Comparisons of Statistical Approaches for Modelling Land-Use Change

by

Bo Sun

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirements for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2018

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Land-use and land-cover change (LUCC) can have local-to-global environment impacts such as loss of biodiversity and climate change as well as social-economic impacts such as social inequality. Models that are built to analyze LUCC can help us understand the causes and effects of LUCC, which can provide support and evidence to land-use planning and land-use policies to eliminate or alleviate potential negative outcomes. A variety of modelling approaches have been developed and implemented to represent LUCC, in which statistical methods are often used in the classification of land use and land cover as well as to test hypotheses about the significance of potential drivers of LUCC. The utility of statistical models is found in the ease of their implementation and application as well as their ability to provide a general representation of LUCC, given a limited amount of time, resources, and data. Despite the use of many different statistical methods for modelling LUCC (e.g., linear models and logistic regression), comparison among more than two statistical methods is rare and an evaluation of the performance of a combination of different statistical methods with the same dataset has not been done before. The presented research fills this gap in LUCC modelling literature using four statistical methods, Markov chain, logistic regression, generalized additive models and survival analysis, to quantify their ability to represent LUCC. The selection of these methods is based on criteria: 1) the popularity of a method, 2) the difficulty level of implementation, and 3) the ability of accounting for different scenarios. Results from this comparison show that generalized additive models outperformed Markov chain, logistic regression and survival analysis in overall accuracy of LUCC but logistic regression performed the best for industrial land-use change, and survival analysis performed the best for low-density residential land-use change. The superiority of generalized additive models is due to its ability to model non-linear LUCC predictors, but there is no absolute favor in generalized additive models over other methods in terms of classification accuracies of specific LU changes and the run time. Markov chain is not competitive with the other three methods in most of the LU change cases but it retains the meaning as a null model (i.e., a model without any predictors) in our study.

Acknowledgements

Finding a thing that I would like to devote my time into and having people who always support me are two of the best things that I can think of. My supervisor, Dr. Derek Robinson, gave me the opportunity to apply my knowledge in such a meaningful way. I want to thank him for the generosity of time and all kinds of other supports he has provided to me. I am glad I have chosen him as my supervisor and I am looking forward for working with him in the future.

I would also like to thank my other committee members, Dr. Chris Fletcher, Dr. Peter Deadman and Dr. Steve Roberts, for inspiring me with their knowledge and experience, and reading my thesis and providing expert opinions.

To whom I love the most, my Mom, Dad, spouse, and other family members, I want to thank you all for the unwavering love and support.

To my friends in the Geospatial Innovation Lab (especially Jenny, Ben, Omar, Alex and Collin), you are the reason for me to enjoy the time in the lab. Thanks for all the talks and jokes, they made my day easier.

To friends who are at Waterloo (especially Zoe, Zhuo and Kanchan) and those who spread all over the world (especially Annetta, Shuxin, Junyi and Li), thank you all for the support.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

Abbreviation	**Term**
ALR	Additive Logistic Regression
AFT	Accelerated Failure Time
ABM	Agent-Based Model
CCA	Canonical Correlation Analysis
CV	Cross Validation
CSM	Combination of Statistical Methods
GLM	Generalized Linear Model
GAM	Generalized Additive Model
GHGs	Greenhouse gases
LR	Logistic Regression
LC	Land Cover
LU	Land Use
LUCC	Land Use and Cover Change
MC	Markov Chain
MELR	Mixed Effects Logistic Regression
MEGAM	Mixed Effects Generalized Additive Model
ML	Machine Learning
PH	Proportional Hazard
RF	Random Forest
SA	Survival Analysis
SVM	Support Vector Machine
SWOOP	Southwestern Ontario Orthoimagery Project

Chapter 1 Primer on land use modelling approaches

1.1 Introduction

Land-use and land-cover change (LUCC) is the result of interactions between humans and their environment. Land use (LU) and land cover (LC) are often used interchangeably but they are very different in definition. The LU of a piece of land is determined by human's interests to describe the function that the land serves. LU change is caused by the change in the activities of humans from one type of LU to another (e.g., from agricultural to residential). LC refers to the biophysical attributes at the surface of the earth and changes due to human or environmental intervention (e.g., from bare ground to impermeable surface; Lambin et al., 2006). The LU and LC often have a relationship with each other and one can help determine the other (e.g., residential LU and buildings). Identification of LU change and LC change requires monitoring and mapping of LU and LC over time. Moreover, LU and LC data provide useful information for applications such as natural resource management and studies of climate change. It is important to distinguish LU and LC since LU and LC data provide distinct information to different applications. For instance, baseline thematic map is created using LC data, and studying social problems such as conflict among different uses of land and developmental pressures incorporate the use of LU data (Natural Resources Canada, 2015). LU change and LC change in combination are called LUCC and usually appear together due to their inseparable effects to a society and an environment. The effects of LUCC span from local alteration to ecosystem services (Quintas-Soriano et al., 2016), land management and planning (Nelson, et al., 2010; Pereira et al., 2012), through to regional and global processes such as weather modification and climate change (Lambin et al., 2006). Because of the relevance of LUCC across different academic disciplines, economic process (e.g., collapsing of agricultural supporting sectors due to a loss of agricultural lands), and government regulations (e.g., meeting UNFCCC, Kyoto, and Paris carbon targets), understanding and modelling LUCC is a priority research area (National Research Council, 2014).

Modelling LUCC can help understand environmental issues, such as increasing greenhouse gases (GHGs) in the atmosphere caused by a variety of activities occurring on the ground, and thus influence human responses to those problems. Studies of LUCC can also help manage natural resources such as land, water and wild animals, which are important to achieve a sustainable development of human society in the long run (Meyer and Turner, 1992). More importantly, revealing patterns of LUCC is critical to future planning and management of the landscape to mitigate associated environmental problems (Foley et al., 2005).

Models are used to analyze causes and effects of LUCC and make predictions on future LU and LC under different scenarios (Verburg et al., 2004). A variety of approaches have been developed to model LUCC, which can be grouped based on different techniques used to construct the model such as mathematical models, statistical models and agent-based models (e.g., Parker et al., 2003). Models can also be grouped to represent similar perspectives, for instance, LUCC models can be grouped as spatial versus non-spatial and object versus field (Verburg et al., 2006). Each approach has its own strength and weakness for modelling different types of LUCC under different scenarios. Many LUCC models (e.g., Forest and Agricultural Sector Optimization Model by Adams et al., 1996; California Urban and Biodiversity Analysis Model by Landis et al., 1994) have been developed to focus solely on a specific sector of the economy or target only one or two LU types. In contrast, statistical models, which are not originally designed for modelling any specific subjects, can detect drivers of any types of LUCC and predict LUCC. Therefore, modelling LUCC with statistical models can provide an overall perspective for all LU changes occurred in an area.

Statistical modelling is one of the most widely used approach to representing LUCC because of its relative simplicity of comprehension and operationalization compared to other approaches (e.g., ABM, Bonabeau, 2002; Systems Dynamics Models, Ford and Ford, 1999). Despite the use of many different statistical methods for modelling LU change (e.g., Markov Chain and logistic regression), to the best of author’s knowledge, no one has investigated the performance of a combination of different traditional statistical methods with the same dataset. Hence, four conceptual approaches (stochastic process, parametric model, non-parametric model, and time series model) to modelling LU change are compared and contrasted in the presented research. These four approaches are operationalized as Markov chain, logistic regression, generalized additive model and survival analysis. The selection of these methods is based on : 1) frequency of use, 2) the difficulty of implementation, and 3) ability to account for different scenarios. The background and mathematical underpinning of these methods are presented prior to concluding the chapter with the overarching goal and research questions of the presented thesis.

1.2 Statistical Approaches

Statistical models usually require distinct sets of drivers that are suitable for different study interests in LUCC modelling. Note that, drivers in LU science context have the same meaning as predictors in the context of statistics. The results from statistical models can be used either as a final product (e.g., a probability of LUCC) or as suitability maps for subsequent allocation of LU across space (Alcamo et al., 2006). In addition, the LU at a location can be classified using the estimated probability of LU change from statistical models. Classification accuracy implies the suitability of a method in LUCC modelling. The classification accuracies of Markov chain, logistic regression, generalized additive model and survival analysis are compared and the analysis is presented in Chapter 2. In this thesis, dependent variable is used interchangeably with response variable; independent variable is used interchangeably with covariates and predictors.

1.2.1 Markov Chain

Markov chain (MC) models incorporate stochasticity in LUCC between states (time steps). A transition probability matrix is used to record probabilities of changes (probabilities of LU changes) between different statuses/events (LU types) occurred over time. MC produces the transition probability between two states as a function of past state. MC has been used for many research interests such as movements of classes of the rental housing in several U.S. cities (Clark, 1965) and consequences of urban growth to agricultural and natural land uses in Niagara Region, Ontario, Canada (Muller and Middleton, 1994). In LUCC studies, it has been applied to quantify the LU changes in a future state (Muller and Middleton, 1994; Iacono et al., 2012).

1.2.1.1 Method

The MC process works slightly differently when time is represented discretely versus continuously. As data are usually collected in a discrete manner in LUCC studies, discrete time MC (DTMC) is more appropriate to model LUCC. DTMC requires a finite number of discrete states with a set of finite events that are mutually exclusive and collectively exhaustive (Stokey and Zeckhauser, 1978). In LUCC context, having mutually exclusive events ensures that only one LU or LC exists at a given location and time. Collectively exhaustive in LUCC context refers to the feature of a land that must be at least one type of LU or LC given all the possibilities.

Together, mutually exclusive and collectively exhaustive guarantee one type of LUCC occurs on a single unit of land at a time over the entire study area. The probability of an event at a given time has a memory-less property, which means the probability only depends on the event occurred in the nearest past. Furthermore, the transition probability of a change between any two specific states is constant over time.

Combining mathematical notations with LUCC context, X_n denotes the LU or LC at time n at a location, where $n = 1\,, 2, \ldots$ and is finite. Any time interval between any two states are assumed to have a uniform length. Given a set of LUs or LCs indexed by i and j, $X_n = i$ or j means that the LU or LC type is i or j at time n at a location. $X_n{}'s$ are rarely independent in LUCC context since the current LU or LC at least depends on the previous LU or LC at a location. Moreover, let P_{ij} denote a transition probability, a fixed probability of going from a current state with status i to a future state with status j. P_{ij} is called a one-step transition probability when $X_n = i$ and $X_{n+1} = j$, which means P_{ij} is the probability of going from state n with status i to the next state $n + 1$ with status j. The mathematical expression of P_{ij} is shown in Equation (1):

$$P_{ij} = P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} \qquad (1)$$

for all statuses $i_0, \ldots, i_{n-1}, i\,, j$ and $n \geq 0$. Equation (1) is read as the conditional probability of status j at time $n + 1$ given all past states. However, the probability of X_{n+1} given all past states only depends on the status of X_n and the probability of $X_n = i$ is conditional on its previous states back to the initial status i_0 at the initial time point. Furthermore, P_{ij} has the following properties: 1) $P_{ij} \geq 0\ \forall i, j \geq 0$ since it is a probability and cannot be negative and 2) $\sum_{j=0}^{\infty} P_{ij} = 1\,, i = 0, 1, \ldots$ since the probabilities of all events occur sum to one. The collection of all $P_{ij}{}'s$, the transitioning probabilities between every two states, can be written in a matrix form in Equation (2):

$$\boldsymbol{P} = \left\Vert \begin{matrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & \vdots & \vdots & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{matrix} \right\Vert . \qquad (2)$$

An *n*-step transition probability should be used when an event takes n steps to occur. Let P_{ij}^{n} denote the *n*-step transition probability of changing from status *i* to status *j* via a period with *n* equal intervals, which can be expressed mathematically as $P_{ij}^{(n)} = P\{X_{n+k} = j \mid X_k = i\}$ where $n, k \geq 0$ and $i, j \geq 0$. The calculation of $P_{ij}^{(n)}$ is done by $(P_{ij})^n$ (i.e., P_{ij} to the *n*th power). Similarly, $\boldsymbol{P}^{(n)}$ is used to denote the *n*-step transition probability matrix.

In addition to one-step and *n*-step transition probabilities, $P_{ij}^{(n+m)}$is used to represent the transitioning probability from status i to status j via state s, where n is the number of steps taken from state i to s, and m is the number of steps taken from state s to j, and is expressed mathematically in Equation (3):

$$P_{ij}^{(n+m)} = \sum_{s=0}^{\infty} P_{is}^{(n)} P_{sj}^{(m)} \qquad \forall n, m \geq 0 \; and \; \forall i, j \tag{3}$$

The transition probability matrix containing all $P_{ij}^{(n+m)}{}'s$ is denoted by $\boldsymbol{P}^{(n+m)}$. The following relationship can be derived: ${P_{ij}}^{(n+m)} = {P_{is}}^{(n)} \cdot {P_{sj}}^{(m)}$. Moreover, time intervals between any two time steps in n and m steps are assumed to be equal.

Knowing the background of transition probabilities is important since LUCC studies with different availability of data require different transition probabilities to conduct MC. When a MC is used to model LUCC between two dates, which is the simplest case of DTMC that only contains a single time interval between the two dates, a one-step transition probability should be used. The *n*-step transition probability is used when there is a need to model LUCC over n equal-length time points. Moreover, P_{ij}^{n+m} is suitable for studying LUCC from one LU or LC type to another via a transient LU or LC when time intervals between all $n + m$ steps are all in the same length.

Furthermore, a transition probability can reach a steady state as time t goes on, which is called a steady state probability. Mathematically, this process can be expressed as $\lim_{t \to \infty} P^t = P$, which means P, the steady state probability matrix, will become steady as time goes to infinity. In fact, the duration that the process takes to become steady can be calculated. Once the steady state has reached, the event will stop changing. A transition probability matrix is a square matrix in which its number of rows or columns should equal the number of elements in the initial state vector that contains possibilities for all events at the initial time point. For example, if an initial vector contains transition probabilities for three LUs, then the transition probability matrix should have a dimension of three by three. The steady states of the three LU types can be determined by multiplying the initial state vector with the steady state probability matrix. This feature of MC is important when the goal of a study is 1) to determine the steady LU or LC for given locations, 2) to verify the steady LU or LC based on theory, and 3) to investigate the duration to steady states.

1.2.2 Logistic Regression

Logistic Regression (LR) is a statistical modelling approach that is used to model categorical dependent variables. There are several types of LR: 1) simple LR that regresses binary responses on a single independent variable, 2) multiple LR that regresses binary responses on a set of independent variables, 3) ordinal LR that requires ordinal responses and 4) multinomial LR that is able to model a dependent variable with more than two categories. Among the statistical approaches found in the literature, LR is a common approach used in LUCC modelling (Brown et al., 2012). In addition to typical LUCC modelling, it has applied to achieve other interests such as modelling of urban development (Landis, 1994; Landis and Zhang, 1998a; 1998b) and modelling of deforestation (Chomitz and Gray 1996; Mertens and Lambin, 1997).

1.2.2.1 Method

LR has gained popularity in modelling LUCC due to the categorical nature of LU and LC data (Muller and Zeller, 2002). Among all types of LR, multiple LR has been used most often since there is typically more than one driver affecting LUCC. Therefore, multiple LR is the focus of this study among all other models in the family of LR. Before going into details of LR, a brief introduction is given to the Exponential family and generalized linear models (GLMs) in order to better understand LR. Exponential family is a class of probability distributions (e.g., Normal distribution, Exponential distribution, and Binomial distribution; Evans et al., 2000) that can be formulated in a general format by re-arranging and transforming parameters (Andersen, 1970). GLMs are a group of models (e.g., linear model and Poisson regression) in which their response variables follow probability distributions from the Exponential family. A link function is required to connect the mean of response variables with a linear combination of covariates in GLMs since the relationship between responses and covariates are not always linear. In summary, LR resides within the broader category of GLMs since its response variables follow a Binomial distribution that belongs to the Exponential family, and require a link function to express the binary feature of responses.

Moreover, a link function connects a linear predictor, denoted by η, and the mean of the response variable, denoted by μ, through the equation $\eta = g(\mu) = \boldsymbol{X\beta}$, where g is a function of μ, $\boldsymbol{X}$ is a set of covariates, and $\boldsymbol{\beta}$ is the corresponding coefficients of $\boldsymbol{X}$. Examples of link functions are identity link (linear regression) and log link (Poisson regression). To understand the use of link function, it is useful to introduce some general notation. Let $\boldsymbol{Y} = (y_1, y_2, \dots, y_n)$ be the vector form of response variables, y_i be the measurement of observation i, where i is the index of n observations (i.e., $i = 1, 2, \dots, n$), and $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)'$ be the design matrix formed by covariates. Each $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$, an element of $\mathbf{X}$, contains a set of p covariates for observation i, where $i = 1, 2, \dots n$. To be more explicit, $\boldsymbol{x}_{ij}$, an element of $\boldsymbol{x}_i$, denotes the measurement of covariate j for individual i where j is the index of p covariates (i.e., $j = 1, 2, \dots, p$). Moreover, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ is a vector of coefficients corresponding to p covariates and $\boldsymbol{\epsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ is a vector of errors corresponding to n regression models.

The formulation of linear regressions is reviewed to better understand the mathematical background of LR. In a linear regression, the linear relationship among $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. The structure of this relationship is expressed in Equation (4):

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{4}$$

$\boldsymbol{Y}$ is assumed to follow a Normal distribution with a mean equals to $\boldsymbol{X\beta}$ and a variance equals to $\boldsymbol{\sigma}^2$. The $\boldsymbol{\epsilon}$ follows a Normal distribution with a mean of 0 and a variance of $\boldsymbol{\sigma}^2$. Moreover, the identity link used in linear regression is expressed as $\eta = g(\mu) = \mu$.

LR replaces the identity link used in linear regression by a logit link that is written generally as $\eta = logit(\mu) = lo\,g\left(\frac{\mu}{1-\mu}\right)$. Binary response variable in a multiple LR follows a Bernoulli distribution with $\pi(\boldsymbol{x}_i)$ that indicates the probability of an observation i with $\boldsymbol{x}_i$. An individual logit link function for observation i is provided in Equation (5) to give a better visualization of the link function in a multiple LR.

$$\eta_i = logit(\pi(x_i)) = log[\frac{\pi(x_i)}{1-\pi(x_i)}] = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_1 x_{ip}. \quad (5)$$

Hence, the logit link function used to represent all observations in a multiple LR can be expressed as $\eta = logit\big(\pi(\boldsymbol{X})\big)$. The linear predictor in LR has the property of being continuous and ranging from $-\infty$ to $+\infty$ as the Normal response variables in a linear regression. The inverse of the logit link function is $\pi(\mathbf{X}) = \frac{e^{X\beta}}{1+e^{X\beta}}$, which can also be expressed as $E(\boldsymbol{Y}|\boldsymbol{X})$ that is read as the expectation of $\boldsymbol{Y}$ given $\mathbf{X}$ (i.e., the mean of $\boldsymbol{Y}$ given $\mathbf{X}$). This feature can be used to calculate the mean parameter of LR when values of all covariates and estimated coefficients are given and can be used to quantify the amount of change in the response variable due to a unit change in one of the covariates.

Equation (5) does not incorporate the LR error term ($\boldsymbol{\epsilon}$) that has a distribution that differs from the Normal distribution of $\boldsymbol{\epsilon}$ in a linear regression. The individual error term equals to $1 - \pi(\boldsymbol{x_i})$ when $y_i = 1$ (i.e., the occurrence of event y_i) with a probability of $\pi(\boldsymbol{x_i})$ and equals to $\pi(\boldsymbol{x_i})$ when $y_i = 0$ (i.e., the absence of eventy_i) with a probability of $1 - \pi(\boldsymbol{x_i})$. In summary, ε_i follows $N(0, \pi(\boldsymbol{x_i})[1 - \pi(\boldsymbol{x_i})])$ in LR. Therefore, $\boldsymbol{Y}, \boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ have the relationship of $\boldsymbol{Y} = \frac{e^{X\beta}}{1+e^{X\beta}} + \boldsymbol{\epsilon}$.

In terms of modelling LUCC with LR, $\boldsymbol{Y}$ is the set of binary variables (0 or 1) that indicate the status of LUCC (unchanged or changed) at a location. $\boldsymbol{X}$ is the set of LUCC predictors and $\boldsymbol{\beta}$ is the set of estimated coefficients corresponding to $\boldsymbol{X}$. Moreover, $\pi(\boldsymbol{x_i})$ represents the probability of a LUCC occurring given a set of LUCC predictors at a location. One approach to interpreting the result of a multiple LR is to use the odds ratio (OR). The OR is a measurement of the likelihood of an outcome in the presence of the effects of some covariates compared to the outcome occurring in absence of the effects of the same covariates. A simple LR is used to illustrate the use of the OR. Let the single covariate x in the simple LR be a binary variable with $x = 1$ meaning a presence of some characteristic and $x = 0$ meaning an absence of the characteristic. Then, OR can be expressed mathematically as $\frac{\pi(x=1)/[1-\pi(x=1)]}{\pi(x=0)/[1-\pi(x=0)]} = e^{\beta_1}$ and is interpreted as the likelihood of an event with presence of x against the absence of x, where β_1 is the coefficient of x. In some cases, the logarithmic form of OR is preferred, which is referred to as the log-odds ratio (LOR). The LOR of the simple LR (i.e., $\log(e^{\beta_1})$) is β_1, which infers to the direct impact on the event caused by different levels of x.

For a multiple LR, it is very often that both continuous and categorical covariates exist at the same time. In general, when x is continuous, the OR equals an exponential of the unit difference of x and is interpreted as the likelihood of an event with a unit increase or decrease of x. When x is categorical, the OR equals an exponential of the level difference of x and is interpreted as the likelihood of an event with x being at a specific level. When both continuous and categorical variables present in a multiple LR, the interpretation of OR needs to account for both the differences in measurements of continuous variables and the differences in levels of categorical variables. Dummy variables or indicator variables are used to account for levels of categorical variables such as gender and treatment groups. In general, if a categorical variable contains K levels, $K - 1$ dummy variables are needed to replace the function of the categorical variable in the model. A level among all K levels is used as the base case and can be set by user's preference and the rest of $K - 1$ levels are represented by $K - 1$ dummy variables to indicate the existence of the corresponding $K - 1$ levels of the categorical variable. The way a continuous variable is used is not affected by categorical variables but the interpretation of it may need to account for the effects of categorical variables at specific levels. Variables in GLMs can be selected by a forward method, a backward method or a step-wise method based on Akaike Information Criterion (AIC; Bozdogan, 1987).

1.2.3 Generalized Additive Models

Generalized additive models (GAMs) extend GLMs by using a series of smoothing splines to represent non-linear relationships between the expected mean of responses and the independent variables (Hastie and Tibshirani, 1990). Predictors in GAMs can have unique and non-linear impacts on the response variable individually, in which an individual non-linear impact of a predictor does not need to follow any probability distributions. GAMs have been used in LU science since drivers of LU change are usually non-linear (Brown, 1994). This advantage of GAMs over GLMs ensures that more realistic situations can be modeled. However, GAMs are used less frequently because they are more difficult to implement and interpret.

1.2.3.1 Method

A GAM with k^{th}-order smoothing splines/functions is referred to as a GAM that uses k^{th}-order piecewise polynomials to represent the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$. A smoothing spline has a continuous property on itself and on the derivatives of all its $k-1$ degree of polynomial functions. A commonly used spline is the cubic spline ($k = 3$).

The smoothing parameter (λ) controls the degree of smoothness of a smoothing spline and determines the complexity of a GAM. In general, a large λ increases the degree of smoothness, lowers the level of complexity of the model and thus can underestimate the real situation; a small λ introduces more variability into the smoothing spline while raising the level of complexity of the model and can cause overfitting.

A GAM can be viewed as a parametric approach when smoothing functions are replaced by parametric variable transformations and parametric functions (Hastie and Tibshirani, 1990). Parametric variable transformations can be done through some functions such as logarithmic function, square-root function, inverse function and polynomials. The set of transformed predictors can also be used to construct non-parametric spline functions. A GAM with smoothing functions is usually seen as non-parametric, given the non-parametric nature of smoothing functions but a GAM can contain a mixture of parametric and non-parametric terms.

Since GAM is an extended version of GLM, the general representation of the linear relationship between the expected mean of responses given a set of predictors with some undefined smoothing functions is very similar to that of a GLM and is shown as follows:

$$E\left(\boldsymbol{Y}\middle|X_1, X_2, \dots, X_p\right) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) = \alpha + \sum_{j=1}^{p} f_j(X_j) \qquad (6)$$

where $\boldsymbol{Y}$ is response variable, $X_1, X_2, \dots, X_p$ are predictors, and j is the index for predictors and $j = 1, 2, \dots, p$. The f_j where $j = 1, 2, \dots, p$ in Equation (6) represents an unspecified smoothing function of predictor X_j; the parameter α is similar to the intercept term, β_0, in a linear regression. Without future restrictions on the model, α is unidentifiable since it can change while smoothing functions change; thus α is not unique. One way to provide an initial guess of α is to set $\sum_{i=1}^{N} f_j(X_{ij}) = 0 \; \forall j$ where N is the total number of observations. This returns $\alpha = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$, where y_i represents the measurement of observation i and $\bar{y}$ represents the mean of $y_i{}'s$.

