling the inhabitants in other formal settlements. However, it was later observed that the solutions made were financially unachievable. Following this, an approach of looking ISs as social and economic entities and searching other mechanisms to upgrade informal settlements has become a way to deal with the issue of informal development(Abbott, 2001). Though existed informal developments can be tackled effectively through upgrading and demolishing, prevention of further informalities cannot be assured through reactive measures. This has created a need to look IS growth based on understanding the local situations and the consequence of such actions and prevent further developments(Abbott, 2002).

Recently, new approaches have been made to address the issue of ISs. These approaches are mainly based on the understanding of the drives (factors) of ISs in a proactive manner. The development of GIS based techniques of modelling, to identify the main factors of ISs and predict the possible future developments of ISs, is of a crucial value in the study of informal developments (Dubovyk, et al., 2011). While Understanding the key factors of informal development would have crucial role in urban planning to deal with the proliferation of ISs, the prediction can be used to support planning and policy making processes, for instance, by developing different scenarios (Dubovyk, et al., 2011; Hill & Lindner, 2010).

As urbanization is a result of numerous factors, the most important component of urban modelling is the definition and identification of the factors. ISs, as manifestations of urban growth in the cities of developing countries, are derived by several factors. Identification of those factors and understanding their relationship with informal developments is the core component in modelling land use changes and in supporting policy and decision-making processes (Dubovyk, et al., 2011). One of the major steps in urban growth modelling is the identification and definition of hypothetical factors of the growth. The selection of the factors is mostly case specific. Though, there are numerous attempts to make categorization of the different sets of factors, researchers agree on the fact that there is no universal way to identify the factors which explain urban growth (Hu & Lo, 2007; Huang, et al., 2009). Three general groups of factors driving urban growth, such as, socio-economic drivers, biophysical drivers and proximate causes (land management variables) would be used in this research(Abebe, 2011b). Another important aspect regarding factors and urban growth can be the theoretical approach to study the relationship between land use change and the drivers. The common bases to this end are the application of theories, physical laws, and expert knowledge to analyse the relationships between IS growth and the factors of the change(Verburg, de Nijs, Ritsema van Eck, Visser, & de Jong, 2004).

2.2.1. Logistic regression models

Logistic regression models (LRM) are types of empirical estimation models which use statistical techniques to model the relationship between land use change and the drivers based on historic time series data (Zhiyong Hu C.P.Lo, 2007).They are used to determine the influence of independent variables (drivers) and provide a degree of confidence about their contribution to the change(Huang, et al., 2009). A logistic regression model generates urban growth probability maps in a way it tells the probability of a cell being urbanized by associating urban growth with demographic, economic, and biophysical drivers of urban growth. Among the several modelling techniques applied in spatiotemporal analysis of urbanization LRMs are well known for their capability to explain urban growth in spatially explicit way (Dubovyk, et al., 2011). However, since LRMs work with basic assumptions, such as, the normal distribution, appropriated error structure of the variables, independence of variables, and model linearity, they hardly ensure high generalization performance for projecting future land use change. 'The future state of a system can be modelled purely on the basis of its immediate preceding state (Markov chain analysis). 'Though, LRMs can be used to examine the changes due to urban growth and can be quite useful to identify the most influential factors of land use change, they lack the capability to explore the casual relationships underlying the transition' (Huang, et al., 2009). Hence, such models are not proficient in predicting future land use changes and developing scenarios.

2.2.2. Cellular automata models

Simulation of urban growth has made advancements as a result of developments in the theoretical concepts and the technical aspects of the modelling world. Spatial modelling needs understanding the phenomenon to be modelled, building strong conceptual background on how the model works and

looking for more efficient and appropriate way of modelling techniques. Currently, a number of studies suggest the need to look into models which use system dynamics to construct stories of the past and possible future than the usual empirical deterministic ways (Guhathakurta, 2002). From this point of view simulation models are quite useful as urban environments are complex in their nature and need dynamic models to represent them.

Models of land use change, such as Cellular Automata (CA), which are based on self-organization or Complex theory approach (Guhathakurta, 2002), are gaining interest because of the fact that: (1) they are based on simple and local relations and behaviour of individual elements; (2) they are capable of representing changes explicitly; (3) they have the capability to generate complex urban phenomenon based on simple local rules and provide the possibility to develop scenarios in order to generate possible future alternatives (Hadwig van Delden, 2005).

2.2.3. Types of CA models

Urban cellular automata models can be classified based on their distinct features in defining the transition rules (TR). Though the classification of CA models can be made based on the major components of the model, the common way of categorization is based on their TR definition. This is because of the fact that TRs are the core components of a CA model. As TRs represent the logic of the process being modelled they determine the spatial dynamics of land use change (White & Engelen, 2000). TRs serve as algorithms to drive the state change of cells (Liu, 2009). Hence, in this section we shall see a brief discussion on the major types of CA models classified based on their definition of TRs.

Constrained CA models, which are started by White and Engelen, take into account the constraints of other factors for the development of each category of land. However, the development of cells in each stage of the standard CA is determined endogenously by TRs (Liu, 2009). The first constrained CA model of White and Engelen was made up of two parts, such as, the macro scale model and micro scale model. The factors such as population density and socio-economic status were the bases for macro scale model to be developed exogenously to the cellular automata model. Constraints with varied land use consumptions were generated by the model to control the amount of cells of a certain state. A set of transition potentials, representing the inherent suitability of a cell from its current state into another state, were estimated at the micro scale. Major factors, such as, accessibility of the cell to the road network, the suitability of the land, the zoning status, and the impact of the neighbourhood on the cell, were used to calculate the transition potentials for a particular land use. Here, constrained CA has got its name because of the fact that the state of each cell will be converted to a state for which it has the highest potential until it reaches the demand and constrained by the macro model (Liu, 2009).

SLEUTH model took the name from the six input data layers, such as, slope, land cover, exclusion, urbanization, transportation and hill shade. The models applies four types of urban land use changes: spontaneous growth which occurs when a randomly chosen cell falls close to enough to an urbanized cell; new spreading centre growth which represents the tendency of a cell to expand outward from existing city centre; edge growth which urbanizes cells that are flat enough to be urbanized; and road-influenced growth which encourages load side development of cells (Liu, 2009).SLEUTH model uses two phases of modelling: Calibration, in which the model is trained with historic development patterns; prediction phase in which historic trends are projected into the future (Liu, 2009).

Fuzzy Constrained CA models use more flexible probability concepts and fuzzy logic in defining TRs. Physical constraints and human decision-making behaviours, which are the causes of urban development, are fuzzy and uncertain by their nature. Here, TRs are defined to integrate the decision-making and the subjective natural language statements to describe certain preconditions of a decision. The application of fuzziness in the model is the reflection of the idea that the TRs of a CA model may not necessarily be restricted to deterministic forms (Liu, 2009).

Stochastic CA models are CA models which use statistical estimations and configure the parameter values as a way to calibrate the models. A stochastic model estimates probability distributions of potential outcomes by allowing for random variation in one or more inputs over time. The random variation is usually based on fluctuations observed in historical data for a selected period using standard time-series techniques (Liu, 2009). These models update the initial probability of the simulation dynamically through local rules of neighbourhood development.
Liu argues that the basis of the advantage of stochastic CA is that factors directly considered by local experts, government authorities and developers, who have the greatest control over land development, are used to estimate the probability of transition of the growth process.

2.2.4. Integrated land use modelling

The future of spatial planning has a lot to do with the integration of different modelling techniques. The basic purpose of modelling as a planning support system (PSS) can be explained in terms of enhancing the capacity of spatial planning strategically for the future (Couclelis, 2005). Couclelis (2005) explains the different roles of land use models in terms of supporting the future-oriented process of planning, such as, scenario writing, visioning and storytelling. The need for integrating different modelling techniques comes crucial for the fact that no single model is likely to fulfil all the basic roles. Hence a well-structured PSS is expected to facilitate seamless integration and mutual reinforcement of the different modelling approaches (Couclelis, 2005). "By becoming better integrated with certain informal techniques designed to promote its normative and future-oriented dimensions, models can help planning recover its true strategic, goal-oriented identity that was lost amidst a host of theoretical, practical, and political doubts" (Couclelis, 2005).

A number of studies have been made with numerous strategies to integrate different types of models. Model integration can be done in various ways with the basic concepts, for instance, improving the future-oriented nature of planning. A number of researches have been made to integrate land use models with other socio-economic and transport models. On the other hand model integration can be done within the land use modelling category having the aim, for instance, increasing the predicting capacity of the models. Therefore, this particular research is has been done by integrating static LR model of informal land use expansion, with dynamic neighbourhood interaction model in a way to improve the predictability, scenario developing capacity and interpretability of the models. The integration of empirical analysis LR with the CA, which is the most powerful visualization technique in urban growth simulation, gives a mutual benefit to both approaches. For the CA models the definition of transition rules and its ability to incorporate global attraction rues have been an issue so far(Wu, 2002). Respective of this, recently a number of researchers have made their focus on integrating differ approaches and techniques of land use models. For instance, the hybrid model of(Poelmans & V.Rompaey, 2010), the CLUE-S model of (Verburg et al., 2002), the stochastic CA model of (Wu, 2002) and the environment explorer of (White & Engelen, 2000). As a basis for the IS expansion modelling in DAR, this research has been based on the previous studies of Hill and Lindner (2010) on CA modelling and Abebe (2010) LR modelling of ISs developments. The short

review of the two researches has been seen (section 2. 13. 2). Hence, from the modelling approaches developed so far the author of this study tries to be developed a comprehensive and structured framework to be applied in DAR in particular and for other similar areas of study in general.

One of the discoveries during this research has been the whole integrated model has been done without demanding sophisticated techniques and software. It can be done with basic GIS software such as Arc map and statistical software packages like SPSS. To this end various models can be downloaded with freely from online sources especially for the LR analysis, for instance, CLUE-s modelling frame work and R software can easily be downloaded and be applied for the analysis.

2.3. Related researches

Urban land use modelling by using LR and CA techniques has been a concern of many researchers. Those techniques have been applied independently as well as in an integrated manner to analyse the phenomena of urban expansion. Urban Hence, for this research the author would provide a brief review of few related researches. The discussions are based on the comparison of LR and CA models application as an integrated technique of land use modelling.

2.3.1. Researches on IS development modelling

The proliferation of ISs is one of the concerns in the sustainable development of the cities of LDC has recently, urged the application of land use modelling techniques, such as, LR and CA, in order to enhance the understanding of IS developments. In this section it is tried to present a brief summary of two studies: the application of LR modelling for spatio-temporal analysis of IS development in Sancaktepe district, Istanbul, Turkey (Dubovyk, et al., 2011); and a CA based land use model for simulating IS expansion in Dar es Salaam, Tanzania (Hill & Lindner, 2010).

The LR model by Dubovyk, et al., (2011) has been applied to see the influence of driving factors of IS expansion in the Sancaktepe district, Istanbul. The model was also used to predict the pattern of IS expansion as a LR model is capable of calculating the probabilities (the likelihoods) of the cells for IS development. The key drivers of the IS expansion of the district, population density, slope, and proportion of ISs in the neighbourhood, were identified based on the parameter values and coefficients by the LR model. Models of different time steps were used to generate probability maps of predicted IS areas and were evaluated for their performance. The results of the model were found to be a role model to be applied in areas of similar IS situations and would have a vital importance in being a tool to support planning and policy making processes. The contribution of the research in applying the technique for other areas of similar IS development, for instance, the IS development modelling of ISs in DAR by Abebe (2011) has been significant.

Hill & Lindner, (2010) CA based IS expansion simulation model has also been one of the recently made studies in DAR in a way to better understand IS drivers and the dynamics and mechanism of the phenomena. The simulation model of Hill & Lindner, (2010) was based on the foundations of CA, which is known for its capacity in capturing neighborhood dynamics, ease of applicability in data poor environments as well as scenario developments. The model has made use of fifferent variables, such as, natural conditions, accessibility, and local-scale dynamics to represent the factors of IS expansion and calculate the transition potentials of cells' for being ISs and make allocation of them. This CA based

model of Hill & Lindner, (2010) has been employed in developing scenarios in order to show the impacts of transport infrastructure projects in the IS expansion dynamics of DAR.

2.3.2. Researches on integrated LR and CA modelling

Several studies have been made based on integrating of LR and CA modelling techniques for modelling the dynamics of land use developments based on the aim of achieving the incorporation of both global and local factors of land use development in the simulation process and hence enhance the interpretability of the models for PSS. In this section two researches have been discuss for their contribution to the concept of CA and LR modelling integration.

One of the first researches's made on integration of LR analysis for the calibration of CA is the study by Wu, (2002). The CA model of Wu, (2002) has calculated what is called a joint probability of cells to be developed to urban land in order to simulate land use development dynamics for the city of Guangzhou, China. The emperical LR amodelling is used to calculate the global b probabilities (prbabilities due to the global factors of land use development) of the cells, while constantly updating the global probabilities through the probabilities from local factors which is defined by another neighborhood function(Wu, 2002). The application of LR analysis for similating land use development dynamics has thoretically accepted by Wu, (2002), except that there will be a challenge in getting a data of each year to the simulatino and a match the time of simulation with time of land use change.

Another important similr research on LR and CA integration based on the idea of adding complexity for better and accurate simulation was the study by Poelmans, et al., (2010)(Poelmans & V.Rompaey, 2010). The application of LR in the research has been in aggrement with Wu, (2002). The LR modelling was first used independently to predict urban development by calculating probabilities of urban development in the Flamders-Brussels region of Belgium, followed by the CA modelling of the same region by considering only simple local probabilities of the cells. The integration of the two modells was eventually done to come up with what is called hybrid model. The evaluation of of the models has show in the study the more accurate and inter[retable results were found from the integrted model.

3. DATA AND RESEARCH METHDOLOGY

The focus of this chapter is to provide an overview of the study area, data; methodological approaches and steps followed in the research in order to achieve the objectives set. Reviewing the methodology is done by logical grouping of the research design into four main parts, Data analysis, modelling, model evaluation, and model comparison. The methodology is fundamentally done by making its basis on integrating logistic regression with CA modelling. Hence the last part of the chapter will try to make a comparison on the two types of models based on their outputs.

3.1. Study area

Dar es Salaam, the commercial capital of Tanzania, has been selected for this research as it can be a good representative of the rapid IS proliferation in the SSA region. According to the official

Figure 4: Location of Dar es Salaam, source: http://en.wikipedia.org/wiki/File:Tz-map.png

2002 census Dar es Salaam is the biggest city of Tanzania with a population of 2.9 million with a growth rate of 4.39% annually. It was founded by Arab Merchants as a coastal city in 1862 being the region's centre of trade. Till then, DAR has been nothing more than a village. However, DAR

has made rapid developmental stages since its establishment. Historically those developments can be traced in three main periods: a stage which DAR has been during establishment in 1962; the colonial period as a European occupation by Germans and British for 74 years; and the post-colonial period. DAR is geographically located at $6^0$48' south, $39^0$17' East (-6.8000, 39.2833) surrounded by a big portion of the Indian ocean on the East (Figure 4). Administratively DAR is divided into three districts: Ilala, Kinondoni and Temeke.

IS development has been the major issue of DAR. Studies show that 70% of the city is covered with ISs(URT, 2006).The development of ISs in DAR which is manifested in expansion and densification is a result of the mismatch between the high residential land use demand and the inefficiency of the formal market to satisfy the need (section 2.1.3).

3.2. Data

Data inputs to the model are basically derived from ITC data base. The data has made a substantial contribution in this study of IS development modelling as most the original documents were made during researches on IS issues (Sliuzas, et al., 2004). For instance, in the data set, ISs were put as a distinct land use types during the land use data preparations. The other important feature of the data set would be similarity in the spatial extent of the data for the years 1982, 1992 and 1998.Although, the available data does not cover the whole extent of Dar es Salaam, the incorporation of the major IS development areas of the city in the existing data set makes the study more important to the overall understanding of IS development. The land use data sets which have been produced by using various images (vertical aerial photographs for 1982 and 1992, SFAP and SPOT for 1998 and more recently IKONOS images for 2002) and other details of data preparation can be found (Sliuzas, et al., 2004).

Shape files, which were used in the factor map preparation with population figures were also important components of the data. The list of the data used in the during the model development are listed on (Table 1)

Data availability and the quality of the existing data are the major issues in modelling IS expansion. The first steps of the research needed the gathering of local knowledge on different factors or drivers of IS expansion through experts opinions and previous studies(Abebe, 2011a). Hence the gathered data has been made to incorporate all factors of IS expansion. However the challenge has been in representing all the factors from the available data set. In this sense the compilation of the independent variables or the factor maps has been restricted to the available data. This would to some extent have an influence to the overall result(Hill & Lindner, 2010). The quality of the existing data, that is, the existence of data error was mentioned but that has not made considerable obstacles in the results.

Table 1: List of data used in the research (adapted from Abebe (2011)

File	Format	Type	Provided by	Description
Municip*	.shp	polygon	ITC	Boundaries of municipalities
populatio*	.shp	polygon	ITC	Population 1992, 1998; area, density, ward name, formal/informal
censusDar2002* dar_wards_2002*	table		ITC	Population/ education/ etc. by ward
landform*	.shp	polygon	ITC	Landform categorised: river valley/ floodplain, swamp, salt pan, quarry, coastal plain, hills, ocean/ estuaries;
dem20*	.tif		ITC	20x20 meter pixel size DEM covering entire study extent
roads*	.shp	line	ITC	Roads, centrelines – categorised: major (1) and minor (2) roads.
dsmlu75_02*	.shp	polygon	ITC	Area; Land use classes 1975, 1982, 1992, 1998, 2002.
cbd_markets*	.shp	point	ITC	Location of CBD and main food markets
rivers*	.shp	line	ITC	Main rivers and streams categorised; ward name
infset92 and infset98	.shp	polygon	Richard Sliuzas	ILWIS data with name informal settlements at 1992 and 1998
landuse2012 and landuse2022	.shp	polygon	Richard Sliuzas	Predicted land use; classes: informal residential, planned residential, other urban

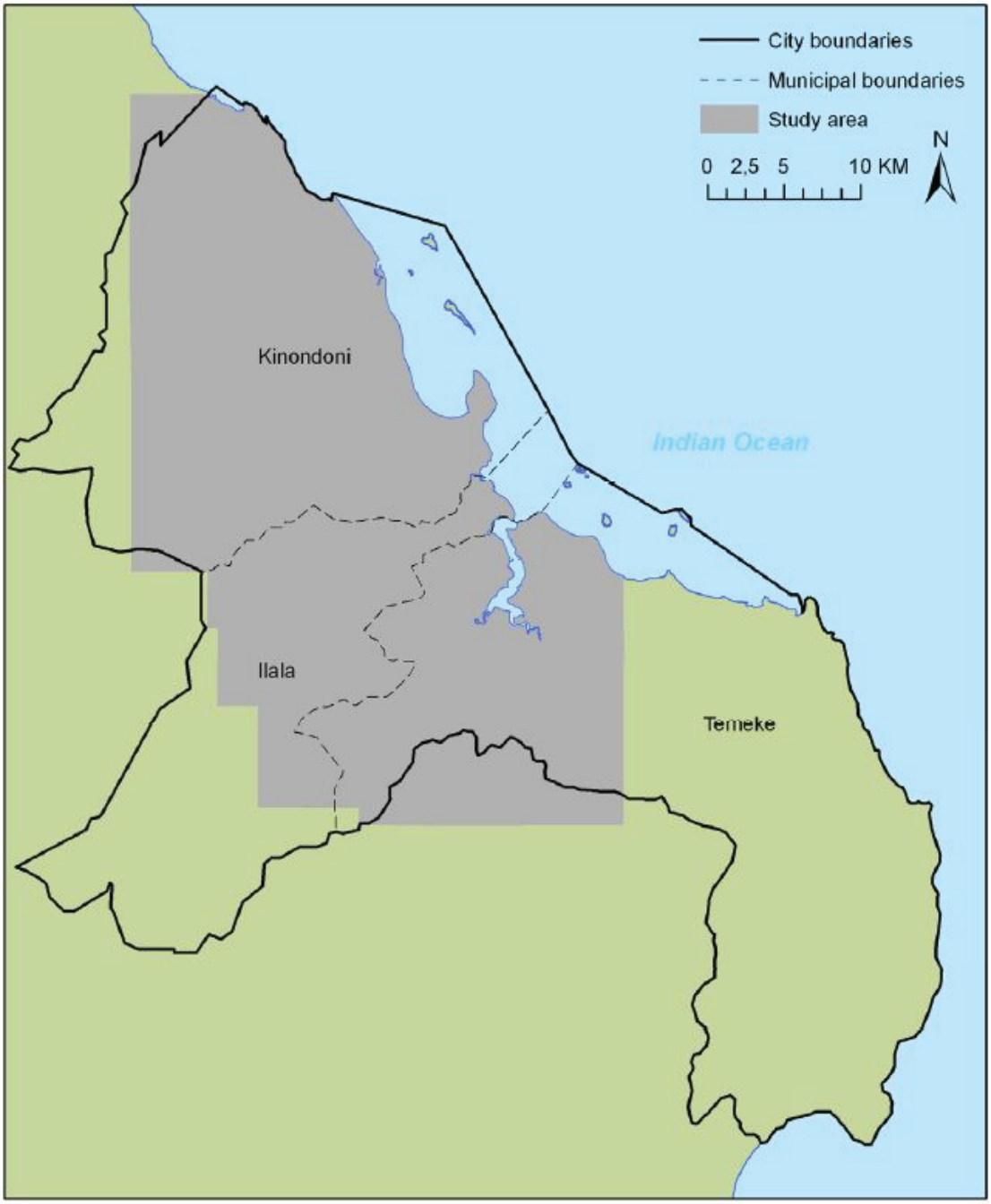
Georeference: Clarke1880 = GCS_Arc_1960 (Geographic Coordinate System), D_Arc_1960 (Datum), Clarke_1880_RGS (Spheroid)

*****File name appears as it does in ITC metadata

3.3. Scale, cell size and spatial extent

In the simulation of IS expansion spatial scale, cell size and spatial extent are important considerations to be made. In CA simulation the state of the cell, the time and the space are discrete. Spatial scale is defined as a window through which the perception of reality is made (Mnard & Marceau, 2005). Cell size is an area of land represented by the state of the cell. It tells how much area of land in reality is covered by an extent of a cell. For instance, a 20m x20m cell size represents a 20m x20m area of land on the ground. Several considerations can be taken in determining the size of a cell, the major ones, however, are the availability of data over consecutive years, the availability resource, and additional information on the spatial unit size (Hill & Lindner, 2010). Spatial extent in land use simulation can be defined as the overall spatial coverage of the simulation. The spatial extent of DAR data covers the vast majority of the urban areas which were covered before 2002. This includes around 97,000 ha of land. The limitations due to the spatial extent have become a constraining issue in the study. However, since the major developments before 2002 of the land use have been included in the dataset, the results of the research will have a valuable contribution in the IS development study in DAR. Cell sizes of 20x20 and 100x100 by the LRM of Abebe (2011) and CA based model of Hill and Lindner (2010) respectively were used in previously made IS modelling researches in DAR. In order to make logical integration when applying LR transition potential for CA calibration, a 100x100 meter cell

size is chosen for ease of processing the data in LR analysis software packages, such as, CLUE-s and Change analyst, which have a limitation in the number of cells to be processed for LR analysis. In addition to that, as the main focus of the research was on having a general overview on the simulation results and the locational drivers of the IS expansion rather than the smaller fragments of cells, the cell size 100x100 was used for the simulation.