Parametric terms in GAMs are estimated the same way as they are estimated in GLMs with defined link functions. Smoothing functions can be estimated using the back-fitting algorithm. The procedure of the back-fitting algorithm was introduced along with the idea of GAMs by Leo Breiman and Jerome Friedman in 1985. The following steps of the back-fitting algorithm were retrieved from Hastie and Tibshirani (1990). They used S to represent an arbitrary scatterplot smoother. The steps are:

1. Initialize: $\alpha = ave(y_i),\ f_j = 0,\ j = 1, \dots, p$
2. Cycle: $j = 1, 2, \dots, p, \dots, 1, 2, \dots, p, \dots$

   $f_j \leftarrow S_j(\boldsymbol{y} - \alpha - \sum_{k \neq j} \boldsymbol{f}_k \mid x_j)$
3. Continue step 2 until the individual functions do not change.

The second step in above process means that a smoothing function f_j is fit by regressing the partial residuals of $f_j(\, x_j)$ on x_j while all other smoothing functions and predictors are fixed. The fitting of a smoothing function stops when the function becomes stable, and every unspecified smoothing function has to go through this process.

The way of selecting a link function for GLMs is extended to GAMs. Due to the nature of LUCC data that is used as response variables in GAM, a logit link should be selected, which specifies the GAM to be an additive logistic regression (ALR). Therefore, a review was done for the fitting of smoothing functions in an ALR using the back-fitting algorithm. Using the notations of $\boldsymbol{Y}$ as a binary response variable and $\boldsymbol{X}$ as the set of predictors, the general form of the linear predictor of an ALR is expressed in Equation (7):

$$g(X) = log\left[\frac{E(Y = 1|X)}{1 - E(Y = 1|X)}\right] = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \tag{7}$$

For ALR, the Newton-Raphson procedure is used along with the back-fitting algorithm to estimate unspecified smoothing functions. The combined algorithm was retrieved from Hastie et al. (2009) and shows as following:

1. Initialize: $\hat{\alpha} = \log[\frac{\bar{y}}{1-\bar{y}}], \hat{f_j} = 0, where\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i\ \forall i, j$
2. Define: $\hat{\eta_i} = \hat{\alpha} + \sum_j \hat{f_j}(x_{ij})$ and $\hat{p_i} = \frac{1}{1+\exp(-\hat{\eta_i})}$

   Iterate: (a) Construct the working target variable $z_i = \hat{\eta_i} + \frac{(y_i - \hat{p_i})}{\hat{p_i}(1-\hat{p_i})}$

(b) Construct weights $w_i = \widehat{p_i}(1 - \widehat{p_i})$

(c) Fit an additive model to the targets z_i with weights w_i, using a weighted back-fitting algorithm. This gives new estimates $\hat{\alpha}, \widehat{f_j}\ \forall j$

3. Continue Step 2 until the individual functions do not change or change less than a pre-specified threshold.

In general, the value of α can be determined once all smoothing functions are stabilized. In a GAM, $\boldsymbol{Y}, \boldsymbol{X}$, $\boldsymbol{\beta}$ (for linear LUCC predictor only) and $\pi(\boldsymbol{x_i})$ (refers to the probability of LUCC given a set of LUCC predictors at a location) are identical to those in a GLM. GAMs and GLMs also use the same variable selection methods (e.g., forward selection, backward selection and stepwise selection). However, the interpretation of smoothing functions of predictors in GAMs requires more attention since their relationship with the response variable cannot simply be explained using linearity.

1.2.4 Survival Analysis

Survival analysis (SA) methods are typically used to analyze longitudinal data through a set of statistical techniques and use both the time length of observations stayed in the experiment and an indicator variable showing the occurrence of an event of interest for all observations as the response variables. Potential factors that influence the occurrence of an event are represented as predictors, which are used to calculate the success/failure rate of the event occurring.

SA models are useful to detect effects of spatial and temporal predictors of LUCC over time (e.g., An et al., 2011). The Cox proportional hazard (PH) model is used to calculate a hazard ratio that represents the risk of an event occurring at a given time. Using a PH model, Irwin and Bockstael (2002) found an evidence of a negative relationship between the share of development within neighborhoods and the hazard of development in residential subdivisions in exurban Maryland, US. SA may have the potential to accurately model LUCC because it can handle time-related variables (An and Brown, 2008) but they are still not widely used in the study of LUCC modelling.

1.2.4.1 Method

SA in LUCC modelling is based on the concept of establishing the survival time of a LU or LC at a specific location given a set of predictors that may influence the occurrence of LUCC. The survival time of a parcel can be characterized using a survival function, a density function, or a hazard function. These functions are mathematically equivalent, which means that one can be derived given any of the other two.

Let T be the survival time，the time that an event of interest (i.e., an occurrence of a particular type of LU or LC change at a parcel) occurs, where $T \geq 0$. The survival function, also known as cumulative survival rate, determines the probability of an object surviving beyond time t. This function is usually denoted by $S(t)$ and can be expressed in terms of probability as

$$S(t) = P(T > t) = P(\mathit{a\ parcel's\ LU\ orLC\ type\ remains\ unaltered\ longer\ than\ t})\ , \quad (8)$$

where $P(\cdot)$ means the probability of some event. Therefore, $S(t)$ calculates the probability of a parcel remaining its initial LU or LC beyond time t. The probability that an event occurs at or before time t is defined by a cumulative distribution function $F(t) = P(T \leq t)$. Hence, $F(t)$ can be used to calculate the probability of a parcel having a LUCC before or at time t. Note that, $S(t)$ is equivalent to $1 - F(t)$. Moreover, $S(t)$ is a non-increasing function of time t and has two properties: 1) $S(0) = 1$ and 2) $S(\infty) = 0$. Survival curve, a graphic presentation of $S(t)$, shows the relationship between survival rate/probability and time. In LUCC context, $S(t)$ can be calculated as

$$S(t) = \frac{\mathit{number\ of\ parcels\ that\ have\ LU/LC\ remained\ unaltered\ beyond\ time\ t}}{\mathit{total\ number\ of\ parcels}}. \quad (9)$$

The survival time T is like any other continuous random variable that has a density function, which can be expressed as

$$f(t) = \lim_{n \to \infty} \frac{P(\mathit{a\ parcel\ fails\ to\ remain\ LU/LC\ unaltered\ in\ the\ interval}\ (t,\Delta t))}{\Delta t}. \quad (10)$$

$f(t)$ also has two properties: 1) $f(t) \geq 0\ \forall t \geq 0$ and $f(t) = 0\ for\ t < 0$, and 2) $\int f(t)dt = 1$. The plot of $f(t)$ is called density curve that shows the relationship between the frequency of failure and time. $f(t)$ can be calculated as

$$f(t) = \frac{\mathit{number\ of\ parcels\ that\ fail\ to\ remain\ LU/LC\ altered\ in\ the\ interval}\ (t,\Delta t)}{\mathit{total\ number\ of\ parcelss} \times \Delta t}\ . \quad (11)$$

$f(t)$ can also be derived from $S(t)$ and $F(t)$ as following:

$$f(t) = -\frac{dS(x)}{dx} = -S'(t) = \frac{dF(x)}{dx} = F'(t). \qquad (12)$$

Hazard function, also known as hazard rate, gives the conditional failure rate and is usually denoted by $h(t)$. It can be considered as a rate of failure per unit of time. Hazard rate is not a probability but a limit of a probability (Equation (13)).

$$h(t) = \lim_{n\to\infty} \frac{P(a\, parcel\, that\, has\, LU/LC\, altered\, falls\, in\, the\, interval\, (t,\Delta t)}{\Delta t}. \qquad (13)$$

In addition, $h(t)$ has the following relationship with $f(t), F(t)$, and $S(t)$: $h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}$, which can be calculated as

$$h(t) = \frac{number\, of\, parcels\, that\, have\, LU/LC\, altered\, in\, the\, intervival\, (t,\Delta t)}{total\, number\, of\, parcels\, having\, LU/LC\, unaltered\, at\, time\, t}. \qquad (14)$$

There is no constraint on the shape of $h(t)$ curve. A cumulative hazard function is defined as $H(t) = \int_0^t h(x)dx$ and has the following relationship with $S(t)$: $H(t) = -\ln S(t)$. Equations (9), (11) and (14) are not applicable to incomplete data. Non-parametric methods and different likelihood functions need to be used to estimate those statistics when incomplete data present.

Similar to a linear model and a LR, the relationship between the responses (i.e., the hazard of an observation at time t) and predictors can be expressed in a linear format in Equation (15):

$$log\, h_i(t) = logh_0(t) + \beta_1 x_{i1}(t) + \beta_1 x_{i2}(t) + \cdots + \beta_1 x_{ip}(t), \qquad (15)$$

where i is the index of parcels and p is the index of predictors. Predictors can be functions of time (i.e., time-varying). The baseline hazard, $h_0(t)$, can also be a function of time but is constant for all observations (i.e., parcels) at time t.

Traditional statistical models (e.g., linear regression and logistic regression) are insensitive to time variables but SA can handle both time-varying and time invariant variables occurred in the same model. In addition, skewness is a feature of survival data due to occurrence of censoring and truncation (Hougaard, 1999; Clark et al., 2003). Therefore, the Normal distribution is usually forgone in favor of other distributions such as Exponential, Weibull, and Gamma.

When only partial information about an observation is known, censoring occurs. Three types of censoring can occur in a LUCC study: right censoring (e.g., the LU type of a parcel is observed at the beginning of the study and remains unchanged at the end of the study), interval censoring (e.g., a LU change of a parcel is observed during the study but the exact time of the event occurred is unknown), and left censoring (e.g., a parcel that is enrolled in a study experienced a LU change prior to the commencement of the study and the time of the event is unknown). Truncation occurs when the failure time of an observation falls at a time that is outside of the study period and is usually caused by the design of a longitudinal study. There are two types of truncations: left and right. Left truncation often occurs with right censoring when age is the time scale of a study. For example, patients who have a certain disease and do not satisfy an age requirement for entering a study are not observed. Right truncation occurs when a patient is infected by some diseases but has not developed the disease in a study period. Therefore, the age and the development of a disease are considered the two milestones to determine a truncation in a longitudinal study. Transferring this idea to a LUCC study, left truncation may occur when a parcel has experienced a LUCC prior to the entry of the study but is not enrolled in the study due to its age. Right truncation may occur when a parcel is observed to have undergone a process of LUCC after the entry but has not completed the conversion before the end of the study.

In LUCC studies, censored observations will be treated differently in the likelihood function of a selected distribution, which is the same as in longitudinal studies. Moreover, left truncation is not a big issue since ages of parcels are usually not an interest of a study, which makes an occurrence of a LUCC considered the only milestone. If age is an interest, left truncation can be taken into account by using a nonparametric product-limit estimator with age being treated as a variable to estimate the distribution of truncated data (Cain et al., 2011). Right truncation in a LUCC study can be solved by classifying developing parcels to a new category such as under-development.

1.3 Agent-based Models and Statistical Approaches

Statistical methods are limited in their ability to account for complicated interactions among many factors including an explicit representation of decision-making process. To accommodate the limitations, agent-based models, which are designed to simulate process-based phenomenon, can be merged with statistical methods to provide improved understanding and representation of LUCC.

1.3.1 Agent-based Models

Agent-based models (ABMs) are composed of agents, an environment, and a set of rules that regulate agents' behavior. The agents in ABMs interact directly with their environment, with each other, or indirectly with each other through their environment. Agents represent real-world decision-making actors, which can be individuals, households, or organizations at all levels. In a decision-making process, an agent can interact and be influenced by other agents and its environment. In many cases, ABMs provide simple proofs of existence that demonstrate how one or more mechanisms taking place at a sub-system level can produce system-level outcomes (Waldrop, 1990). For example, using a simple ABM that simulates household movement within an urban area, Schelling (1969) demonstrated that society has the potential to be segregated based on a relative preference (e.g., languages and races) of individuals to be adjacent to other individuals who are similar to them. Alternatively, ABMs could be developed to incorporate large amounts of data, and simulate highly detailed and complicated processes (e.g., simulation of traffic network; Nguyen and Ho, 2016).

Compared to traditional statistical methods, ABMs have the advantage of representing human decision-making, which may be influenced by interaction among social, economic and environment factors at different levels, about how a piece of land is used (Matthews et al., 2007) in the context of LUCC modelling. ABMs are also flexible in adding, removing, and exchanging the components in the model (e.g., changing rationales of agents and creating a new environment for agents to react) outside the simulation period, which adds strength to the wide application of ABMs. However, the ability of ABMs to represent the complexity found in human systems and human decision-making can be influenced by decentralized agents that face limited local information while acting in a parallel fashion (Huigen, 2003).

On the other hand, as computer simulation models, ABMs face challenges associated with model validation. Compounding the challenges of validation is the availability of appropriate data. While quantitative techniques can be used to assess the validation (e.g., statistical measurements of linear similarity; Huigen, 2003) of model structure and outputs against real-world processes and observations, our ability to represent seemingly stochastic, chaotic, and irrational human decisions is difficult and requires extensive amounts of data. Due in part to the challenges associated with representing human decision-making behaviour, most ABMs have been case based and applied at small spatial extents in LUCC studies. While some are attempting the application of ABMs across large spatial extents (e.g., Murry-Rust et al., 2014), data limitations remain a predominant constraint. One solution could be to construct hybrid models of ABMs that incorporate statistical representations that 1) can act as agents to produce outcomes repeatedly in a relatively short time period, 2) can be validated, and 3) can be applied across large spatial extent.

1.3.2 Hybrid Models of Statistical Approaches and Agent-based Models

Statistical models can produce probabilities of LUCC relatively efficiently in terms of cost and time, and can be constructed to cover a large spatial extent with the support of remote sensing data. In contrast, obtaining agent characteristic data and defining practical rules that regulate agents behavior (how they interact with other agents and their environment) in ABMs can be a costly and labor intensive undertaking that typically involves the use of survey data which may not be feasible to scale up to large spatial extents. If there are insufficient data being collected using a sampling method, the validity of the model and the representativeness of the agents at a larger scale might be questioned. When empirical data about actor characteristics and decision-making are not available or scarce, probabilities of LUCC can be used as substitutes for the decision-making process used by agents. In other words, statistical models can be situated within an agent-based framework and represent the individual decisions of agents driving LUCC 1) in lieu of behavioral data about the actors represented by the statistical models, 2) as a placeholder for more characteristic and behavioral data about real-world actors making LU decisions, and 3) to provide a mechanism, which one can get a representative model up running quickly, and can provide a range of insights and findings that can be extended when behavior data become available.

1.5 Thesis Overview

To contribute to current scientific efforts in LUCC modelling across large spatial extents, four statistical methods (MC, LR, GAM and SA) have been conducted and compared for their predicting powers of LUCC to fulfill the gap of a lack of formal comparison between their relative performances. These methods span a range in 1) frequency of use in existing research (from frequent to rare) and 2) statistical approach (e.g., probabilistic versus time-series analysis). The result of this study can also help develop hybrid models that could underpin a provincial agent-based model. The presented research answers the following research questions:

1) What is the overall accuracy of different types of statistical methods in representing LUCC?
2) What is the distribution of accuracies for different types of statistical methods by LU type?

In order to answer the research questions stated above, each of the four methods was used to model changes among pre-defined LU types. The prediction accuracy of each model was calculated and recorded by method type and LU change type.

The structure of the remaining portion of the thesis is as following: Chapter 2 is structured as a manuscript that situates the research questions in the context of LUCC modelling literature and then describes the study area, data used, and results. The broader implications of the research are then discussed and conclusions are provided. Chapter 3 highlights the contribution of the presented research, discusses broader applications of these methods, and identifies future research directions that can be built based on the presented research.

Chapter 2 Comparisons of Statistical Models in Modelling Land-use Changes

2.1 Introduction

Land use (LU) describes the use/purpose of a piece of land, which is defined by human interests and can be altered by human activities. Land cover (LC) is the biophysical characteristic of a piece of land. Even though LU types appear in various patterns across different parts of the world, they generally tend to sacrifice the natural environment in exchange for providing for human needs (Foley et al., 2005). For example, LU practices such as clearing forest and grassland for farming cause changes in soil carbon storage (Bolin and Sukumar, 2000). Moreover, the change of LU can cause a change of LC, for instance, from deciduous to crops. More carbon is released into the atmosphere as the capacity of soil absorbing carbon reduces thus enhancing problems such as climate change. Other consequences of LU change are, but not limited to, a loss of biodiversity and impacts on ecosystem services (e.g., Foley et al., 2005; Pereira et al., 2012).

Land-use and land-cover change (LUCC) models can help understand causes of LUCC through detecting drivers of LUCC. Drivers of LUCC come from various aspects, such as social, economic and biophysical, and can interact with other drivers to influence LUCC. Models can also help reveal patterns and impacts of LUCC, which provides evidence and support for LU policies and planning. LUCC models that incorporate different disciplines also imply the diverse influences that LUCC can produce.

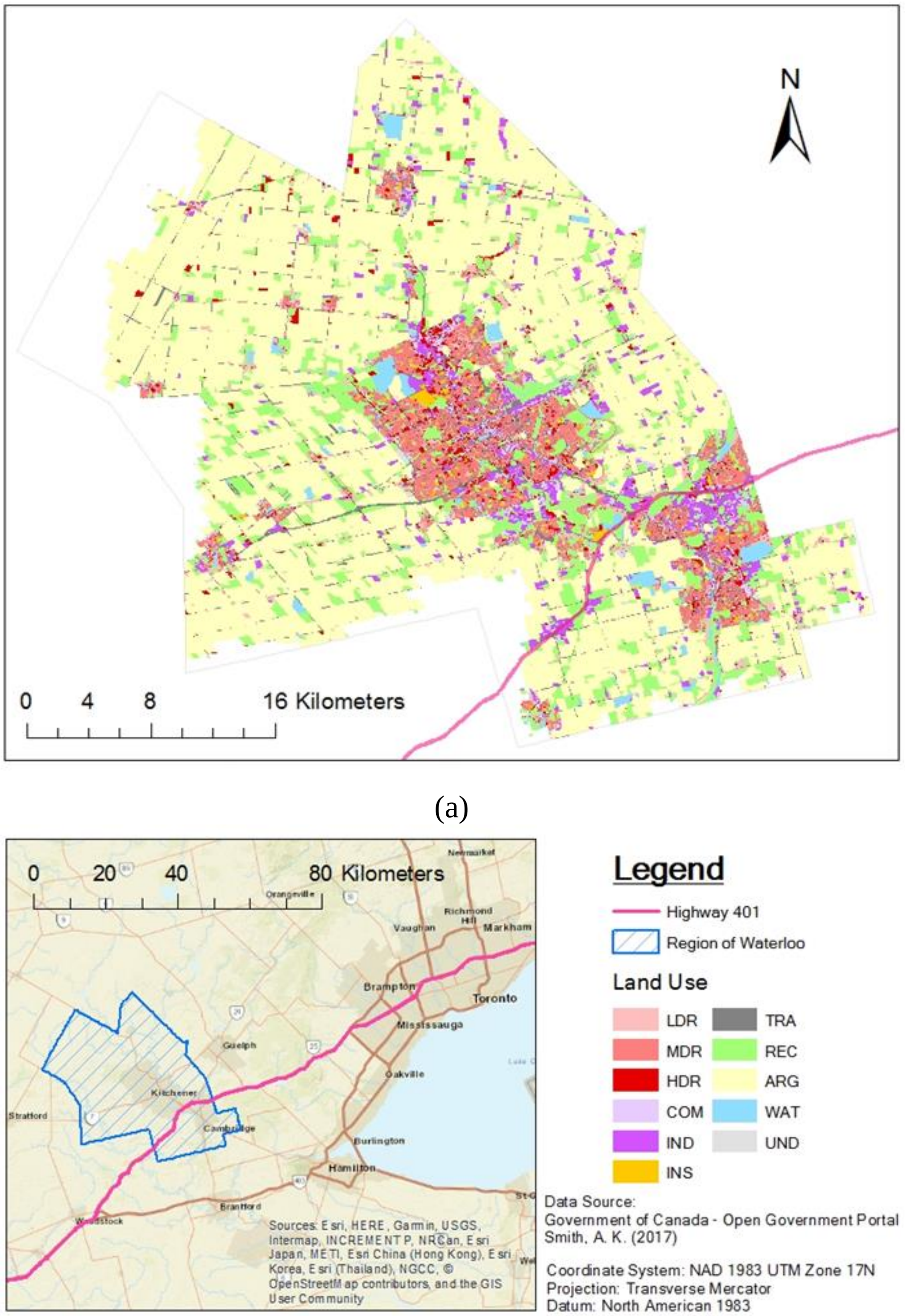
A variety of models have been developed and implemented to represent and improve our understanding of LUCC (e.g., FASOM by Adams et al., 1996; CLUE Model by Veldkamp and Fresco, 1996a). Among the methods used to model LUCC, empirical statistical models are often used to test hypotheses about the significance of potential drivers of LUCC, which can be seen as complementary to the development of process-based models (Veldkamp and Lambin, 2001). To date, many statistical models have been used to model LUCC, of which logistic regression and linear regression are the most frequent (Aspinall, 2004). The utility of statistical models is found in the ease of their implementation and application as well as their ability to provide a general representation of LUCC, given a limited amount of time, resources, and data. The trade-off in the use of statistical modelling approaches is their limited ability to represent the explicit processes associated with human decision-making (e.g., farmers' planting decision on agricultural lands) which can be complemented by process-based models such as agent-based models.

Despite the utility and widespread use of statistical methods in LUCC modelling, there is a lack of review or assessment of the performance of more than two different statistical methods (or different combinations) with the same dataset at the same location in the field of LUCC modelling. As a step toward filling this gap, four statistical approaches (Markov chain, logistic regression, generalized additive models and survival analysis), which were selected based on popularity, difficulty of implementation and ability to account for different scenarios (e.g., availability of data and structure of data), were conducted to model LUCC in the Region of Waterloo. Their performance of predicting LUCC was quantified in terms of prediction accuracy. The study of modelling LUCC with these methods can answer the questions: what is the overall accuracy of different types of statistical methods in representing LUCC and what is the distribution of accuracies for different types of statistical methods by LU type?

2.2 Methods

2.2.1 Study Area

The presented research is situated in the Region of Waterloo, which compresses 1369 km^2 and is located in southern Ontario, Canada (Figure 1). The region is composed of three cities (Kitchener, Waterloo and Cambridge) and four townships (Wellesley, Woolwich, Wilmot and North Dumfries), with which exists a mixture of residential, commercial, agricultural, and other LU types. The Region of Waterloo has been experiencing above average population growth and subsequent LUCC due in part to the employment opportunities in high-tech research and development, low cost housing relative to Toronto, its location along the highway 401 (the busiest highway in North America; Maier, 2007), and close proximity to the City of Toronto (the 5th largest North American financial centre; Yeandle, 2017).



(a)

Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, © OpenStreetMap contributors, and the GIS User Community

Data Source:
Government of Canada - Open Government Portal
Smith, A. K. (2017)

Coordinate System: NAD 1983 UTM Zone 17N
Projection: Transverse Mercator
Datum: North American 1983

(b)

Figure 1: (a) Land use map of the Region of Waterloo in 2010. (b) The location of the Region of Waterloo associated with Toronto and Higyway 401. [Notes: low-density residential (LDR), medium-density residential (MDR), high-density residential (HDR), commercial (COM), industrial (IND), institution (INS), transportation (TRA), protected area and recreation (REC), agriculture (AGR), water (WAT), under-development (UND); white areas within the boundary of the Region of Waterloo but beyond the colored LUs in (a) are areas without available data of ownership property parcels.]

Waterloo Region had the 7th largest population in Ontario and the 13th largest population in Canada according to the 2011 Census data (Region of Waterloo, 2011). The population of the region increased from 478,121 to 507,096 from 2006 to 2011, which was a 6.06 percent increase in population that exceeded the 5.7 percent provincial and the 5.9 percent national population growth rates recorded in 2011. Moreover, the total urban areas of the three cities together ranked as the 10th largest in Canada in 2011. The fast growth rate for the region has resulted in urban sprawl (Figure 2), which influences the types of LU and LC transitions occurring in the region.

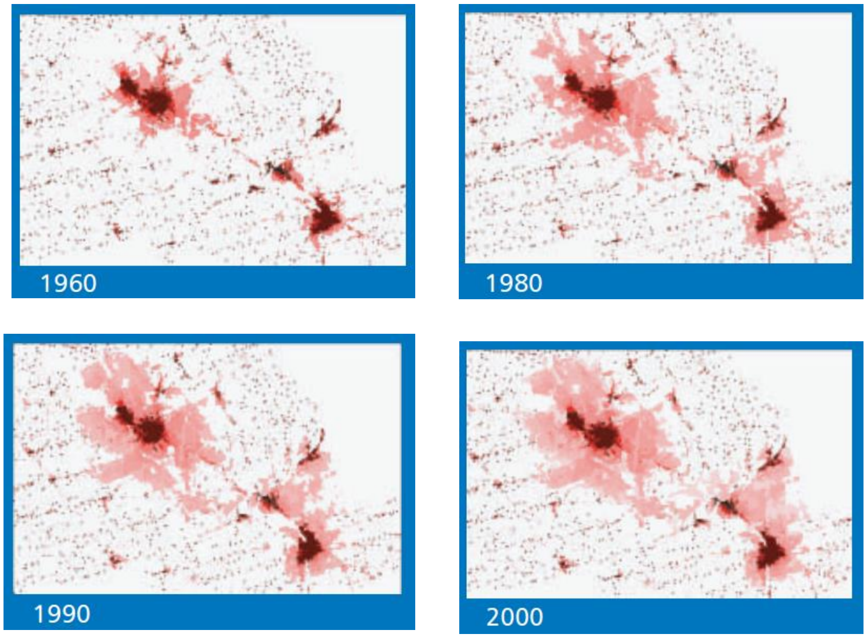


Figure 2: Suburban development in the Region of Waterloo, 1960 – 2000. Reprinted from Planning Our Future: Regional Growth Management Strategy (2003) (p. 2).

According to the 2010 and 2015 LU data used in this study, which were classified by Smith (2017), 10,606 parcels have experienced LU changes (Figure 3), in which approximately 67 percent of the parcels have converted to medium-density residential LU (51 percent) and commercial LU (16 percent). Other noticeable LU changes are LU change to transportation LU (9 percent), high-density residential LU (8 percent) and under-development LU (7 percent). The proportions of other LU changes are all less than 4 percent.

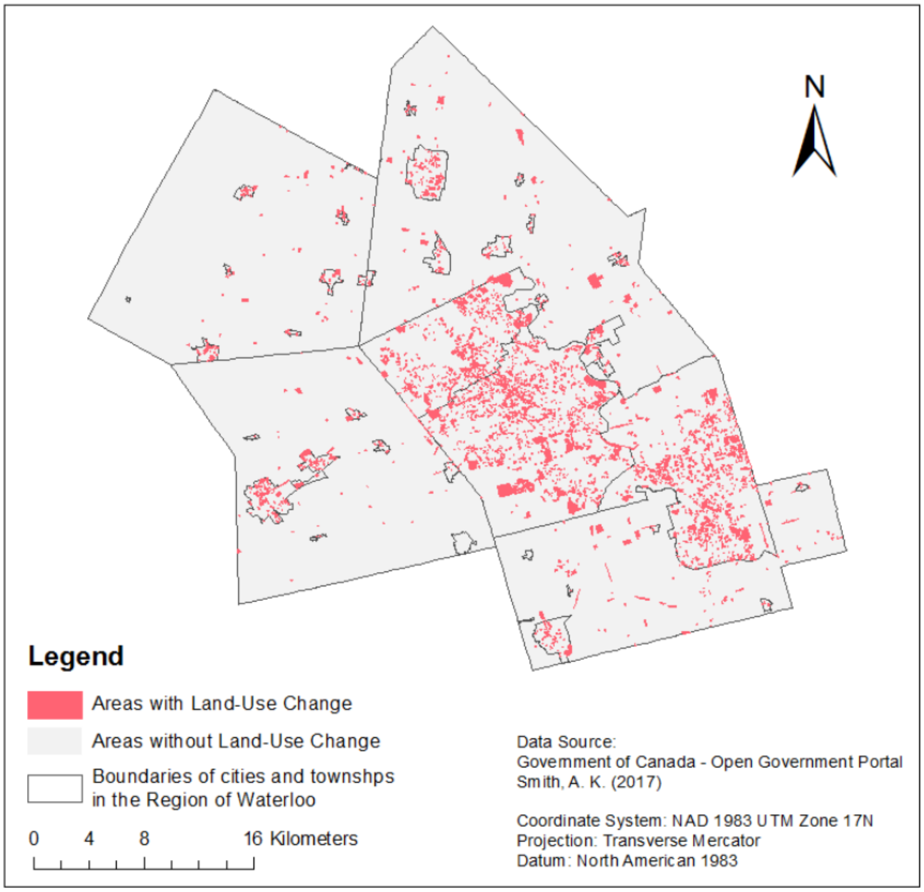

N
Legend
Areas with Land-Use Change
Areas without Land-Use Change
Boundaries of cities and townshps in the Region of Waterloo
0 4 8 16 Kilometers
Data Source:
Government of Canada - Open Government Portal
Smith, A. K. (2017)
Coordinate System: NAD 1983 UTM Zone 17N
Projection: Transverse Mercator
Datum: North American 1983

Figure 3: Land use change in the Region of Waterloo from 2010 to 2015.

Moreover, within the study area, approximately 0.7 percent, 0.8 percent and 6 percent of protected areas and recreational parcels have been converted to residential LUs, commercial and industrial LUs, and transportation LU from 2010 to 2015, respectively. Agricultural parcels have lost about 1 percent due to expansion of residential, commercial, industrial, transportation and under-development LUs from 2010 to 2015. In addition, approximately 66 percent, 9 percent and 4.7 percent of parcels classified as under-development LU in 2010 completed their transition to residential LUs, commercial LU and transportation LU by 2015, respectively. These statistics were obtained by comparing 2010 LU data and 2015 LU data that were used to construct statistical models in this study in the same region.

In addition to the statistics mentioned above, historical statistics about agricultural LU changes show that the total number of farms decreased from 1,444 to 1,398 (4 percent) from 2006 to 2011 (Region of Waterloo, 2011). The net loss of agricultural land from 2006 to 2011 was around 5,000 acres. Despite the decline in the total number of farms and total acres of farming land, the dominant LU type across the region remains in agricultural land, which accounts for 65 percent of total land area. Another characteristic of agricultural LU change between census years 2006 and 2011 is that the average size of farms had increased. The study area has not only experienced farm loss, expansion of residential LU but also a set of LU changes such as conversion from low-density residential LU to high-density residential LU and conversion from medium-density residential LU to commercial LU according to the data used in this study.

2.2.2 Data

LU raster data were generated for the years 2006, 2010 and 2015 for the Region of Waterloo by a member of the Modelling and Spatial Analysis Lab at the University of Waterloo (Smith, 2017). The original LU raster data consist of ten LUs and one LC: low-density residential (LDR), medium-density residential (MDR), high-density residential (HDR), commercial (COM), industrial (IND), institution (INS), transportation (TRA), protected area and recreation (REC), agriculture (AGR), water (WAT), under-development (UND), in which WAT is the only LC in the dataset. LC raster data were also generated for the same years for the Region of Waterloo by Smith (2017) and were used as LU change predictors. Even though the LU raster data contain one LC type, the name, LU data, is specifically referred to the data that contain ten LUs and one LC and all elements in the LU data are considered LUs including WAT in the following context in order to distinguish from the real LC data. In addition to the LU data and LC data, 2010 parcel data (Ownership property parcels) for the Region of Waterloo, which contain boundary of parcels, were acquired from Teranet. LU and LC data were extracted to parcel data based on parcel units. The LU of a parcel was determined by the majority of LU within the parcel boundary.

LU change drivers were selected based on LU literatures and were categorized as geometric variables (i.e., parcel perimeter and area), site variables (i.e., slope and DEM), demographic variables (i.e., population density), distance variables (e.g., distance from a parcel to the nearest highway ramp) and spatial variables (Table 1). Geometric variables were calculated using ArcGIS. Demographic variables were acquired from the 2011 Census data at the Dissemination Areas (DAs) level (Canadian Census Analyzer, 2011). The Canadian Census is taken every five years and 2011 provided the closest year to 2010. Moreover, spatial variables were created to account for some spatial autocorrelation among parcels (e.g., the proportion of commercial LU parcels around a target parcel). All these variables were treated as potential LU change drivers and were evaluated during model construction.

In addition to the LU change predictors mentioned above, zoning is an important factor that can influence the types and locations of LUs in a municipality. Zoning policy regulates LUs in defined zones (e.g., residential and commercial). In other words, a specific LU in a zone can be restricted by zoning policies; thus, a change of LU can be prohibited to occur in a particular zone (Maser et al., 1977). Zoning may vary across different municipalities and can change over time at a given location. Therefore, rules that regulate the change of a parcel's LU may vary spatially and temporally. An exception of LU may occur in addition to the LUs permitted by the zoning (i.e., zoning variance; Cohen, 1994), which increases the difficulty of making predictions of LU change. Moreover, zoning regulations and LU plans are made at local levels (i.e., municipalities) and are not collated across broader spatial scales, which limits their use in models that can be applied across large spatial extents. Thus, when zoning information of an area is used as drivers in a statistical model to predict future LU changes, the model becomes non-transferable, which means that the model is restricted to predict local LU changes and contribute to local LU planning. The goal of this thesis is to construct statistical models that can be applied widely across the world and can provide a general representation of LU change pattern. Therefore, the consideration of zoning effect to LU changes was excluded from this study.

In general, LU modelling was conducted through analyzing the changes between 2010 and 2015 LU types and all potential drivers. The detailed steps of variable creation and data processing are documented in Appendix A. The full list of variables used in model building can be found in Appendix C. In addition to variable creation, variable transformation was performed to unify the units of distance variables from meters to kilometers, units of population density variables from person/m^2 to person/km^2, units of parcel geometry variables from m^2 to km^2, and units of elevation from m^2 to km^2. The measurement transformations ensure the gradients of values of all predictors are on a similar scale.

Table 1: Name and source of original data.

Name	Source
LU data (2006, 2010, 2015) 80cm resolution	Smith, 2017
LC data (2006, 2010) 80cm resolution	Smith, 2017
Ownership property parcels in 2010	Teranet Inc.
Census Dissemination Areas Divisions	Statistics Canada,2011
Population	Canadian Census Analyzer, 2011
Ontario road network	Ontario Ministry of Transportation
Highway Access Point	Ontario Ministry of Transportation
Rivers	Ontario Ministry of Natural Resources
Water Bodies	Ontario Ministry of Natural Resources
Wooded Areas	Ontario Ministry of Natural Resources
Digital Elevation Model (DEM)	Ontario Ministry of Natural Resources
Slope10m resolution	Ontario Ministry of Natural Resources

Note: All spatial datasets were projected to the NAD_1983_UTM_Zone_17N throughout the study.

2.2.3 Analysis

For LR, GAM and SA, parcels with the same LU types at 2015 were grouped together as one full dataset for each type of LU change. This indicates that parcels in a full datasets may have different LUs at 2010 but surely have the same LU at 2015. When 2010 LU type and 2015 LU type of a parcel are different, the parcel is considered experienced a LU change. Otherwise, it is considered unchanged. For instance, the full dataset for LDR LU change contains all parcels that have changed from anything to LDR from 2010 to 2015 and all parcels that have remained as LDR during the study period. Eleven full datasets were created corresponding to the eleven defined LUs in 2015. Binary response variables, *Y*, were created to indicate the status of a LU change for each parcel in each full dataset (i.e., 1=changed, 0=unchanged) and were attached to full datasets.

Based on full datasets, two more datasets were created for each of the eleven types of LU change. One is named the full and balanced (FB) dataset and another one is named the reduced and balanced (RB) dataset. In general, a balanced dataset refers to a dataset that contains the same amount of observations for all existing levels of a categorical independent variable (Batista et al., 2004). In this study, the word 'balanced' was used to describe the structure of dependent variable. The dependent variable in each of the FB datasets and RB datasets was constructed to contain an equal number of changed and unchanged parcels.

The purpose of creating balanced Y is to use 0.5 as the threshold for grouping probabilities into two categories in the prediction phase. For LR, GAM, and SA, probabilities that are greater than or equal to 0.5 were classified as changed and otherwise classified as unchanged. While LR and GAM produce probabilities, SA produces a hazard function that can be converted to a survival function that determines the probability of an object surviving beyond a given point in time. Subtracting the survival probability from a value of one produces a death probability that was used in the same manner as the probabilities for LR and GAM.

In FB datasets, the number of $Y = 1$ is equal to the total number of changed parcels. In RB datasets, the number of $Y = 1$ is fixed at 500; therefore, the total number of parcels is fixed to 1,000. The size of a RB dataset (i.e., 1000) is chosen to 1) reduce the potential of over-fitting, yet maintain enough data to construct a model with many predictors and 2) access the minimum number of samples required to yield similar or the same results as the FB. When the number of Y that equals to 1 is less than 500 for one type of LU change, the FB dataset and the RB dataset for this LU change are the same. The size of FB datasets ranges from 10 to 10,816 points (Table 2). In our data, no parcel had changed to agricultural LU from 2010 to 2015. Moreover, only five and thirty-six parcels had changed from some LUs to institution and water, respectively. Thus, there are insufficient data for LR, GAM and SA to model agricultural, institutional, and water LU changes with many predictors.

The format of all datasets (Figure 4) includes the binary response variable Y that indicates the status of LU change for each parcel and the LU change predictors X_i*'s*, where i is the index of predictors and ranges from 1 to *p* where *p*>=1.

Variable Parcel	Y	X_1	X_2	……	X_p
1	Y_1	X_{11}	X_{12}	……	X_{1p}
2	Y_2	X_{21}	X_{22}	……	X_{2p}
⋮	⋮	⋮	⋮	⋮	⋮
n	Y_n	X_{n1}	X_{n2}	……	X_{np}

Figure 4: Format of datasets.

Table 2: Sample sizes of full balanced (FB) and reduced balance (RB) datasets by land-use type.

LU Change (To)	FB	RB
LDR	870	870
MDR	10816	1000
HDR	1698	1000
COM	3320	1000
IND	530	530
INS	10	10
TRA	1892	1000
REC	520	520
AGR	500*	500*
WAT	72	72
UND	1484	1000
* The samples of agricultural parcels are only for MC.		

The traditional hold-out method (e.g., Kohavi, 1995) was used to partition FB datasets and RB datasets into training and test datasets. A 10-fold cross validation (CV; e.g., Kohavi, 1995) was applied to the training data to produce an averaged CV accuracy for each LU change type and an overall CV accuracy for each method and dataset (Figure 5). Spatial CV (Brenning, 2012) with 10-fold was performed and compared with the conventional CV to reveal the effect of spatial autocorrelation in spatial data to MC, LR, GAM and SA in this study. Spatial autocorrelation is a common problem in LU modelling, which causes violation of independence among observations in many statistical techniques. With conventional CV, parcels that have been randomly selected to form training data and test data may be neighbors and can cause over-fitting of statistical models due to spatial autocorrelation. Spatial CV alleviates over-fitting by partitioning dataset using k-means clustering (Hartigan, 1975) based on parcels' spatial coordinates. The notations of CCV and SCV are used to denote conventional CV and spatial CV in the rest of this thesis. Predictor coefficients were estimated, and only significant predictors with p-values less or equal to 0.1, which was chosen with an intention to expand the range of significant predictors, were retained in final LR, GAM and SA models. Final models were further adjusted since originally identified significant predictors can become insignificant when fitting a new model The FB and RB test data were then used to calculate the classification accuracy of the final LR, GAM, and SA models.

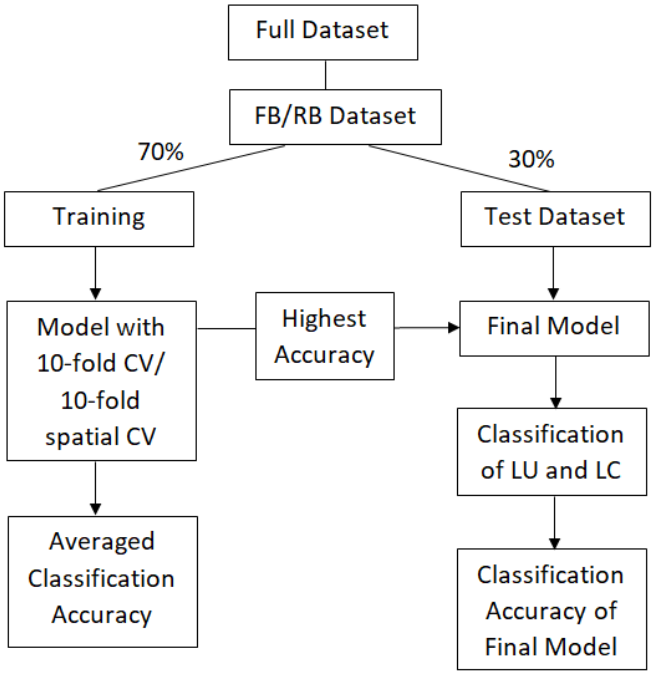

Full Dataset
FB/RB Dataset
70%
30%
Training
Test Dataset
Model with 10-fold CV/ 10-fold spatial CV
Highest Accuracy
Final Model
Classification of LU and LC
Averaged Classification Accuracy
Classification Accuracy of Final Model

Figure 5: Methodology for developing full balanced (FB) and reduced balance (RB) training and testing data, model selection, and land use (LU) and land cover (LC) classification accuracy assessment.

For MC, the transition probability matrix, which contains the probabilities of changing from any LU to another, requires data that contain all types of LU changes. Therefore, the ten LU datasets (Table 2) are combined with 500 agriculture parcels to form the MC FB and RB dataset, of which 70 percent was randomly selected for training models with 10-fold CCV and 10-fold SCV and the rest was used for testing final models.

The averaged CV accuracy (i.e., the average of ten accuracy values produced by fitting models with ten folds of data) for predicting LU change is the criterion used to compare the performance of methods since 1) the most direct future implication of these methods is to make prediction of future LU changes and 2) result from a first-order DTMC cannot be evaluated by other criteria (e.g., R^2; Taylor, 1990). The overall CV accuracy of a method was calculated by averaging averaged CV accuracies for defined LU changes. Thus, methods that produce the highest overall CV accuracies and the combination of methods that produces the highest averaged final model accuracies by LU change type were determined. Furthermore, the distribution of accuracy values for different types of statistical methods by LU change type was determined.

2.2.4 Implementation of Statistical Approaches

Each of the following subsections consists of a short description of the implementation of a statistical method used in LUCC modelling. Among many available software, R statistical software (R Core Team, 2017) was used to implement these methods due to its open source format and widespread document and use. For MC, LR, GAM and SA, the partition of subsets of data for 10-fold CCV and 10-fold SCV was done using *creatFolds* function from the *caret* package (Kuhn, 2018) and *partition.kmeans* function from the *sperrorest* package (Brenning, 2012), respectively. The fitting of 10-fold CCV models and 10-fold SCV models were done manually for MC, LR and SA and wad one using *train* function from the *caret* package (Kuhn, 2018) for GAM. Other fitting procedures are described in the following subsections for the implementation of MC, LR, GAM and SA.

2.2.4.1 Markov Chain

Markov Chain (MC) is a statistical method that incorporates stochasticity in the process of changes between states. Discrete time MC (DTMC) requires a countable set of states and events that are mutually exclusive and collectively exhaustive (Stokey and Zeckhauser, 1978). Moreover, uniform length is required between any two time points. In this research, a DTMC method was used to model LU changes occurred between 2010 and 2015 LU at a parcel level, which is considered a first-order DTMC that the status at a given time only depends on the status occurred at the nearest past state.

Using the current LU data (2015) and the past LU data (2010), a transition probability matrix that contains only one-step transition probabilities was calculated. The probability matrix was obtained by observing the frequencies of LU changes occurred between two years. To test the performance of MC, a random value between 0 and 1 was assigned to each parcel in the test set to represent the LU transition probability from 2010 to 2015. A roulette-wheel-selection approach (Lipowski and Lipowska, 2012) was used to determine the LU type a parcel will be converted to according to its starting state LU type and the transition probability matrix. The creation of the transition probability matrix was done in R using *prop.table* function which is applied to a contingency table of LU classes. The roulette-wheel-selection approach was created to suite this specific study in R.

2.2.4.2 Logistic Regression

Logistic regression (LR) is a type of statistical method used to model categorical variables and is a member of generalized linear models. Its response variable follows a binomial distribution and connects to the linear combination of all covariates though a logit link function. LR can be used to simulate and predict categorical LUCC outcomes (Trexler and Travis, 1993). Multiple LR, which contains more than one independent variable, has often been used to model LC change (Muller and Zeller, 2002). Since the idea of modelling LC change with multiple LR is the same as modelling LU change, multiple LR was used in the presented research among all other models in the family of LR (e.g., ordinal LR and multinomial LR).

When predicting parcels with unknown LU types, either 0 or 1 was assigned to each parcel based on the estimated probability and a threshold. If the LU change status at a location was determined to be 1, it means the model predicted the parcel converts from one LU type to a target LU type; otherwise, it means no change occurred..

2.2.4.3 Generalized Additive Model

Generalized additive models (GAMs) extend generalized linear models (GLMs) by using a series of smoothing splines to express the non-linear relationship between the expected mean of responses and a set of predictors (Hastie and Tibshirani, 1990). Therefore, GAM has been implemented in LU science in addition to GLM (e.g., Brown, 1994). The advantage of GAM over GLM is that it has the ability to represent non-linear relationships, which ensures that more realistic situations can be represented. The presented research used the additive logistic regression (ALR) model among all other GAMs.

ALR was chosen to conduct the modelling because responses variables (i.e., statuses of LU changes) are binary. The fitting of CCV and SCV models was conducted using the *train* function that adopts the algorithm of GAM from the *mgcv* package (Wood, 2003, 2004, 2011 and 2017; Wood et al., 2016). The *train* function with *gam* from *mgcv* package fits predictors with default smoothing functions, the thin plate regression splines that are considered robust smoothers regardless of the dimension of basis functions, which cannot be modified. Similar to LR, a value of 0 or 1 was assigned to each parcel to determine the status of predicted LU change.

2.2.4.4 Survival Analysis

SA analysis (SA) is used mostly in health and clinical studies to predict the mortality rate or recovery rate (e.g., recovery rate from injury or diseases). SA can handle both time-varying and time invariant variables and can take into account incomplete data. SA models use both the duration of each observation in the experiment and an indicator variable showing the occurrence of the event of interest as response variables, and all other potential factors that influence the occurrence of the event as covariates to calculate the success/failure ratio of the event at the time.

In the context of LUCC modelling, the event of interest is the change from one LU to another between two time points and the failure time of a parcel would be the time that a LU change occurs. The LU data were derived from remotely sensed images for years 2010 and 2015. In this sense, each parcel was observed twice in the five year time period. However, multiple measurements of LU for each parcel in the study period are required to determine the failure time. Therefore, a *time* variable was created by randomly generating integers in the range of 1 to 6 to represent the failure time of each parcel. The year is 2010 when *time=1* and is 2015 when *time=6*. Other values of *time* represent years between 2010 and 2015 in an ascending order. The reason for generating discrete times instead of continuous times is to keep consistent for the unit of time (year) since the LU raster data created by Smith (2017) were generated from SWOOP data that are considered to represent the LU for a year.

The *coxph* function in the *survival* package (Therneau, 2015) in R was used to construct a Cox proportional hazards (PH) model (Cox, 1972), which is a type of SA technique. The Cox PH model was selected because the PH assumption assumes constant effects of covariates on hazards over time, which is consistent with the assumptions of predictor effects in other models (i.e., LR and GAM). Since the *coxph* function in R can only handle right-censored data, all time measurements are considered precise and only right-censored data (i.e., parcels have not experienced LUCC at the end of study) exist in this study.

2.3 Results

2.3.1 Conventional Cross Validation and Final Models

Results from the 10-fold CCV by method demonstrated that the overall CCV accuracy was highest for GAM, followed by LR, SA, and MC for both the FB and RB training datasets (Table 3). GAM achieved 85.17 percent and 82.39 percent overall accuracies for FB and RB training datasets, respectively. For FB training datasets, the overall accuracy of GAM is 4.2 percent, 4.26 percent and 42.53 percent higher than the overall accuracies of LR, SA and MC, respectively. For RB training datasets, the overall accuracy of GAM is 2.15 percent, 2.8 percent and 35.03 percent higher than the overall accuracies of LR, SA and MC, respectively. This infers that sample size positively influences the differences between overall accuracies of GAM and any one of MC, LR and SA.

Excluding MC, the increase in overall CCV accuracy by method and averaged CCV accuracy by LU change due to use of the FB dataset over the RB dataset was at most 2.78 percent (i.e., GAM FB and GAM RB) and 6.98 percent (i.e., GAM FB and GAM RB for HDR LU change) among LR, GAM, and SA, respectively, which implies that the size of sample dataset is not a critical factor that would influence the overall accuracy but surely has more impact on averaged CCV accuracy for LR, GAM and SA. In contrast, MC RB performed better than MC FB by approximately five percent in terms of overall accuracy. Moreover, averaged MC RB accuracy performed better than averaged MC FB accuracy for all LU changes except MDR and WAT, which may due to the mechanism of MC observing frequencies of changes based on given sample datasets. For example, the rise of RB accuracy over FB accuracy is the highest for INS LU change, which is approximately 52 percent. This large difference was caused by a relatively large decrease in sample sizes of other LU changes while keeping all parcels with INS LU change (i.e., five) in MC RB dataset.