3.4. Research Methodology

In this section the study would cover the major modelling stages. Aiming at model integration that would basically be seen as putting the major out puts of LR analysis to CA simulation of IS expansion. The overall methodology section of the study would be classified in two major parts, the part which mainly explains the data preparation and LR modelling process and the part which discusses how the integration and simulation is worked in the CA based model.

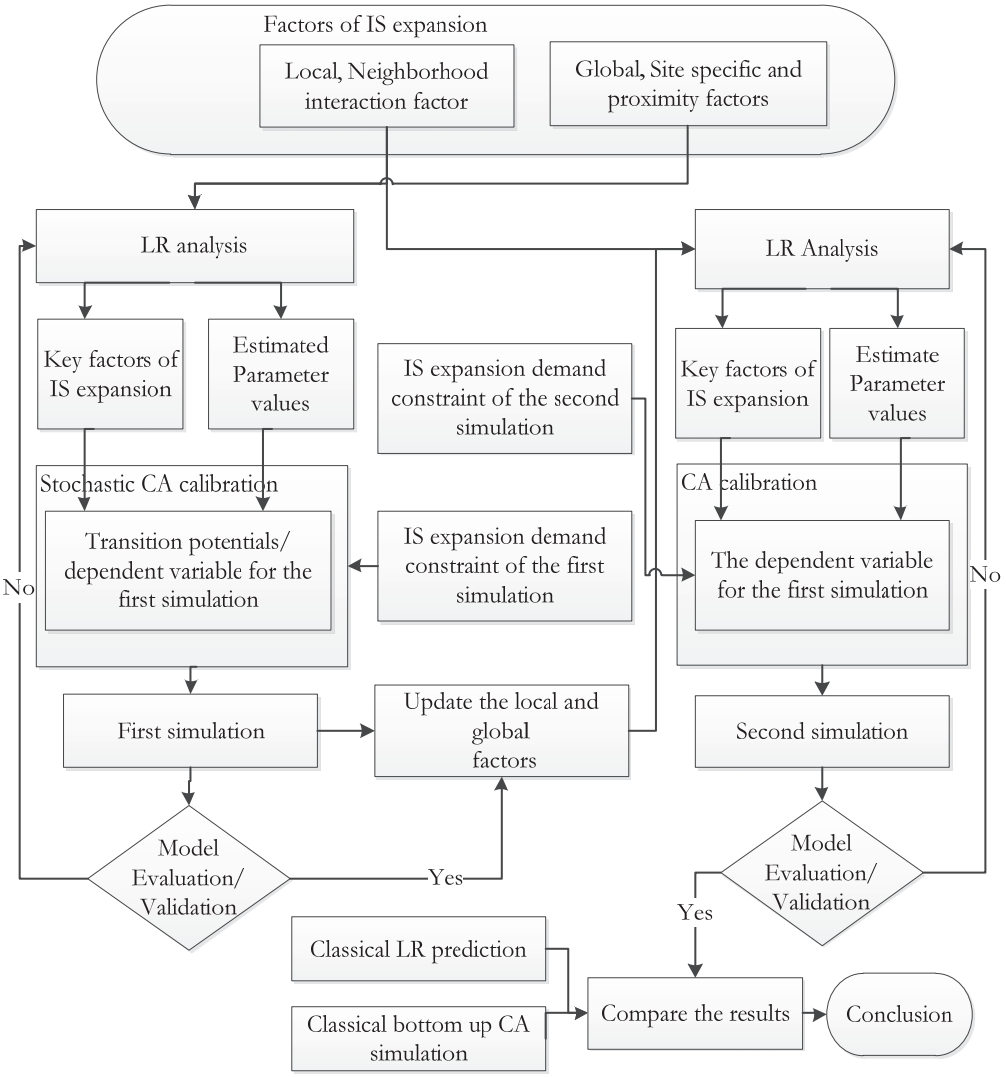


Figure 6: flow chart showing the methodological steps of the research.

3.5. LR modelling

The application of LR modelling needs identification of the probable drivers of the development and developing a model which should be robust to represent the ground data and predict the probabilities of the cells to be IS expansions.

3.5.1. Compilation of factors of IS expansion

As urban expansion is a result of factors of global attraction and local neighbourhood interactions (section 3. 5. 1. 1). Previous researches of IS expansion have classified the drivers of IS development into three classes, site specific, proximity and neighbourhood interaction factors (Abebe, 2011a; Dubovyk, et al., 2011; Sliuzas, et al., 2004), in this particular research, however, those factors are regrouped into two, global attraction factors and local neighbourhood facts. Therefore, in this section of the research those probable factors grouped under the two groups of factors, global and local of DAR IS expansion, would be discussed.(section 3.5.2.) In the discussion, while site specific and proximity characteristics were grouped under global factors of IS expansion, the neighbourhood interaction factors were classified under local factors of IS expansion.

3.5.2. Input Data Preparation

The input data preparation stage of binary LR modelling incudes the preparation of independent variables and binary IS maps (dependent variables) and of two discrete time steps. Compilation of all inclusive probable drivers of IS expansion and the dependent variable has previously been done by Abebe (2011) through a web survey with selected key informants and from literature review (Abebe, 2011b). Both types of input maps were prepared with a raster file format in ArcGIS environment.

The raster input maps of the dependent variables has been prepared for the years 1982, 1992 and 2002 having a dichotomous values of 0 and 1 showing being non-informal and being informal of the cell respectively. On the other hand, the independent variables which represent the drivers of IS expansion have been prepared as input factor maps for the years 1982 and 1992. Figures 8, 9 and 10 show those factor maps which are applied for the years 1982, 1992 and factor maps commonly applied for both years respectively. Input data preparation in land use modelling is needed to consider two types of drivers, local and global factors of land use development as urban land use development consists of two processes (Wu, 2002). The first type of process takes place is independent of the changes from local interaction, and the sequential land use change. In this process land use change takes place by the propensity of the development. The second process of development is a development that results from the development in its neighbourhood (Wu, 2002). Accordingly the factors of IS development are divided in to two, such as, factors of local interaction (local factors) and those factors independent of local interaction (global factors).and hence, the factors which were compiles as all-inclusive (site specific, proximity and neighbourhood factors) are regrouped to represent the two land use processes. This has been also to show the potential of the LR integrated stochastic CA model in incorporating both processes and its capacity to represent realistic simulation of IS expansion.

Table 2: Global probable drivers of IS expansion in Dar es Salaam.

The nature of global Input factor maps of IS expansion			
Number	**Category**	**Name of variable**	**Nature of cell value**
1	Proximity characteristic	Distance to CBD	Continuous
2	>>	Distance to existing ISs	Continuous
3	>>	Distance to food markets	Continuous
4	>>	Distance to hills	Continuous
5	>>	Distance to informal sub-centres	Continuous
6	>>	Distance to major rivers	Continuous
7	>>	Distance to major roads	Continuous
8	>>	Distance to minor rivers	Continuous
9	>>	Distance to minor roads	Continuous
10	>>	Distance to planned residential	Continuous
11	>>	Distance to river valleys	Continuous
12	>>	Distance to satellite centres	Continuous
13	>>	Distance to Ocean	Continuous
14	>>	Distance to other urban	Continuous
15	Site specific characteristic	Population density[person/km2]	Continuous
16	>>	Environmental Hazard	Dichotomous
17	>>	Slope[%]	Continuous

Table 3: Local interaction drivers of IS expansion in Dar es Salaam.

Local neighborhood interaction Independent Input factor maps of IS expansion			
Number	**Category**	**Name of variable**	**Nature of cell value**
1	Neighbourhood characteristic	Proportion of urban land in a surrounding	Continuous
2	>>	Proportion of IS in a surrounding	Continuous
3	>>	Proportion of undeveloped land in an area	Continuous

Table 4: Dependent variables which made inputs to the LR analysis

Dependent variable input maps			
Number	**Variable in LRM**	**Name of variable**	**Nature of cell value**
1	IS expansion	1982 IS map	Dichotomous
2	>>	1992 IS map	Dichotomous
3	>>	2002 IS map	Dichotomous

3.5.2.1. Global factors of IS expansion

As LR modelling technique is basically static it can be used to reflects the global distribution of IS changes. Inherently, it does not reflect the self-organisation nature of land use development (Hu & Lo, 2007). From In land use modelling, land use development factors such as IS expansions drivers are represented by urban spatial structures such as road networks, public facilities and the use of the land. This relationship is expressed by bid-rent theory in urban economics Wu, (2002), that is, different land users have their own varied utility functions to buy a land for a particular use due to preferences and their willingness to pay(Wu, 2002). In reality, in giving a value to the land a user makes a preference between various attributes such as physical characteristics of the development sites, accessibility and transport costs and make a trade of between the attributes, for instance land rent and transport costs and the biggest bid gets the land. These attraction factors which are independent of local scale land use clustering and can be grouped as global factors of IS expansion are site specific factors and proximity factors. The detail of the data organisation and compilation of the site specific and proximity characteristics can be read in Abebe (2011).

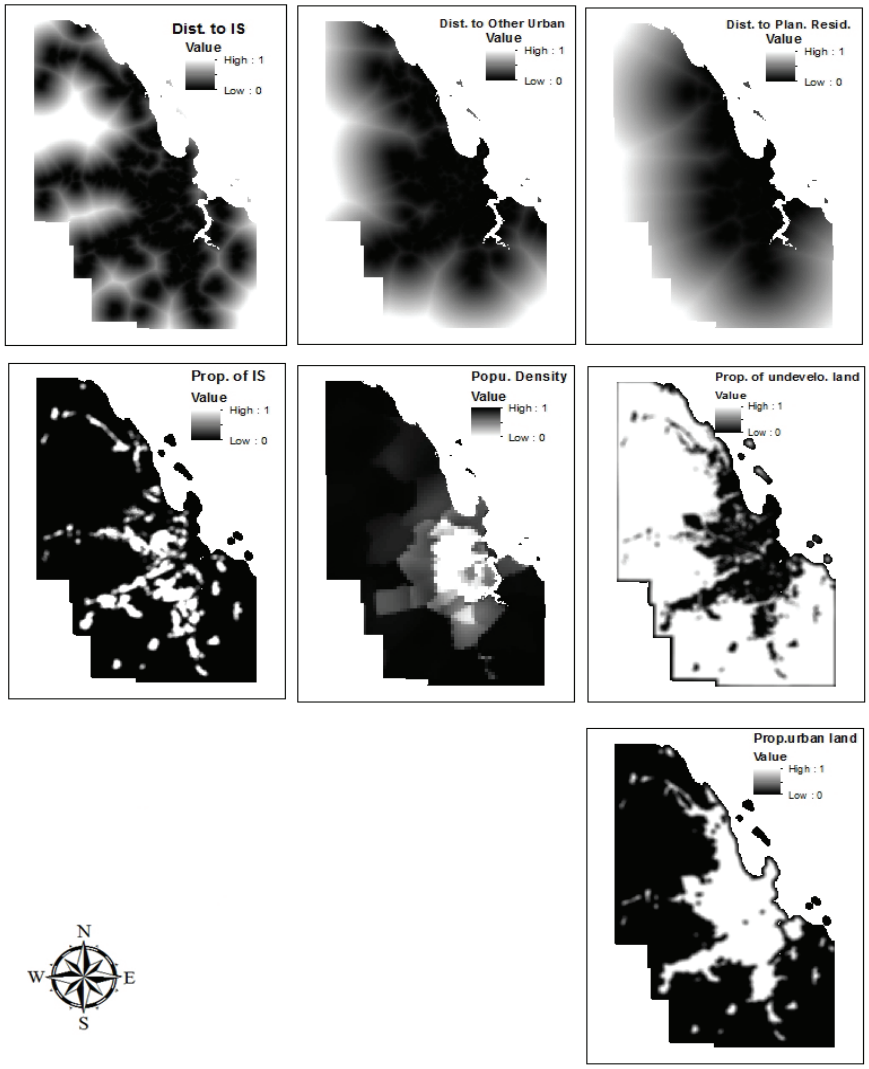


Figure 7: Factor maps 1992

Figure 8: Factor maps 1982

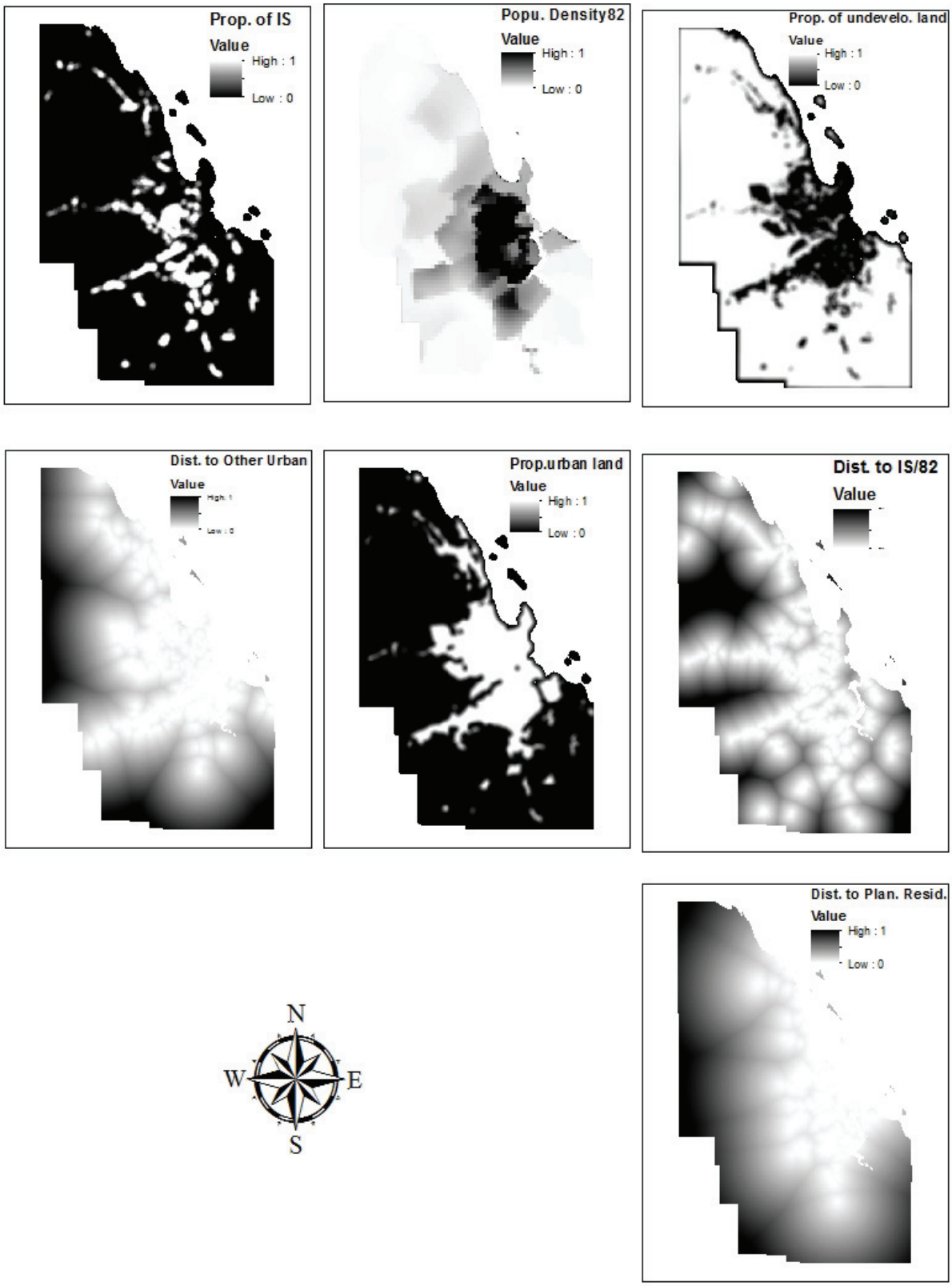
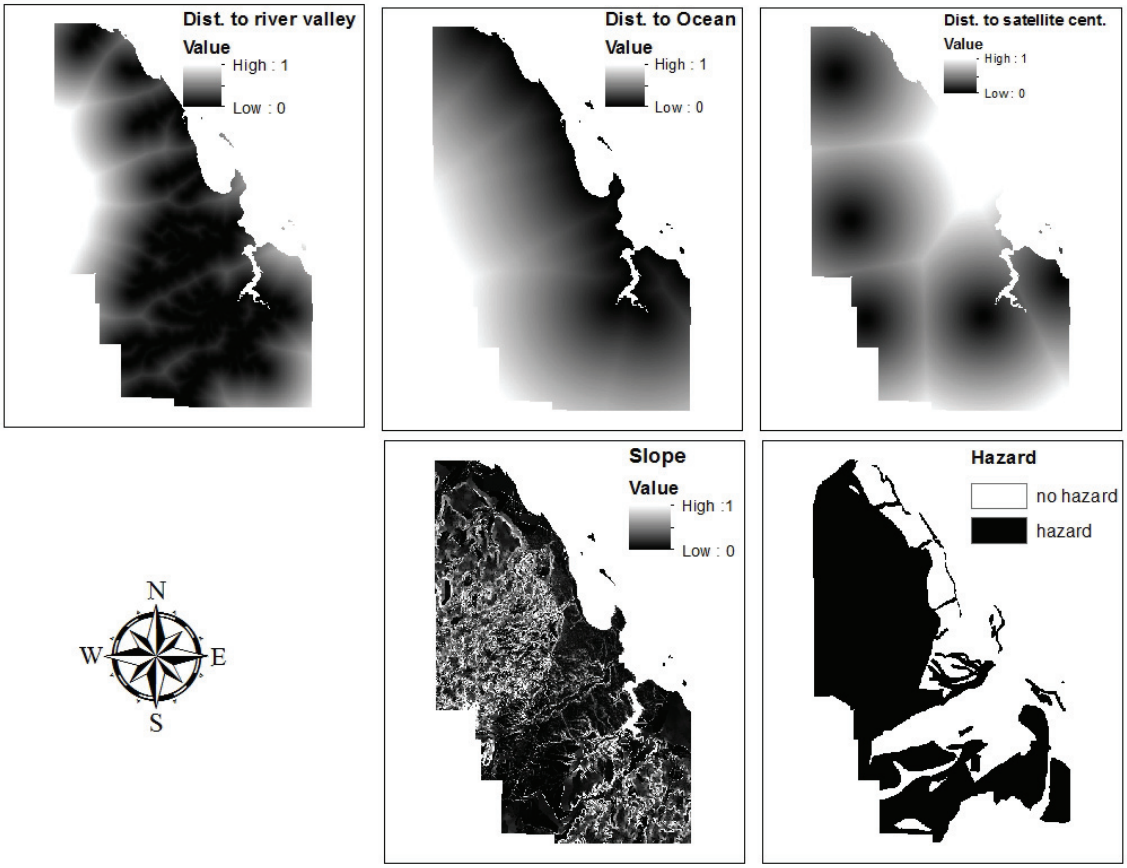


Figure 9: Factor maps common to both 1982 and 1992.



The site specific which can be categorised under global factors of IS expansion in DAR have been population density, environmental hazard and slope. The global drivers compiled as proximity characteristic probable drivers are road layers, different land uses, public facilities and the sub-centres. Distance map for both minor and major roads have been prepared for all years. However, as the available data has not been updated for the all input years the road network data has been the same. This has also be a case in reality in Dar es Salaam for the major roads, except some changes on the minor roads networks. The attraction of different land use classes, such as industrial, commercial, institution, and recreational, which are all grouped under other urban land uses except for the 1982 land use data were also global drivers of IS expansion in Dar es Salaam. Distance maps to natural land scape elements, satellite centres and central business districts have been produced as a proximity inputs to the LR model. The distance map to informal sub-centres has been integral parts of the probable drivers of IS expansion and crucial input to the LR analysis as the informal users mainly relay on the informal sub-centre rather than the CBD. The distance map of the informal sub-centres has been produced from a point data digitalized by Abebe (2010) from a map by Hill and Lindner (2010).

3.5.2.2. Local factors of IS expansion

As land development follows not simply geographical attractiveness but it is also decided by the sequence of local land use development by neighbourhood effect(Wu, 2002). The application of LR in order to

calculate the probabilities of the cells' ISs from local factors for CA based modelling simulation has only been theoretically accepted. In this study, however logical approaches would be made to incorporate LR modelling for calculating local probabilities of cells to be ISs for stochastic CA based simulation IS expansion in DAR. According to the survey to experts by (Abebe, 2011a), IS expansion is locally influence by the concentration of the three major land use classes in the neighbourhood, such as, proportion of urban land in a neighbourhood, proportion of IS in a neighbourhood, and proportion of undeveloped land in a neighbourhood.