Given the close performance of LR, GAM, and SA, it is worth noting that the run time for computing 10-fold MC, LR, and SA results were less than 20 seconds (Appendix D-1).

However, performing the 10-fold CV for GAM took over eighty-five minutes for both FB and RB datasets due to the iteration in back-fitting of smoothing functions.

GAM outperformed the other modelling approaches in overall accuracy but it did not achieve the highest average accuracy for LU changes of LDR and IND. Among the eight LU changes modeled by LR, GAM and SA, GAM performed best for six LU changes with FB datasets (MDR, HDR, COM, TRA, REC, UND) and five LU change with RB datasets (HDR, COM, TRA, REC, UND), and LR (IND). SA performed the best one type of LU change (LDR), LR performed the best for LU changes of LDR and MDR (RB) (Table 3). The MC approach performed best for the three LU types (INS, AGR, WAT) that were not modeled by the other methods and performed surprisingly well for AGR due to the observed low frequency of AGR LU change that reflects the reality in the data.

Differences in averaged accuracies between methods were much greater than the differences between overall accuracies for some LU changes (e.g., GAM FB and MC FB for UND; GAM FB and LR FB for COM) and much lower for some other LU changes (e.g., LR FB and SA FB for TRA; GAM FB and LR FB for IND). Excluding MC, the differences in averaged accuracies across the other three models were lowest for FB TRA (0.08 percent) and greatest for FB COM (13.16 percent). These results suggest that, model choice is critical to gaining an accurate representation of pattern for some LU change types. In addition, our results suggest that greatly increasing the sample size from RB to FB for LU changes of MDR, HDR, COM and UND has little effect on the within LU accuracy for LR, whereby the difference is less than 1.18 percent for all LU types except transportation (3.53 percent). The difference is around 7 percent or less for all LU types within GAM, and the difference is 3.54 percent or less within SA (Appendix D-2).

The final LR, GAM, and SA models that derived from 10-fold CCV were models that produced the highest accuracy in the 10-fold CCV process for specific types of LU changes and showed the significance of some predictors. Each final model was evaluated against partitioned test datasets. Since MC does not have a specific form and does not contain any predictor variables, it was excluded from this comparison.

Table 3: The averaged and overall 10-fold CCV accuracy for FB and RB training datasets.

Model / LU Change (To)	Averaged accuracy (%)							
	MC		LR		GAM		SA	
	FB	RB	FB	RB	FB	RB	FB	RB
LDR	36.67	46.91	68.62	68.62	65.85	65.85	**70.95**	**70.95**
MDR	66.41	28.02	89.62	88.44	**93.98**	87.00	89.56	**89.18**
HDR	43.29	43.36	71.10	69.88	**79.46**	**77.71**	71.64	70.26
COM	23.90	24.56	80.89	81.80	**91.61**	**86.53**	78.45	76.25
IND	48.56	53.99	**90.05**	**90.05**	89.95	89.95	88.84	88.84
INS	**10.00**	**62.50**	n/a	n/a	n/a	n/a	n/a	n/a
TRA	56.43	57.38	79.68	76.15	**83.08**	**79.72**	79.76	76.71
REC	24.02	36.70	87.24	87.24	**89.80**	**89.80**	85.52	85.52
AGR	**93.45**	**96.05**	n/a	n/a	n/a	n/a	n/a	n/a
WAT	**57.45**	**47.45**	n/a	n/a	n/a	n/a	n/a	n/a
UND	8.88	24.08	82.93	82.13	**87.66**	**82.56**	82.54	79.00
Overall	42.64	47.36	80.97	80.24	**85.17**	**82.39**	80.91	79.59

Note: FB means full and balanced and RB means reduced and balanced. For the definition and difference between FB and RB, please refer to section 2.4 Analysis. Bold values indicate highest accuracy by land-use and land-cover type.

The ranking of overall final models' accuracies by method with RB test datasets is the same as the ranking of the overall CCV accuracies by method with either FB or RB training datasets (i.e., GAM > LR > SA). The overall final models' accuracies with FB test datasets are again slightly higher than the overall final models' accuracies with RB test datasets. For FB test datasets, the overall accuracy of GAM is 4.37 percent and 4.11 percent higher than the overall accuracies of LR and SA, respectively. For RB test datasets, the overall accuracy of GAM is 2.41 percent and 3.07 percent higher than the overall accuracies of LR and SA, respectively. This implies that the advantage of GAM predicting overall LU changes with FB dataset over RB dataset is increased with final models.

While the best GAM outperformed the best LR and SA models in overall accuracy, variation was observed among LU change types. Of the eight LU types tested, final GAM performed best for seven LU changes (LDR, MDR (FB), HDR, COM, IND, TRA, UND), final SA performed best for two LU changes (MDR (RB) and IND), and final LR performed best for one LU change (REC, Table 4). The differences in accuracies by LU change type between methods were much greater than the difference in overall accuracies between methods for some LU change types (e.g., GAM FB and LR FB for COM; GAM RB and LR RB for MDR) and smaller than for some others (GAM FB and SA FB for REC; GAM RB and SA RB for LDR).

Table 4: The individual and overall accuracy for final models derived from 10-fold CCV with FB and RB test datasets.

Method / LU Change (To)	Accuracy (%)					
	LR		GAM		SA	
	FB	RB	FB	RB	FB	RB
LDR	67.05	67.05	**70.11**	**70.11**	69.73	69.73
MDR	89.83	87.63	**93.78**	89.30	90.01	**89.90**
HDR	69.22	67.33	**78.24**	**76.67**	69.35	68.67
COM	78.82	83.67	**92.15**	**90.17**	80.32	82.33
IND	91.19	91.19	**91.82**	**91.82**	**91.82**	**91.82**
INS	n/a	n/a	n/a	n/a	n/a	n/a
TRA	79.05	79.67	**83.98**	**80.00**	79.05	76.59
REC	**90.32**	**90.32**	88.39	88.39	88.24	88.24
AGR	n/a	n/a	n/a	n/a	n/a	n/a
WAT	n/a	n/a	n/a	n/a	n/a	n/a
UND	83.86	**82.33**	**85.87**	82.00	82.95	76.67
Overall	81.17	81.15	**85.54**	**83.56**	81.43	80.49

2.3.2 Spatial Cross Validation and Final Models

The ranking of overall SCV accuracies of MC, LR, GAM and SA is similar to the ranking of overall CCV accuracies of these four methods. The overall accuracies of 10-fold SCV are highest for GAM, followed by SA, LR and MC for FB training dataset, and by LR, SA and MC for RB training datasets (Table 5). GAM achieved 79.03 percent and 79.58 percent overall SCV accuracies for FB and RB training datasets, respectively. The overall SCV accuracy of GAM is 2.79 percent, 3.13 percent and 32.54 percent higher than the overall SCV accuracies of SA, LR and MC for FB training datasets, and is 0.72 percent, 0.74 percent and 30.67 percent higher than the overall accuracies SCV of LR, SA and MC for RB training datasets. This infers that sample size positively influences the differences between overall SCV accuracies of GAM and any one of MC, LR and SA.

The differences in overall SCV accuracies within methods are within 2 percent (0.01 percent for MC, 0.55 percent for LR, 1.86 percent for GAM, and 0.19 percent for SA). Therefore, sample sizes of designed 10-fold SCV datasets for different types of LU changes are not a critical factor that influenced the difference in overall SCV accuracies within methods. Moreover, the differences in overall SCV accuracies within methods are all lower than the differences in overall CCV accuracies within methods (i.e., 4.72 percent for MC, 0.73 percent for LR, 2.78 percent for GAM, and 1.32 percent for SA), which infers that 10-fold SCV reduced these differences.

In general, GAM was the best for modelling six LU changes with FB training dataset (MDR, HDR, COM, IND) and three LU changes with RB training dataset (HDR, IND, TRA); LR performed best for LU changes of COM and UND with RB training dataset, and REC with training dataset; SA performed best for LU changes of LDR regardless of sample sizes, and MDR with RB training dataset. MC was the best for modelling LU changes of INS, AGR and WAT. Differences in averaged accuracies between methods were much greater than the differences between overall accuracies for some LU changes (e.g., GAM FB and LR FB for COM; GAM RB and SA RB for HDR) and much lower for some other LU changes (e.g., GAM FB and SA FB for REC; GAM RB and LR RB for UND). This agrees with the conclusion made in Section 2.3.1 that model choice is critical to gaining an accurate representation of pattern for some LU change types.

For final models derived from SCV, GAM performed the best overall as well as for modelling LU changes of MDR, HDR, COM, TRA and UND (FB) (Table 6). LR produced the second highest overall final model accuracies and best models LU changes of LDR, IND, REC, and UND (RB). SA produced the same accuracy value, which is the second highest, for UND (RB) as LR did. The overall final models' accuracies with FB test datasets are again slightly higher than the overall final models' accuracies with RB test datasets in most cases. For FB test datasets, the overall accuracy of GAM is 3.14 percent and 4.41 percent higher than the overall accuracies of LR and SA, respectively. For RB test datasets, the overall accuracy of GAM is 2.18 percent and 3.57 percent higher than the overall accuracies of LR and SA, respectively. The differences between FB and RB overall accuracies within methods are 0.64 percent, 1.4 percent, and 0.56 percent for LR, GAM and SA, respectively. The largest difference in averaged 10-fold SCV accuracies by LU change type occurred for COM, which is 12 percent. These results confirm the finding made by analyzing results from CCV and CCV-final models that model choice is a more critical factor than sample size is for making predictions of LU changes.

Table 5: The averaged and overall 10-fold SCV accuracy for FB and RB training datasets.

Model / LU Change(To)	Averaged accuracy (%)							
	MC		LR		GAM		SA	
	FB	RB	FB	RB	FB	RB	FB	RB
LDR	35.37	50.76	67.43	67.43	66.06	66.06	**69.15**	**69.15**
MDR	68.35	28.20	89.38	89.67	**92.87**	86.78	89.25	**90.35**
HDR	41.09	40.80	66.72	69.50	**73.92**	**78.19**	68.86	72.03
COM	25.41	28.36	79.82	**78.92**	**88.29**	78.59	78.99	77.70
IND	45.56	45.57	89.37	89.37	**90.52**	**90.52**	90.03	90.03
INS	**80.00**	**80.00**	n/a	n/a	n/a	n/a	n/a	n/a
TRA	57.82	60.32	76.77	76.91	**82.52**	**78.80**	79.40	76.54
REC	27.78	45.57	**85.68**	**85.68**	84.51	84.51	84.32	84.32
AGR	**90.71**	**97.36**	n/a	n/a	n/a	n/a	n/a	n/a
WAT	**63.78**	**41.65**	n/a	n/a	n/a	n/a	n/a	n/a
UND	9.90	27.29	77.09	**78.98**	**78.58**	78.92	74.98	76.34
Overall	49.62	49.63	79.03	79.58	**82.16**	**80.30**	79.37	79.56

Note: FB means full and balanced and RB means reduced and balanced. For the definition and difference between FB and RB, please refer to section 2.4 Analysis. Bold values indicate highest accuracy by land-use and land-cover type.

Table 6: The individual and overall accuracy for final models derived from 10-fold SCV with FB and RB test datasets.

Method	Accuracy (%)					
	spLR		spGAM		spSA	
LU Change (To)	FB	RB	FB	RB	FB	RB
LDR	**70.88**	**70.88**	69.35	69.35	67.82	67.82
MDR	88.20	89.67	**93.71**	**92.98**	89.89	90.00
HDR	69.61	67.33	**77.65**	**79.33**	70.00	67.33
COM	81.33	80.27	**91.37**	**82.94**	80.62	79.67
IND	**97.48**	**97.48**	94.34	94.34	91.19	91.19
INS	n/a	n/a	n/a	n/a	n/a	n/a
TRA	79.58	77.00	**83.98**	**83.33**	78.70	77.00
REC	**89.30**	**89.30**	87.74	87.74	87.74	87.74
AGR	n/a	n/a	n/a	n/a	n/a	n/a
WAT	n/a	n/a	n/a	n/a	n/a	n/a
UND	82.96	**82.33**	**84.75**	81.67	81.61	**82.33**
Overall	82.42	81.78	**85.36**	**83.96**	80.95	80.39

Moreover, by comparing results from Table 3 and Table 4, it can be found that the overall SCV accuracies were reduced about 1.94 percent and 0.66 percent for LR with FB and RB datasets, 3.01 percent and 2.09 percent for GAM with FB and RB datasets, and 1.54 percent for SA with FB dataset, which implies that SCV only contributed to alleviate over-fitting by a small amount in terms of the reduction in overall SCV accuracies. In terms of the number of averaged SCV accuracies by LU change type being reduced, the problem of over-fitting was most serious for LR, followed by GAM, SA and MC. On the other hand, GAM was the method that suffered the most from over-fitting in terms of the averaged reduction in averaged SCV accuracies (4.15 percent), in which HDR (FB), COM (RB) and UND (FB) contributed 5.54 percent, 7.94 percent and 9.08 percent, respectively. Therefore, it can be concluded that spatial autocorrelation existed in our data did not cause severe over-fitting of statistical models in terms of overall accuracies but had more impact on individual LU changes and single methods.

2.3.3 The Combination of Statistical Methods

Given the above best models and their different performance by LU change type, a LU change model should take advantage of the methods that perform best for specific LU changes. A theoretical LU change model was constructed by obtaining methods that produced the highest final model accuracies by LU change type (Table 7). Since SCV-final models generally alleviated over-fitting caused by spatial autocorrelation, final models derived from SCV were selected to form the theoretical method. Moreover, MC derived from CCV was selected to model a LU change if the accuracy value was higher than the corresponding accuracy value regardless of the amount since MC only accounted for frequencies of LU changes. In the following context, the extensions "-CCV" and "-SCV" followed by a method's name indicate the type of CV a final model of the method has derived from.

Table 7: The combination of final models that produces the highest accuracy by LU type with FB and RB test datasets.

LU Change (To)	FB		RB	
	Method	Accuracy	Method	Accuracy
LDR	LR-SCV	70.88	LR-SCV	70.88
MDR	GAM-SCV	93.71	GAM-SCV	92.98
HDR	GAM-SCV	77.65	GAM-SCV	79.33
COM	GAM-SCV	91.37	GAM-SCV	82.94
IND	LR-SCV	97.48	LR-SCV	97.48
INS	MC-CCV/MC-SCV	25	MC-CCV/MC-SCV	40
TRA	GAM-CCV/GAM-SCV	83.98	GAM-SCV	83.33
REC	LR-SCV	89.30	GAM-SCV	89.30
AGR	MC-SCV	96.58	MC-SCV	96.03
WAT	MC-CCV	55	MC-SCV	53.85
UND	GAM-SCV	84.75	LR-CCV/LR-SCV/ SA-SCV	82.33
Overall[1]		78.70		78.95
Overall[2]		86.14		84.82
Note: Overall[1] is the overall accuracy of all LU changes. Overall[2] is the overall accuracy excluding LU changes of INS, AGR and WAT.				

The theoretical LU change model produces an overall accuracy that is 1.26 percent and 0.86 percent higher than the best performing GAM models derived from SCV with FB and RB datasets, respectively. When IND, AGR and WAT are included (as represented by the MC model) the overall accuracy drops from 85.36 percent (overall GAM-SCV) to 78.70 percent for FB datasets and from 83.96 percent (overall GAM-SCV) to 78.95 percent for RB datasets; however, it is only through this mixed approach that all types of LU changes can be modelled.

2.3.4 The Effect of Land-use Change Predictor

LU change predictors are important for constructing statistical LU change models. The coefficients of significant predictors in final LR, final GAM and final SA models with RB and FB test datasets can be found in Appendix E. Different methods (i.e., LR, GAM and SA) chose different sets of predictors to construct final models. Moreover, the sets of predictors selected by final models with RB test dataset and FB test dataset may also be different for a single method. Coefficients of regular predictors in LR and GAM are both log-odds ratios and are not directly comparable to coefficients of predictor in SA that are hazards in terms of magnitudes. However, the signs of coefficients in LR, GAM and SA can provide useful information for interpreting the effects of predictors. A positive sign of a coefficient in LR and GAM will result an odds ratio (exp(log-odds ratio)) that is greater than 1, which also indicates an increase in the odds of experiencing a LU change. A positive sign of a coefficient in SA will result a hazard ratio (exp(hazard)) that is greater than 1, which also indicates an increase in the hazard (risk) of experiencing a LU change. Therefore, the impacts of common LU change predictors in final LRs, final GAMs and final SAs were analyzed by comparing signs of coefficients. Moreover, the discussion of effects of LU change predictors is focus on predictors other than LU type and LC type in 2006 and LC type in 2010. In the following context, a smoothed term refers to a variable that was fit using a smoothing function to represent the non-linear relationship between it and the response variable in GAM.

Final LR, final GAM, and final SA derived from both CCV and SCV for predicting LDR LU change all recognize the variable *MRoad_dist* (distance to the nearest main road) as a significant LU change predictor even if final GAM sees it as a smoothed term. The positive coefficients of *MRoad_dist* in final LR-CCV and final SA-CCV (i.e., 0.67 and 0.26), and in final LR-SCV and final SA-SCV (i.e., 0.76 and 0.23) indicate that an increase in the unit of this predictor will increase the odds and the hazard of a parcel experiencing a LDR LU change while holding all other predictors constant, respectively. Hence, both final LR and SA agree with the positive effect of this predictor. Moreover, final LR-SCV, final GAM-SCV and final SA-SCV for predicting LDR LU change also agree with the fact that the odds and the hazard of a parcel experiencing a LDR LU change will increase as the size of DA increases (i.e, positive effect of *DA_Area*) and will decrease as the distance to its nearest commercial parcel increases (i.e., negative effect of *lu4_dist*).

The negative effect of *lu4_dist* was considered significant for LU change of MDR by final LR, final GAM, and final SA derived from both CCV and SCV regardless of sample sizes. Moreover, final GAM-CCV and final GAM-SCV see *lu4_dist* as a significant smoothed term. The negative effects of *lu4_dist* indicate that one kilometer increase in distance will reduce the odds and the hazard of a parcel experiencing a MDR LU change while holding all other predictors constant, respectively. Final LR-CCV, final GAM-CCV, and final SA-CCV for predicting MDR LU change with FB test datasets all recognize variables *Wood_dist* (distance to the nearest wooded area), *Water_dist* (distance to the nearest water area), *LRoad_dist* (distance to the nearest local road), *lu8_dist* (distance to the nearest recreational parcel) and *DA_Popn_Density* (DA population density) as significant LU change predictors while final GAM-CCV estimated them as smoothed terms. Except for *Water_dist*, final LR-CCV and final SA-CCV agree with signs of all coefficients of significant predictors mentioned above. Final LR-SCV and final SA-SCV with FB test datasets also agree with the significance and the effects of *Wood_dist*, *Water_dist*, *LRoad_dist*, and *lu8_dist*. Final GAM-SCV estimated them as significant smoothed terms. In addition, the significance and negative effects of *Residential_Popn_Density* (population density calculated by dividing population by total residential areas) and *Change_Popn* (change in population) were also identified by final LR-SCV and SA SCV.

The significance and negative effects of *ParcelArea* (an area of a parcel), *Wood_dist*, and *lu4_dist* regardless of sample sizes, and the significance and negative effects of *Residential_Popn_Density* and *DA_Popn_Density* with FB test datasets were all identified by final LR-SCV and final SA-SCV for modelling LU change of HDR. Final GAM-SCV estimated them as significant smoothed terms except for *ParcelArea* that was estimated to have negative effect.

Final LR-CCV, final GAM-CCV and final SA-CCV recognized the significance of *Parcel_Area* and *lu4_dist* for modelling LU change of COM regardless of sample sizes. They also agree with the significance of *LRoad_dist* and *MRoad_dist* with FB test datasets. In addition to the four predictors, final LR-SCV, final GAM-SCV and final SA-SCV also revealed the significance of *lu9_dist* to LU change of COM. Final LR and final SA derived from both CCV and SCV agreed with the effects of significant predictors.

Final LR-CCV, final GAM-CCV and final GAM-SCV did not reveal the significant effect of any LU change predictor for LU change of IND. Final LR-SCV and final SA-SA agreed with the significance and effects of *lu4_dist* and *lu5_dist*. Final LR-SCV, final GAM-SCV and final SA-SCV also successfully revealed significant effect of *ParcelArea*, *LRaod_dist* and *lu4_dist* with FB datasets and the significant effect of *lu8_dist* regardless of sample sizes.

For LU change of TRA, final LR-CCV, final GAM-CCV and final SA-CCV agreed with the significance of *ParcelArea* and *MeanDEM* (the mean elevation value within a parcel) with FB test datasets, the significance of *lu4_dist* with RB test datasets, and the significance of *lu8_dist* regardless of sample sizes.Both final LR and final SA derived from both CCV and SCV agreed with the significance and negative effect of *ParcelArea*, the significance and positive effects of *River_dist* and *lu8_dist* for modelling LU change of REC. In contrast, both final GAM-CCV and final GAM-SCV did not return any significant predictors.

ParcelArea was identified as a significant LU change predictor for modelling LU change of UND with FB test dataset by final LR, final GAM and final SA derived from CCV and with RB test dataset by final LR, final GAM and final SA derived from SCV. Moreover, *MRoad_dist*, *lu4_dist*, and *DA_Popn_Density* showed significant influence to LU change of UND with RB test datasets in final LR_CCV, final GAM-CCV and final SA-CCV. In addition to *ParcelArea*, final LR, GAM and SA derived from SCV only identified *MeanSlope* (mean slope within a parcel) and *lu4_dist* as significant predictors with RB test datasets.

2.4 Discussion

The study investigated the performance of MC, LR, GAM and SA in predicting eleven types of LU changes occurred during 2010 and 2015 in the Region of Waterloo. For most of the types of LU change, LR, GAM and SA have very similar results. MC is not competitive with LR, GAM and SA in predicting LU change except for INS, AGR and WAT. Thus, there is no absolute favor in LR, GAM and SA in terms of the prediction accuracies. The advantages and limitations of each method are addressed in this section. With the consideration of different aspects of each method, MC, LR, GAM and SA should be selected accordingly to model each type of LU change. Moreover, future work can be done to improve these models with addressed potential solutions that are given based on reviews of similar studies and experiences. As the result of improving each method, the combination of methods that is expected to give the optimal result of LUCC modelling can also be improved.

2.4.1 Opportunities and Challenges

2.4.1.1 MC

MC is the only method used to model institutional LU change in this study. In total, there are only ten parcels found with institutional LU change. For training datasets with 10-fold CV, MC produced 55 percent accuracy with FB and 65 percent accuracy with RB for institutional LU change, which means MC performed slightly better than a random classification result (i.e., 50 percent chance being accurate). However, the accuracies dropped to 0.4 percent with FB and 0.25 percent with RB when MC was tested with test datasets. In contrast, all accuracies of predicting agricultural LU change by MC are greater than 93 percent and the difference among FB and RB accuracies is at large 2.6 percent. Above findings suggest that 1) MC does not provide robust performance for predicting LU changes with limited sample data (i.e., small sample set) and 2) MC can give better prediction performance for a LU with sufficient data but rarely changes to another LU type which indicates that MC performs better with unbalanced datasets.

Iacono et al. (2012) investigated the ability of MC to predict ten LU types over a medium to long-term time scale with five time periods in Minneapolis-St. Paul region, Minnesota, US. Their results, using equal time intervals based on two different time periods, achieved 70 percent and 84.4 percent, which are substantially higher than the overall 10-fold CV MC accuracy (i.e., 46.73 percent for FB and 47.59 percent for RB) of this study. This superiority could be caused by the larger sample dataset (610,000 cells) used by Iacono et al. (2012). In comparison, the MC FB dataset contains 21,712 parcel data and the MC RB dataset contains 7,502 parcel data in our study. This finding agrees with the conclusion in Clark (1965) that “Markov Chain Analysis is most effectively applied when there are a large number of time periods (year to year for example) and a large number of observations". Moreover, our sample datasets are all balanced, which further reduces the ability of MC to make better predictions.

In conclusion, MC cannot sense the trend of LU changes driven by exterior predictors and account it in a transition probability, and does not consider the effect of spatial autocorrelation among spatial objects. In reality, LUCC models typically do not follow all the assumptions of MC (Turner, 1987), in which the uniformed length assumption of time interval in DTMC is often violated when several states are presented in LUCC modelling since spatial data typically lack consistent intervals between dates of acquisition. For instance, Iacono et al. (2012) showed that the prediction accuracy drops noticeably when the available data was used to make long-term prediction, which means the accuracy decreases as the time interval of the forecasted year becomes larger than the period that the probability matrix was constructed. Furthermore, the time period that each LU change takes may be different. Hence, the time periods used in a study may not best reflect the rate of LU conversion. However, it needs to be tolerated by many researchers since the availability of data is the key to solve the problem. MC would be a good choice to model LUCC when LUCC drivers are unavailable. It can also serve as a null model (i.e., a model without any predictors) that performs similar to random classification in our study, which means it should not perform better than other statistical models with predictors. Moreover, MC has been coupled with other techniques such as LR (Arsanjani et al., 2013), Cellular automata (CA) (Huang et al., 2015; Ebrahimipour et al., 2016), and genetic algorithm (GA) (Tang et al., 2007) to better perform LUCC modelling. Therefore, results of MC can be improved by having a larger sample dataset, more time steps with equal lengths and coupling with other methods.

2.4.1.2 LR

LR is a relatively popular statistical method for modelling LUCC. It has been used in some studies to detect drivers of LUCC (e.g., Serneels and Lambin, 2001), and compare its performance in detecting LUCC and making prediction of LUCC with some other methods (e.g., Lin et al., 2011; Wang et al., 2013). One worth noting is that Wang et al. (2013) compared the performance for detecting spatial predictors of LU changes between LR and SA and concluded that SA performed better than LR due to the ability of SA accounting for temporal variables. In our study, SA does not perform better than LR in terms of making prediction of LU changes. This could due to 1) the limited time steps (two time steps), 2) the randomization of five failure times between the two time steps in our study since the simulated failure times may not represent the reality of the duration of LU conversion, and 3) the absence of temporal variables. However, the difference between accuracies produced by LR and SA in our study is within 1.7 percent for the overall 10-fold CV accuracies and 0.86 percent for the final model accuracies, which does not provide an evidence for having a favor of anyone of them.

In general, LR is relatively simple to implement and is especially made for categorical response variables. The relationship between the linear predictor, usually denoted by η, and the linear combination of predictors can be assessed by estimating coefficients of LU change predictors in this relationship. Due to the linearity presented between η and the linear combination of predictors, interpretation of its results can be made relatively easily by using odds ratio or log-odds ratio. One limitation of using LR in modelling LUCC is that LR often ignores the spatial aspect of data (Zeng et al., 2008). This problem could be mitigated by choosing spatially independent observations (Serneels and Lambin, 2001) and potentially be alleviated by using predictors that can account for spatial autocorrelation. Moreover, the LR used in our study contains only fixed effects, which means the effect of each predictor is constant for all observations. In reality, some predictors are time-varying and have different effects by groups. To solve the problem, random effects can be introduced to regression models to allow predictors being varying for different reasons (e.g., group and time). A model that contains both fixed effects and random effects is recognized as a mixed effects models. Future study can be done to investigate the ability of a mixed effects LR for modelling LUCC.

2.4.1.3 GAM

The GAM used in this study is the additive logistic regression (ALR) that is a non-linear version of LR and allows capturing non-linear relationship between predictors and response variables. In our study, the ALR performed the best in terms of prediction accuracies. Brown et al. (2002) achieved 87 percent prediction accuracy for modelling LU change from non-forest to forest and 90 percent prediction accuracy for modelling LU change from forest to non-forest with the use of GAMs. Similar LU change exists in our study (i.e., LU change from any LUs to REC) and the corresponding prediction accuracies are above 90 percent for both the 10-fold CV and final model regardless of sample sizes. The gap between the studies' results could be caused by the disparity between sample sizes, in which the number of samples in our study (i.e., 520 parcels) is nearly half of the amount used by Brown et al. (2002) (i.e., 1,014 cells). A GAM with a quasi-binomial distribution was used to "account for spatial autocorrelation and the boundedness of the percent woody cover variable" in a study done by Eitzel et al. (2016) to model historical land cover change. The quasi-binomial distribution is not a real distribution. It infers to the quasi-likelihood estimation used in binomial distribution to allow for over-dispersion.

Compared to generalized linear models (e.g., LR), GAMs have the advantages of having a relaxation of variable assumptions and allowing non-linear relationships, but the advantages can also be considered as limitations. These advantages can increase the overall complexity of models, make interpretation of results more complicated and cause a need of more computational power. However, because many LU change drivers interact non-linearly to influence the future LU, GAM can be helpful to determine the complex relationship between drivers when they interact heavily in a non-linear fashion. Moreover, the quasi-likelihood estimation and smooth function of coordinates could help reduce spatial autocorrelation without resampling data for GAM (Eitzel et al., 2016).

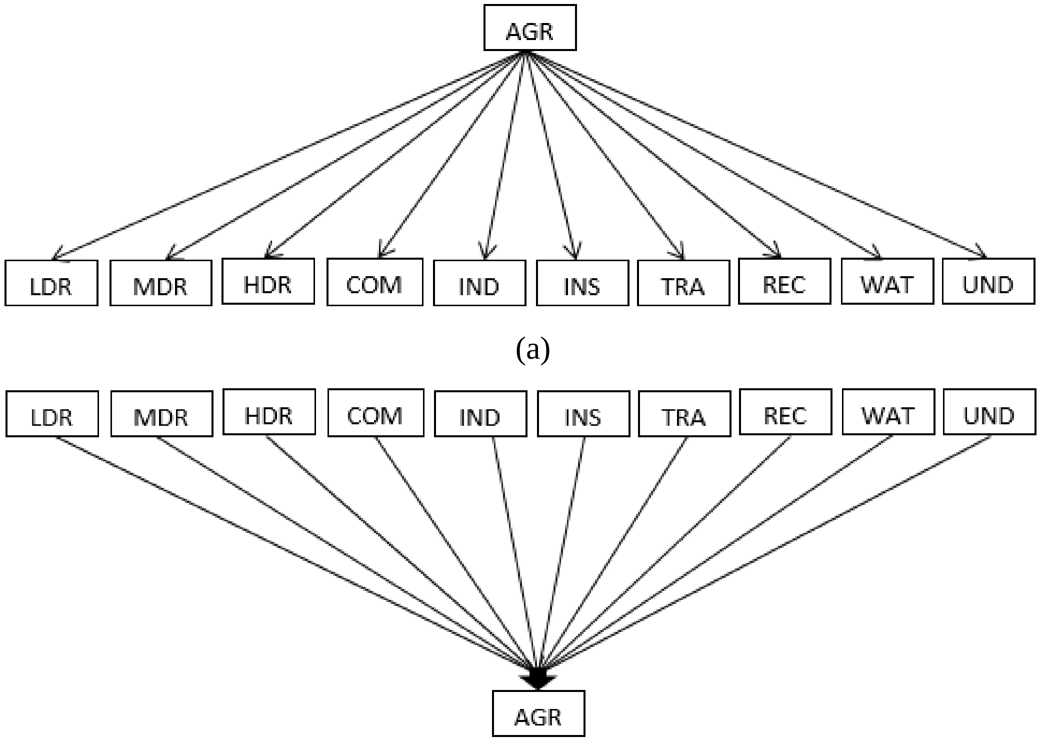
2.4.1.4 SA

In general, SA can model time-varying predictors, and capture behavior of response variables and predictors at different time steps. The implementation of SA requires the input of a time variable, which means at least two years' data are required. SA would be considered the same as LR when data comprising both complete observations and censored observations are only available for two time points. Therefore, it is better to have more than two measurements per observed parcel to distinguish SA from LR. Furthermore, the effects of covariates on hazard may increase over time with more information involved and causes competition of land and resources, which can cause violation of the PH assumption that effects of predictors are constant on hazard over time (Wang et. al., 2013).

In our study, the prediction accuracies of LR are greater than the accuracies of SA for both the overall 10-fold CV accuracy and the final model accuracy regardless of the sample sizes. This infers that even though SA took less time to process, LR may be preferred over SA under the condition that no time varying predictors present, and only complete data and right censored data exist. The comparison result between LR and SA is expected to change when time-varying predictors are present. In a study of detecting spatial predictors done by Wang et al. (2013), which includes one time-independent variable and two time-varying variables, a Cox model outperformed a logistic model by 16 percent. Therefore, the performance of SA with presence of time-varying predictors requires further investigation.

2.4.3 Modelling One-to-One LUCC vs. Many-to-One LUCC

LR and GAM that have been used to model LUCC are usually designed to represent one-to-one LUCC (e.g., non-urban to urban, Wang et al., 2013; Braimoh and Onishi, 2006; non-forest to forest and forest to non-forest, Brown et al., 2002). SA has been used to model one-to-one LU change from non-urban to urban (Wang et al., 2013) and LU changes from farm to three types of subdivisions (An and Brown, 2008). In this way, focusing on modelling one-to-one LUCC could help reveal effects of predictors on a specific LUCC and help better understand causes of a specific LUCC. However, modelling one-to-one LUCC can be time-consuming when many LU classes are involved in a study. Moreover, the availability of LU data could also restrict the ability of statistical methods for modelling LU changes. For example, there are eleven LU classes in our study. If LR, GAM and SA were designed to model one-to-one LUCC, datasets would be constructed by including parcels that experienced a specific LU change and a matching number of unchanged parcels. In this case, each unique LU is treated as a starting LU for modelling ten types of LUCC. Figure 5(a) shows an example of the ten possibilities of modelling one-to-one LUCC from AGR to all other LUs. Hence, each of the four methods (MC, LR, GAM and SA) would need to obtain 110 final models in order to model all unique LU changes. In contrast, each unique LU is treated as an end LU in the case of modelling many-to-one LUCC. Therefore, each method would only have eleven final models for all eleven possible LU changes. Figure 5(b) shows an example of modelling many-to-one LUCC from any LU to AGR.



(a)

(b)

Figure 6: (a) The ten possibilities of modelling one-to-one LUCC from AGR to all other LUs. (b) Modelling many-to-one LUCC from any LU to AGR. [Notes: low-density residential (LDR), medium-density residential (MDR), high-density residential (HDR), commercial (COM), industrial (IND), institution (INS), transportation (TRA), protected area and recreation (REC), agriculture (AGR), water (WAT), under development (UND)]

Furthermore, preparing data for modelling unique LUCC would result in samples with only a few observations and even an empty set of samples (i.e., LU change of AGR) in our study. Therefore, shifting the focus of LUCC models from modelling one-to-one LUCC to many-to-one LUCC helped us increase sample observations in each designed dataset and reduce spatial autocorrelation by having samples with non-monotone LUCC type. Our results prove that the many-to-one modelling technique performed well in terms of prediction accuracy for LR, GAM and SA and performed fairly well compared to prediction results from models modelling one-to-one LUCC.

2.4.4 Operationalizing the Combined Statistical Model

An application of this study is to predict future LU with MC, LR, GAM and SA in a given area. When a method is applied to an area, each parcel in the area will be given a set of probabilities of changing to different LUs, which determines the future LU of the parcel by the highest probability. The four methods are competitors with each other when they are applied to the same parcel. In other words, final models will be constructed for each of LR, GAM and SA when all LUCCs are present. Probabilities will be produced for a single parcel and compete for the highest probability of LUCC when a single method is used to map the LU in the whole study area. The highest probability produced by each method will further compete for the highest value to determine the LU type when all methods are used to map the LU in the whole study area (detailed procedures can be found in Appendix F).

In addition to provide a reference for future LU, statistical LUCC models built in this study can help allocate predicted LU changes to appropriate locations in the Region of Waterloo. Since the statistical LUCC models built in this study do not account for government LU policies and zoning regulation, the feasibility of predicted LU changes needs to be verified when MC, LR, GAM and SA are used to predict future LUs in this region. For example, if a projection of a residential development is made on a piece of protected natural land, the local government may not grant a permit to the development. However, the prediction of new residential development may be driven by population growth in this area. It raises the necessity of re-allocating the predicted amount of residential development to appropriate locations. Moreover, the amount and types of residential development (LDR, MDR and HDR) can be allocated by residential house developers according to criteria such as people's preference about the location and LU policy (Robinson et al., 2012).

Furthermore, future population can be projected as a derivative of statistical LUCC models. In this study, population density and population change rate were created using population and were used as predictors to model LU changes. Once the predicted amount and types of residential LU changes have been determined, population can be projected based on these data. For instance, when a residential development of a hundred MDR parcels is predicted, the population is expected to grow for approximately 300 people based on an average amount of three people per MDR.

2.5 Conclusions

MC, LR, GAM and SA have been used to model LU changes and a comparison of their predicting powers has been conducted. To the best of the author's knowledge, this is the first study to formally compare the relative performance of these statistical methods for LUCC modelling. The goal of determining the overall accuracies of the four methods for 10-fold CCV, 10-fold SCV and final models derived from CCV and SCV in representing eleven types of LU changes illustrated that GAM performed the best in making prediction of future LUs but the superiority of it is not obvious when we consider the difficulty of its implementation and interpretation, and its run time relative to LR and SA models. For both CCV and SCV results, it can be concluded that SCV did reduce the averaged accuracy by LU change type and overall accuracy by method. The effect of SCV alleviating over-fitting caused by spatial autocorrelation among spatial parcels is minor in terms of overall accuracy but is substantial for some methods modelling parcels with specific LU changes.

Moreover, a decrease in sample size causes a reduction of overall accuracy for LR, GAM and SA and reduces the difference in overall accuracies of LR, SA and MC from the overall accuracy of GAM. However, the reduction in overall accuracy between FB and RB training datasets with 10-fold CCV and 10-fold SCV is minor. For both 10-fold CCV and 10-fold SCV, GAM is the most sensitive method to the reduction of sample sizes since it experiences the largest difference between overall accuracies. GAM is also the best method in modelling training datasets with 10-fold CCV and 10-fold SCV, and test datasets with final models derived from CCV and SCV. In contrast, LR and SA are less sensitive to results of CV models and final models. Moreover, SA has the shortest run time for conducting 10-fold CCV compared to MC, LR and GAM for FB training datasets and has the second shortest run time for conducting 10-fold CCV compared to MC, LR and GAM for RB training datasets. LR only took a few seconds more compared to SA, and overall accuracies of LR rank the second place for both 10-fold CV and final models regardless of sample sizes. In contrast, the decrease in sample size has an adverse effect to MC. The overall accuracy is 4.72 percent higher for the RB training dataset than the overall accuracy for the FB training dataset with CCV and is 0.01 percent higher for the RB training dataset than the overall accuracy for the FB training dataset with SCV.

Chapter 3 Contribution and Future Work

3.1 Summary

The presented analysis and comparison of statistical methods sought to evaluate the prediction power of four different statistical methods (MC, LR, GAM and SA) for representing LU change. The LU of a parcel was classified into one of eleven LU classes (ten LUs and one LC) in the Region of Waterloo. LU change among the pre-defined LU classes between 2010 and 2015 were modeled by the four statistical methods. The preparation of the study involves an identification of potential LUCC drivers from literature, a creation of LUCC drivers with available data, a review of mathematical background of the four statistical methods and a review of related studies. An analysis of the modelling results was conducted, which quantified the relative performance of each method and revealed the distribution of overall accuracy and final model accuracy of MC, LR, GAM and SA by LU type.

Among the four methods, GAM performed slightly better than LR and SA in terms of overall prediction accuracy by method type due to its ability of modelling non-linear relationship between responses and LUCC predictors. However, non-parametric smooth functions in GAM can increase the difficulty of implementation and interpretation at the same time. The estimation of smooth functions also requires a large amount of time, which is caused by iterations of finding fitted smooth functions for all LUCC predictors in the back-fitting algorithm. LR and SA produced similar overall accuracies and run times. Furthermore, even though GAM yielded the highest overall accuracy, it did not produce the highest accuracy for all LU changes. Moreover, MC was not competitive for modelling most LU changes. However, when the amount of data is scare or predictors are unavailable, it performs better than standard null models (e.g., pure persistence of LU and LC, Pontius Jr. and Spencer, 2005) and can outperform GAM, LR and SA under these data constraints.

Perhaps it is not surprising that no single statistical method achieved the highest accuracy for all LU changes. Different methods have different strengths and weaknesses for capturing the different underlying processes of LUCC. Therefore, a combination of methods should be used to make a more accurate prediction instead of using a single method for modelling all LU changes. Surprisingly, to the best of the author’s knowledge this is rarely done, with a few exceptions (e.g., Robinson et al. 2012). The theoretical relationship between the overall accuracy and model complexity for the MC, LR, GAM and SA methods individually and in combination (i.e., the combination of statistical methods (CSM)) is shown in Figure 6(a). Figure 6(b) further illustrates the superior performance of a LU change model that uses different statistical approaches for different LU types in terms of overall accuracy and overall time consumption. In this context, the model complexity refers to the author's consideration of complexities in terms of learning and implementation, thus is a relative scale that cannot be quantified. The overall time consumption is also presented as a relative scale since the point is not to show the exact run time of each method. The overall accuracy of each method is based on the real value of overall accuracy of final model with FB test dataset.

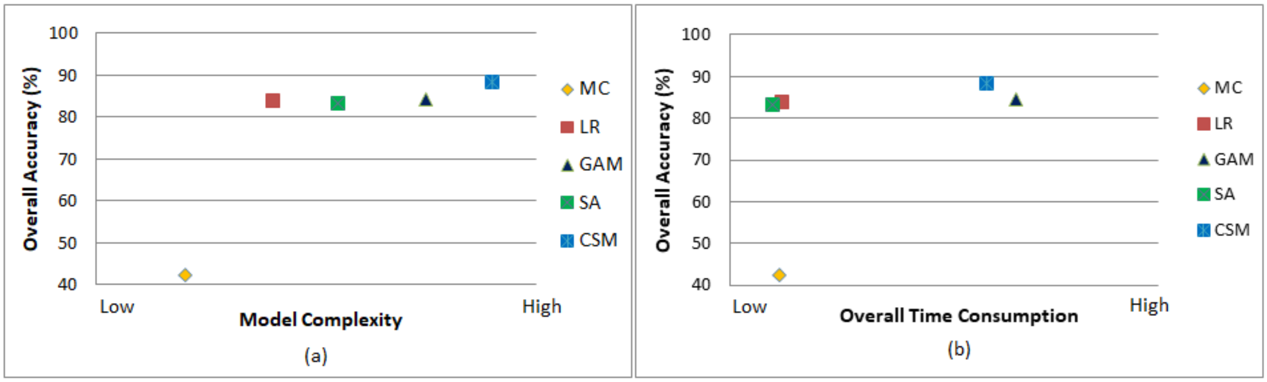


Figure 7: (a) The theoretical relationship between the overall accuracy and model complexity for MC, LR, GAM, SA and the combination of statistical methods (CSM). (b) The theoretical relationship between the overall accuracy and overall time consumption for MC, LR, GAM, SA and CSM.

In addition, the four statistical methods used in this study can be categorized as stochastic process (MC), parametric model (LR), non-parametric model (GAM) and time series analysis technique (SA)). However, the four tested statistical methods are not restricted to the classes of models listed here. For instance, a GAM with both linear terms and smoothing functions can be considered semi-parametric; Cox PH model, the SA technique used in this study, is a semi-parametric model; an accelerated failure time (AFT) model, a type of SA technique, is parametric; Kaplan-Meier estimator, a type of SA technique, is non-parametric. Moreover, LR and ALR are classification models but MC and SA are not. MC and SA can handle time series data but LR and ALR cannot. Hence, statistical methods cannot simply be grouped into a few categories. In conclusion, the selection of the four statistical methods covers a wide range of statistical techniques that target different interests. This study provides a general understanding of modelling LU change with MC, LR, GAM and SA, in which ALR and Cox PH model are specifically chosen for GAM and SA, respectively.

Furthermore, it is worth investigating the performance of mixed effects models (Pinheiro et al., 2007), especially mixed effects LR (MELR) and mixed effects GAM (MEGAM), and some machine learning (ML) techniques such as random forest (RF; Liaw and Wiener, 2002) and support vector machine (SVM; Suykens and Vandewalle, 1999). MELR and mixed effects additive logistic regression (MEALR) are expected to perform better than LR and ALR, respectively, since random effects in mixed effects models could take account of time-varying variables and spatial autocorrelation but would increase model complexity. ML approaches are more flexible than traditional statistical methods since they are spared from general assumptions of statistical methods such as linearity, independency and an underlying distribution of data. Moreover, ML can benefit from a large amount of input data, both observations and predictors, and is less affected by multi-collinearity among predictors. RF is a popular ML method, which has been used in LC classification (Rodriguez-Galiano et al., 2012; Liu et al., 2016). SVM is another popular ML method used for classification purpose and has been used to classify LC (Kavzoglu and Colkesen, 2009; Huang et al., 2002). Based on literature and knowledge about MELR, MEALR, RF and SVM, the conceptual performance of MELR, MEGAM, RF and SVM with MC, LR, GAM, SA and CSM can be constructed against model complexity and time consumption (Figure 7). Both model complexity and overall time consumption are relative scales that designed to show the conceptual relationship among methods. Moreover, the overall accuracy of each method in Figure 7 is a conceptual value that was created based on the consideration of author’s experience and results of this study.

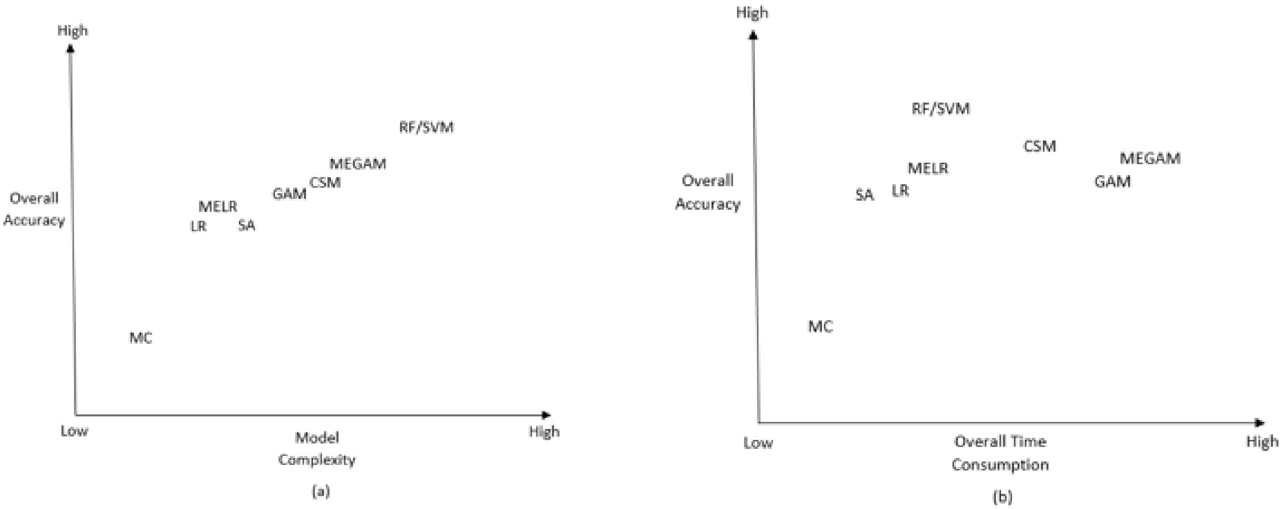


Figure 8: (a) The theoretical relationship between the overall accuracy and model complexity for MC, LR, GAM, SA, MELR, MEGAM, RF, SVM and CSM. (b) The theoretical relationship between the overall accuracy and overall time consumption for MC, LR, GAM, SA, MELR, MEGAM, RF, SVM and CSM.

3.2 Contribution and Future Work

Prediction results of the four statistical methods for modelling LU changes can be used to support decision-making associated with solving real world problems since human's decision-making on LU have clearly linked LU science and policy (Aspinall, 2007). The population in the Region of Waterloo has been projected to reach 742,000 in the year 2031 (Ministry of Municipal Affairs, 2017). This indicates an approximately 46 percent increase in population with the reference year 2011. The Region of Waterloo will experience a variety of social problems caused by LU practices that are made to satisfy human needs such as building residential houses and industries. Then, there will be environmental problems followed by social problems after LU and LC have been altered (Bell, 2009). Therefore, understanding how different statistical approaches model different LU changes can help draw a big picture that represents real-world patterns. Moreover, moelling LU change with statistical methods not only contributes to a historical and contemporary area of scientific investigation, insight from a comparison of these methods is also essential to the sustainable development of social, environmental and scientific aspects of society.

3.2.1 Social Aspect

Results from statistical LUCC models can influence LU policies and LU planning by providing improved understanding of LUCC in an area. The prediction results of statistical LUCC models can provide estimated LU types at each potential location. Then, an estiamted amount of each

LU change can be quantified. It is also important to study the effect of LUCC predictors since knowing why LUCC occurs is as critical as knowing where and how it occurs. The effect of a LUCC predictor on a specific LUCC, the value and the magnitude, can be revealed by studying coefficients of significant LUCC predictors. Therefore, results from our models can provide suggestions and evidences for making LU policies and LU planning.