3.5.3. Multicollinearity analysis

LR models of several explanatory variables may encounter multicollinearity effects among the variables. Multicollinearity is "correlations among predictors making it seem that no one variable is important when all the others are in the model"(Alan Agresti, 2002). Multicollinearity is a condition in which one predictor variable is highly correlated to another predictor variable in the LR model such that the requirement for independence is not met(Kutner, Nachtsheim, & Neter, 2004). High spatial multicollinearity, between the derivers of IS, is identified by high values standard deviation in the LR coefficients. Multicollinearity between variables would lead to a wrong decision and the acceptance of the null hypothesis which assumes that all the coefficients be zero(Owen, 1988). Multicollinearity would also create a wrong signs and magnitudes of regression coefficients resulting in mislead conclusions. In this study LR analysis has been done on SPSS software package. Multicollinearity was detected using a stepwise regression with different combinations of independent variables in the LR analysis and the last step coefficients are taken as deleting such redundant is helpful. Independence between variables is a precondition in LR modelling (Cheng, Masser, & Ottens, 2003; Kok & Veldkamp, 2001). The value of the variance inflation factor (VIF) indicates whether there is multicollinearity between predictors or not. Any predictor with VIF of greater than 10 eliminated at each step of the diagnostics(Field, 2009). Hence, according to their vale of VIF factors would be either allowed or rejected from further analysis in the LR model.

3.5.4. Sampling scheme

As Correlation statistics show relationships between variables, and autocorrelation statistics were to show correlations within variables, spatial autocorrelation shows the correlation within variables through space (Arthur, 2007). The basic idea of dealing with Spatial autocorrelation is that since it can make vague the results of regression analysis as regression coefficient and significance level of individual variables are sensitive to its presence (Kok & Veldkamp, 2001), and misled conclusions on the hypothesis tests (Irwin & Geoghegan, 2001).Therefore sampling has to be made prior to LR analysis in order to minimize effects of. Spatial autocorrelation within variables would be due data measurement errors, omitted variables or spatial interactions of observations (Irwin & Geoghegan, 2001). In logistic regression modelling certain sampling schemes, such as stratified random sampling and systematic sampling has been frequently used. While random sampling works efficient in representing population and performs low on reducing spatial dependence, systematic sampling works efficient in spatial dependence reduction but may lose data on isolated sites if population is not homogeneous(Cheng, et al., 2003).I this study, a random sampling scheme has been applied in order to minimize the effects of spatial autocorrelation.

3.6. Stochastic CA based modelling

The technique of LR integrated CA modelling has been used for the simulation IS expansion dynamics in Dar es Salaam. The LR analysis has been made in order to estimate the parameter values and the coefficients of the variables as well as the transition probability maps as a core analysis in the stochastic calibration of CA. Several researchers have used LR modelling for estimating the parameter values and calculating the global attraction probabilities of the cells in definition in the calibration of land use CA models(Hill & Lindner, 2010; Poelmans & V.Rompaey, 2010; Wu, 2002).

3.6.1. Calibration of CA by LR predictions

Calibration is an important issue to be addressed in developing CA models as a reliable procedure for urban growth simulation (Wu, 2002). Urban development is not purely a local process; it is also a global process with a complex pattern. The Simulation of IS dynamics is required to consider that the IS expansion can be best represented with the combination of global and local factors (section 3.6.3). The aim of model calibration is to reproduce the land use of a reference year. without calibration it will be impossible to correctly describe the behaviour of the system and predict the pattern of the future development (Wu, 2002). Calibration determines the parameter value from observed process of state change. In this process numerical values are assigned to the model parameters to ensure that the model simulation reproduces the phenomenon observed in reality realistically.

The cumulative probability of a particular type of development, for instance IS expansion, occurring in a particular cell can be estimated by calculating the transition probability of the cells due to all the possible drivers of the development.

So far, calibration of land use development in CA modelling has been done through: (1) Intensive computation in which repetitive running of the same model with different combinations of parameter value (Clarke & Gaydos, 1998). (2)Through automatic training by NN (Li & Yeh, 2001). Though the meaning of the parameter values might be difficult to interpret, it was able to automatically retrieve the parameter value.

In this study LR model was developed in order to calibrate CA based model in such a way that the LR model is used to calculate the probability of the cells in definition. This would help to formalize the calibration of CA so as to reduce the complexity of the model development. From this point of view, the purpose of calibration would be to extract the coefficients or parameter values of the rules from the observation of land use pattern at time t and t+1 an predicting the probability potential of the cells in definition. This can be expressed as the estimation of the probability of a particular cell at location (i, j) mathematically through a function of independent variables (development factors) (x_1, $x_{2,}$,…x_n). According to the logistic model, the probability of a site experiencing land conversion can be computed as:

$$Pex(sij = informal) = \mathrm{a} + \frac{\exp(Z)}{\exp(Z)} = \frac{1}{1+\exp(-Z)} \quad (1)$$

Where, p_{ex} is the observed IS expansion probability of a cell, S_{ij} is the state of the cell i,j, $\mathbf{z}$ is a vector that describes the development features of the site:

$$Z = \mathrm{a} + \sum bkXk \quad (2)$$

Where a is a constant, b^k are coefficients of the regression model; x^k is a set of site attributes.

Here, what is to be known is that 'the probability is estimated as a result of local and global factors of urban expansion but without time denomination (although the probability is estimated from sequential data).'
In the simulation of IS expansion dynamics by Logistic regression integrated CA model can be used to interpretation of the urban dynamics in terms of the drivers involved in the change of spatial pattern. The LR model outputs of probability maps can be used as transition potentials on which the transition rules rely. This is because of the fact that LRMs can calculate (estimate) the parameter values and specific parameter coefficients (Wu, 2002). LR model has been used to model the relationship between land use changes and the driving forces based on historic data. In this particular research LRM is used to identify the influence of independent variables and also to

provide a degree of confidence about their contribution (Hu & Lo, 2007). Hence, the result of The LRM is used, in the study, to structure the calibration process of stochastic CA models and assist the simulation of CA in defining transition rules. Dependent variables of IS expansion and a number of independent variables that would represent the probable driving forces of the expansion have been used in the simulation. The statistical mode of logistic regression can be given by the equation:

$$P(Y) = \frac{1}{1+e^{-(bo+b1X1i+b2X2i+\cdots\ldots+bnXni)}} \quad (3)$$

Where *P(Y)* stands for the probability of *Y* occurring (i.e. the probability that a case belongs to a certain category), *e* is the base of natural logarithms, b_0 is a constant, b_n is coefficient (or weight) attached to a predictor, and X_{ni} is a predictor. The resulting value from the equation ranges from 0 to 1. A value close to 0 means that Y is very unlikely to have occurred in the next time step and a value close to 1 means that Y is very likely to have occurred in the time step (Field, 2009). For instance, the dependent variable – expansion of IS– is binomial, either 1 or 0 which indicates the presence of expansion or no expansion, respectively. When Logistic regression is expressed by the logit function:

$$\log\left(\frac{\rho}{1-\rho}\right) = \beta_0 + \beta_1 x \quad (4)$$

Where, ρ is a binomial proportion and x is the explanatory variable. The parameters of the logistic model are **β0** and **β1**.

3.6.1.1. Parameters of Logistic regression

The different parameters of LR and their interpretation in land use modelling shall be discussed in the following. By looking at the association of independent and dependent variables LR brings an outcome of different parameters which can tell the association of the variables. The following are some of the most common parameters in LR analysis.
Chi-square statistic: is used to evaluate the goodness of fit of the LRMs. It tests the association of the variables. The equation applied to calculate the chi-square statistics is:

$$x^2 = \sum \frac{(observed\ count - model\ count)^2}{model\ count} \quad (5)$$

Where, x^2 is the chi-square, observed count is the value of the cell, and model count is expected/ predicted values of the cells.

The values indicate how the influence of the variable would be on the outcome of the model. In the chi-square test p-value estimated is compared with the significant level (α), and when the estimated p-value is smaller than α the null hypothesis H_o gets rejected. In the null hypothesis variables are assumed to have odds ratio value 1 so that no one variable will have any use in describing the dependent variable(Abebe, 2011a; Dubovyk, et al., 2011).

Odds ratio (Exp (β)): is a comparison of the ratio of the odds of an event in different groups. It indicates the increase or the decrease of a predicator by looking at the values(Field, 2009). The values odds ration starts from 0 and can go up to positive infinity. The value 1 has a threshold effect, that is, odds ratio value is below 1 it indicates that the predicator increases and hence the odds of the outcome occurring decreases while odds ratio greater than one indicates with increase of the predicator the odds of the outcome increases. The odds ratio 1, hence, indicates that altering a group makes no effect on the outcome or amount in the predicator.

T-Wald Statistic: is used to assess the significance of independent variables in LR analysis(Hu & Lo, 2007)

3.6.2. Stochastic perturbation (randomness effect)

As recent studies have tried to show, it is possible to include the effects of randomness, stochastic perturbation in the LR integrated CA model while calculating the probabilities of the cells. The randomness effect would be added at each predictor of during the LR analysis(Agresti, 2003). Those LRMs which include randomness effect in their analysis are categorised under Generalised Linear Mixed Models (GLMM)(McCulloch & Neuhaus, 2005). GLMM represents a group of regression models, such as, linear regression, logistic regression, and Poisson regression of continuous, dichotomous, and count data and include both fixed effects and the randomness effects in their analysis. The randomness effects in GLMM can represent a heterogeneity which might be caused by the absence of certain explanatory variables in addition to addressing cluster effects.
The mixed-effect LRM which is the most commonly applied GLMM which is applied for analysis of multilevel dichotomous data. It is represented by the equation:

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \log\left[\frac{\mu_{ij}}{1+\mu_{ij}}\right] = \eta_{ij} \quad (6)$$

Where the conditional expectation, $\mu_{ij} = E(Y_{ij}|(v_i\ , x_{ij})$ equals $\text{P}(Y_{ij}|(v_i\ , x_{ij})$, namely, the conditional probability of a respose given the random effects (and covariate values). The models equation can be rephrased as:

$$\text{P}(Y_{ij} = 1|v_i\,, x_{ij}\,, z_{ij}\,) = g^{-1}(\eta_{ij}) = \psi(\eta_{ij}), \quad (7)$$

Where the inversed link function $\psi(\eta_{ij})$ is the logistic cumulative distribution function (cdf), namely $\psi(\eta_{ij}) = [1 + \exp(-\eta_{ij})]^{-1}$
A nicety of the logistic distribution that simplifies parameter estimation is that the probability density function (pdf) is related to the cdf in a simple way, as $\psi(\eta_{ij}) = \psi(\eta_{ij})[1 - \psi(\eta_{ij})]$.

3.6.3. Neighbourhood definition

The neighbourhoods of CA model are sets of cells around a specific cell with which it interacts. Most two dimensional grids apply the common two types of neighbourhood definitions: The Van Neumann

neighbourhood (four cells), which includes the cells on the North, South, East and West of the cell in question; and the Moore neighbourhoods (eight cells), in this type of neighbourhood definitions in addition to the Van Neumann cells cell in the North-west, North-east, South-west and South east directions will be added in the neighbourhood of the cell in definition(Liu, 2009).
In reality a cell of land use is not influenced only by its immediate four or eight neighbouring cells rather it is also affected by the cells beyond its immediate neighbouring cells with the effect being reduced by distance decay. Hence, In this research a neighbourhood definition in which a neighbourhood with a circle of a six cell distance around the cell in question is be used to define the neighbourhood interactions(White, Engelen, & Uljee, 1997). Looking into the only adjacent neighbouring cells of a cell in definition neglects the influence of the cells which are around the cell in definition but not with immediate contact to it. Previous studies of urban land use simulation models, such as by Hill and Lindner (2010) have been using six cells distance around the cell in question (Hill & Lindner, 2010).

3.6.4. Transition rules

In CA modelling the transformation of land uses is based on the transition potentials of the cells and the application of Transition rules. The rule definition of CA modelling relies on the intuitive understanding of the process of cell state change, through some relationships between independent variables (drives of the change) and cell states. The number of transition rules is unlimited and difficult to select among many of them. Except few starts of researches, there is no standard procedure to define transition rules in CA models of urban dynamics (Wu, 2002). Hence, the application of the LR model to identify key drivers (independent Variables) of IS expansion would help to make a framework to the definition of the transition rules.

The transition potential of a cell to be developed as IS expansion will be calculated (estimated) by using LRMs (section 3.5.9). TP shows the specific potential of a cell to be converted to an informal residential. Here, a set of distinct rules shall be applied to make the process of land use allocation based on the cells' TP and the overall demand for IS expansion. The rules define show the state of a cell should be changed being dictated by the current state of the cell and the neighbouring cells state. The LR analysis estimated the transition potential of the cells by analysing the influence of both the Global and local factors of IS expansion. Therefore, the transition rule which converts the cells with the highest probability of being informal based on the exogenous demand of the year. Transition rules are essential components of a good CA modelling practice as they decide the nature of the process of the system being modelled (White & Engelen, 2000).Though, the definition of the transition rules in CA is based on the idea of creating simple local interactions between the cells in definition and the neighbourhood, it gives a rise to the complex pattern at the macro level.

3.6.5. Informal Land use demand calculation

Estimating exogenous informal land use demands has been one of the integral parts of this research. Various demand calculation approaches can be applied to estimate the amount of residential land use demand based on different assumptions made and the availability of data. In this research the average population density of residential land is used to estimate the informal land use demand. Land use demand at a time t in land use development is taken as the sum of planned residential and informal residential land area. In residential land use demand population growth is the main driver. Hence, theoretically the land use demand for residential land can be seen as the land area required accommodating the increased

population at the time t. The calculation of population density which is the population over the area of land would be the linking equation to calculate the informal land use demand. This basic assumption would lead to the fact that the total population over the total area would give the population density, hence, the product of the population density and the overall residential land use demand would give the total population of the time t. This is shown on the (equation 8).

Population increase at the time t= the average population density X the total residential land use demand.

$$Pt=D*L \quad (8)$$

Where Pt is the population increase at the time t, D the average population density and L the residential land use demand.

Therefore as the total residential land use is the sum of the informal residential land and the planned residential land (equation 9) can be rephrased as:

Population increase at the time t= (the average population density) x (Informal residential land at a time t + planned residential land of the same time the total residential land use demand.

$$Pt=D*(IL+PL) \quad (9)$$

Where Pt is the population increase at the time t, D the average population density, IL informal residential land at a time t and PL is the planned residential land of the same time. The application of the equations applied in the calculation of the informal land use demand would have been nothing without a logical technique of population projection and population density calculation. Therefore the next section would explain how the population calculations were made in the research.

3.6.5.1. Population density estimation

In the estimation of future ISs population figures of DAR different previous researches have been used for the prediction of the population density. Even though contrasting opinions were there on the fact that the population in ISs will continue to grow with increasing rate or not, the majority of the opinions have a consensus that population increase in ISs is anticipated to be increasing(Sliuzas, et al., 2004). The proportion of the IS residents has been increasing between 1978 and 2002 by 10%. In 1978, 60% of the city's population was living in unplanned settlements. This figure has grown to 70% in 2002(HABITAT, 2010). According to this figures almost within 24 years of time 10% increase was observed. For extrapolation and interpolation of the informal population density projections there are no enough data or estimated figures so far on the proportion of ISs for different years. Hence for population density calculations possible assumptions were made based on literature reviews. The IS population proportion estimations of two different years by UN Habitat and the areal increase from spatial data as well as the population projection by the Tanzanian office of Statistics were used to calculate the population density over the years(NBS TANZANIA, 2006; UN HABITAT, 2009). Assuming uniform growth, the percentage distribution of the population in the ISs was uniformly distributed for the years in between 1978 and 2002, that is, around 0.4percent increase for every year.

For the projection of the population in ISs since 2002, the proportion of the population in ISs was made to be constant, that is, 70% of the whole population. This assumption is made for the fact that IS expansion is DAR is expected to be a continuing problem in DAR, however, as Tanzania is one of the signifiers of the UN declaration of millennium development goals (MDG) and Habitat Agenda of adequate shelter for all(Kyessi & G. Kyessi, 2007) at least the increasing population of unplanned settlements will be kept constant by the upgrading and formalisation programs.

3.6.5.2. Population projection

The Malthus model of exponential growth model is used to estimate the future population projection of DAR. The population projection by the Tanzanian National Bureau of Statistics (TNBS) and the UN

(2008c) projection is used to validate the population projection made based on Malthusian exponential population growth model and scenario development. As can be seen on (section …Figure 3) of the world's the population projection trend LDCs and the African population growth trend (section…Figure X) follow a graph similar to the Malthus exponential model(Seidl & Tisdell, 1999). Figure 13 shows the nature of Malthusian exponential curve which most likely represents the population growth in LDCs. The annual growth rate of the year 2002 which was estimated to be 4.3% was applied for projecting the population(UN HABITAT, 2009).The NBS of Tanzania is applied for evaluating and scenario development as many researchers including UN Habitat have cited the population figures estimated by the bureau(HABITAT, 2010).

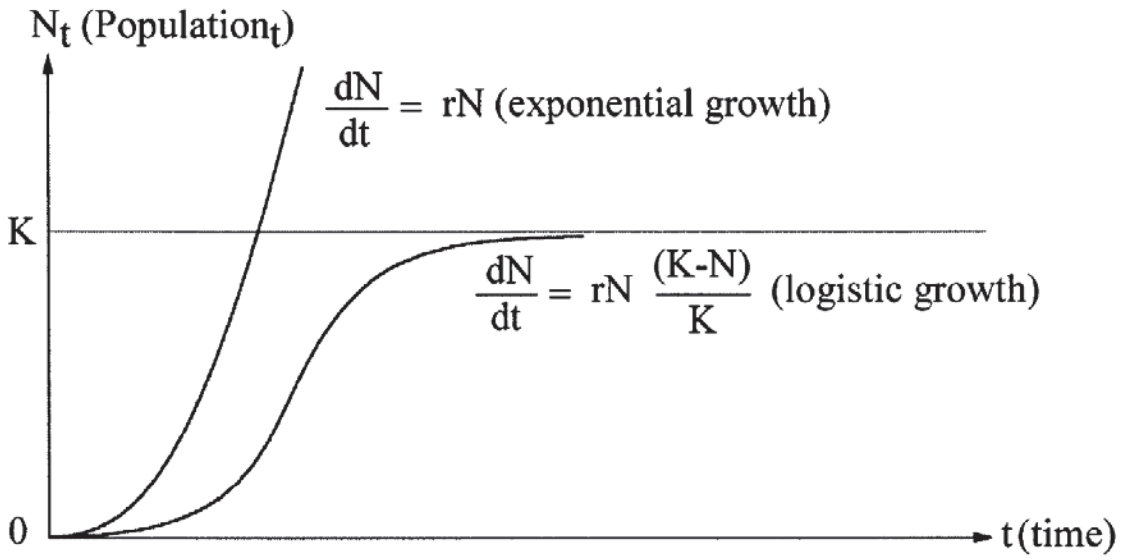


Figure 10: Malthusian Exponential and logistic population growth (Seidl & Tisdell, 1999)

The Malthusian equation is:

$$\frac{dN}{dt} = rN \tag{10}$$

Where N is the population and the Malthusian parameter r is the difference between birth rate (b) and death rate (d)
Hence the Malthusian equation can be rephrased as:

$$N_t = N_o(1 + r)^t \tag{11}$$

Where N_t the population at time t of year is, N_o the initial population size, and r the growth rate of the population.

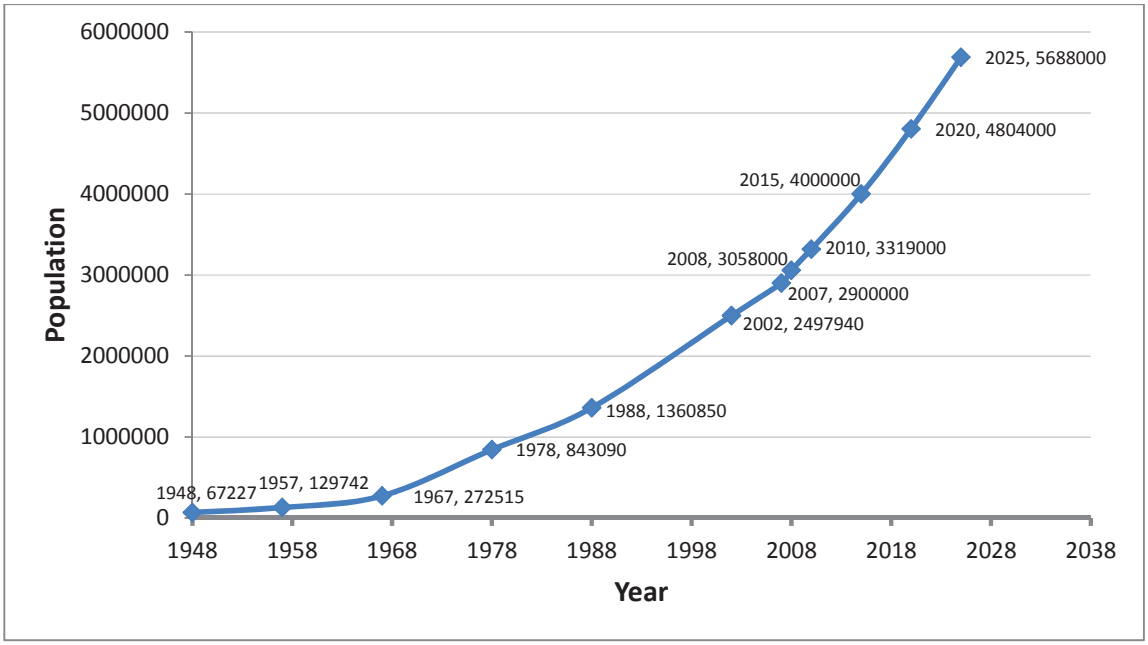


Figure 11: Dar es Salaam population projection (own source)(UN-HABITAT, 2008)

3.6.6. Constraint areas

Constraint areas, where no land use developments are expected to take place, need to be considered in land use models. Hence, land developments are prohibited from taking place in those areas which are constrained either due to naturally conditions or restricting regulations. However due to the nature of land use development in DAR, which is overwhelmed by the development of ISs, formal regulations do not that power to make planning restrictions(Hill & Lindner, 2010). Natural features which were constrained from being areas of IS expansion in other IS models in DAR are river valleys, swamp areas, and forest reservation. However in this study except the ocean all the land uses were included in the LR analysis to see the probability of the IS expansion in those areas. This was done because while developing the LR model the calibration has been done based on the IS expansion trend, which is the observed IS expansion (dependent variable) versus the drivers (independent variable), that is, if in the observed trend there is an IS expansion on those constraint areas it is so natural to include that reality in predicting the future IS expansion modelling. Therefore, for not interrupting with the model predictions, all the land use types are in the LR analysis. However, the results displayed are only for the simulations on vacant agricultural land.

3.1. Simulation of IS

For the simulation of the IS expansion the probabilities of the cells calculated as an out of the LR model, which as a GLMM model has included randomness on the calculation of the probabilities of the cells) would be constrained by the predicted demand of each year based calculated based on the observed demand of IS expansion, the population projection, and projected percentage of IS population in DAR. The simulation of the IS expansion for time t (year), would be applied to calibrate and predict the simulation at time, (t+1) as it updates the factor maps for the next simulation. This nature of self-organising has shown the incorporation of the basic characteristics of CA model in the general characteristic of the integrated model. However from the nature that the model generates its probability values not only from local factors but also global ones makes its inherent nature different from ordinary bottom-up CA models.

3.2. Model Evaluation

Assessing how robust a model would be in predicting the future land use development by using different methods of evaluation is an important component of land use modelling. The fact that evaluation of CA based models demands a data for evaluation after every iteration of land use simulation adds a challenge due to the unavailability of such data. The available data for evaluating the IS simulations were only for the years 1998 and 2002. Therefore in the assessment of the models two levels of evaluation by the integration of observable and non-observable model evaluation with subjective and objective approaches. Observable evaluations are mainly concerned with individual behaviour of the model with our referring to other models, while non-observable evaluations try to compare between different models. The two levels of evaluating the model were done during the development of the LRMs and after the simulation of the IS expansion of DAR. The central statistical (objective) evaluation approach used to evaluate the LRMs and the simulation of 2002 was ROC statistics. However, since due to the lack of land use data for the other years Google earth data was used to evaluate the simulation for 2011. The evaluation based on Google earth was based on a random sample data of the simulation taken on ArcGIS environment.

3.2.1. ROC statistics

The ROC curve which is known as the Receiver operational Characteristic curve is a sensitivity which is the true positive rate verses (1-specificity) or false positive graph on y-axis and on x-axis respectively for a possible threshold. ROC curve plot has mostly a concave shape and it connects (0, 0) and (1,1) points.

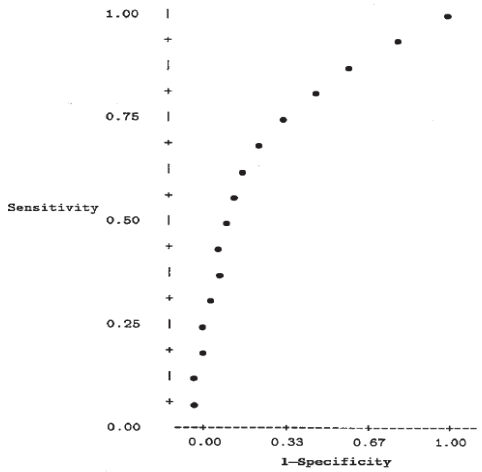


Table 5: ROC curve (source Agresti, (2003))

Conceptually the model is expected to predict most of the true positives and few false positives. The closer the curve is to the y-axis (left-hand border) and the top border of the graph (ROC space), the better the accuracy of the prediction. This is because of the high sensitivity and specificity even at low thresholds. On the other hand the closer the curve is to the diagonal shows the lower accuracy of models. The diagonal line in the graph is for random case. In that case the prediction will be at random of chances true positive being equal to that of false positive at ant threshold. Therefore predicting accuracy measure depends on the area under the curve (AUC). The higher the AUC the more accurate the prediction will be(Alan Agresti, 2002).

3.3. Error propagation and sources of error

In recent studies the validation of spatial models shows that spatial models contain a certain degree of uncertainty which can be because of error in the original data or the error of the model(Van Rompaey &

Govers, 2002). The errors from both the source data and the error propagates through the simulation process according to the type and the nature of the model(Poelmans & V.Rompaey, 2010). In this particular study errors may be caused and propagate through the different stages of the models. For instance, the preparation of the original data and the model input data can cause error as it uses different assumptions and operations applied. In the integrated modelling the propagation of error could have a decreasing effect as relative to LR potential errors can be reduced in CA simulation (Yeh & Li, 2006) due to the fact that the neighbourhood functions implemented in CA would have an averaging effect. However, due to the defining elements, such as, the neighbourhood, the transition rule, the cell size and the time of computation, CA models will have inherent uncertainties.

3.4. Softwares employed

During the different stages of the modelling process, such as, the data preparation, data analysis and simulation the following software environments were used: ArcGIS 10 software package, SPSS 16:00 from authorized licence of ITC; and additionally SPSS 19, R 2.12.1, Microsoft office software package.

4. RESULTS

This chapter of the research would present the main results of the study which were the findings of the research based on the methodologies discussed (section 3.5) in order to achieve the research objectives and questions of the research. Three main sections are discussed in this chapter: the conceptual integration of the LR and the CA model; the LR analysis and calibration of the CA model results; the informal expansion dynamics and based on the projected IS expansion land use demand. By presenting this results the chapter tries to show the logicality and advantage of integrating LR and CA based simulation of IS development in DAR.

4.1. LR integrated stochastic CA conceptual model

Finding logical ways to integrate LR and CA land use models has been an integral part of this research (Figure 1 and Figure 13). The bottom up modelling environment of CA needs to define cell to cell interaction rules within a defined neighbourhood, which are the prime aspects to predict the next time step state of the cell, or in this case the land use of an area in definition. The next state of a cell is hence decided by the transition rules. The attempt of the research was to facilitate and structure the rule definition of CA models by using the empirical LR model capacity to predict the probabilities of cells to be developed, in this case the cells of vacant agricultural cells. Probability values of the vacant agricultural land use calls were estimated based on the local and global drivers of IS expansion (section 3.5.1). Based on the interaction rule created by the LR model, the rule for the, a cell of vacant or agricultural land use to be informal would be 'cells with the highest probability value would be converted to IS use till the exogenous informal land use demand is satisfied'. Table shows the transition potentials of the first ten vacant agricultural calls for IS expansion in the next time step in a descending order, based on the IS expansion demand for the year 2002.

Figure 12: Conceptual model showing the simulation of IS expansion with LR and CA integration

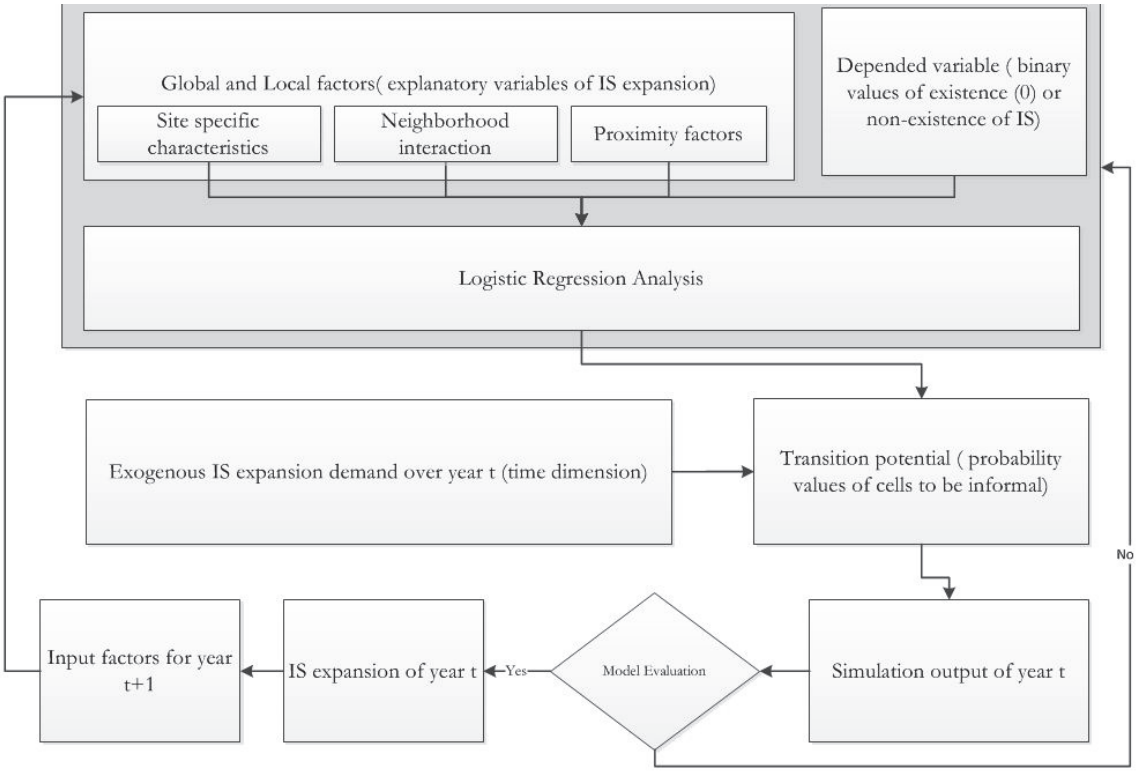


Table 6: A table showing the TP of sample cells in a decending order

		coordinates of cells	
Land use	Cell probability potential	X-coordinate	Y-coordinate
Vacant agricultural	2.30E-07	508874.6563	9263918
Vacant agricultural	2.13E-07	508874.6563	9264018
Vacant agricultural	1.76E-07	514574.6563	9261118
Vacant agricultural	1.68E-07	508774.6563	9264718
Vacant agricultural	1.67E-07	514574.6563	9261018
Vacant agricultural	1.66E-07	514574.6563	9260818
Vacant agricultural	1.64E-07	508874.6563	9264818
Vacant agricultural	1.63E-07	514574.6563	9260918
Vacant agricultural	1.52E-07	509274.6563	9264618
Vacant agricultural	1.48E-07	519274.6563	9229318

4.2. LR modelling for IS dynamics in Dar es Salaam

Three different LRMs were built for the calibration, calculating the transition potential, of the IS expansion dynamics model. The three models were built based on the IS expansion from the land use data of the years 1982, 1992, and 2002. The models A, B, and C were built on IS expansion from 1982 to 1992, 1982 to 2002, and 1992 to 2002 respectively. Those variables which have passed the multicollinearity assessment were used to build the three models. The method applied during running the LRM for identifying those independent variables with considerable influence for the model to predict was backward stepwise. The results for the estimation of the model parameters such as t-wald statistics were summarised in (Table 9,11,12) spatial autocorrelation was checked for residuals of the model so that those models with spatial autocorrelation would be removed from further analysis. The evaluation with ROC statistics was eventually made for the remaining models. Sample data for validating the prediction of the 2011 IS expansion simulation was taken from Google earth.

4.2.1. Dependent Variables for the developing LRM

A categorical data with two categories, binary, for existence of IS (denoted by value 1) and non-existence of IS (denoted by value 0) has been used to represent the dependent variables, IS expansion. The IS expansion change has been taken from the land use data of 1982, 1992, and 2002. Hence, three models have been developed based on the IS expansion change from 1982 to 1992, 1982 to 2002, and 1992 to 2002. This has been done following a similar step done by Abebe (2011), a previously done research on IS expansion.

4.2.2. Global and local, predictors for LRM of DAR IS expansion

After the preparation of the input data the global and local predictors, independent variables, have been applied to LR modelling in two major stages. The first stage of the application was during the model calibrated,(model development) where the dependent variables and the predictors of IS expansion, were exposed to LR analysis to see if there is, any association which can be described in LR equation and a number of parameters of association where checked next to the diagnosis for multicollinearity. The

second application stage of predictors has been during the calibration of the stochastic CA model, that is, to predict the future probabilities of vacant agricultural land use cells for IS expansion based on the coefficients and the constant estimated during the model development. Hence, the predictors were used to calibrate both LR and the stochastic CA models. Out of the twenty predictors listed by Abebe (2011), (Table 1, and 2) seventeen were applied in the LR analysis after being assessed for multicollinearity. (Figures 8, 9, and 10) show the independent variables used during model development.

4.2.1. Sampling

Minimising the effect of spatial autocorrelation in the LR modelling results, random sampling has been applied to the input data before the LR modelling was done. A random sample of 90,000 cells out of the total 97, 998 cells of the DAR data extent was taken during the sampling process in ArcGIS 10 environment. The sample size was made larger so that simulations made would be large enough to be visualized for possible result interpretation. Then those sampled data were exported to SPSS in a single attribute table for LR analysis.

4.2.2. Multicollinearity evaluation

For the compiled list of explanatory independent variables multicollinearity evaluation was done for the factors of all the three models, that is, for the variables of the years 1982 and 1992. The results of the multicollinearity diagnostic were significant for independent variables X_{18}, X_{19}, and X_{20}, which are the distance from food markets, distance to planned residential and proportion of urban land in an area respectively. The variables were eliminated from list and further application in the LR analysis as their VIF value is greater than 10 as their presence in the analysis would make biased and misinterpretation of the influence of the variables in the IS expansion.

The VIF values for the eliminated independent variables, X_{18}, X_{19}, and, X_{20} from the year 1982 were found to be 11.72, 36.124, and 21.784 respectively. In a similar manner independent variables of the same description were eliminated for having VIF values of 11.725, 35.88, and 28.274 respectively. The results of multicollinearity diagnostics to the compiled probable all-inclusive (global and local interaction) explanatory variables of IS expansion is summarised on (Table 23).

Table 7: a summary of the VIF for all predictors from 1982 and 1992

No	**Variable in LRM**	Description	**1982 (models A & B) VIF**	**1992(Model B) VIF**
1	X_1	Proportion of undeveloped land in a neighbourhood	2.848	3.411
2	X_2	Slope[%]	1.133	1.135
3	X_3	Population density[person/km2]	2.321	2.120
4	X_4	Proportion of IS in a neighbourhood	1.760	1.877
5	X_5	Environmental Hazard	1.423	1.456
6	X_6	Distance to satellite centres	1.593	1.644
7	X_7	Distance to river valleys	3.124	3.325
8	X_8	Distance to other urban	6.292	
9	X_9	Distance to Ocean	3.381	3.249
10	X_{10}	Distance to minor rivers	1.895	1.862
11	X_{11}	Distance to minor roads	1.381	1.409
12	X_{12}	Distance to major rivers	1.932	1.938
13	X_{13}	Distance to major roads	3.473	3.275
14	X_{14}	Distance to existing ISs	3.980	3.566
15	X_{15}	Distance to informal sub-centres	2.531	2.477
16	X_{16}	Distance to hills	2.560	2.390
17	X_{17}	Distance to CBD	3.368	3.313
18	X_{18}	Distance to food markets	eliminated	eliminated
19	X_{19}	Distance to planned residential	eliminated	eliminated
20	X_{20}	Proportion of urban land in a neighbourhood	eliminated	eliminated

4.2.3. LRMs and model parameters

A number of parameters estimation and coefficient estimations and evaluation were made during the LR modelling. By combining the IS expansion data from the land use and the independent variables of the years 1982 and 1992 three LR models, model-A,B and C were developed and evaluated to find the best fit model for the calibration process of the CA model. Model-A and Model-B were built on the IS expansion from 1982 to 1992 and 1982 to 2002, and independent variables from 1982, while the third model C was based on IS expansion from 1992 to 2002 and independent variables from 1992 data set provided that the factors from the base year are causes of the expansion.

4.2.3.1. LR Model- A (based on IS expansion between 1982 and 1992)

A number of model parameters were used to assess the model-A, built based on the IS expansion between 1982 and 1992. The overall model summary from SPSS shows that, the model is significant with 0.00, standard error 0.18, Wald of 35022, and chi-square of 19.458.

To test the significance of the predictors -2loglikelihood statistics has been verified. The summary (Table 8) has shown the addition of independent variables (predictors) to the model has dropped the -2loglikelihood statistics from 69148.667 to 55190.714 showing that the added predictors have increased the predicting capacity of the model. The R2 statistics on the same summary table has also reinforced the same idea by increasing from 0.101 to 0.337.

To determine the associated ***p-value*** the change the -2logliklihood has been transformed in SPSS and found out that a decrease of 13957.95 has shown $p<0.001$ proofing that the added predictors have significantly improved the model.

The Hosmer-Lemeshow test on the (Table 7) has shown the ***chi-square*** statistics of the model. The chi-square was computed by comparing the frequencies observed with the null hypothesis, of a linear relationship between the independent variables and the log odds of the dependent variable. This has been seen on (Table 8), that a comparison between the observed frequencies with the null hypothesis was made to compute the chi-square statistics. A 0.013 significant chi-square value of 19.458 was found showing that the data fits the model well.

The results from the (Table 9) of the variables in the equation, provide the summary of the LRMs coefficients and the odds of Model-A. The parameter values and the coefficients evaluate then individual contribution of each predictor in the context of others. This has been done by making other predictors constant so as to eliminate any overlapping effect between then predictors. As it is shown on (Table 9) except three predictors population density, environmental hazard and distance to CBD with significance of 0.476, 0.486, and 0.71 respectively all the remaining thirteen predictors were significant with less than 0.017 which much better than the conventional 0.05 significance level.

Table 8: overall model parameters of Model-A

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-3.415	.018	35022.814	1	.000	.033

Table 9: Chi-square statistics of Model-A.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	19.458	8	.013

Contingency Table for Hosmer and Lemeshow Test					
	IS 1982_1992		IS_1982_1992		Total
	Observed	Expected	Observed	Expected	
1	9745	9744.983	0	.017	9745
2	9745	9744.497	0	.503	9745
3	9735	9741.554	10	3.446	9745
4	9734	9733.016	11	11.984	9745
5	9723	9715.629	22	29.371	9745

6	9682	9680.308	63	64.692	9745
7	9610	9607.485	135	137.515	9745
8	9445	9454.751	300	290.249	9745
9	9141	9097.942	604	647.058	9745
10	7785	7824.836	1957	1917.164	9742

Table 10: Model-A fitness measures

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	23537.110[a]	.040	.162
2	22783.820[a]	.047	.192
3	22216.155[a]	.053	.214
4	21342.515[a]	.061	.249
5	20762.789[a]	.067	.271
6	19998.789[a]	.074	.301
7	19876.198[a]	.075	.306
8	19611.595[a]	.078	.316
9	19511.562[a]	.079	.320
10	19438.813[a]	.079	.323
11	19396.799[a]	.080	.324
12	19375.545[a]	.080	.325
13	19360.074[a]	.080	.326
14	19354.751[a]	.080	.326
a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.			

Table 11: Model-A (IS 1982-1992) model parameters and variable coefficients

No	Variable in the LRM	Description	B	S.E.	Wald	df	Sig.	Exp(B)
1	X_1	Proportion of undeveloped land in the neighborhood	3.501	.129	731.656	1	.000	33.157
2	X_2	Slope (%)	.823	.346	5.646	1	.017	2.277
3	X_3	Population density 1982	1.257	1.762	.509	1	.476	3.515
4	X_4	Proportion of IS in the neighborhood	1.472	.158	87.141	1	.000	4.356
5	X_5	Environmental hazard	.032	.046	.485	1	.486	1.033
6	X_6	Distance to satellite centers	-2.885	.164	310.482	1	.000	.056
7	X_7	Distance to river valley	-5.815	.266	476.113	1	.000	.003
8	X_8	Distance to other urban	-3.519	.234	226.859	1	.000	.030
9	X_9	Distance to ocean	3.755	.143	687.750	1	.000	42.739
10	X_{10}	Distance to minor river	-2.856	.275	108.112	1	.000	.058
11	X_{11}	Distance to minor roads	-12.049	.554	472.259	1	.000	.000
12	X_{12}	Distance to major rivers	3.458	.214	260.159	1	.000	31.756
13	X_{13}	Distance to major roads	-1.813	.239	57.640	1	.000	.163
14	X_{14}	Distance to IS 1982.	-11.861	.366	1049.029	1	.000	.000
15	X_{15}	Distance to informal sub-centers	.744	.166	20.005	1	.000	2.104
16	X_{16}	Distance to hills	.669	.168	15.832	1	.000	1.953
17	X_{17}	Distance to CBD	.374	.207	3.264	1	.071	1.453
		Constant	-4.046	.193	438.588	1	.000	.017

4.2.3.2. LR Model- B (based on IS expansion between 1982 and 2002)

The results of the model parameters used to assess model-B, which was built based on the IS expansion between 1982 and 2002 and predictors from 1982 are summarized in this section. The overall model summary from SPSS shows that, the model is significant with 0.00, standard error 0.10, Wald of 40048.48, and chi-square of 244.675.

To test the significance of the predictors -2loglikelihood statistics has been verified. The summary (Table 13) has shown the addition of independent variables (predictors) to the model has dropped the -2loglikelihood statistics from 69148.667 to 55190.714 showing that the added predictors have increased the predicting capacity of the model. The R2 statistics on the same summary table has also reinforced the same idea by increasing from 0.101 to 0.337.

To determine the associated ***p-value*** the change the -2logliklihood has been transformed in SPSS and found out that a decrease of 13957.95 has shown $p<0.001$ proofing that the added predictors have significantly improved the model.
The Hosmer-Lemeshow test on (Table 13) has shown the ***chi-square*** statistics of the model. The chi-square was computed by comparing the frequencies observed with the null hypothesis, of a linear relationship between the independent variables and the log odds of the dependent variable. This has been seen on (Table 13), that a comparison between the observed frequencies with the null hypothesis was made to compute the chi-square statistics. A chi-square value of 244.675 which is significant at 0.000 was found showing that the data fits the model well.

The results from the (Table 11) of the variables in the equation, provide the summary of the LRMs coefficients and the odds of Model-B. The parameter values and the coefficients evaluate the individual contribution of each predictor in the context of others. As it is shown on (Table 11) except three predictors population density, and distance to river valley with significance of 0.611, and 0.174 respectively all the remaining thirteen predictors were significant with less than 0.028 which much better than the conventional 0.05 significance level.