Since the study area is the Region of Waterloo, results from the four statistical methods can provide insight about future LU in the region. In fact, building permits worth $670 million were issued for the residential sector in the Region of Waterloo in 2017, in which the largest proportion of permits (47%) were given to development of HDR (apartments) (Region of Waterloo, 2018). Other permits were issued to industrial, commercial and institutional development. Among the eight building permits issued to institutional sector in 2017, six of them were granted to the addition and renovation of existing institutional facilities and two were for the development of new elementary schools. Residential, commercial and industrial LU changes can all be modeled using statistical LUCC models constructed in this study but it is difficult to model institutional LU change since the amount of such LU change is minor. Even though it is difficult to predict institutional LU changes with limited data, understanding how other LU changes work can provide clues and insight toward understanding the conditions that lead to institutional LU changes.

As previously mentioned, by studying the coefficients of significant predictors driving LUCC may add decision making capacity to the Region of Waterloo when conducting LU planning. Using the final LR model and conversions to HDR as an example, we illustrate the use of coefficients of three significant predictors: the distance from a target parcel to its nearest COM parcel, the distance from a target parcel to its nearest AGR parcel, and the proportion of neighborhood REC parcels of a target parcel. In the context of statistics, the coefficients of LR are log-odds ratio that can be converted to odds ratio (i.e., exp(log-odds ratio)). The odds ratio is more regularly used to interpret results than the log-odds ratio. The odds ratio is a relative measurement of the odds of some event occurring given some covariates and the odds of the event not occurring without the same set of covariates. The odds of an event is interpreted as the likelihood of the event occurring. In LUCC context, the odds ratio measures the strength of LUCC predictors to the presence or absence of a LUCC. Higher values of odds ratio indicate higher strength of LUCC predictors to the occurrence of some LUCC and vice versa. Therefore, the estimated coefficients for these three predictors can be interpreted as the likelihood of a parcel experiencing a change from any LU to HDR increases by a factor of 0.0622 and 0.8050 for each kilometer increases comparing to the likelihood of the parcel not experiencing any LUCC for the first two predictors in our example, and an increase by a factor of 51.5215 with a percent increase comparing to the likelihood of the parcel not experiencing any LUCC for the third predictor. When interpreting the odds ratio of a predictor, all other predictors are held constant. Interpreting these results suggest that a parcel is less likely to convert to HDR as the distance between the parcel and its nearest COM parcel (e.g., shopping mall, grocery store and small business) or AGR parcel increases, and it is more likely to convert to HDR as the proportion of parcels in its neighborhood classified REC (e.g., green areas, trails and protected areas) increases.

The above interpretation of coefficients would suggest that new HDRs would be built 1) near urban fringe, the transitioning area between urban and rural, where commercial LU (e.g., plaza that contains grocery store, restaurant and bank) has already existed nearby in the Region of Waterloo and 2) at places inside the city where green spaces and business services can be easily accessed. Therefore, local government could expect residential housing developers to seek permits to build high-rise residential buildings in satisfied areas, which are the areas identified by our models for having HDR LU change, when the region experiences urban intensification and sprawl. Similarly, the trend of other residential development (LDR and MDR) can be revealed by studying corresponding model coefficients. By knowing the trend of residential LU change in advance, the local government can regulate the development of this LU change with considerations of other criteria such as projected population, the availability of existing public services and zoning regulations. Other types of LU changes would be influenced by residential LU changes in response to the growing demand of some LU types (e.g., commercial LU and institutional LU). The development of other LUs and the consequence of such development in the Region of Waterloo can be determined and revealed by investigating through the coefficients. Therefore, LU planning and LU policies can be made to manage and regulate LUCC with the help of statistical LUCC models when the effects of LUCC predictors have been estiamted.

3.2.2 Environment Aspect

Protecting farmlands and sensitive natural areas is an objective of the regional government of the Regional of Waterloo (Ministry of Municipal Affairs, 2017). Lands not only retain social values, they also possess ecological value. LUCC can directly affect our environment and ecosystem function that refers to the underpinning processes conducted by ecosystems that often provide goods and services to humans. Therefore, understanding the potential patterns of LUCC can help estimate the ecological consequences of different policies and development plans. Actions can then be taken to alleviate the negative impacts. Furthermore, the aggregated effects of LU practices at small scales can influence global environment and global climate. Hence, the study of modelling LUCC with statistical methods at a smaller scale (i.e., the Region of Waterloo) can provide meaningful insight to reveal the impacts of LU practices at the provincial level.

With statistical LUCC models, the effects of LU changes can be estimated when the scale and the area of LU changes have been determined. For instance, predicted urban sprawl with mixed LUs, such as residential, commercial and transportation, can take place on lands previously defined as farmlands (AGR) or protected natural areas (REC) in the Region of Waterloo. The replacement of these LUs by urban LUs will create impermeable surface, which would exacerbate problems of climate change by reducing the intake of carbon by vegetation and soil (Watson et al., 2000), reduce evapotranspiration and increase local heat island effects (Trenberth, et al., 2007), as well as contribute to surface runoff and eutrophication of local waterways (Shi et al., 2007; Huang et al., 2013). Moreover, human activities occurred on newly developed LUs (e.g., driving and heating) may have broader scale impacts such as increase energy consumption causing the efflux of greater concentrations of greenhouse gases (GHGs) being produced and released into the atmosphere. Furthermore, features on land can also cause environmental issues. For instance, high-rise buildings covered by massive amount of light-reflective glasses can cause light pollution (Horváth et al., 2009). All these consequences need to be taken into consideration when a LUCC plan is being made. The Region of Waterloo has already experienced a loss of AGR lands and REC lands over time. Therefore, urban sprawl needs to be curbed when the speed of expansion predicted by statistical LUCC models causes unaffordable consequences to the environment.

3.2.3 Science Aspect

In addition to contributions made to the society and environment, this study also has several scientific contributions. In LU science, statistical methods are more often used to detect spatial predictors of LUCC instead of directly modelling LUCC. The presented research seeks to achieve and contribute to both our understanding of the drivers of LUCC and our ability to model LUCC. Moreover, GAM and SA models are rarely used to represent LUCC. The presented comparison of these two methods to MC and LR provides insight about their overall and specific performance for modelling LU changes. The comparison of methods remains rare in the LUCC modelling literature.

A comparison among different statistical methods is challenging for many reasons, including maintaining an adequate amount of knowledge about the mechanism of these methods and knowing underlying assumptions. Therefore, this study is important for discovering and understanding the potential of MC, LR, GAM and SA in modelling LUCC under the same circumstances. This study also reveals limitations of each method in modelling LU change with current data. The advantages and limitation of each method are discussed (Section 2.4.1) and further improvement can be made with a support from similar studies.

Furthermore, canonical correlation analysis (CCA) was reviewed for its ability to model LU change in addition to MC, LR, GAM and SA. CCA is a multivariate statistical method and is often used to explore the relationships between two sets of variables (Härdle and Simar, 2007). It has been used to identify relationships between LU patterns and influential factors. However, to the best of the author’s knowledge, only one publication of its use for modelling LC changes exists (Lee et al., 1999). In an effort to include CCA along with the four methods presented in this thesis, an experiment of modelling LU changes with CCA was conducted. Results showed that CCA is not an appropriate method to make prediction of future LUs with a set of LUCC predictors. The accuracy of classification achieved was extremely low. This implies that CCA is not able to capture the relationship between LU changes and associated LUCC predictors. In this study, LU change status was used as one set of variables in CCA models since all other methods were constructed to model LU change instead of classifying LU. In future studies, it is worth investigating the performance of CCA for classifying LUs or LCs instead of LU changes. A detailed review of the mathematical background of CCA can be found in Appendix G.

In conclusion, given the nature of statistical models, the methods used in this study can be deployed across large spatial extents with high-resolution data and be implemented with relative ease when data are available. In addition, statistically significant drivers of LUCC can provide insights to impacts of geographical, demographical and social-economic factors on specific LU changes. Considering all the facts mentioned here, statistical models built in this study can be used as guide for future studies of modelling LUCC and can provide a reference to screen a set of variables for modelling each type of the defined LU changes. Moreover, statistical models can be integrated with many other methods in LU science. For instance, statistical models can produce probabilities, which can be used in agent-based models (ABMs). An ABM is under the risk of failure when parameters are not properly calibrated, and a failure of an ABM will cause a waste of time and resources. Meanwhile, statistical methods alone have to face limitations for accounting complicated interactions among many factors and having minimal ability to account for explicit decision-making processes. Therefore, a hybrid model that combines statistical LUCC models and ABMs can offset both limitations of statistical methods and ABMs due to the ability of ABM to simulate process-based phenomenon. This can be achieved by using statistical models as agents in ABMs and using probabilities as decision roles of stakeholders in a decision-making process when empirical data are scarce. Hence, statistical LUCC models can help reduce the risk of wasting resources such as data and human labor due to a relatively efficient cost and time of producing data (i.e., probabilities) compared to some data collecting methods such as traditional survey especially when the scale of the study area is large.

Reference

Adams, D. M., Alig, R. J., Callaway, J. M., Winnett, S. M. and McCarl, B. A. (1996). The forest and agricultural sector optimization model (FASOM): model structure and policy applications. *DIANE Publishing*.

Alcamo, J., Kok, K., Busch, G., Priess, J. A., … and Heistermann, M. (2006). Searching for the Future of Land: Scenarios from the Local to Global Scale. In E. &. Lambin, *Land-use and land-cover change: Local processes and global impacts* (pp. 137-155). Berlin: Springer.

An, L. and Brown, D.G. (2008). Survival Analysis in Land Change Science: Integrating with GISceience to Address Temporal Complexities. *Annals of the Association of American Geographers*, 98:2, 323-344.

An, L., Brown, D. G., Nassauer, J. I. and Low, B. (2011). Variations in development of exurban residential landscapes: timing, location, and driving forces. *Journal of Land Use Science*, 6, 13-32.

Andersen, E. B. (1970). Sufficiency and expoenntial families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331), 1248-1255.

Arsanjani, J., Helbich, M., Kainz, W., & Boloorani, A. (2013). Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. *International Journal of Applied Earth Observation and Geoinformation*, 21, 265-275.

Aspinall, R. (2004). Modelling land use change with generalized linear models - a multi-model analysis of change between 1860 and 2000 in Gallatin Valley, Montana. *Journal of environmental management*, 72(1-2), 91-103.

Batista, G., Prati, R., & Monard, M. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.

Bell , M. (2009). Environmental problems and society. In M. Bell, *An invitation to environmental socieology. Edition 3.* (pp. 1-29). Sage Publication.

Bolin, B., & Sukumar, R. (2000). Global perspective. In R. T. Watson, I. R. Noble, B. Bolin, N. H. Ravindranath, D. J. Verardo, & D. J. Dokken, *Land Use, Land-Use Change, and Rorestry* (pp. 23-52). Cambridge, UK: Cambridge University Press.

Bonabeau, E. (2002). Agent-based modelling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Science*, 99(3), 7280-7287.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.

Braimoh, A. K., & Onishi, T. (2006). Spatial determinants of urban land use change in Lagos, Nigeria. *Land Use Policy*, 24(2007), 502-515.

Brenning, A., Schratz, P., & Herrmann, T. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package.

Brown, D. G. (1994). Prediting vegetation at treeline using topography and biophysical disturbance variables. *Journal of Vegetation Science*, 5(5), 641-656.

Brown, D. G., Goovaerts, P., Burnicki, A., & Li, M. Y. (2002). Stochastic Simulation of Land-Cover Change Using Geostatistics and Generalized Additive Models. *Photogrammetric engineering and remote sensing*, 68(10), 1051-1062.

Brown, D. G., Walker, R., Manson, S. and Seto, K. (2012). *Modeling land use and land cover change.* Dordrecht: Springer.

Cain, K., Harlow, S., Little, R., Nan, B., Yosef, M., Taffe, J., et al. (2011). Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *American journal of epidemiology*, 173(9), 1079-1084.

Canadian Census Analyser. (2011). Retrieved from CHASS: http://dc1.chass.utoronto.ca/cgi-bin/census/2011/displayCensus.cgi?year=2011&geo=da#

Chomitz, K. M. and Gray, D. A. (1996). Roads, land-use, and deforestation: A spatial model applied to Belize. *The World Bank Economic Review*, 10(3), 487-512.

City of Waterloo. (2012, August 23). Zoning by-law No. 1108.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British journal of cancer*, 89(2), 232.

Clark, W. (1965). Markov chain analysis in geography: an application to the movement of rental housing areas. *Annals of the Association of American Geographers*, 55(2), 351–359.

Cohen, J. (1994). A Constitutional Safety Valve: The Variance in Zoning and Land-Use Based Environmental Controls. *Boston College Environmental Affairs Law Review*, 22, 307.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B, Methodological*, 34, 187-220.

Ebrahimipour, A., Saadat, M., & Farshchin, A. (2016). Prediction of urban growth through cellular automata-Markov chain. *Bull. Soc. R. Sci. Liège*, 85, 824-839.

Eitzel, M., Kelly, M., Dronova, I., Valachovic, Y., Quinn-Davidson, L., Solera, J., et al. (2016). Challenges and opportunities in synthesizing historical geospatial data using statistical models. *Ecological informatics*, 31, 100-111.

Environmental System Research Institute (ESRI). (2016). ArcGIS Desktop Release 10.4. Redlands, CA: ArcGIS Desktop: Release 10.

Evans, M., Hastings, N. and Peacock, B. (2000). *Statistical Distributions Third Edition.* New York: Wiley.

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., ... and Helkowski, J. H. (2005). Global consequences of land use. *Science*, 309, 570-574.

Ford, A., & Ford, F. (1999). Modelling the environment: an introduction to system dynamics models of environmental systems. *Island press*.

Härdle, W. and Simar, P. (2007). *Applied multivariate statistical analysis 2nd edition.* Berlin; New York: Springer.

Hartigan, J. (1975). *Clustering Algorithms.* New York, NY: Wiley.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320-338.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London; New York: Chapman and Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction.* New York: Springer.

Horváth, G., Kriska, G., Malik, P., & Robertson, B. (2009). Polarized light pollution: a new kind of ecological photopollution. *Frontiers in Ecology and the Environment*, 7(6), 317-325.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55(1), 13-22.

Huang, C., Davis, L. S., & Townshend, J. R. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4), 725-749.

Huang, J., Wu, Y., Gao, T., Zhan, Y., & Cui, W. (2015). An Integrated Approach based on Markov Chain and Cellular Automata to Simulation of Urban Land Use Changes. *Applied Mathematics & Information Sciences*, 9(2), 769.

Huang, J., Zhan, J., Yan, H., Wu, F., & Deng, X. (2013). Evaluation of the impacts of land use on water quality: a case study in the Chaohu Lake basin. *The Scientific World Journal*, Vol.13.

Huigen, M. (2003). *Agent Based Modeling in Land-Use and Land-Cover Change Studies.* Laxenbury, Austria: International Institute for Applied System Analysis.

Iacono, M., Levinson, D., El-Geneidy, A. and Wasfi, R. (2012). *A Markov Chain Model of Land Use Change in the Twin Cities, 1958-2005.* St. Louis: Federal Reserve Bank of St Louis.

Irwin, E. and Bockstael, N. (2002). Interacting agents, spatial externalities, and the endogenous evolution of residential land-use pattern. *Journal of Economic Geography* , 2, 31-54.

Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11(5), 352-359.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, Vol. 14, No. 2, 1137-1145.

Kuhn, M. Contributions from Wing, J.,Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt, T. (2018). *caret: Classification and Regression Training. R package version 6.0-78.* Retrieved from https://CRAN.R-project.org/package=caret

Lambin, E. F., Geist, H. and Rindfuss, R. R. (2006). Introduction: Local Processes with Global Impacts. In E. F. Lambin, *Land-use and land-cover change: Local Processes and global impacts* (pp. 1-8). Berlin: Springer.

Landis, J. D. (1994). The California Urban Futures Model: a new-generation of metropolitan simulation-models. *Environment and Planning B*, 21(4), 399-420.

Landis, J. D. and Zhang, M. (1998a). The second generation of the California urban futures model. Part 1: model logic and theory. *Environment and Planning A*, 25, 657-666.

Landis, J. D. and Zhang, M. (1998b). The second generation of the California urban futures model. Part 2: specification and calibration results of the land-use change submode. *Environment and Planning A*, 25, 795-824.

Lee, J., Park, M., & Kim, Y. (1999). An application of canonical correlation analysis technique to land cover classification of LANDSAT images. *ETRI* , 21(4), 41-51.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.

Lin, Y. P., Chu, H. J., Wu, C. F., & Verburg, P. H. (2011). Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling - a case study. *International Journal of Geographical Information Science*, 25(1), 65-87.

Lipowski, A. and Lipowska, D. (2012). Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications*, 391(6), 2193-2196.

Liu, J., Feng, Q., Gong, J., Zhou, J., & Li, Y. (2016). Land-cover classification of the Yellow River Delta wetland based on multiple end-member spectral mixture analysis and a Random Forest classifier. *International Journal of Remote Sensing*, 37(8), 1845-1867.

Maier, H. (2007, October 9). *Chapter 2: Pavement Selection Strategies in Long-life Concrete Pavements in Europe and Canada*. Retrieved 01 30, 2018, from Office of International Programs: https://international.fhwa.dot.gov/pubs/pl07027/llcp_07_02.cfm

Maser, S., Riker, W., & Rosett, R. (1977). The effects of zoning and externalities on the price of land: An empirical analysis of Monroe County, New York. *The Journal of Law and Economics*, 20(1), 111-132.

Matthews, R. B., Gilbert, N. G., Roach, A., Phlhill, J. G. and Gotts, N. M. (2007). Agent-based land-use models: a review of applications. *Landscape Ecology*, 22(10), 1447-1459.

Mertens, B. and Lambin, E. F. (1997). Spatial modelling of deforestation in Southern Cameroon: spatial disaggregation of diverse deforestation processes. *Applied Geography*, 17, 143-168.

Meyer, W. B., & Turner II, B. L. (1992). Human population growth and global land-use/cover change. *Annual review of ecology and systematics*, 23(1), 39-61.

Ministry of Municipal Affairs. (2017). *Growth Plan for the Greater Golden Horseshoe.* Queen's Printer for Ontario.

Muller, D. and Zeller, M. (2002). Land use dynamics in the central highlands of Vietnam: a spatial model combining village survey data with satellite imagery interpretation. *Agricultural Economics*, 27, 333-354.

Muller, M. R. and Middleton, J. (1994). A Markov model of land-use change dynamics in the Niagara Region, Ontario, Canada. *Landscape Ecology*, 9(2), 151-157.

Murray-Rust, D., Brown, C., van Vliet, J., Alam, S. J., Robinson, D. T., Verburg, P. H. and Rounsevell, M. (2014). Combining agent functional types, capitals and services to model land use dynamics. *Environmental modelling & software*, 59, 187-201.

National Research Council (NRC). (2014). *Advancing land change modelling: opportunities and research requirements.* Washington, D.C., USA: National Academies Press.

Natural Resources Canada. (2015, Nov. 20). *Land Cover & Land Use*. Retrieved from Natural Resources of Canada: https://www.nrcan.gc.ca/node/9373

Nelson, E., Sander, H., Hawthorne, P., Conte, M., Ennaanay, D., Wolny, S., et al. (2010). Projecting global land-use change and its effect on ecosystem service provision and biodiversity with simple models. *PLos One*, 5(12), e14327.

Nguyen, M. H. and Ho, T. V. (2016). An agent-based model for simulation of traffic network status: Applied to Hanoi city. *Simulation*, 92(11), 999-1012.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., & Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the association of American Geographers*, 93(2), 314-337.

Pereira, H. M., Navarro, L. M. and Martins, I. S. (2012). Global biodiversity change: the bad, the good, and the unknown. *Annual Review of Environment and Resource*, 37,25-50.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2007). Linear and nonlinear mixed effects models. *R package version*, 3(57), 1-89.

Pontius Jr, R., & Spencer, J. (2005). Uncertainty in extrapolations of predictive land-change models. *Environment and Planning B: Planning and design*, 32(2), 211-230.

Quintas-Soriano, C., Castro, A. J., Castro, H. and García-Llorente, M. (2016). Impacts of land use change on ecosystem services and implications for human well-being in Spanish drylands. *Land Use Policy*, 54, 434-548.

R Core Team. (2017). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* Retrieved from https://www.R-project.org/

Region of Waterloo. (2003). *Planning Our Future: Regional Growth Management Strategy*. Retrieved from Region of Waterloo: https://www.regionofwaterloo.ca/en/resources/RegionalGrowthManagementStrategy.pdf

Region of Waterloo. (2011). *Census Bulletin: Agriculture*. Retrieved from Region of Waterloo: https://www.regionofwaterloo.ca/en/resources/Census/CensusBulletin-FinalAgri.pdf

Region of Waterloo. (2011). *Census Bulletin: Population and Dwelling Counts*. Retrieved from Region of Waterloo: https://www.regionofwaterloo.ca/en/resources/Census/CensusBulletin-PopDwellFINAL.pdf

Region of Waterloo. (2018). *2017 Building Permit Activity and Growth Monitoring.* Planning, Developmen and Legislative Services, Community Planning. Region of Waterloo: File Code: D07-40(A).

Robinson, D., Murray-Rust, D., Rieser, V., Milicic, V., & Rounsevell, M. (2012). Modelling the impacts of land system dynamics on human well-being: Using an agent-based approach to cope with data limitations in Koper, Slovenia. *Computers, Environment and Urban Systems*, 36(2), 164-176.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classificaiton. *Journal of Photogrammetry and Remote Sensing*, 67, 93-104.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12-24.

Schelling, T. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2), 143-186.

Schelling, T. C. (1969). Models of segregation. *The American Economic Review*, 59(2), 488-493.

Serneels, S., & Lambin, E. F. (2001). Proximate causes of land-use change in Narok District, Kenya: a spatial statistical model. *Agriculture, Ecosystems & Environment*, 85(1-3), 65-81.

Shi, P., Yuan, Y., Zheng, J., Wang, J., Ge, Y., & Qiu, G. (2007). The effect of land use/land cover change on surface runoff in Shenzhen region, China. *Catena*, 69(1), 31-35.

Smith, A. K. (2017). *An evaluation of high-resolution land cover and land use classification accuracy by thematic, spatial, and algorithm parameters (Master's thesis)*. Retrieved from UWSpace: http://hdl.handle.net/10012/12506

Stokey, E. and Zeckhauser, R. (1978). *A primer for policy analysis.* New York: W. W. Norton.

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3), 293-300.

Tang, J., Wang, L., & Yao, Z. (2007). Spatio-temporal urban landscape change analysis using the Markov chain model and a modified genetic algorithm. *International Jounal of Remote Sensing*, 28(15), 3255-3271.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 5(1), 35-39.

Therneau, T. (2015). A Package for Survival Analysis in S. R package version 2.38. https://CRAN.R-project.org/package=survival.

Trenberth, K., Jones, P., Ambenje, P., Bojariu, R., Easerling, D., Tank, A., et al. (2007). Observations: Surface and Atmospheric Cliamte Change. In S. Solomon, D. Qin, M. Manning, M. Marquis, K. Averyt, M. Tignor, et al., *Climate Change 2007 The Physical Science Basis* (pp. 239-336). Cambridge University Press.

Trexler, J. C. and Travis, J. (1993). Nontraditional regression analyses. *Ecology*, 74(6), 1629-1637.

Turner, M. G. (1987). Spatial simulation of landscape changes in Georgia: a comparison of 3 transition models. *Landscape Ecology*, 1(1), 29-36.

Veldkamp, A. and Lambin, E.F. (2001). Predicting land-use change. *Agriculture, Ecosystems and Environment*, 85(1), 1-6.

Veldkamp, A., & Fresco, L. O. (1996a). CLUE: a conceptual model to study the conversion of land use and its effects. *Ecological Modelling*, 85" 253-270.

Verburg, P. H., Kok, K., Pontius R. G. J. and Veldkamp, A. (2006). Modeling Land-Use and Land-Cover Change. In E. G. Lambin, *Land-use and land-cover change: Local processes and global impacts* (pp. 117-135). Berlin: Springer.

Verburg, P. H., Schot, P. P., Dijst, M. J. and Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *GeoJournal*, 61(4), 309-324.

Verburg, P., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., & Mastura, S. (2002). Modelling the spatial dynamics of regional land use: the CLUE-S model. *Environmental management*, 30(3), 391-405.

Waldrop, M. (1990). Asking for the moon; the moon-Mars initiative may or may not fly on Capitol Hill, but NASA wants the scientists on its side. *Science*, 247(4943), 637-639.

Wang, N., Brown, D.G., An, L., Yang, S. and Ligmann-Zielinska, A. (2013). Comparative performance of logistic regression and survival analysis for detecting spatial predictors of land-use change. *International Journal of Geographical Information Science*, 27:10, 1960-1982.

Watson, R., Noble, I., Bolin, B., Ravindranath, N., Verardo, D., & Dokken, D. (2000). Global Perspective. In R. T. Watson, I. R. Noble, B. Bolin, N. H. Ravindranath, D. J. Verardo, & D. J. Dokken, *Land Use, Land-Use Change, and Forestry* (pp. 23-51). Cambridge: Cambridge University Press.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95-114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673-686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1): 3-36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R (2nd edition).* Chapman and Hall/CRC.

Wood, S. N., Pya, N. and Saefken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111, 1548-1575.

Yeandle, M. (September 2017). *The Global Financial Centre Index22.* Long Finance.

Zeng, Y. N., Wu, G. P., Zhan, F. B. and Zhang, H. H. (2008). Modeling spatial land use pattern using autologistic regression. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37(B2), 115-118.

Appendices

Appendix A – Data Processing, Variable Creation and Problems with the Data

The need of a manual classification arose when misclassification found in computer simulated 2010's LU raster data created by Alexander Smith in 2016. The manual classification of 2010's LU was conducted for the Region of Waterloo with the support of 2010's SWOOP data and ownership parcels data. The LU raster data were extracted to ownership parcels based on the main features (e.g., grass, house and road) and functions (e.g., agricultural and residential) of the land within parcels. The rules of manual LU classification were the result of discussion between us, which can be found in Smith (2017).

After the manual classification, three small areas in the Region of Waterloo were randomly chosen and compared with the classification result from computer simulated LU data. The result of comparison has achieved an overall accuracy of 90 percent. Later, an overall accuracy of the computer simulated 2010 LU data in the whole study area was computed using manually classified LU data as reference, which is about 88 percent. Thus, the classification of computer simulated 2010 LU data was considered satisfactory. Similarly, computer simulated 2006 and 2015 LU data were also considered satisfactory since all LU data were all classified using the same rules, methods and technology. Therefore, computer simulated LU data for 2006, 2010 and 2015 were used as ground truth data toward modelling LU changes with the proposed statistical methods.

Before creating any variables, DAUIDs (i.e., unique IDs of DAs) in the Region of Waterloo in 2010s were assigned to parcels according to the location of parcels in the DA in order to relate DA's information to parcels. Geometries (i.e., perimeter and area) of the parcel polygons and DA polygons were calculated in ArcGIS and attached to the parcel data. Geographic variables (i.e., mean slope and mean DEM) were created using slope and DEM data. Demographic data from 2010 Census data (e.g., population) associated with the DA in the Region of Waterloo retrieved from Statistics Canada were merged with the parcel data in excel format in R by the common variable DAUID. The parcel data were then imported back to ArcGIS since spatial variables need to be created using the Polygon Neighbors tool in ArcGIS. The tool created a table that contains IDs of source polygons, IDs of neighbor polygons and records of LU types for neighbor polygons. The table was then used to calculate the percentage of each LU type around each source polygon by manually programmed code in R. The percentage of neighbor LU was merged with the parcel data by unique parcel ID (i.e., ID of source polygon) in R. Furthermore, the parcel data were converted to point data (i.e., centroids of parcels) in order to calculate Euclidean distances from parcel centroids to other features (e.g., highway ramp and commercial parcel) in ArcGIS. During the process of variable creation, some parcels have been removed from the full dataset due to the lack of data for creating drivers associated with the parcels (e.g., a lack of census data in some areas).

Unfortunately, misclassification was found in some rare cases of LU change during the process of creating status for LU change (i.e., binary response variables). Since rare cases of LUCC (e.g., from industrial to water) are a relatively small amount of data compared to the total, second round of manual classification was conducted by myself to increase the classification accuracy of these cases. During this round, SWOOP data for all three years (i.e., 2006, 2010 and 2015) were used to classify LU types that associated with parcels being found with the occurrence of rare cases of LU change. The manual classification was also done to parcels that found with the occurrence of some other LU changes that have a relatively small amount of data compared to the total due to the consideration of data accuracy. Meanwhile, some rules of manual LU classification have been modified based on Smith's work. The complete and modified rules are presented in Appendix B. Moreover, spatial variables were re-created using 2010 as the reference year since the neighborhood parcels may be changed.

Furthermore, some other problems have raised. As mentioned in Section 2.2.4 in Chapter 2, each full dataset is consisted of all parcels that had their LU converted to a specific LU in 2015 and parcels that had remained the specific LU during the study period. After conducting an explanatory analysis on full datasets, it has been found that the majority types of LU change in a full dataset came from the part of parcels that has not been manually verified or modified in the second round. Therefore, it is reasonably to suspect the accuracy of the data. However, the study has moved forward with the current set of data due to the high overall accuracy of computer simulated LU data and the time constraint of this project. Another finding is that the classification accuracy of computer simulated 2010 LU is about 68 percent for the approximately 6000 parcels that have been verified or modified in the second round of manual classification. This infers that the computer classification approach performed differently for different LU types.

In addition, the season that SWOOP data have been taken is another factor that can cause misclassification of the same LU in different years by computer simulation methods.

Appendix B – Rules of Manual Land Use Classification

Table B-1: Manual land use classification of parcels in the Region of Waterloo.

#	Name	Classification description based on perceived uses and services
1	Low-Density Residential	Parcels which appear to contain a single dwelling for a single family on a large property. These parcels typically appear outside the urban core in suburbs or rural areas. While houses tend to be larger than medium density residential, it is not a requirement for the classification.
2	Medium-Density Residential	Average sized parcels containing a single dwelling for a single family, which may or may not be attached to adjacent dwellings. This class contains the majority of residential parcels within subdivisions and the urban core. In most parcels, the house and driveway cover most or all of the width of the parcels, with yards in the front and back. Townhouses are usually classified as medium-density residential.[1]
3	High-Density Residential	Parcels containing buildings with multiple dwellings or units, and therefore multiple families within the parcel. Typically in two forms, apartment or condo buildings, and townhouses where one parcel contains multiple units. Parcels may contain green space and parking lots in addition to the buildings.
4	Commercial	Parcels containing business where customers visit to obtain products and services, or office buildings which may not receive customers. Larger parcels, such as malls or box stores, will contain large parking lots for customers. These parcels do not contain large outdoor storage areas, although garden and home improvement stores may have some outdoor storage.
5	Industrial	Parcels which contain a business with an outdoor storage area such as a factory or a car scrapyard. These business typically do not receive customers although there may be parking lots for employees and areas for incoming materials and outgoing products.
6	Institutional	Manually classified parcels for schools (private and public) and hospitals. Schools and hospitals can appear as a variety of classes but provide different services from these misclassifications (e.g.

[1] This is an additional clarification to the rule of manual land use classification.

		Commercial or Protected Areas and Recreation). Manually classifying these parcels allows for them to be included in the landscape without large amounts of misclassification.
7	Transportation	Parcels which represent roads and railways. These parcels often include the boulevard and sidewalks. Highway interchange parcels include all the land which is owned and managed by the managing government.
8	Protected Areas and Recreation	Areas which have a primary purpose of recreation, such as parks, or protected areas such as forests. Commercial forests and private forests are included in this class as they appear very similar, or even identical to the natural forests.
9	Agriculture	Parcels which are primarily used for raw food production. This includes fields for crops and pastures. Some parcels will have barns and/or a farm house, while others may have neither. Parcels may also include a portion which is forested, sometimes referred to as “the back forty”.
10	Water	Parcels which have a main purpose of outlining waterbodies such as rivers. Lakes are included when the lake occupies a majority of the parcel. The rest of the parcel may include sections which would otherwise be classified as Protected Areas and Recreation.
11	Under Development	Properties where construction has not been completed and no residents or business has moved in. These parcels may become many different classes when complete, but the class cannot be guaranteed at the time of the imagery. Depending on the progress of a development project, residential areas and big box stores or shopping complexes may appear similar as the area is represented by only a single parcel.

Table B-2: Clarifications between similar land use classes.

First Class	Second Class	Problem	Solution
Low-Density Residential	Medium-Density Residential	Parcel size is a continuous variable and it is difficult to define the exact separation between the two classes.	In many cases where there is confusion, the house is the same size as the surrounding properties which are either low or medium density and is a similar distance from the road. The parcel in question will usually have its additional size added through its backyard. If the backyard visually occupies two thirds of the property, it can be easily called low density, if less, medium density. If the parcel has a backyard smaller than two thirds, but the front yard and house are large, then it can also be classified as low density. If an absolute value of size is needed, 2000m2 should be used as the minimum size for Low Density Residential.
Low-Density Residential	Protected Area and Recreation	Household in a large parcel is surrounded by forest or green land with no appearance of backyard/garden. Sometimes the parcel could contain a small portion of backyard/garden relative to the total of the parcel.[2]	Even the size of the house and the maintained portion of the property is very small compared to the area of the forest, the parcel should be classified as low-density residential.[3]
Medium-Density Residential	Under Development	A house is visible in the parcel that is under development	If there is a completed house with grass on the property it should be considered complete and classified as Medium Density Residential. If the house does not appear complete or there is no grass where there should be, it should be classified as Under Development.

[2] This is an additional clarification to the problem.

[3] This is the change in the rule of the original manual land use classification.

Table B-3: Exemptions and special cases in land use classification.

Example	Class	Reasoning
Airport	Commercial	Airports provide services similar to Commercial parcels, where people are constantly visiting the parcel. Visually they are similar as they both include large paved areas such as parking lots and a large building.
Fire station	Commercial	Although functionally different from Commercial parcels, they are very similar in the imagery.
Graveyard	Protected Areas and Recreation	Graveyards and cemeteries are visually similar to parks, where there are paths for people to walk and grass fields. The only visual difference is that there are pieces of stone (headstones) scattered across the fields and there is no sports equipment.
Water Tower	Protected Areas and Recreation	Water towers can be visually similar to parks as they can have large grassy areas surrounding the tower. If the water tower is in a parcel without much grassed area, it may be classified as Commercial instead.
Commercial Forest – Post-Harvest	Various	If the harvested forest appears to be converted into agriculture, classify as Agriculture. If it shows signs of urban development, it should be classified as Under Development. If it appears to be replanted and is still being used as a commercial forest, classify as Protected Areas and Recreation.
Catwalk	Transportation	The paths between houses, or catwalks, are similar to roads, although a little smaller. A path through a park or green space would not be considered transportation.
Walking paths	Protected Areas and Recreation	Walking paths in the area can often be found under large electrical transmission lines. The transmission lines and towers account for a small portion of the parcel, and therefore simply appear as grassy corridors through subdivisions, similar to parks.
Church	Commercial	Churches are visibly similar to Commercial parcels because they are a building which has a parking lot and

		some property. Functionally they are also similar as people will visit a church for a relatively short period of time, similar to a business.
Artifacts	N/A	The parcel data is not perfect and has artifacts from either previous versions, or mistakes during creation. Some artifacts have little impact on the data, while others have large impacts. The most frequent example is a single parcel being divided into multiple parcels by the artifacts.
Artifacts – Splits	N/A	When a parcel is divided by artifacts all segments should be classified as the original type if suitable. If a segment can clearly be classified as another land use type it should be done. For example, if a Low Density Residential parcel is divided into three pieces, two covering the house and one covering a forest at the back of the property, the two on the house should be Low Density Residential and the one on the forest should be Protected Areas and Recreation.
Artifacts – Slivers	N/A	Another form of artifact is a sliver. These sliver parcels are very thin and long. Examples can be a few centimeters wide but almost a kilometer long. Sliver parcels should be ignored and not classified if noticed.
Mixed Parcels	N/A	Occasionally parcels will contain multiple land use types other than the previously mentioned scenarios. For example a parcel may contain a house and land on one half and part of a waterbody on the other half. In these scenarios where there is no clear majority of land use type the following order of priority should be used: Medium Density Residential > High Density Residential > Low Density Residential > Commercial > Industrial > Institution > Transportation > Under Development > Agriculture > Protected Areas and Recreation > Water
Future Development	N/A	In the scenarios where parcels have been created but no development has begun, classify the parcel based on the

	currently present land use type. If the imagery shows evidence of development, then classify as Under Development.

Note: The original tables were created by Smith (2017) and can be found in the Appendix section in his thesis paper. Footnotes are used to indicate modifications made to the original content.

Appendix C – Predictors of Land-Use and Land-Cover Change

Table C-1: Names and Description of Predictors

Name	Description (unit)
lu2015	2015 land-use of parcels in the Region of Waterloo
lu2010	2010 land-use of parcels in the Region of Waterloo
lu2006	2006 land-use of parcels in the Region of Waterloo
lc2010	2010 land-cover of parcels in the Region of Waterloo
lc2006	2006 land-cover of parcels in the Region of Waterloo
DA_Area	Area of a DA (km2)
Parcel_Area	Area of a parcel (km2)
MeanDEM	Mean DEM of a parcel (km)
MeanSlope	Mean slope of a parcel
Ramp_dist	Distance from the centroid of a parcel to the nearest fixed highway ramp (km)
River_dist	Distance from the centroid of a parcel to the nearest river (km)
Water_dist	Distance from the centroid of a parcel to the nearest water body (km)
Wood_dist	Distance from the centroid of a parcel nearest wooded area (km)
LRoad_dist	Distance from the centroid of a parcel to the nearest local road (km)
MRoad_dist	Distance from the centroid of a parcel to the nearest main road (km)
lu4_dist	Distance from the centroid of a parcel to the nearest commercial parcel (km)
lu5_dist	Distance from the centroid of a parcel to the nearest industrial parcel (km)
lu8_dist	Distance from the centroid of a parcel to the nearest protected area/recreational parcel (km)

lu9_dist	Distance from the centroid of a parcel to the nearest agricultural parcel (km)
DA_population_density	Population density in a DA (population in DA/DA area) in 2010 (person/km2)
Residential_population_density	Residential population density in a DA (population in DA/total residential areas in DA) in 2010 (person/km2)
Change_Population	The rate of change of population from 2006 to 2011 based on the DA a parcel resides
Change_AveIncome	The rate of change of average income from 2006 to 2011 based on the DA a parcel resides
F_lu1	Proportion of low-density residential parcels around a parcel
F_lu2	Proportion of median-density residential parcels around a parcel
F_lu3	Proportion of high-density residential parcels around a parcel
F_lu4	Proportion of commercial parcels around a parcel
F_lu5	Proportion of industrial parcels around a parcel
F_lu6	Proportion of institution parcels around a parcel
F_lu7	Proportion of transportation parcels around a parcel
F_lu8	Proportion of protected area/recreation parcels around a parcel
F_lu9	Proportion of agricultural parcels around a parcel
F_lu10	Proportion of water parcels around a parcel
F_lu11	Proportion of developing parcels around a parcel

Note: Variables listed in this table are the variables actually being used to construct models in this study. Some variables have been created were excluded from model building since they are highly correlated with some variables listed in this table.

Appendix D – Additional Analysis Results

Table D-1: Running time of methods with 10-fold CCV by land-use type.

Model / LUC (To)	Running Time (second)							
	MC		LR		GAM		SA	
	FB	RB	FB	RB	FB	RB	FB	RB
LDR	n/a	n/a	2.93	2.93	152.06	152.06	1.14	1.14
MDR	n/a	n/a	6.21	3.56	2468.95	2816.7	4.56	1.24
HDR	n/a	n/a	2.92	1.02	282.7	199.04	1.41	1.2
COM	n/a	n/a	1.55	1.03	1127.83	3408.59	2	1.27
IND	n/a	n/a	0.83	0.83	5	5	1.16	1.16
INS	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
TRANS	n/a	n/a	1.07	0.98	311.44	229.72	1.5	1.25
REC	n/a	n/a	0.89	0.89	17.28	17.28	1.14	1.14
AGR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
WAT	n/a	n/a	0.65	0.65	4.42	4.42	0.79	0.79
UD	n/a	n/a	1.06	1.03	779.71	434.11	1.47	1.24
Total	17.42	8.7	18.11	12.92	5149.39	7266.92	15.17	10.43

Table D-2: Absolute difference between averaged 10-fold CCV accuracies and overall accuracies of FB and RB training datasets in percentage (%).

Model / LUC (To)	Difference in averaged accuracy (%)			
	MC	LR	GAM	SA
LDR	10.24	0	0	0
MDR	38.39	1.18	5.04	0.38
HDR	0.07	1.22	6.98	1.38
COM	0.66	0.91	1.75	2.20
IND	5.43	0	0	0
INS	52.5	n/a	n/a	n/a
TRANS	0.95	3.53	3.36	3.05
REC	12.68	0	0	0
AGR	2.6	n/a	n/a	n/a
WAT	10	n/a	n/a	n/a
UND	15.2	0.80	5.10	3.54
Overall	4.72	0.73	2.78	1.32

Table D-3: Absolute difference between overall accuracies of final models derived from 10-fold CCV with FB and RB test datasets in percentage (%).

Model / LUC (To)	Difference in overall accuracy (%)		
	LR	GAM	SA
LDR	0	0	0
MDR	2.20	4.48	0.11
HDR	1.89	1.57	0.68
COM	4.85	1.98	2.01
IND	0	0	0
INS	n/a	n/a	n/a
TRANS	2.83	3.98	2.46
REC	0	0	0
AGR	n/a	n/a	n/a
WAT	n/a	n/a	n/a
UD	1.53	3.87	6.28
Overall	0.02	1.98	0.94

Table D-4: Absolute difference between averaged 10-fold SCV accuracies and overall accuracies of FB and RB training datasets in percentage (%).

Model LUC (To)	Difference in averaged accuracy (%)			
	MC	LR	GAM	SA
LDR	15.39	0	0	0
MDR	40.15	0.29	6.09	1.10
HDR	0.29	2.78	4.27	3.17
COM	2.95	0.90	9.70	1.29
IND	0.01	0	0	0
INS	0	n/a	n/a	n/a
TRANS	2.5	0.14	3.72	2.86
REC	17.79	0	0	0
AGR	6.65	n/a	n/a	n/a
WAT	22.13	n/a	n/a	n/a
UND	17.39	1.89	0.34	1.36
Overall	0.01	0.55	1.86	0.19

Table D-5: Absolute difference between overall accuracies of final models derived from 10-fold SCV with FB and RB test datasets in percentage (%).

Model / LUC (To)	Difference in overall accuracy (%)		
	LR	GAM	SA
LDR	0	0	0
MDR	1.47	0.73	0.11
HDR	2.28	1.68	2.67
COM	1.06	8.43	0.95
IND	0	0	0
INS	n/a	n/a	n/a
TRANS	2.58	0.65	1.70
REC	0	0	0
AGR	n/a	n/a	n/a
WAT	n/a	n/a	n/a
UD	0.63	3.08	0.72
Overall	0.64	1.40	0.56

Appendix E – Coefficients of Significant Land-use Change Predictors

In the following context, a smoothed term refers to a variable that was fit using a smoothing function to represent the non-linear relationship between it and the response variable in GAM.

Table E-1: Coefficients of significant LU change predictors in final LR derived from CCV with RB test datasets.

Method	Final LR							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.21					4.04	3.37	
lu2006_3	2.63		-1.36					
lu2006_4	1.11			-2.51				
lu2006_5			-1.13					-3.57
lu2006_7								-2.71
lu2006_8	1.53		2.75					-1.96
lu2006_9								-2.07
lu2006_11								-3.24
lc2006_2	1.82	-4.23						
lc2006_3		-2.99						
lc2006_5		-3.74						
lc2006_6								-5.08
lc2006_7		-3.77						
lc2006_8	2.32	-2.74						
lc2010_2		1.69	-1.42					1.47
lc2010_3		3.47				-1.28		-1.03
lc2010_5	1.06	2.33				0.97	4.40	
lc2010_6							4.45	
lc2010_7	2.62	5.01					6.51	-2.65
lc2010_8	2.89					1.91		2.39
ParcelArea			-48.15	-209.63		-17.55	-41.70	12.70
DA_Area				-54.93				
MeanSlope								-0.59
Wood_dist			-0.69					-0.67
River_dist							2.11	
LRoad_dist		9.73		-6.63				
MRoad_dist	0.67							-1.42
Ramp_dist		0.17						
lu4_dist	-0.41	-6.01	-2.72	19.15		-1.16		-3.04
lu5_dist								
lu8_dist	2.03					-3.44	15.49	
lu9_dist			-0.27					
Residential_ Popn_Density		0.03						
DA_Popn_ Density			$-1.43\text{x}10^{-4}$					$2.42\ \text{x}10^{-4}$
Change_ AveIncome				-1.02				
F_lu1								3.10
F_lu2		-1.37						
F_lu7		-2.56						
F_lu9				-2.69				
F_lu11								-1.77

Table E-2: Coefficients of significant LU change predictors in final LR derived from CCV with FB test datasets.

Method Predictor	Final LR							
	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.21	-4.43				2.71	3.37	
lu2006_3	2.63	-2.59	-1.49					
lu2006_4	1.11	-2.93	-0.52	-2.47		1.78		
lu2006_5			-1.27					-3.69
lu2006_7		-1.81				-1.29		-3.55
lu2006_8	1.53							-2.80
lu2006_9				1.75				-2.18
lu2006_11		-2.07	-1.05					-3.39
lc2006_2	1.82	-1.65						
lc2006_3		-1.50						
lc2006_4		-1.17						
lc2006_5		-1.56						
lc2006_6				-2.16				
lc2006_7		-1.92						
lc2006_8	2.32	-1.52	-1.74					
lc2010_2		0.78						2.06
lc2010_3		2.39		-0.98		-1.85		-0.90
lc2010_5	1.06	1.97		0.61		0.52	4.40	-1.47
lc2010_6		-1.05					4.45	
lc2010_7	2.62	3.91		0.61			6.51	-2.86
lc2010_8	2.89	1.21				2.09		1.51
ParcelArea			-0.01	-48.91		-40.99	-41.70	5.19
DA_Area				-28.73				
MeanSlope								-3.52
MeanDEM		-3.54	5.94			-9.62		-9.18
Wood_dist		0.33	-0.70					-0.72
River_dist							2.11	
Water_dist		0.32						0.52
LRoad_dist		6.47		-9.23		2.41		
MRoad_dist	0.68			1.26				
Ramp_dist						0.08		
lu4_dist	-0.41	-10.42	-2.78	8.69		-0.86		-1.33
lu5_dist								
lu8_dist	2.03	-1.71	1.05			-3.80	15.49	
lu9_dist			-0.22	0.18				
Residential_ Popn_Density	-0.04		-0.04	0.03				
DA_Popn_ Density		-1.31x10^{-4}	-1.49x10^{-4}					2.74x10^{-4}
F_lu1								2.48
F_lu2								1.46
F_lu7		0.83						1.25
F_lu8			3.94			3.00		
F_lu9								2.30
F_lu11		0.84				0.76		

Table E-3: Coefficients of significant LU change predictors in final GAM derived from CCV with RB test datasets.

Method	Final GAM							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.22					4.56		2.16
lu2006_3	2.67		-1.52					
lu2006_4	0.98		-0.84	-3.79				
lu2006_5			-1.51					
lu2006_8	1.54							
lu2006_9								-1.39
lu2006_11								-2.07
lc2006_2	1.76							2.48
lc2006_3								3.54
lc2006_5								2.59
lc2006_7	0.66							2.18
lc2006_8	2.61							2.28
lc2010_2			-2.18	4.85				
lc2010_3		3.28	-2.05	3.80				-2.41
lc2010_5	0.96	2.85	-0.82	3.52		1.45		-1.25
lc2010_6						2.14		
lc2010_7	2.66	4.51		5.52		1.76		-3.69
lc2010_8	2.37	-2.34		4.51				
ParcelArea			S	S		S		S
DA_Area			S	S				
MeanSlope								S
MeanDEM			S					S
Wood_dist			S					S
River_dist								S
LRoad_dist	S			-9.98		S		
MRoad_dist	S					S		S
Ramp_dist		S						
lu4_dist		S	S	S		S		S
lu5_dist				S				
lu8_dist						S		S
Residential_Popn_Density	S		S	S				S
DA_Popn_Density								S
F_lu2		-1.81		-1.14				
F_lu11								-2.09
Note: The symbol “S” in the table indicates that the predictor is considered significant as a smoothed term.								

Table E-4: Coefficients of significant LU change predictors in final GAM derived from CCV with FB test datasets.

Method Predictor	Final GAM							
	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.22	-3.76		-1.24		2.08		
lu2006_3	2.67	-2.66	-1.65					
lu2006_4	0.98	-2.74	-1.05	-4.51				
lu2006_5			-1.81					-3.68
lu2006_7				-1.46		-2.00		-3.99
lu2006_8	1.54							-3.54
lu2006_9			-3.44					-3.52
lu2006_11			-1.32	-2.00				-4.41
lc2006_2	1.76	-2.11						
lc2006_3		-1.94						
lc2006_4		-1.87						
lc2006_5		-1.88						
lc2006_7	0.66	-1.93						
lc2006_8	2.61	-1.92						
lc2010_2		-0.41	-2.23	2.28				
lc2010_3		1.17	-1.09	1.09		-0.98		-2.53
lc2010_5	0.96	1.90		1.18		0.82		-2.63
lc2010_7	2.66	4.32		2.65		0.88		-4.76
lc2010_8	2.37		-1.20	2.59		1.61		
ParcelArea			-6.83	S		S		S
MeanSlope		S						
MeanDEM			S			S		S
Wood_dist		S	S					S
River_dist								S
Water_dist		S		S				
LRoad_dist	S	S		S		S		
MRoad_dist	S			S		S		S
Ramp_dist		S		S				S
lu4_dist		S	S	S		S		S
lu5_dist				S				
lu8_dist		S				S		S
lu9_dist			S	S				
Residential_Popn_Density	S		S					
DA_Popn_Density		S	S					
Change_AveIncome			S	S				
Change_Popn		S						S
F_lu2								S
F_lu11		0.69						
Note: The symbol "S" in the table indicates that the predictor is considered significant as a smoothed term.								