Table 12: Model parameter for 1982-2002 model

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.919	.010	40048.483	1	.000	.147

Figure 13: Chi-square statistics showing the fitness of model-B

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	244.675	8	.000

Table 13: Parameter values and estimated coefficients of model-B 1982-2002

Variables in the Equation								
No	Variable in LRM	Description	B	S.E.	Wald	df	Sig.	Exp(B)
1	X_1	Proportion to undeveloped land in a neighborhood	4.413	.079	3143.197	1	.000	82.478
2	X_2	Slope (%)	-.441	.200	4.841	1	.028	.644
3	X_3	Population density	-.618	1.213	.259	1	.611	.539
4	X_4	Proportion to IS in a neighborhood	1.034	.108	91.819	1	.000	2.812
5	X_5	Environmental hazard	.107	.026	16.596	1	.000	1.113
6	X_6	Distance to satellite center	-.665	.084	63.151	1	.000	.514
7	X_7	Distance to river valley	-.149	.110	1.845	1	.174	.862
8	X_8	Distance to other urban	-4.409	.125	1250.307	1	.000	.012
9	X_9	Distance to ocean	2.140	.081	690.592	1	.000	8.500
10	X_{10}	Distance to minor river	-1.328	.131	103.415	1	.000	.265
11	X_{11}	Distance to minor road	-6.763	.204	1103.350	1	.000	.001
12	X_{12}	Distance to major river	.946	.103	84.192	1	.000	2.576
13	X_{13}	Distance to major roads	.398	.120	10.910	1	.001	1.488
14	X_{14}	Distance to IS 1982	-6.106	.132	2127.896	1	.000	.002
15	X_{15}	Distance to informal sub-centers	-1.197	.095	160.242	1	.000	.302
16	X_{16}	Distance to hills	.940	.090	108.255	1	.000	2.560
17	X_{17}	Distance to CBD	-1.544	.104	222.295	1	.000	.213
		Constant	-3.145	.112	793.139	1	.000	.043

Table 14: Model fitness measures of Model-B

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	69148.667[a]	.054	.101
2	65011.219[a]	.093	.174
3	61890.366[b]	.122	.228
4	59528.889[b]	.143	.267
5	57294.713[b]	.162	.303
6	56775.887[b]	.167	.312
7	55821.874[b]	.175	.327
8	55573.157[b]	.177	.331
9	55458.874[b]	.178	.333
10	55372.170[b]	.179	.334
11	55307.001[b]	.179	.335
12	55226.927[b]	.180	.336
13	55210.455[b]	.180	.337
14	55196.303[b]	.180	.337
15	55190.714[b]	.180	.337

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

4.2.3.3. LR Model- C (based on IS expansion between 1992 and 2002)

The results of the model parameters used to assess model-B, which was built based on the IS expansion between 1992 and 2002 and predictors from 1992 are summarized in this section. The overall model summary from SPSS shows that, the model is significant with 0.00, standard error 0.11, Wald of 42502.103, and chi-square of 244.675.

To test the significance of the predictors -2loglikelihood statistics has been verified. The summary (Table 17) has shown the addition of independent variables (predictors) to the model has dropped the -2loglikelihood statistics from 69148.667 to 55190.714 showing that the added predictors have increased the predicting capacity of the model. The R2 statistics on the same summary table has also reinforced the same idea by increasing from 0.101 to 0.337.

To determine the associated ***p-value*** the change the -2logliklihood has been transformed in SPSS and found out that a decrease of 13957.95 has shown $p<0.001$ proofing that the added predictors have significantly improved the model.

The Hosmer-Lemeshow test on the (Table 15) has shown the ***chi-square*** statistics of the model. The chi-square was computed by comparing the frequencies observed with the null hypothesis, of a linear relationship between the independent variables and the log odds of the dependent variable. This has been seen on (table 15), that a comparison between the observed frequencies with the null hypothesis was made to compute the chi-square statistics. A chi-square value of 244.675 which is significant at 0.000 was found showing that the data fits the model well.

The results from the (Table 16) of the variables in the equation, provide the summary of the LRMs coefficients and the odds of Model-B. The parameter values and the coefficients evaluate the individual

contribution of each predictor in the context of others.. As it is shown on (Table 16) except three predictors population density, and distance to river valley with significance of 0.611, and 0.174 respectively all the remaining thirteen predictors were significant with less than 0.028 which much better than the conventional 0.05 significance level.

Table 15: Model parameter 1992-2002

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	-2.236	.011	42502.103	1	.000	.107

Table 16: A table of the Chi-square statistics of Model-C

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	120.250	8	.000

Contingency Table for Hosmer and Lemeshow Test					
	IS_1992_2002 =0		IS_1992_2002 = 1		Total
	Observed	Expected	Observed	Expected	
1	9744	9738.300	1	6.700	9745
2	9659	9698.466	86	46.534	9745
3	9626	9631.826	119	113.174	9745
4	9602	9551.650	143	193.350	9745
5	9517	9450.428	228	294.572	9745
6	9357	9290.562	388	454.438	9745
7	9016	8997.954	729	747.046	9745
8	8335	8467.199	1410	1277.801	9745
9	7399	7571.371	2346	2173.629	9745
10	5779	5636.242	3963	4105.758	9742

Table 17: Parameter values and estimated coefficients of the LR model-C

No	Variable in LRM	Description	B	S.E.	Wald	df	Sig.	Exp(B)
1	X_1	Proportion of undeveloped land in a neighborhood	5.175	.089	3365.112	1	.000	176.777
2	X_2	Slope (%)	-1.016	.222	21.031	1	.000	.362
3	X_3	Population density	-1.150	1.533	.563	1	.453	.317
4	X_4	Proportion of IS in a neighborhood	.602	.111	29.265	1	.000	1.825
5	X_5	Environmental Hazard	-.044	.030	2.131	1	.144	.957
6	X_6	Distance to satellite centres	.743	.098	57.028	1	.000	2.102
7	X_7	Distance to river valleys	2.173	.125	300.465	1	.000	8.786
8	X_8	Distance to other urban	-5.566	.140	1570.960	1	.000	.004
9	X_9	Distance to ocean	1.012	.090	126.770	1	.000	2.750
10	X_{10}	Distance to minor river	-1.615	.143	127.224	1	.000	.199
11	X_{11}	Distance to minor road	-6.323	.221	817.523	1	.000	.002
12	X_{12}	Distance to major rivers	-.277	.112	6.105	1	.013	.758
13	X_{13}	Distance to major roads	-.599	.133	20.390	1	.000	.549
14	X_{14}	Distance to existing ISs	-6.020	.150	1607.909	1	.000	.002
15	X_{15}	Distance to informal sub-centres	-1.281	.110	136.675	1	.000	.278
16	X_{16}	Distance to hills	1.220	.098	155.666	1	.000	3.388
17	X_{17}	Distance to CBD	-2.081	.116	320.171	1	.000	.125
		Constant	-3.605	.123	860.496	1	.000	.027

Table 18: -2Log likelihood and R2 evaluation of model-C

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	57671.461[a]	0.042	0.09
2	55823.818[a]	0.06	0.128
3	51978.813[b]	0.097	0.206
4	48943.660[b]	0.124	0.265
5	47826.948[b]	0.134	0.286
6	46831.384[b]	0.143	0.305
7	46534.397[b]	0.146	0.31
8	46379.343[b]	0.147	0.313
9	46309.051[b]	0.148	0.314
10	46218.736[b]	0.149	0.316
11	46172.163[b]	0.149	0.317
12	46148.486[b]	0.149	0.317
13	46128.610[b]	0.149	0.318
14	46109.680[b]	0.149	0.318
15	46104.585[b]	0.15	0.318

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

4.3. Population projection and Exogenous IS expansion demand

By interpolating the population estimated by UN-Habitat the population for the time of simulation were found. (Table 18) shows the interpolated population figures for the simulation time.

As it can be seen from the trend on table 4 the population density in unplanned areas has got decreased while proportion of population has increased. For the year 2002 almost 100 people were living in a hectare of land. As per the assumption made on this the population density shall remain either constant or will be decreasing. Therefore by taking the assumption that the proliferation of ISs shall remain be a continuing problem in DAR the extrapolation of the future IS land use demand was done. Table 5 shows the IS land demand calculated based on the UN-Habitat projection and by the assumption of 100 people per hectare of land.

Year	Population	Informal Land demand (ha)	IS population (%)	IS population	IS population density	population increase
1982-1992	1675000	3058	65.6	1098800	133.139464	
1993	1720000	245	66	1135200	148	36400
1994	1800000	367	66.4	1195200	164	60000
1995	1900000	414	66.8	1269200	179	74000
1996	2000000	385	67.2	1344000	194	74800
1997	2100000	361	67.6	1419600	209	75600
1998	2175000	265	68	1479000	225	59400
1999	2260000	346	68.4	1545840	193	66840
2000	2310000	268	68.8	1589280	162	43440
2001	2400000	548	69.2	1660800	130	71520
2002	2,497,940	886	70	1748558	99	87758
2003	2,600,000	714	70	1820000	100	71442
2004	2,680,000	560	70	1876000	100	56000
2005	2,750,000	490	70	1925000	100	49000
2006	2,810,000	420	70	1967000	100	42000
2007	2,900,000	630	70	2030000	100	63000
2008	3,058,000	1106	70	2140600	100	110600
2009	3,200,000	994	70	2240000	100	99400
2010	3,319,000	833	70	2323300	100	83300
2011	3,450,000	917	70	2415000	100	91700
2012	3,600,000	1050	70	2520000	100	105000
2013	3,710,000	770	70	2597000	100	77000
2014	3,880,000	1190	70	2716000	100	119000
2015	4,000,000	840	70	2800000	100	84000
2016	4,150,000	1050	70	2905000	100	105000
2017	4,300,000	1050	70	3010000	100	105000
2018	4,470,000	1190	70	3129000	100	119000
2019	4,620,000	1050	70	3234000	100	105000
2020	4,804,000	1288	70	3362800	100	128800
2021	5,000,000	1372	70	3500000	100	137200
2022	5200000	1400	70	3640000	100	140000

Table 19: Interpolated population figures for the years 1982, 1992, 1998 and 2002.

Table 20: Population increase and IS land use demand

Year	Population	Informal Land demand (ha)	IS population (%)	IS population	IS population density	population increase
1982-2002	2,497,940	17650	70	1748558	99.0684419	
2003	2,600,000	714.42	70	1820000	100	71442
2004	2,680,000	560	70	1876000	100	56000
2005	2,750,000	490	70	1925000	100	49000
2006	2,810,000	420	70	1967000	100	42000
2007	2,900,000	630	70	2030000	100	63000
2008	3,058,000	1106	70	2140600	100	110600
2009	3,200,000	994	70	2240000	100	99400
2010	3,319,000	833	70	2323300	100	83300
2011	3,450,000	917	70	2415000	100	91700
2012	3,600,000	1050	70	2520000	100	105000
2013	3,710,000	770	70	2597000	100	77000
2014	3,880,000	1190	70	2716000	100	119000
2015	4,000,000	840	70	2800000	100	84000
2016	4,150,000	1050	70	2905000	100	105000
2017	4,300,000	1050	70	3010000	100	105000
2018	4,470,000	1190	70	3129000	100	119000
2019	4,620,000	1050	70	3234000	100	105000
2020	4,804,000	1288	70	3362800	100	128800
2021	5,000,000	1372	70	3500000	100	137200
2022	5200000	1400	70	3640000	100	140000

Year	**Population**	**Informal Land demand (ha)**	**IS population (%)**	**IS population**	**IS population density**
2002	2,497,940	17650	70	1748558	99.0684419
2012	3,600,000	25200	70	2520000	100
2022	5200000	36400	70	3640000	100

4.4. GLMM A

In order to incorporate unforeseen conditions, the effects of omitted variables and also further minimize the correlation effects GLMM models was analysed on R 2.12.1 environment, and the new version of SPSS. The results of the GLMM parameter values are presented on (table 20). The parameter values show a similar response with the with the simple LRM that the key drivers of The IS expansion in Model-A are distance to minor roads (X_{11}) and distance to IS (X_{14}) with coefficients of negative relationship (-12.51929) and (-11.49880) respectively.

Table 21: Parameter values of GLMM

No	Variable in the LRM	Description	B	S.E.	Sig.
1	X_1	Proportion of undeveloped land in the neighborhood	1.61680	0.08636	2e-16 ***
2	X_2	Slope (%)	0.46908	0.35535	1.36e-10 ***
3	X_3	Population density 1982	-0.27372	2.16474	2e-16 ***
4	X_4	Proportion of IS in the neighborhood	1.12740	0.12546	2e-16 ***
5	X_5	Environmental hazard	0.03333	0.05221	2e-16 ***
6	X_6	Distance to satellite centers	-3.14996	0.17661	2e-16 ***
7	X_7	Distance to river valley	-5.75318	0.25458	2e-16 ***
8	X_8	Distance to other urban	-3.13450	0.25529	4.69e-16 ***
9	X_9	Distance to ocean	3.54611	0.15871	2e-16 ***
10	X_{10}	Distance to minor river	-2.76149	0.29994	1.09e-07 ***
11	X_{11}	Distance to minor roads	-12.51929	0.55422	2e-16 ***
12	X_{12}	Distance to major rivers	3.19631	0.22544	5.82e-05 ***
13	X_{13}	Distance to major roads	-1.89770	0.23150	3.06e-14 ***
14	X_{14}	Distance to IS 1982.	-11.49880	0.35776	2e-16 ***
15	X_{15}	Distance to informal sub-centers	0.72591	0.17509	0.2151
16	X_{16}	Distance to hills	0.41701	0.18319	2.54e-06 ***
17	X_{17}	Distance to CBD	0.34208	0.24189	0.0936
		Constant	-3.41789	0.16288	2e-16 ***

Significance codes 0 *** 0.001 ** 0.01 * 0.05 0.1 1

4.5. Model Evaluation

Based on the two levels of the evaluation of the models both the LRMs, 1982-1992 (model-A), 1982-2002 (model-B) and (1992-2002) and the simulation from the stochastic CA , simulation 2002, model were evaluated by using ROC statistics, while the simulation of 2011 was evaluated by using sample data from Google earth.

4.5.1. Evaluation of the LRMs

The predicting capacity and model performance of the LRMs were evaluated by using visual comparison and ROC statistics, which is one the best ways to evaluate land use models as it compares both true positive and false positive predictions(Alan Agresti, 2002). During the evaluation the individual models A, B, and C were assessed to see how robust the models would be in predicting the future. The summary (Table 21) shows the ROC evaluation for the three models which all of them were significant at 0.000. In the ROC evaluation Model-A has got the highest area under curve of 0.905 which is the measure of the predicting capacity of the model. The more near the area under the curve is to 1 the stronger the predicting capacity of the model is. The (Figures 15, 16, and 17) as well as (Table 21) show the ROC curve evaluation results of the three models

Table 22: A summary table showing the area under the ROC curve, for the models A, B and C

Area Under the Curve, summary					
Test Result Variable(s):Predicted probability					
Model	Area	Std. Error a	Asymptotic Sig.b	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
1982-1992	.905	.002	.000	.900	.909
1982-2002	.853	.002	.000	.850	.856
1992-2002	.857	.002	.000	.854	.861

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

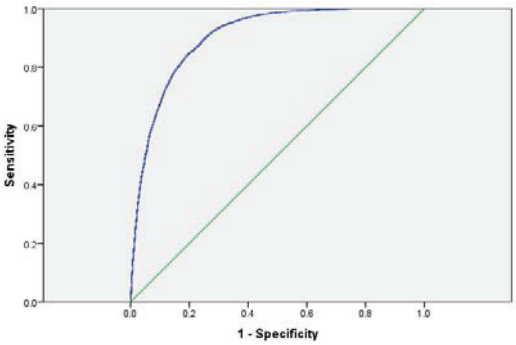


Figure 15: ROC curve of model 1982-1992

Figure 14:ROC curve of model 1992-2002

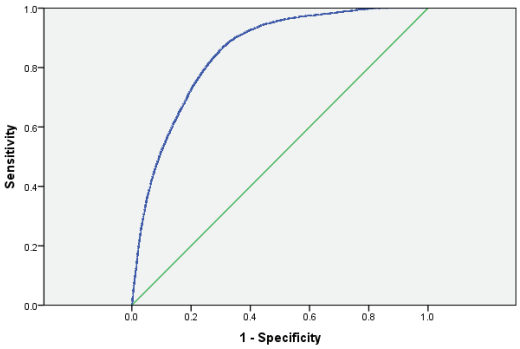
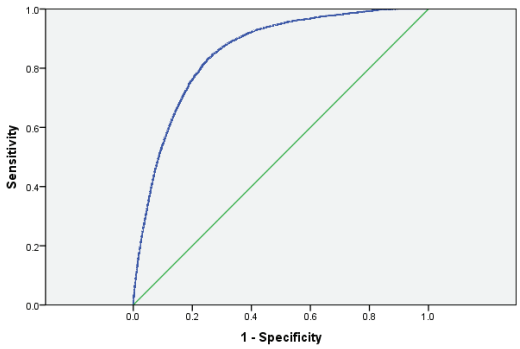


Figure 16: ROC curve of model 1982-2002

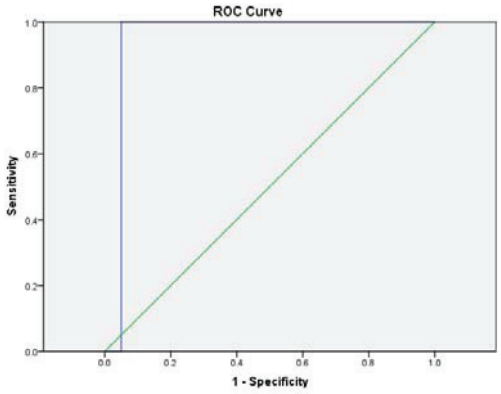


4.5.2. Evaluation of the CA simulations

The CA based simulation results have been evaluated twice in the whole process of the model first in the LR modelling stage and second after simulation. However, as the evaluation of CA simulation results is data intensive only those years, 2002 with ROC evaluation. In the yearly IS expansion simulation, only the land use data of 2002 was available to evaluate models. The simulation based on the combined LRMs was evaluated by using ROC and has shown an area under the curve value of 0.95 (Appendix A) Which is by far higher than the ROC of 0.905 before the integration. For evaluating the remaining years since the simulation based on the combined integrated model was found robust the results from combined integrated model were evaluated by taking 30% random sample points for 2011from Google earth. 30% sample was taken out of 917 predicted simulation points in ArcGIS. That is, 275 points out of the total prediction (Appendix B).

By a validation made on the new IS growths on Google earth, 222 points were correctly predicted on areas where it is possible see visually new IS expansions. Hence the evaluation on Google earth shows 80.72% of the random sample points were on new IS expansion areas, while 53, that is, 19.28% points were incorrectly predicted on areas where there is no IS expansion.

Figure 17: The ROC curve of the simulation based on the combined LRMs



4.6. Model interpretation

Based on global and local predictors three LR models were developed for the calibration of the CA based model. As mentioned on (Table 1&2), 17 predictors were applied after three out of the 20 were eliminated for having FIV more than 10 in all the three models. The three predictors which were similarly eliminated in all the three models were distance to food markets (x3), distance to planned residential (x10), and proportion of urban land in an area. Model-A, which is based on the IS expansion between 1982 and 1992, however was found to have the highest area under the ROC curve, that is 0.905 which is significant at 0.00 and of standard error 0.002. In addition to this during LR modelling after checking the significance of the individual predictors additional predictors were checked out.. From model-A predictors, environmental hazard and distance to CBD were eliminated from the model for having insignificant levels of 0.476, 0.486, and 0.071 respectively. This implies that the expansion of IS was not manifested near highly populated areas, environmental hazard areas and near the CBD of DAR at least the model development area. On the other hand, in model-B, which is based on 1982 to 2002 IS expansion, predictors such as, population density and distance to a river valley were additionally removed from the model development for having insignificant levels 0.611 and 0.174 respectively. The implication would be in addition to population density, river valleys have had no significant contribution in the development of IS expansion over the years between 1982 and 2002. Eventually the significance level of predictors in model-C(1992-2002) for the variables population density and environmental hazard were 0.453 and 0.144 respectively.

4.6.1. Transition rule definition and interpretation of key predictors of IS expansion

The transition potential of a cell being IS is decided by the nature of influence that each predictor, either global or local, has on the expansion of ISs. The nature of influence and the influence of the predictors on the IS expansion is interpreted from the estimated coefficients and odds ratio of the predictors. This way the probability/ transition potential of cells is defined and as a rule those vacant agricultural cell s with the highest probability will be converted to IS until the exogenous demand is satisfied.
From the global predictors of IS expansion such as, slope (X_2) population density (X_3), and environmental hazard (X_5), only slope (X_2) has a significant influence in Models of 1982-1992 and 1992-2002, while model 1982-2002 has slope (X_2) and environmental hazard (X_5) with significant influence on the IS expansion of the models. Slope (X_2) has odds ratio of 2.277, 0.644, and 0.000 for the models 1982-1992, 1982-2002, and 1992-2002 respectively which is positive (direct) relationship with IS expansion for all the three models. The positive relationship implies that the expansion of ISs has increases with increasing slop and hence IS expansion occupying high land areas. The interpretation of the odds ratio, for instance, of model 1982-1992 which is 2.277 implies that there would be 227.7% loss in the odds in the IS expansion for every 1% increase in slope.

Another category of global attraction predictors, proximity characteristics have inversely (negatively) defined relationship with IS expansion. In the models the most influential predictors (key drivers) of IS expansion are found in the category of proximity characteristics predictors. The estimated coefficient and the value of the odds ratio for key drivers, distance to minor roads (X_{11}) and distance to ISs (X_{14}) of 1982-1992 model is 0.000 (significant at $\propto$=0.000) implying a unit distance increase will result in 0% no decrease in the odds of having IS expansion. For models of 1982-2002, and 1992-2002 the same predictors, distance to minor roads (X_{11}) and distance to ISs (X_{14}) are the influential predictor for the IS expansion. The predictors as many of the proximity drivers do, they have negative relationship with the expansion of ISs. Both, the 1982-2002 and 1992-2002 models have the same odds ratio of 0.002 for the two predictors implying a unit distance increase in both distance to minor roads (X_{11}) and distance to existing ISs (X_{14}) will result in 0.2% decrease in odds of having IS expansion. The combination of land use types industrial, commercial, institutional and recreational land uses which are included under other urban (X_8) land uses have odds ratio of 0.03, 0.012, and 0.004 for the models of 1982-1992, 1982-2002, and 1992-2002 respectively with coefficients of -3.519, -4.409, and -5.566 for the respective years. The strong inverse relationship is seen from the negative sign and high value coefficients. From the odds ratio it is observed that the probability of an area with one unit distance less from other urban land uses would only be increased by 3%, 1.2%, and 0.4% for the respective models. The interpretation of other predictors can be defined in a similar fashion except that for this section we are interested in showing how influential predictors are interpreted for their effect in the simulation IS expansion models of DAR. The odds ratio and the coefficients of each predictor are summarised on (Tables 10, 12, 16) for each predictor with the corresponding significant levels.

The influence of local (neighbourhood) predictors have been the strongest influential drivers of IS expansion in all the three models, next to the proximity global predictors. However the nature of the influence from the two categories, proximity and neighbourhood, have opposite influences on the expansion of ISs. While proximity global predictors have negative strong influence on the expansion of ISs the local neighbourhood predictors influence positively. Two neighbourhood predictors, proportion of ISs (X_4) in the area and proportion of undeveloped (X_1) land in an area have significant influence in the development of ISs as can be seen on (Tables 10, 12, 16) of the three models.
The odds ratio for the proportion of undeveloped land in an area (X_1) for models A, B and C were 33.175, 82.478, and 176.777 respectively. The interpretation of the odds ratio figures implies that The addition of

one undeveloped land use cell would result in a probability to be IS 33.175 times for Model A; 82.478 times for model B and 176.78 for model C as compared to one cell less IS neighbourhood.

4.6.2. LRMs predictions for stochastic calibration of CA modelling

The LR models built are used to calculate the probabilities of vacant agricultural land uses, on which it is assumed IS expansion takes place, which is a calibration of the CA model so that IS expansion will be allocated based on the probabilities of the vacant cells being IS expansion land uses and the exogenous demand (Table 18) for IS for the given time, for ten years' time and on a yearly basis.. The simulations were made in at ten years intervals as the both dependent and independent variables from data available are of ten years interval. On the other hand, as the key drivers which affect the expansion of IS are some of them naturally constant, like distance to major and minor rivers, distance to river valley, environmental hazard, distance to oceans, while other predictors such as, distance to minor and major roads do not have significant changes for many years in DAR (Abebe, 2011a; Hill & Lindner, 2010) and the main driver distance to existing IS can update itself after every simulation so that the application of LR modelling on a yearly basis was found to be logical.

For the calibration of the stochastic CA of the simulation for every ten year, the IS expansion for 2002, 2012 and 2022 were predicted as shown on (Figure 19, 20, 21). The model from the 1982-1992 was found to the best model to do the prediction as the area under ROC curve of this model is the highest with a value of 0.905 as compared to 0.853 and 0.857 areas under ROC curve, which a better way of measuring the predicting capacity of a land use model. One of the challenges for the validation of the simulation results was the availability of land use data. The only data available for the ROC evaluation of the simulations was the 2002 land use data. The simulation for the year 2011 was evaluated by taking sample points from Google earth map of 2011. (Figure 18) in the appendix shows the evaluation for the 2011 IS simulation, sample points. In ArcGIS environment a random sample size of 30% was retrieved out of 917 cells., that is, 275 cells. Based on the spatial characteristics of the new developments on Google earth, 2011, 222 cells were found to be relatively correctly predicted. That is 80.72%of the simulation in 2011 (Appendix B).

4.6.3. Simulation of IS dynamics

The simulation made for every year was calibrated by combining the three models so that the land use data would be updated during the modelling processes years with land use data. In addition to that the year by year IS expansion modelling is more representative of the self-organization nature of IS dynamics as the simulation of the yearly IS expansion becomes input to the next time step IS expansion. In this way through the modelling process the model updates it predictors for the every next iteration(Wu, 2002). (Figure 19, 20, 21) shows the IS expansion simulation results from 1992 to 2012. The combined model for the yearly IS expansion simulation is expected to be updated with a yearly data of the predictors and is ideally assumed to be validated after every iteration. The IS predictions of the year t are used to update at least three predictors at every t+1 iteration, such as, distance to ISs, proportion of IS in an area, and proportion of undeveloped land in an area. This way every (t+1) year iteration was updated by the year t output. However, predictors like distance to other urban are not updated for every year except for the years with additional land use data, such as, in 1998 and 2002.

Figure 18:IS expansion simulation by model-A.

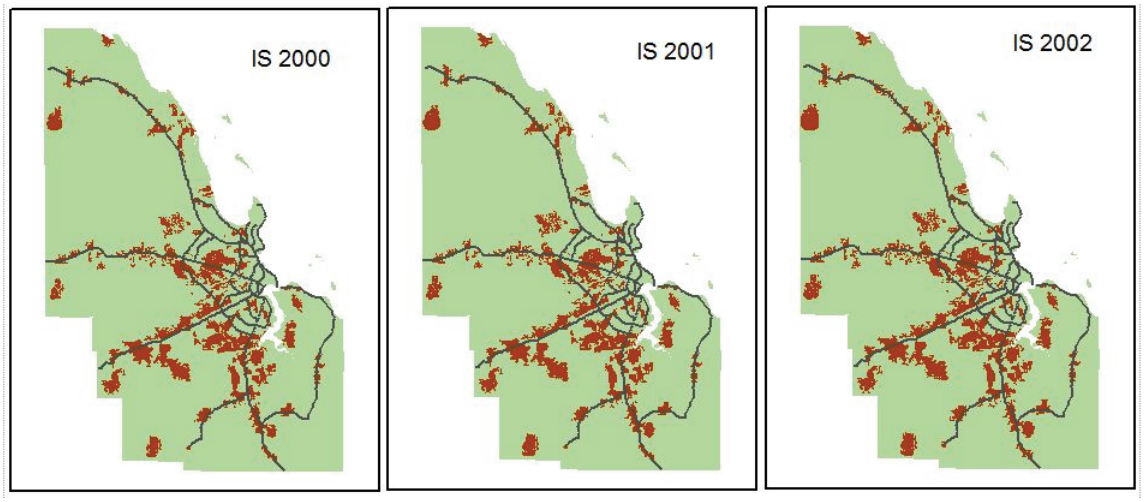


Figure 19: Comparison of the IS dynamics of simulation with the observed IS expansion 2002

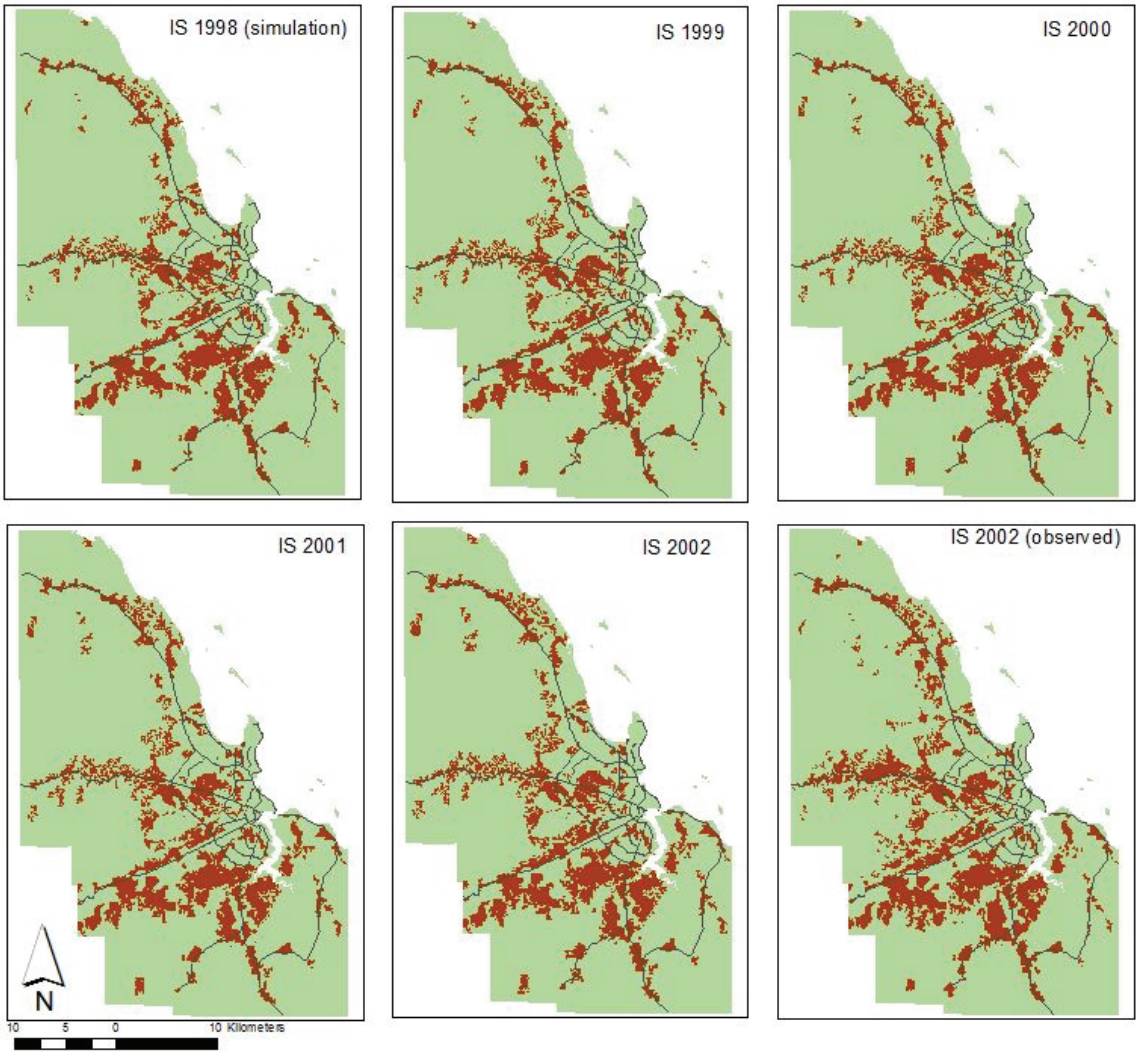


Figure 20: simulation of IS expansion 2003-2022 based on combined LR integrated CA based model

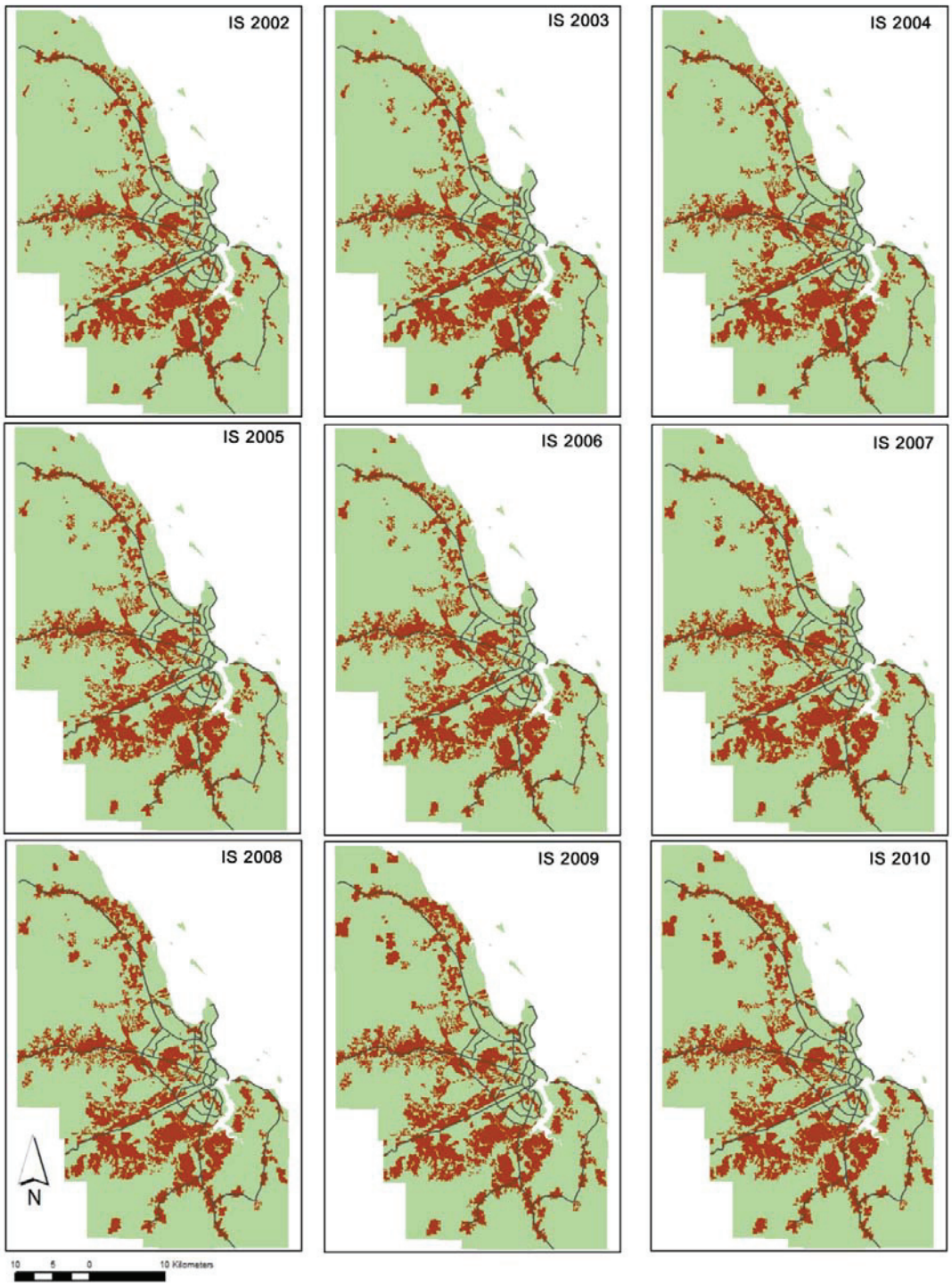


Figure 21: Simulation for 2011, 2012, and 2022 based on combined LR integrated CA based model

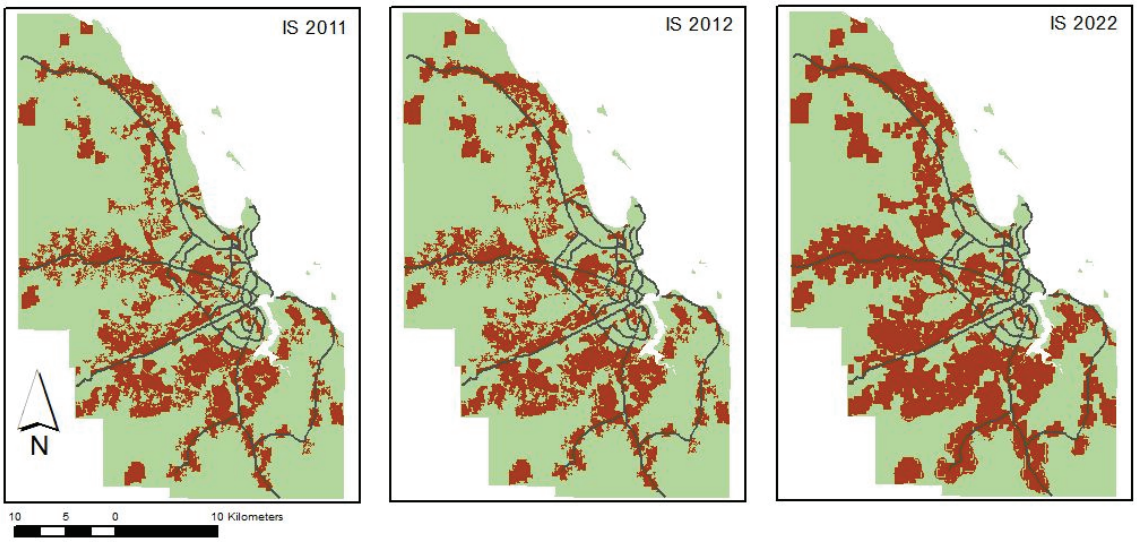


Figure 22: The IS expansion simulation for 2002, 2012, and 2022 versus Key factors

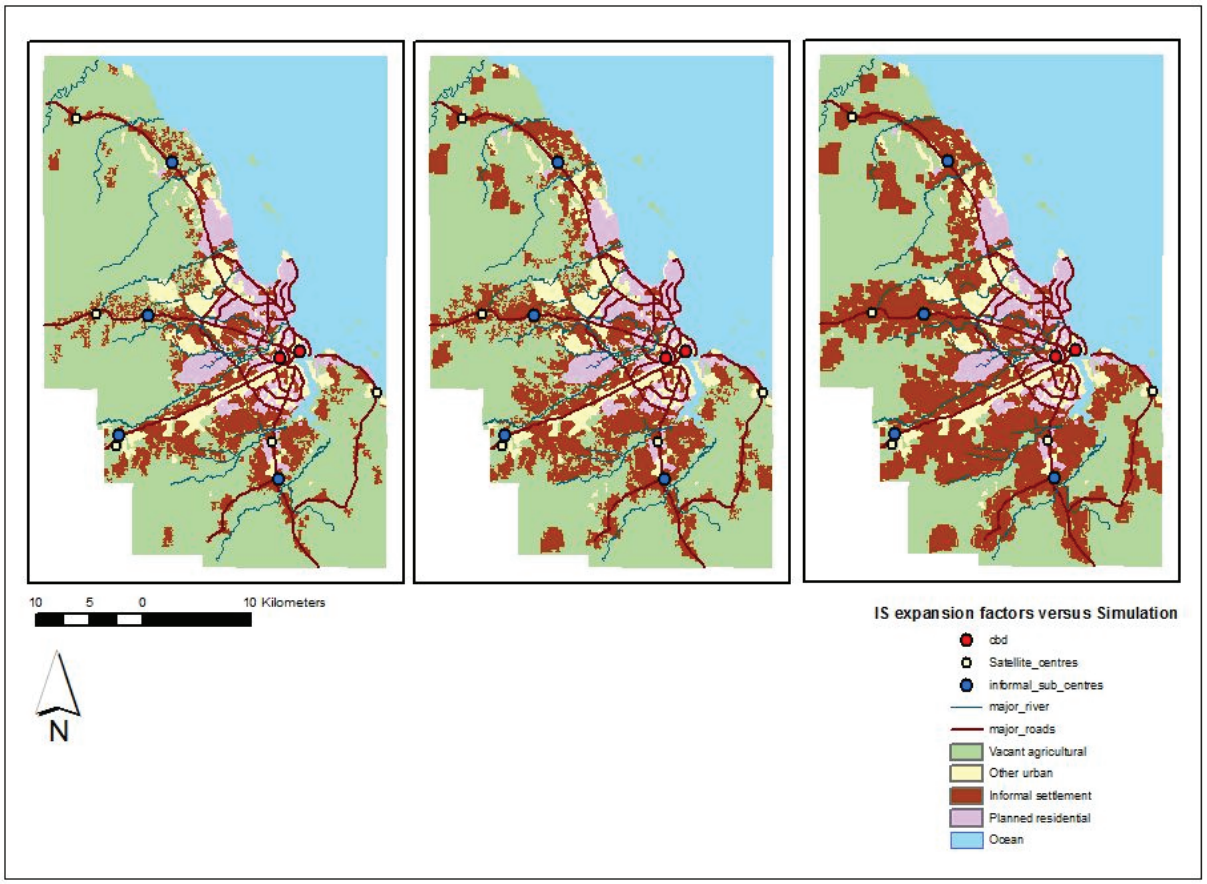


Figure 23: Probability map of the 1982-1992, and LR prediction (allocation), (Model-A), left and Right respectively.

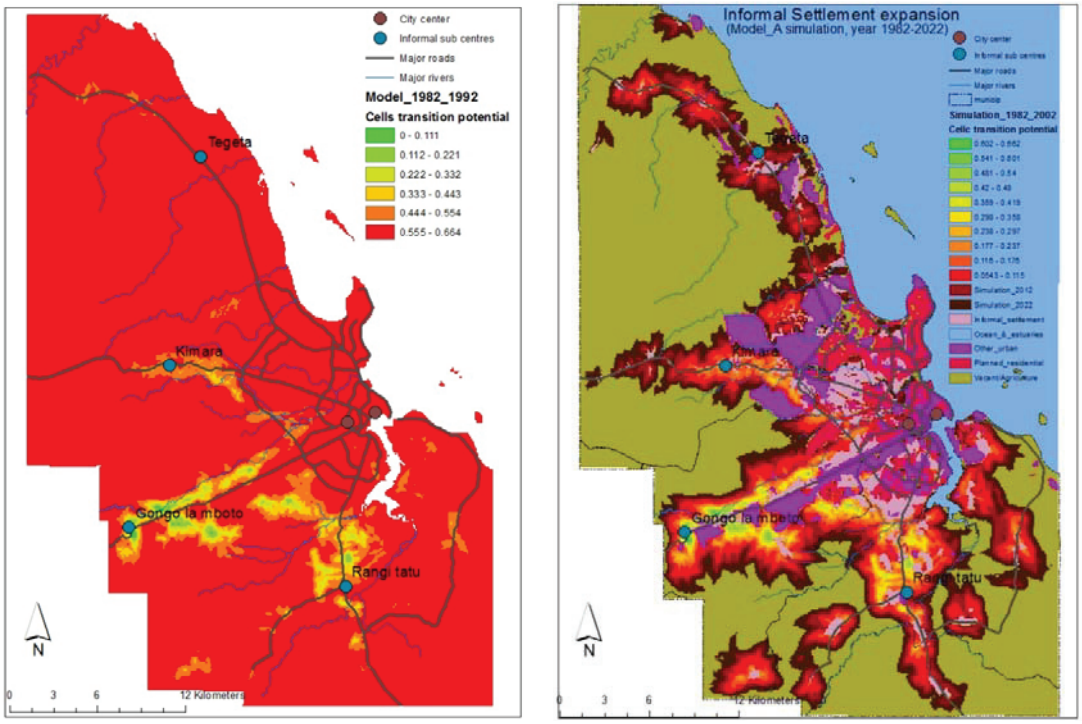


Figure 24: Probability map and LR prediction (allocation) of IS, (Model-B), left and right respectively.

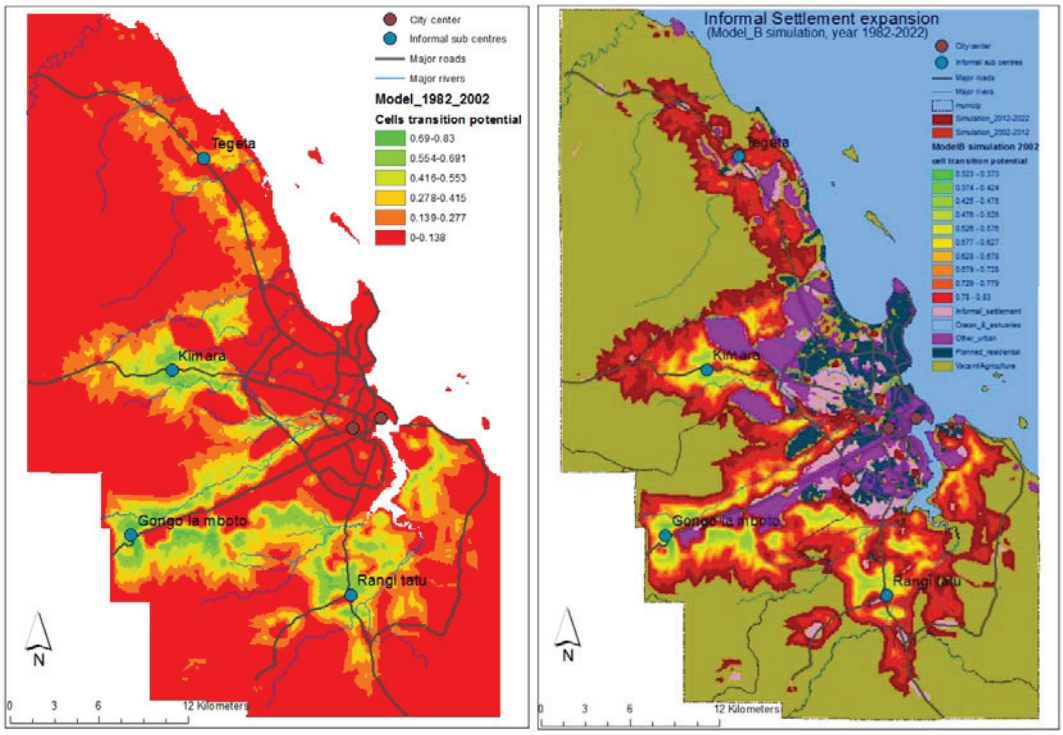
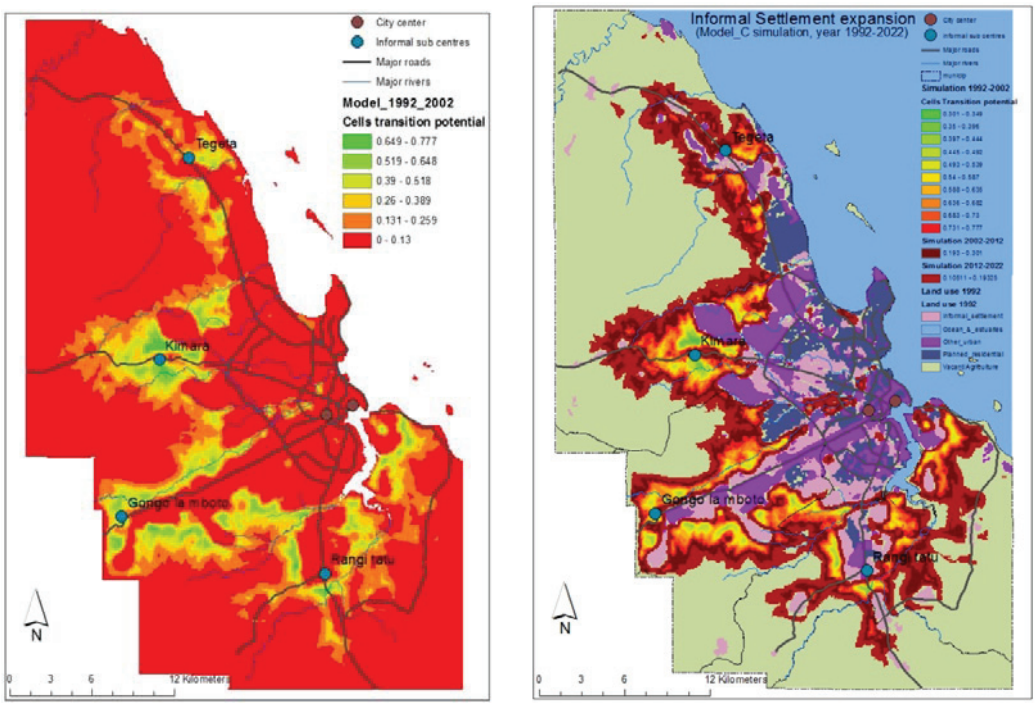


Figure 25: Probability map and LR prediction (model-C), left & right respectively



5. DISCUSSIONS

The focus of this particular chapter would be discussing the main findings of the research results, which were based on the methodological approaches. The chapter tries to explain the importance of the research by relating to the general realm of urban planning. The main discussion points of the chapter would be land use model integration and the logical application of LR modelling for the rule definition and calibration of CA based simulation model of IS expansion. Discussions based on the model evaluation results and comparison of the model simulation with observed IS expansion in DAR and similar previous works is also the concern of the chapter.

5.1. LR integrated CA modelling of IS expansion DAR

The integration of LR approach to support CA modelling has been the focus of this study. Numerous recent researches have been made on integration of different land use models with the idea of combining the individual benefits of the models for better understanding and interplay between land use activities(Poelmans & V.Rompaey, 2010; White & Engelen, 2000).Model integration benefits have also been measured in terms of model complexity for the performance of the model. Even though model complexity relies on the particular nature of the application and data quality the more the complex the models are the accurate the model would be. Hence model integration increases the accuracy of the model by increasing complexity in the process.

Therefore the conceptual grounds of the model integration in this study were based on those mentioned facts and the expert knowledge during the development of the research. The research tries to develop different techniques which have conceptually been verified so as to integrate empirical LR modelling approach to support dynamic CA modelling of IS expansion in DAR.

Theoretically, it was supported by various researches that LR modelling can be used to calibrate stochastic CA models (Hill & Lindner, 2010; Poelmans & V.Rompaey, 2010; Verburg, et al., 2004; Wu, 2002). In the calibration process, the probabilities of the cells were calculated for those probabilities from the global attraction factors (global probability) and probability from local neighbourhood interaction factors (local probability). However, the application of LR modelling so far has only been for calculating global probabilities of cells in definition. This is due to the fact that including local factors directly in the LRMs to calculate the local probabilities of the cells would be difficult because of the unavailability of data at a fine time scale. The data of land use change are mostly found mainly in ten years of time hence this is not in agreement with simulation time which is in a yearly basis(Wu, 2002).

Two fundamental strategies were chosen to tackle the problem. (1) To much the time scale of simulation and time scale of the land use conversion, the simulation of the models will be constrained by exogenous land demand. (2) The simulation was made in two time scale, one at every ten years of time and the second at a year by year basis.

The assessment for multicollinearity and minimizing the effect of spatial autocorrelation have been important aspects when modelling with the empirical LR method(Cheng, et al., 2003). For the LR modelling independent explanatory variables were compiled by Abebe (2010) in a similar research of IS development modelling in DAR. A list of 20 probable drivers was assessed for multicollinearity and after three independent variables were eliminated for having VIF>10. Those variables with high VIF and eliminated from further analysis in the model are expected to be replaced with a variable in the analysis(Field, 2009). Therefore, 17 of them were applied during the first stage of the LR model development. However, during the first stage of the model development additional drivers were checked out for their influence being not significance for the models. The predictor population density has been

removed during the first LRM development for having 0.476, 0.611, and 0.45 significance values, which is by far higher than the standard 0.05 value from models A, B, and C respectively. Additionally, predictors, distance to CBD, distance to river valley, and environmental hazard were found insignificant for the models A, B, and C respectively. Hence, all the three models were eventually developed by using the dependent variable IS expansion and 15 independent variables each. Spatial autocorrelation has been minimised by taking a random sample of 90,509 out of the 97,998 call total extent of the study area making the number of dependent variables with equal and 0 and 1 observations. The sample size has been made large to see the spatial distribution of IS expansion clearly at every simulation.

Model performance has been assessed for the three models developed so far, through the application of ROC statistics and visual evaluation. According to the comparison of the area under the curve of ROC statistics results Model-A of the IS expansion between 1982 and 1992 was found to be the best with the area under the curve 0.905 showing that 90.5% of the model output would be correctly predicted. The remaining two models, Models B and C, which are based on 1982-2002 and 1992-2002 IS expansion have had 0.853 and 0.857 respectively of ROC statistics for the area under the curve as seen on (Table 21) and (Figures 15, 16, and 15) All the three models have got a satisfactory ROC evaluation result. Hence for the calibration process the three models altogether were combined to make a single mode so as to increase the complexity of the model in a way to have a more accurate simulations of the IS expansion. The way of the combination was made by updating the input factor maps and the dependent variables during the iteration after the simulation for the years where there is input land use data. This has been visually validated by comparing those simulations without the combination of the three LRMs and those simulations from the combined LRMs. (Figure 19, 20, 21) of the simulation result map shows the visual evaluation results of the combined integrated CA model and the model-A integrated CA model.

Model integration concepts, especially between empirical LR and dynamics self-organising CA has been the focus of this study for the fact that LR integrated CA model can minimise the limitations of the independent approach and shall bring in a more structured, more interpretable, dynamic and more accurate prediction results of the IS expansion in DAR. LRMs are used to determine the influence of independent variables (drivers) and provide a degree of confidence about their contribution to the change(Huang, et al., 2009). In this case, a logistic regression model generates IS expansion probability maps in a way it tells the probability of a cell being informal by associating urban growth with explanatory variables of site specific, proximity characteristic, and neighbourhood interaction of both global and local in nature. However, due to the reason that LRMs lack to incorporate temporal dynamics, and tell the spatial patterns of land use changes but not when that change will happen, in addition to its incapability to develop different urban development scenarios(Hu & Lo, 2007), CA modelling techniques were coupled to the empirical LR modelling technique. This is also an advantage to the dynamic CA modelling that the LR would be used to ease the CA modelling calibration process in a stochastic way and would help to incorporate explanatory variables in the model and enrich the interpretability of the model(Wu, 2002).

5.2. Drivers of IS expansion and CA rule definition.

The probability of the cells' in definition, in the LR integrated CA modelling, are calculated by transition rules(Poelmans & V.Rompaey, 2010). The probability of the cell in definition to be informal at each (1 year) iteration is a base for the dynamic CA model. The whole bunch of driver of IS expansion compiled from literature review and expert opinions are the once which affect the rule definition for calculating the transition potential (probabilities) either in a negative or positive way.

In the three models, the influence of some factors varies and the others remain the same. Proximity factors, such as, distance to other urban (X_8), distance to minor rivers (X_{10}), distance to minor roads (X_{11}), and distance to existing ISs (X_{14}) have negative relationship with the expansion of IS ,and hence, affect the predictions of the with inverse relationship, that is cells, for instance, near to other urban or existing ISs will be converted to ISs in the next iteration. The coefficients of the factors (-3.519,-4.409, -5.566) for distance to other urban (X_8), (-2.856, -1.328, -1.615) for distance to minor river (X_{10}), (-12.049, -6.763, -6.323) for distance to minor road (X_{11}), and (-11.861, -6.106, -6.020) for distance to existing ISs(X_{14}) for the models A, B, and C respectively show the strength of the influence of the factors in negative way. Other predictors such as, proportion of undeveloped land in a neighbourhood (X_1), proportion of IS in the neighbourhood (X_4), distance to ocean (X_9), and distance to hills (X_{16}), have all positive relationship with expansion ISs with corresponding coefficients of (3.15, 4.43, 5.175), (1.472, 1.034, 0.602), (3.755, 2.14, 1.012), and (0.669, 0.94, 1.22) for the models A, B, and C respectively. The remaining eight significant predictors proximity factors such as distance to satellite centers (X_6), distance to river valley (X_7), distance to major roads (X_{13}) have a negative relationship with expansion of ISs in model A, while out these only distance to major roads (X_{13}) in model-B and distance to satellite centers (X_6), and distance to river valley (X_7) have positive relation with IS expansion. The remaining five predictors, such as, slope (X_2), environmental hazard (X_5), distance to major rivers (X_{12}), distance to informal sub-centers (X_{15}), and distance to CBD (X_{17}) have positive relation with expansion of ISs for model-A while slope (X_2), distance to informal sub-centers (X_{15}), and distance to CBD (X_{17}) of Model-B and all, slope (X_2), environmental hazard (X_5), distance to major rivers (X_{12}), distance to informal sub-centers (X_{15}), and distance to CBD (X_{17}) of model-C have negative relationship with expansion of ISs.

The positive relationship of the factors with the expansion of ISs refers to the relation that when the values of the factor for a particular cell increases the probability of the cell to be IS, in the next time step (t+1), also increases. The value of the coefficients, on the other hand, refers to the strength of the influence of the factors to the cell's being IS. The higher the value of the coefficient the more the influence of the factor in the rule definition so that cell shall be influenced to be IS. The negative coefficients, however, affect the expansion of ISs with their decrease in value to the cell in definition. That is, when the value of the coefficients is higher and negative, for instance, the factor distance to existing ISs (X_{14}) the smaller value of the factor to the cell in definition yields higher probabilities for the cell.
Eventually, in order to calculate the probability of the cells the cumulative influence of the factors is analysed by LR method. The constant (model intercept), the coefficients, and the values of the drivers from the developed models A, B, and C will be applied to calculate the probabilities of the cells for the first iteration (t+1) . The rule definition selects the cells with the highest probability (transition potential) based on the exogenous IS expansion demand calculated (Hill & Lindner, 2010; Poelmans & V.Rompaey, 2010).
`

5.3. IS Expansion Simulation and areas of expansion

The IS expansion simulation by the LR integrated CA model has been done in two ways, strategically to see the difference of the prediction by simple LR and though self-organization step of CA. The first way of simulation has been done by calculating the IS expansion demand the ten years' time and constraining the probabilities of the cells which is calculated by the LRM. In the process those cells with the highest probabilities have been selected until the demanded number of cells was selected. The cell size has been 100x100 meters. The expansion simulation was done on vacant agricultural as most IS expansion has been on vacant agricultural land uses(Hill & Lindner, 2010). Constraint areas and other land uses were excluded from the simulation by selecting only vacant agricultural areas from the attribute table containing the

probability values of all other land uses. The first simulation made at every ten years was allocation of the vacant agricultural cells based on their probability value and the demand of the first ten years. This allocation can give a general overview on the spatial pattern of IS expansion except that it lacks to show the yearly IS expansion dynamics, and hides the effect of self-organising nature of land use change (Dubovyk, et al., 2011; Wu, 2002). This limitation of LRM makes its capacity for land use prediction quite limited(Hu & Lo, 2007). Hence, in order to make a more representative prediction of the IS expansion dynamics the second way of simulation which is based on a year by year and self-updating process of land use change modelling was done. For the year by year simulation, the three LRMs were combined after their evaluation of their ROC statistics. The ROC statistics for the 1982-1992 model (model-A) was found to be the highest with 0.905 value of the area under the curve, while the 1982-2002 model (model-B) and 1992-2002 model (model-C) have value of the area under the ROC curve 0.852, and 0.857 respectively. This implies the high predicting capacity of the models. Directly from the values it model-A has the highest prediction capacity. However, since it is based on the only the IS expansion from 1982-1992 its accuracy would be less in predicting further IS expansion dynamics. The other models, on the other hand, have included dependent variable, IS expansion until 2002, and predictors from 1992 land uses. This would increase the prediction power of the two models in the further years. The combined model, therefore, will have the power to correctly predict the IS expansion dynamics of both the near future and the further future.

5.4. Model integration

Model integration in this study has been made within different LRMs and between empirical LRM and dynamic CA model. In the research three LRMs based on the IS expansion between three consecutive years have developed and assessed for their fitness. That brought the idea of model integration between the models of the LR approach. On the other hand, the combination of the LRM and the dynamic CA modelling techniques was made based on the notion of explaining IS expansion by relating the effects of explanatory variables and the dynamics of IS expansion in DAR.

5.4.1.1. Combination of the LRMs

The year by year simulation of IS expansion in DAR has been strategically made by combining the three LRMs, so that better accurate predictions and model interpretations can be made. This has been a way of adding complexity to the model as adding complexity to a model can be a means to increase accurate predictions(Poelmans & V.Rompaey, 2010).

The LR modelling combination (integration) between the three models was designed in way that the recent predictions were made by model 1982-1992 (model-A) that is until 2002 while, after the evaluation, the data was updated to incorporate the IS expansion and input factor maps from the Models-B and C for further predictions. Accordingly, the first simulation (for 1993) was predicted by taking the probability map from model-A, which is shown on (Figure 25), and constraining by the IS expansion demand of 245 (ha) of land that is 245 cells as the cell size is 100X100 meters. The estimation of IS expansion demand has been discussed on (section 4.3) and the demand until 2022 was shown on (Table 18). The higher the probability values the first it is allocated until the demanded number of cells are satisfied (Poelmans & V.Rompaey, 2010; Wu, 2002). After the first iteration (of the 1993) the simulation for the year 1994 was done by self-updating the predictors of model-A with the simulation of 1993. The simulation provides with the IS expansion of the year. Hence, the three factors: distance to existing ISs (X_{14}), proportion of IS in a neighbourhood (X4), and the proportion of undeveloped land in a neighbourhood(X_1) were updated. Other predictors, in the model except for the factor, other urban (X_8), remain constant for two reasons. Firstly, many of the factors were constant as they were derived from natural features such as slop, environmental hazard, river valley, and distance to ocean, distance to minor river, distance to Major River,

and distance to hills. Secondly, the other remaining factors which are related to infrastructure , such as, distance to satellite centers (X_6), distance to minor roads (X_{11}), distance to major roads(X_{13}), distance to informal sub-centers (X_{15}), and distance to CBD (X_{17}) were known for not being frequently updated during the time of simulation(Hill & Lindner, 2010). The factor, distance to other urban (X_8) for which we do not have data available to update except for 1998 and 2002 remains constant for all the simulations. Therefore the LRM output, the probability of the cells for vacant agricultural cells were calculated based on a self-updating manure.

5.4.1.2. Integration of LRM and with the dynamic CA

The integration of LRM and the self-organising CA was basically in the rule definition as discussed on section (5.2.) and the calibration of the modelling (estimating cell probabilities) for the simulation to come. The basic elements for a land use model to CA, which are discussed on (section 2.2.2), such as, the cell, the state of the cell, the neighbourhood, transition rules definition and stochastic calibration processes were clearly defined during model integration. Hence the whole integration of the LR approach with the elements of CA has been the point in discussion so far in the study.

5.5. Comparing Simulation results with prediously done CA based and LRM predictions

Predictions of the IS expansion in DAR has been recently done by Hill and Linder (2010) and Abebe

(2011) with CA based modelling and LR modelling respectively. The CA based model of Hill and Lindner (2010) has been briefly discussed (section 2.12.1), while the data compiled by Abebe (2010) has been used to develop the LRMs of the study. Similar procedures, with Abebe (2010) were followed in order to develop the LRMs and calculate the probabilities of cells. Subjective comparison would be made based on the predictions of the three models, the LR integrated CA based model simulation, the LRM of Abebe (2010) and the CA based simulation (Hill and Lindner (2010)) by looking at the major model seminaries and differences. The fact that all the three models share a common grounds such as, the data used, the modelling technique applied, that is LR, and the predictions being to the same years, 2012 and 2022 makes the subjective comparison more logical. The comparison can also be interpreted based on the key drivers and the pattern of the prediction made, even though the basic differences on the assumptions made in order to make various inputs and the time of simulation are also expected make certain variation.

The visual comparison made between the three models based on the predictions of 2012 and 2022 show that the 2012 simulation of the author's LR integrated CA based model shows relatively dispersed pattern of IS compared to the prediction of Abebe (2011) and CA based simulation. This can be related to the year by year simulation done and the self-organization (self-updating) nature of simulation as the self-organisation nature enables the model to update itself after each iteration and accentuates a bottom-up nature of CA that is the interaction of the cell with its neighbourhood. Allocations based on land use demand were made based on the probabilities of the cells but since it was based on a ten years data it tries to eliminate the bottom up nature of IS expansion (Figures 24, 25, 26).
The amount of prediction however looks similar for the 2012 model among all the three models while LR prediction of Abebe (2010) for the year 2022 shows a relatively lesser amount of prediction. This could be a difference on the assumptions in calculation in land use demand. The attraction of the drivers such as , the once displayed on the maps, distance to rivers, distance to roads, distance to informal sub centers , undeveloped land CBD looks similar in all the three models with exception of certain locations, such as the North Western vacant agricultural areas, in the authors model showing high attraction of IS expansion.

This would be the influence that after every iteration the updating of certain drivers such as proportion of undeveloped land in a neighbourhood (section 5.4.1.1).

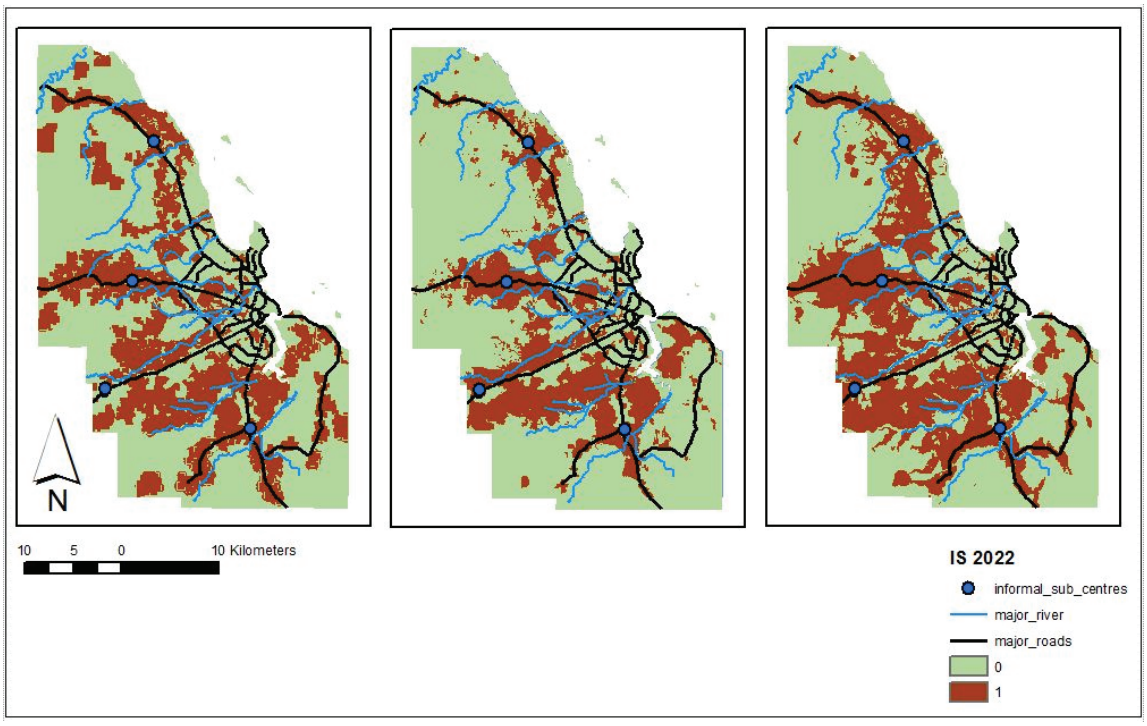


Figure 26: Comparison of IS expansion in DAR at 2022 Author model (left), LRM (Abebe (2011)) (middle), and CA model (Hill and Lindner (2010)) (right)

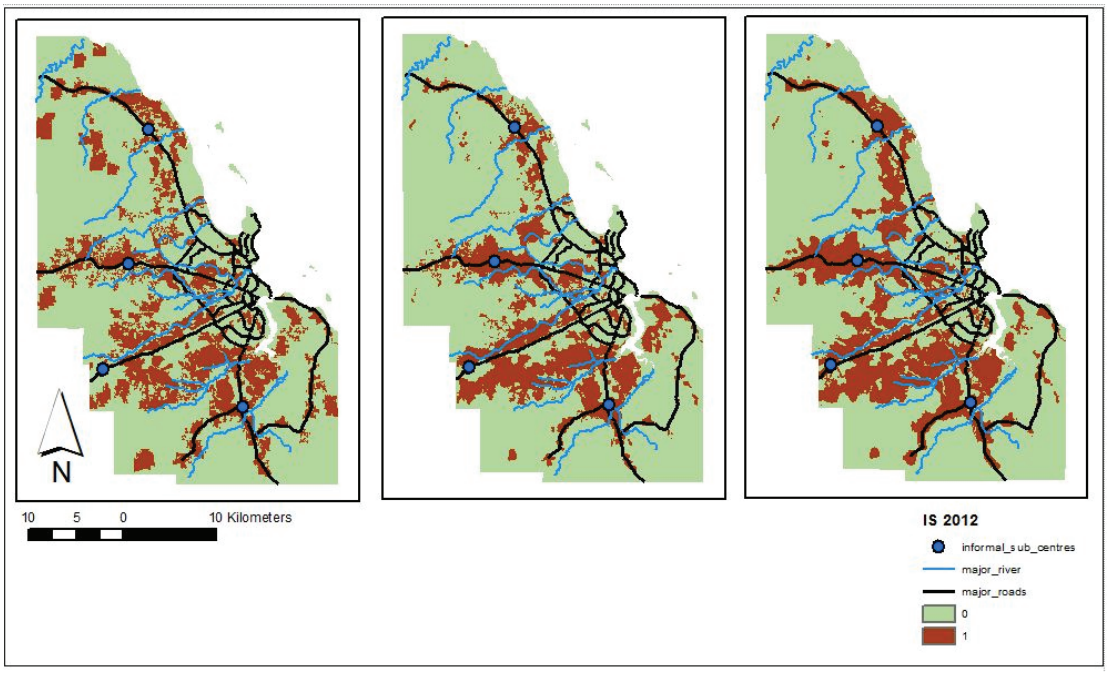


Figure 27: Comparison of IS expansion in DAR at 2012 Author model (left), LRM (Abebe (2011)) (middle), and CA model (Hill and Lindner (2010)) (right)

6. CONCLUSION AND RECOMMENDATION

This research has been done based on the integration of two land use modelling approaches, LR and CA modelling, for IS expansion in DAR. The research objectives and questions which have been formulated and designed under the umbrella of the basic objective were attempted to be addressed in the research. Hence, the conclusion chapter is articulated to show how the research three sub-objectives: (1) to develop a conceptual method to integrate LR and CA modelling approaches for analysing IS expansion in DAR; (2) to revise the LR model already applied in DAR to support CA modelling of IS expansion; (3) To simulate IS expansion in DAR using LR integrated CA model; were addressed in the study.

6.1. To develop a conceptual method to integrate LR and CA modelling approaches for analysing IS expansion in DAR

The scientific justifications, made on the logical conceptual model integration between LR and the dynamic CA, have shown the theoretical possibility of the integration. It appeared that there are two major outputs of LR analysis to be applied for CA modelling, those are key derivers of the IS expansion and probability values of the cells in definition (Hill & Lindner, 2010; Poelmans & V.Rompaey, 2010; Wu, 2002). However, the challenge which remained as a research question is the application of LR modelling for calculation local probabilities of cells. Some of the constraints for LR to be applied in CA modelling were the unavailability of a year by year data to make the simulations, and the fact that time of simulation and time of land use change may not coincide. This has been strategically been addressed in the study by (1) self-updating the input factors after every iteration (2) applying exogenous IS expansion demand in order to constrain the quantity of IS expansion land conversion (Poelmans & V.Rompaey, 2010; Wu, 2002).The LR analysis provides with parameter values and estimated coefficients of the independent variables of the IS expansion model. The value of coefficients shows the strength of the relationship between the drivers and the expansion of the ISs while their sign (positive or negative) shows the type of the influence. This defines the transition rule and structures the calibration of the CA based model. The probability(transition potential) of the cells to be informal was also calculated from the LRM (Poelmans & V.Rompaey, 2010).

6.2. To revise the LR model already applied in DAR to support CA modelling of IS expansion

The LRMs used to calibrate (calculate the probabilities and estimate the parameter values of the factors) the CA model were built on the 1982-1992, 1982-2002, and 1992-2002 IS expansion of DAR as dependent variable; and 20 initial all inclusive independent variables from 1982 and 1992 datasets. The initial independent variables were checked for multicollinearity and significance and 15 of them were used to build the models. Variables, distance to food markets, distance to planned residential , and proportion of urban land in a neighbourhood were eliminated in the multicollinearity diagnostics for having VIF of greater than 10.The driving factors of IS expansion were analysed in LR model and based on this, drivers such as distance to minor roads (X_{11}), distance to existing ISs (X_{14}), distance to other urban (X_8) and distance to minor rivers (X_{10}) with all negative coefficients with distance to minor roads (X_{11}), and distance to existing ISs (X_{14}) being the most influential drivers in all the three LRMs.

6.3. To simulate IS expansion in DAR using LR integrated CA based model; were addressed in the study.

More than 20 iterations were made to simulate the IS dynamics of until 2022. The year by year simulation has updated the independent variables by the self-organisation nature of CA. Therefore, three factors: distance to existing ISs (X_{14}), proportion of IS in a neighbourhood (X_4), and the proportion of undeveloped land in a neighbourhood(X_1) were self-updated after every iteration. Other predictors, in the model except for the factor, other urban (X_8), remain constant as many of the factors were constant for being derived from natural features such as slop, environmental hazard, distance to river valley, distance to ocean, distance to minor river, distance to major river, and distance to hills, while the other remaining factors, such as, distance to satellite centers (X_6), distance to minor roads (X_{11}), distance to major roads(X_{13}), distance to informal sub-centers (X_{15}), and distance to CBD (X_{17}) were known for not being frequently updated during the time of simulation(Hill & Lindner, 2010). The only factor, for which was not updated at each iteration has distance to other urban (X_8). However, effects due to the data unavailability were tried to minimize by updating the simulations at 1998 and 2002. Based on the integrated LRMs were made by first calculating the probability of the cells and constraining the quantity with the yearly exogenous demand calculated based on the population projections made in DAR and the proportion of IS residents

The ROC evaluation of, has given 0.905, 0.853, and 0.857 values for the area under the ROC curve of the three models, 1982-1992, 1982-2002, and 1992-2002 respectively. As can be seen directly from the ROC evaluation, model-A (1982-1992) has the highest prediction capacity. However, since this model is based on the only the IS expansion from 1982 to 1992 its accuracy would be less in predicting further IS expansion dynamics. As the remaining two models, on the other hand, have included dependent variable, IS expansion until 2002, and independent variables from 1992 land uses, combination among those LRMs would increase the prediction power of the two models in the further years. The ROC statistics made on the simulation of 2002 IS expansion has also proved this hypothesis. The combined model, therefore, has had a relatively better predicting the IS expansion dynamics of both the near future and the further future.

6.4. Further research direction

- To repeated LR integrated CA based modelling by incorporating stochastic perturbation by using GLMMs instead of using simple LRMs.
- To use GLMMs models to deal with spatial autocorrelation effects.
- To look for other ways of land use model integration for better model performance and increased interpretability
- The repeated the LR integrated model and compare the simulation by modelling IS expansion with simple bottom-up CA modelling of IS expansion and compare the and evaluate objectively as well as subjectively the results of the models.
- To use the potentials of web data such as Google earth for validation of perditions in data poor environments.
- To test the structure of the integrated model in other area of similar IS expansion situation.

7. APPENDEX

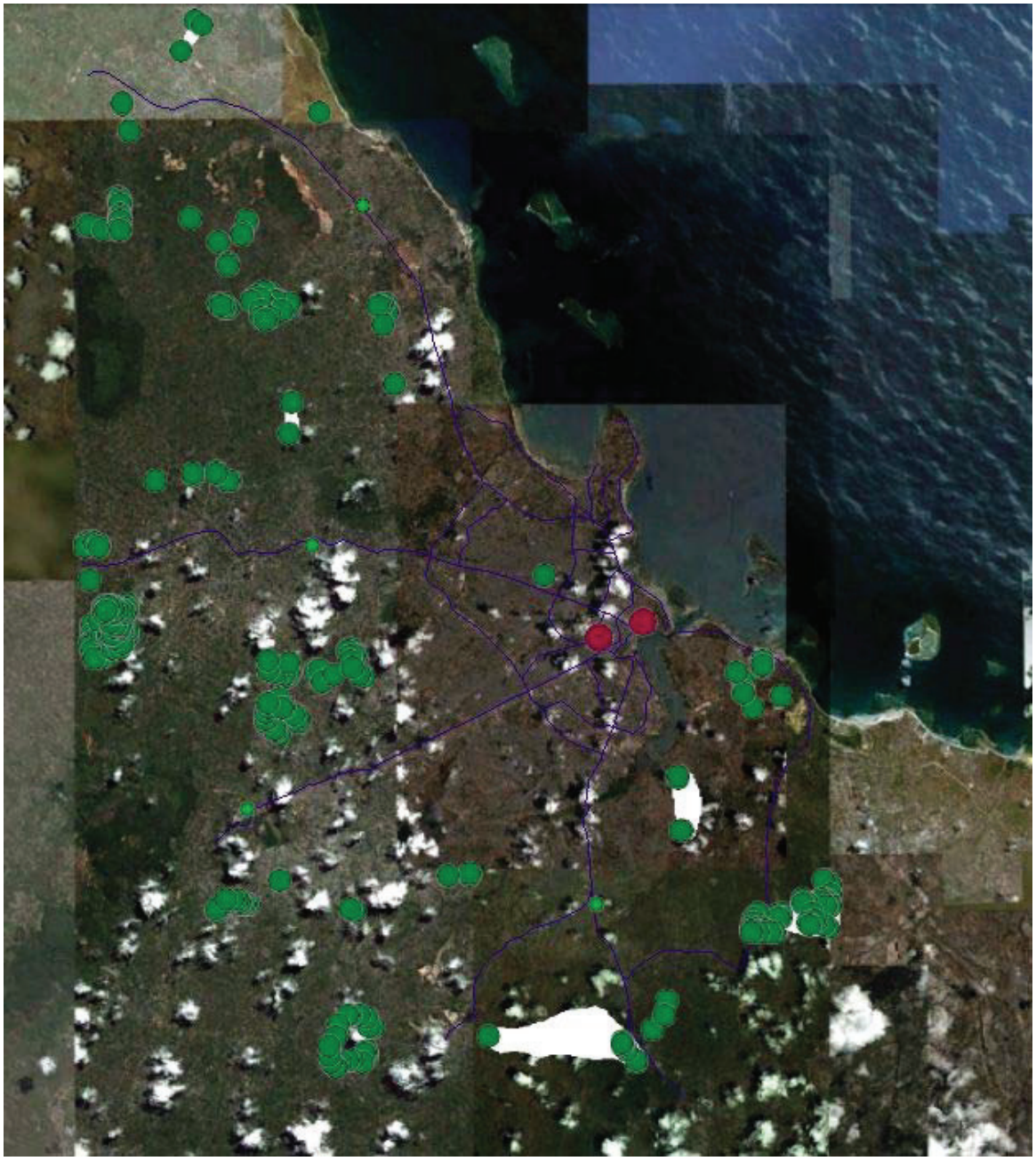
Appendix A: Area under ROC curve of simulation 2002.

Table 23: the ROC statistics of the combined LRMs simulation for 2002

Area Under the Curve
Test Result Variable(s):pre_1
Area
.950

Appendix B: Sample points for evaluation on Google Earth.

Figure 28: Sample points exported to Google earth to visually validate correctly predicted.



LIST OF REFERENCES

Abbott, J. (2001). Use of spatial data to support the integration of informal settlements into the formal city. *International Journal of Applied Earth Observation and Geoinformation, 3*(3), 267-277.

Abbott, J. (2002). An analysis of informal settlement upgrading and critique of existing methodological approaches. *Habitat International, 26*(3), 303-315.

Abebe, F. K. (2011). *Modelling informal settlement growth in Dar es Salaam, Tanzania.* University of Twente Faculty of Geo-Information and Earth Observation ITC, Enschede.

Agresti, A. (2003). Frontmatter *Categorical Data Analysis* (pp. i-xv): John Wiley & Sons, Inc.

Alan Agresti. (2002). Categorical Data Analysis. Second Edition. University of Florida.

Arthur, G. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics, 37*(4), 491-496.

Cheng, J., Masser, I., & Ottens, H. F. L. (2003). *Modelling spatial and temporal urban growth.* Utrecht University ; ITC, Utrecht ; Enschede.

Clarke, K. C., & Gaydos, L. J. (1998). Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. [Article; Proceedings Paper]. *International Journal of Geographical Information Science, 12*(7), 699-714.

Couclelis, H. (2005). ?Where has the future gone?" Rethinking the role of integrated land-use models in spatial planning. *Environment and Planning A, 37*(8), 1353-1371.

Dubovyk, O., Sliuzas, R. V., & Flacke, J. (2011). Spatio - temporal modelling of informal settlements development in Sancaktepe district, Istanbul, Turkey. *JournalISPRS journal of photogrammetry and remote sensing, 66*(2), 235-246.

Field, A. P. (2009). *Discovering Statistics using SPSS* (Third ed.): SAGE Publications Ltd.

.

Guhathakurta, S. (2002). Urban Modeling as storytelling: using simulation models as a narrative.

HABITAT, U. (2010). Informal Settlements and Finance in Dar es Salaam, Tanzania. Avilable online: http://www.unhabitat.org/pmss/getElectronicVersion.aspx?nr=2935&alt=1

Hadwig van Delden, P. L., Guy Engelen,. (2005). Integration of multi-scale dynamic spatial models of socio-economic and physical processes for river basin management.

Hill, A., & Lindner, C. (2010). Modelling informal urban growth under rapid urbanisation, A CA-based land-use simulation model for the city of Dar es Salaam, Tanzania. Avilable online http://eldorado.uni-dortmund.de:8080/bitstream/2003/27283/1/Dissertationsschrift_Hill_Lindner_Juni_2010.pdf

Hu, Z., & Lo, C. P. (2007). Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems, 31*(6), 667-688.

Huang, B., Zhang, L., & Wu, B. (2009). Spatiotemporal analysis of rural-urban land conversion. *International Journal of Geographical Information Science, 23*(3), 379-398.

Irwin, E. G., & Geoghegan, J. (2001). Theory, data, methods: developing spatially explicit economic models of land use change. *Agriculture, Ecosystems & Environment, 85*(1-3), 7-24.

Kironde, L. J. M. (2006). The regulatory framework, unplanned development and urban poverty: Findings from Dar es Salaam, Tanzania. *Land Use Policy, 23*(4), 460-472.

Kok, K., & Veldkamp, A. (2001). Evaluating impact of spatial scales on land use pattern analysis in Central America. *Agriculture, Ecosystems & Environment, 85*(1-3), 205-221.

Kombe, W. J. (2005). Land use dynamics in peri-urban areas and their implications on the urban growth and form: the case of Dar es Salaam, Tanzania. *Habitat International, 29*(1), 113-135.

Kutner, M., Nachtsheim, C., & Neter, J. (2004). Applied Linear Regression Model. Available online:http://www.amazon.com/Applied-Linear-Regression-Models-Student/dp/0073014664

Kyessi, S. A., & G. Kyessi, A. (2007). Regularisation and Formalisation of Informal Settlements in Tanzania:Opportunities and Challenges

A Case of Dar es Salaam City: Avilable online: http://www.fig.net/pub/fig2007/papers/ts_8d/ts08d_01_kyessi_kyessi%20_1210.pdf

Li, X., & Yeh, A. G. O. (2001). Calibration of cellular automata by using neural networks for the simulation of complex urban systems. *Environment and Planning A, 33*(8), 1445-1462.

Liu, Y. (2009). *Modelling urban development with geographical information systems and cellular automata*. Boca Raton: CRC.

McCulloch, C. E., & Neuhaus, J. M. (2005). Generalized Linear Mixed Models *Encyclopedia of Biostatistics*: John Wiley & Sons, Ltd.

Mnard, A., & Marceau, D. J. (2005). Exploration of spatial scale sensitivity in geographic cellular automata. *Environment and Planning B: Planning and Design, 32*(5), 693-714.

NBS TANZANIA, N. B. o. S. M. o. P., Economy and Empowerment. (2006). Dar es Salaam regional and distnict projections. Avilable online: http://www.nbs.go.tz/projections/dsm_projections.pdf

Owen, J. G. (1988). On Productivity as a Predictor of Rodent and Carnivore Diversity. *Ecology, 69*(4), 1161-1165.

Poelmans, L., & V.Rompaey, A. (2010). Complexity and performance of urban expansion models. *Computers, Environment and Urban Systems, 34*(1), 17-27.

Seidl, I., & Tisdell, C. A. (1999). Carrying capacity reconsidered: from Malthus' population theory to cultural carrying capacity. *Ecological Economics, 31*(3), 395-408.

Sliuzas, R. V., Ottens, H., & Kreibich, V. (2004). *Managing informal settlements : a study using geo - information in Dar es Salaam, Tanzania*. ITC, Enschede.

UN-HABITAT. (2008). UN-Habitat. United Nations Human Settlement Programme:The state of African Cities-A framework for addressing urban challenges in Africa. Avilable online http://www.unhabitat.org/pmss/listItemDetails.aspx?publicationID=2574

UN-HABITAT. (2010). The state of African cities 2010. Avilable online: http://www.unhabitat.org/documents/SOAC10/SOAC-PR1-en.pdf

UN HABITAT. (2009). UN HABITAT Tanzania,Dar es Salaam city Profile. Avilable online: http://www.unhabitat.org/pmss/getElectronicVersion.asp?nr=2726&alt=1

United Nations. (2006). World Population Prospects The 2006 Revision. Avilable online http://www.un.org/esa/population/publications/wpp2006/WPP2006_Highlights_rev.pdf

URT. (2006). *Analytical Report of 2002 population Census*. Avilable online: www.nbs.go.tz

Van Rompaey, A. J. J., & Govers, G. (2002). Data quality and model complexity for regional scale soil erosion prediction. [Article]. *International Journal of Geographical Information Science, 16*(7), 663-680.

Verburg, P. H., de Nijs, T. C. M., Ritsema van Eck, J., Visser, H., & de Jong, K. (2004). A method to analyse neighbourhood characteristics of land use patterns. *Computers, Environment and Urban Systems, 28*(6), 667-690.

Verburg, P. H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., & Mastura, S. S. A. (2002). Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. *Environmental Management, 30*(3), 391-405.

White, R., & Engelen, G. (2000). High-resolution integrated modelling of the spatial dynamics of urban and regional systems. *Computers, Environment and Urban Systems, 24*(5), 383-400.

White, R., Engelen, G., & Uljee, I. (1997). The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. [Article]. *Environment and Planning B-Planning & Design, 24*(3), 323-343.

Wu, F. (2002). Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International Journal of Geographical Information Science, 16*(8), 795-818.

Yeh, A. G.-O., & Li, X. (2006). Errors and uncertainties in urban cellular automata. *Computers, Environment and Urban Systems, 30*(1), 10-28.

Zhiyong Hu C.P.Lo. (2007). Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems, 31*(6), 667-688.