Table E-5: Coefficients of significant LU change predictors in final SA derived from CCV with RB test datasets.

Method	Final SA							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	0.79	-2.09				0.91	1.87	-0.67
lu2006_3	1.36		-0.85					
lu2006_4	0.61	-1.04		-1.27				
lu2006_5			-1.08					-2.58
lu2006_7			0.88		2.60	-0.99	1.30	
lu2006_8	0.80		0.92					-0.66
lu2006_9	1.56				1.90		1.01	-2.11
lu2006_11	0.54							-1.96
lc2006_2	1.15				2.77			
lc2006_3					2.51			
lc2006_5					2.32		-1.48	
lc2006_6								-1.63
lc2006_7							-1.16	
lc2006_8	0.85						-1.42	
lc2010_2		1.42	-1.19		1.40			
lc2010_3		2.42		-0.79		-0.76		-0.62
lc2010_4								
lc2010_5	0.58	1.86				0.62	2.43	
lc2010_6							1.88	
lc2010_7	1.01	2.12			-1.33		3.25	-0.91
lc2010_8	1.11	0.96	-1.00			0.57		
ParcelArea			-49.28	-1.05x10^{2}	6.42		-19.21	
MeanSlope			-0.13					
MeanDEM						-3.85		
Wood_dist			-0.35		0.78			
River_dist		-0.42					0.88	
MRoad_dist	0.26							-1.16
lu4_dist		-3.13	-1.22	1.66	-24.36	-1.10		0.47
lu5_dist								
lu8_dist	0.68					-1.36	3.07	
lu9_dist			-0.27					
Residential_Popn_Density			-0.02					
DA_Popn_Density								1.21x10^{-4}
F_lu7								0.81
F_lu9					0.90			

Table E-6: Coefficients of significant LU change predictors in final SA derived from CCV with FB test datasets.

Method	Final SA							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	0.79	-1.92		0.43			1.87	-0.95
lu2006_3	1.36	-0.60	-0.81					
lu2006_4	0.61	-0.73		-1.20				-1.04
lu2006_5			-0.76					-2.59
lu2006_7					2.60	-1.11	1.30	-2.39
lu2006_8	0.80							-1.94
lu2006_9	1.56			0.57	1.90		1.01	-1.17
lu2006_10		-0.46						
lu2006_11	0.54	-0.34						-2.02
lc2006_2	1.15	-0.64			2.77			
lc2006_3		-0.40			2.51			
lc2006_5		-0.52			2.32		-1.48	
lc2006_6		-0.92						-1.87
lc2006_7		-0.68					-1.16	
lc2006_8	0.85	-0.38					-1.42	
lc2010_2		0.80	-1.26		1.40			
lc2010_3		1.66		-0.65		-1.03		-0.70
lc2010_5	0.58	1.56		0.29		0.30	2.43	-0.68
lc2010_6		-1.07					1.88	
lc2010_7	1.01	1.85		0.29	-1.33		3.25	-2.01
lc2010_8	1.12	1.22				0.59		
ParcelArea			-75.41	-18.69	6.42	-23.15	-19.21	2.12
DA_Area				-26.90				
MeanSlope								-0.16
MeanDEM		-2.13		2.66		-4.14		
Wood_dist		0.28	-0.45		0.78			
River_dist							0.88	
Water_dist		-0.15						
LRoad_dist		0.576		-2.46				
MRoad_dist	0.26	-0.26		0.28				-0.80
Ramp_dist						0.04		
lu4_dist		-4.65	-1.90	1.47	-24.36			
lu5_dist								
lu8_dist	0.68	-0.59				-1.93	3.07	
lu9_dist								
Residential_Popn _Density		-0.02	-0.02	0.02				
DA_Popn_Density		$-0.65x10^{-4}$	$-1.16x10^{-4}$					$0.98x10^{-4}$
Change_AveIncome		-0.20						
Change_Popn		-0.03						
F_lu3						0.93		
F_lu7			0.54	0.49				
F_lu9					0.90			
F_lu11						0.37		

Table E-7: Coefficients of significant LU change predictors in final LR derived from SCV with RB test datasets.

Method	Final LR							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.49	-3.45			4.04		3.51	
lu2006_3	3.87		-1.30					
lu2006_4	1.74			-2.73				
lu2006_5			-1.11					-3.21
lu2006_7								-2.50
lu2006_8	1.57							-1.89
lu2006_9					1.79			-2.00
lu2006_11	1.49				1.34		1.66	-2.89
lc2006_6								-4.75
lc2006_8	1.06							
lc2010_2		1.34	-1.49	1.10				1.39
lc2010_3		3.49			-1.37			-1.16
lc2010_5	0.77	1.89			0.88		4.35	
lc2010_6	0.59						4.37	
lc2010_7	1.08	3.56		0.93			6.39	-2.72
lc2010_8	1.67			1.39				2.41
ParcelArea	24.30		-46.73	-45.64	-12.75		-3.88	12.38
MeanSlope	-0.21			0.41				-0.63
MeanDEM				9.70				
Wood_dist			-0.92					
River_dist							2.30	
LRoad_dist				-7.23				
MRoad_dist	0.76							
lu4_dist	-0.42	-4.87	-3.02	13.83	-38.12	-2.05		-3.27
lu5_dist					24.05			
lu8_dist						-3.06	15.36	
lu9_dist				0.33				
Residential_ Popn_Density						-0.04		
DA_Popn_ Density			-1.64x10^{-4}					2.44x10^{-4}
Change_ AveIncome				-1.19				
F_lu1								3.91
F_lu2	-1.52							
F_lu8				-3.91				
F_lu11								-1.85

Table E-8: Coefficients of significant LU change predictors in final LR derived from SCV with FB test datasets.

Method	Final LR							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	1.49	-4.36	-1.54			2.89	3.51	
lu2006_3	3.87	-2.43	-0.62			1.90		
lu2006_4	1.74	-2.80	-1.14	-2.43				
lu2006_5								-4.07
lu2006_7		-1.55				-1.27		-3.75
lu2006_8	1.57							-3.08
lu2006_9				2.23				-1.81
lu2006_11	1.49	-2.00	-1.17				1.66	-3.68
lc2006_2		-1.65	1.41					
lc2006_3		-1.42						
lc2006_4		-1.11						
lc2006_5		-1.50						
lc2006_6							4.35	
lc2006_7		-1.82					4.37	
lc2006_8	1.06	-1.38					6.39	
lc2010_2		0.81						2.00
lc2010_3		2.48		-0.98		-1.84	-3.88	-0.90
lc2010_5	0.77	2.04		0.54		0.53		-1.47
lc2010_6	0.59	-1.00		1.90				
lc2010_7	1.08	3.88		0.61				-2.69
lc2010_8	1.67	1.33				2.16	2.30	1.12
ParcelArea			-110.10	-60.91		-43.29		5.03
DA_Area	24.30			-30.89		30.50		
MeanSlope	-0.21							-0.34
MeanDEM						-9.38		-10.43
Wood_dist		0.37	-0.86				15.36	-1.12
Water_dist		-0.42						0.72
LRoad_dist		6.21		-1.00		2.67		
MRoad_dist	0.76			1.38				
Ramp_dist						0.07		
lu4_dist	-0.42	-10.28	-3.07	10.34	-38.12	-1.14		
lu5_dist		0.45			24.05			
lu8_dist		-1.79	1.46			-3.90		
lu9_dist				0.26				0.54
Residential_ Popn_Density		-0.03	-0.03	0.03				
DA_Popn_ Density		-1.32x10^{-4}	-1.62x10^{-4}					2.40x10^{-4}
Change_Popn		-0.18						
F_lu1		-0.77						1.77
F_lu2								1.24
F_lu7		0.78						
F_lu8			-4.02			2.79		
F_lu11						0.94		

Table E-9: Coefficients of significant LU change predictors in final GAM derived from SCV with RB test datasets.

Method	Final GAM							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2		-4.56	-1.45			5.00		
lu2006_3	2.90		-0.89					
lu2006_4	1.00	-4.81	-1.51	-3.95				
lu2006_5								-3.36
lu2006_7								-2.44
lu2006_8	1.74			-2.50				
lu2006_9								-2.55
lu2006_11	1.41							-3.38
lc2006_2	1.50							
lc2006_3	1.46							
lc2006_4								
lc2006_5	0.91							
lc2006_6								-5.49
lc2006_7								
lc2006_8	1.28							
lc2010_2			-2.06	3.08				
lc2010_3		2.68	-1.98	1.46				
lc2010_5	0.92	2.18	-0.68	1.63				-1.83
lc2010_6	0.58							-1.08
lc2010_7	1.11	4.27		2.66		1.96		-3.49
lc2010_8				3.12		2.89		2.18
ParcelArea			S			S		S
DA_Area	35.23		S					
MeanSlope								-0.64
MeanDEM			S					
Wood_dist			-2.01					
River_dist								S
Water_dist		S				S		
LRoad_dist	S					S		
MRoad_dist	S					S		
Ramp_dist								
lu4_dist	-0.39	S	S	S				S
lu5_dist								
lu8_dist				S		S		S
lu9_dist								
Residential_Popn_Density			S			S		S
DA_Popn_Density								S
Change_AveIncome				S				
Change_Popn								
F_lu1		S						
F_lu2								
F_lu11								

Note: The symbol “S” in the table indicates that the predictor is considered significant as a smoothed term.

Table E-10: Coefficients of significant LU change predictors in final GAM derived from SCV with FB test datasets.

Method	Final GAM							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2		-3.96	-1.87	-1.54				
lu2006_3	2.90	-3.13	-1.19					
lu2006_4	1.00	-2.91	-1.98	-4.79				
lu2006_5								-2.87
lu2006_7						-2.77		-3.32
lu2006_8	1.74							-2.28
lu2006_9			-3.52					-2.42
lu2006_11	1.41		-2.06	-2.32				-3.03
lc2006_2	1.50	-1.86						
lc2006_3	1.46	-1.58						
lc2006_4	0.91	-1.55						
lc2006_5		-1.64						
lc2006_7		-1.70						
lc2006_8	1.28	-1.60						
lc2010_2		-0.69		2.43				
lc2010_3		0.90		1.11		-1.02		-2.31
lc2010_5	0.92	1.80		1.17		0.78		-2.38
lc2010_7	0.58	4.11		2.59		0.78		-3.98
lc2010_8	1.11	-0.63	-1.22	2.75		1.58		
ParcelArea			-73.74	S		S		
DA_Area	35.23	S						
MeanSlope		S						
MeanDEM						S		
Wood_dist		S	S	S				S
Water_dist		S						
LRoad_dist	S	S		S		S		S
MRoad_dist	S			S		S		S
Ramp_dist				S				
lu4_dist	-0.39	S	S	S		S		S
lu5_dist				S				
lu8_dist		S				S		S
lu9_dist				S				
Residential_Popn_Density		S	S					
DA_Popn_Density			S	S				
Change_AveIncome			S	S				
Change_Popn		S		S				S
F_lu2								S
Note: The symbol “S” in the table indicates that the predictor is considered significant as a smoothed term.								

Table E-11: Coefficients of significant LU change predictors in final SA derived from SCV with RB test datasets.

Method	Final SA							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	0.90	-2.09			2.36	1.05	1.91	
lu2006_3	1.33		-0.75					
lu2006_4	0.55	-1.04		-1.29				
lu2006_5	1.30							-1.88
lu2006_6					7.33			
lu2006_7			1.10		2.99	-0.88	1.36	-1.47
lu2006_8	0.63		1.78					-1.00
lu2006_9	1.56		3.04		1.66		1.07	-1.13
lu2006_11	0.48						0.80	-1.47
lc2006_2	1.28				2.83		-1.72	
lc2006_3	0.96				2.44		-1.35	
lc2006_5					2.23		-1.58	
lc2006_7	0.45						-1.25	
lc2006_8	0.87				2.03		-1.57	
lc2010_2		1.42	-1.04					
lc2010_3		2.42		-0.77		-0.73	3.75	-0.49
lc2010_5	0.67	1.86				0.60	2.37	-0.39
lc2010_6	-0.50						1.83	
lc2010_7	1.24	2.12			-1.10		3.22	-1.82
lc2010_8	1.18	0.96						
ParcelArea	8.13		-39.48	-101.00	6.73		-20.11	6.14
DA_Area	1.56							
MeanSlope								-0.20
Wood_dist			-0.60		0.92			
Water_dist			0.26					
River_dist		-0.42					0.91	
LRoad_dist								0.84
MRoad_dist	0.23							-0.79
lu4_dist	-0.35	-3.13	-2.20	2.06	-25.33	-1.46		-1.29
lu5_dist				0.53	0.48			-1.18
lu8_dist	0.77				-1.69	-1.14	2.68	
DA_Popn_Density			-1.21×10^{-4}					
F_lu1				-1.70				
F_lu2					-0.39			
F_lu7	1.00							
F_lu11								-2.10

Table E-12: Coefficients of significant LU change predictors in final SA derived from SCV with FB test datasets.

Method	Final SA							
Predictor	LDR	MDR	HDR	COM	IND	TRA	REC	UND
lu2006_2	0.90	-1.92	-0.84		2.36		1.91	-1.16
lu2006_3	1.33	-0.60	-0.25					-1.15
lu2006_4	0.55	-0.73	-0.76	-1.37				-2.56
lu2006_5	1.30							-2.62
lu2006_6					7.33			-1.96
lu2006_7					2.99	-1.17	1.36	-0.99
lu2006_8	0.63							-2.08
lu2006_9	1.56			0.51	1.66		1.07	
lu2006_10		-0.46						
lu2006_11	0.48	-0.34					0.80	
lc2006_2	1.28	-0.64			2.83		-1.72	
lc2006_3	0.96	-0.40			2.44		-1.35	
lc2006_5		-0.52			2.23		-1.58	
lc2006_6		-0.92						-1.90
lc2006_7	0.45	-0.68					-1.25	
lc2006_8	0.87	-0.38			2.03		-1.57	
lc2010_2		0.80	-1.42					
lc2010_3		1.66		-0.68		-1.01	3.75	-0.71
lc2010_5	0.67	1.56				0.34	2.37	-0.66
lc2010_6	-0.50	-1.07					1.83	
lc2010_7	1.24	1.85	-0.31	0.24	-1.10		3.22	-1.88
lc2010_8	1.18	1.22	-0.75			0.63		
ParcelArea	8.13		-6.77	-20.99	6.73	-24.66	-20.11	
DA_Area	1.56			-21.53				
MeanSlope								-0.18
MeanDEM		-2.13						
Wood_dist		0.28	-0.51		0.92			
River_dist							0.91	
Water_dist		-0.15						
LRoad_dist		0.58		-3.73		1.45		
MRoad_dist	0.23	-0.26		0.43				-0.62
Ramp_dist						0.04		
lu4_dist	-0.35	-4.65	-2.35	2.53	-25.33	-0.52		
lu5_dist					0.48			
lu8_dist	0.77	-0.59		-0.85	-1.69	-1.96	2.68	0.81
lu9_dist				0.14				
Residential_Popn _Density		-0.02	-0.02	0.02				
DA_Popn_Density		-0.65x10 $^{-4}$	-1.05x10 $^{-4}$					9.30x10 $^{-5}$
Change_AveIncome		0.20						
Change_Popn		-0.03						
F_lu2					-0.39			0.31
F_lu3						1.07		
F_lu7	1.00							
F_lu11			-0.82			0.37		

Appendix F – Procedures of Mapping Land Use with Statistical Methods

When all final models of a single method, LR, GAM or SA, have been applied to all parcels in a study area, each of the final models will produce a probability of a certain type of LU change for each parcel. Thus, a parcel will have eleven probabilities produced by eleven final models of a method that correspond to eleven types of LU change. Among eleven final models applied to a parcel, those that produce LU change probabilities that exceed a predefined threshold (e.g., 0.5) will enter the competition of determining LU for a parcel. If none of the final models could produce a probability that is higher than a predefined threshold for a parcel, then the parcel will remain its previous LU. If only one final model among all eleven final models can produce a probability that is higher than a predefined threshold for a parcel, then the LU of the parcel will be determined by the final model. When there are several final models qualified for the entry and no exterior forces (e.g., LU planning and policies), the one with the highest probability determines the LU change of a parcel and others would provide options of LU change for the parcel in case the first priority has been withdrawn due to interference by exterior forces. For example, if the threshold of LU change is 0.5 and three final models produce probabilities that exceed the threshold for a parcel, then the LU of the parcel will be determined by the final model with the highest probability among the three under the circumstances of no intervention and restrictions from LU policies and government. The other two will kept as options of LU change.

Every parcel in the study area will experience the above process to determine if its LU will change or remain unchanged. For MC, it will directly produce transition probability matrices instead of final models. If prediction of LUs is entirely based on probabilities in the matrix, LU changes will be predicted perfectly since there is no randomness involved. Therefore, a random probability will be generated for each parcel and the Roulette Wheel selection approach will be used to set ranges for probability of each type of LU change. If the random value generated for a parcel resides in a range of a LU change, the LU type will be determined to be the specific LU type. Prediction accuracy of each individual method for the entire study area can be calculated using the real LU data. Then, the method that makes the best prediction of LU changes can be determined. In addition, a map can be constructed for each of the methods to show predicted LUs in the study area. These maps can help visually identify locations of LU changes occurred and the trend of LU changes.

In addition to the LU competition within final models of a method, a competition also exists among the four methods, MC, LR, GAM and SA. MC is usually uncompetitive to LR, GAM and SA in modelling LU change. However, it has the advantage of modelling LU changes of rare cases. For instance, there are insufficient LU change data for institution and agriculture in our study. Hence, only MC can model these two LU changes.

As mentioned in Section 2.2.4, the LU type of a parcel will be determined by the highest probability, either exceeds or does not exceed a predefined threshold, produced by a final model for a single method. For convenience, let us call the final model of LR, GAM and SA with the largest probability within a method the ultimate model. Each parcel will have three ultimate models if LR, GAM and SA are all applied to the parcel. If none of ultimate models produces a probability that is greater than a threshold, the LU of the parcel will remain unchanged. If only one ultimate model produces a probability that is higher than the threshold, the LU type of the parcel will be determined by the ultimate model. If two or more ultimate models produce probabilities that exceed a predefined threshold for a parcel, the LU type of the parcel will be determined by the ultimate model with the highest probability. Let us call the ultimate model that determines the LU type of a parcel the end model. Moreover, the probability of an end model will be compared to probabilities produced by MC. If the probability of an end model is greater than a threshold and all probabilities produced by MC, the LU of a parcel will be determined by the end model. If any probability produced by MC is higher than a threshold and the probability of an end model for a parcel, the higher probability from MC determines the LU of the. Otherwise, the parcel remains its previous LU type. Attention needs to be paid to parcels classified as rare LU classes since they may not be able to be modeled by LR, GAM and SA. Therefore, LU types classified by MC alone for these parcels may have a greater chance to be misclassified even with high probabilities of changing. For example, parcels classified as institution by MC in our study may require additional visual inspection since 1) this LU change rarely happens, 2) institutions may require special locations (e.g., elementary schools usually locate near to or in the residential area), and 3) the 10-fold CV prediction accuracy of LU change to institution from MC is above 50 percent, which means a parcel has 50 percent chance being classified as institutions. Moreover, the real probability of converting LU to agriculture is very low. Therefore, rational decision of whether agree or disagree with a LU change to these two types of LUs needs to be made carefully. Overall, parcels may be modeled by different methods.

The same process goes through every parcel to produce a LU map for a study area. If MC, LR, GAM and SA are used to project future LU change, which means future LU changes are unknown at this moment, the CSM produced until this point would be considered the optimal set of methods that best predicts LU changes in the area. If the four methods are used to predict known LUs, the predicted LUs will be compared to real LU data to assess the prediction accuracy of modelling LU change by the CSM. Furthermore, if an end model fails to make correct prediction under the condition that MC has lower probability than the end model does, the place that the best method for predicting LU change at a parcel goes to the next method that contains the ultimate model with the second highest probability. If all three ultimate models fail to model the LU change at a parcel, the chance of being the best method goes to the method that contains the final model with the highest probability excluding the ultimate model. This process continues until a method is found to produce a probability that is greater than a threshold and correctly predicts the LU change. Moreover, we can conclude which method can best model which type of LU change after finding the best method that models each parcel. Thus, a CSM that produces the highest accuracy of predicting LU change for all parcels in the study area can be determined. At the end, a LU map could be made to show the result of the CSM and be compared with LU maps produced by individual statistical methods. LU patterns can be revealed visually through LU maps. The misclassified LUs can be revealed by comparing a LU map constructed by real LU data and the LU map constructed by a CSM.

Appendix G – Review of Canonical Correlation Analysis

Canonical correlation analysis (CCA), a multivariate method, is used to explore the relationships between two sets of variables (Härdle and Simar, 2007). Variables in the same set should be related and the two sets of variables should come from the same observations. CCA can work with both quantitative and qualitative data. It has been used to identify relationship between LU patterns and influential factors. However, it has rarely been used to classify LUCC with the exception of Lee et al. (1999).

To understand how CCA works, let $\boldsymbol{X} = (x_1, x_2, \dots, x_p)$ and $\boldsymbol{Y} = (y_1, y_2, \dots, y_q)$ be two sets of variables measured from the same observations where $p \leq q$.The goal of CCA is to find a relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ through linear combinations of $\boldsymbol{U} = \boldsymbol{a'X}$ and $\boldsymbol{V} = \boldsymbol{b'Y}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors chosen to maximize the correlation (ρ) between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (i.e., $\rho = corr(\boldsymbol{U}, \boldsymbol{V})$). In addition, $\boldsymbol{U}$ and $\boldsymbol{V}$ are called canonical variates. $\boldsymbol{U}$ and $\boldsymbol{V}$ contain p elements and q elements, respectively. The number of canonical pairs of $\boldsymbol{U}$ and $\boldsymbol{V}$ equals p. An CCA requires the following constraints to be satisfied: 1) $Var(U_i) = Var(V_i) = 1$ where $i = 1, 2, \dots, p$; 2) $Cov(U_i) = Cov(V_j) = 0$ where $= 1, 2, \dots, p$ and $i \neq j$.

In CCA the symbol $\sum$ is used to represent the variance-covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$. After conducting a series of tests, the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be determined. In an example of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ containing two variables, $\sum$ can be expressed as

$$\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$$

where $\sum_{11}$ and $\sum_{22}$ are variances of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively and $\sum_{21}{}' = \sum_{12}$ is the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The standard deviations of $\boldsymbol{U}$ and $\boldsymbol{V}$are $(\boldsymbol{a}'\sum_{11}\boldsymbol{a})^{1/2}$ and $(\boldsymbol{b}'\sum_{22}\boldsymbol{b})^{1/2}$, respectively. The term $\boldsymbol{a}'\sum_{12}\boldsymbol{b}$ represents the covariance of $\boldsymbol{U}$ and $\boldsymbol{V}$. Therefore, the correlation between U and V can be expressed as

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}'\sum_{12}\boldsymbol{b}}{(\boldsymbol{a}'\sum_{11}\boldsymbol{a})^{1/2}(\boldsymbol{b}'\sum_{22}\boldsymbol{b})^{1/2}} .$$

Lagrange multipliers are used to solve $\boldsymbol{a}$ and $\boldsymbol{b}$. The following equation can be constructed:

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a}'\sum_{12}\boldsymbol{b} - \frac{1}{2}\lambda(\boldsymbol{a}'\sum_{11}\boldsymbol{a} - 1) - \frac{1}{2}\gamma(\boldsymbol{b}'\sum_{22}\boldsymbol{b} - 1).$$

Steps for solving this equation are as follows:

1) Differentiating and equating to $\mathbf{0}$.

$$\frac{\partial\rho}{\partial\boldsymbol{a}} = \sum_{12}\boldsymbol{b} - \lambda\sum_{11}\boldsymbol{a} = \mathbf{0} \tag{F.1}$$

$$\frac{\partial\rho}{\partial\boldsymbol{b}} = \sum_{21}\boldsymbol{a} - \gamma\sum_{22}\boldsymbol{b} = \mathbf{0} \tag{F.2}$$

2) Multiplying Equation (F.1) by $\boldsymbol{a}'$ and Equation (F.2) by $\boldsymbol{b}'$.

$$\boldsymbol{a}'\sum_{12}\boldsymbol{b} - \lambda\boldsymbol{a}'\sum_{11}\boldsymbol{a} = \mathbf{0} \tag{F.3}$$

$$\boldsymbol{b}'\sum_{21}\boldsymbol{a} - \gamma\boldsymbol{b}'\sum_{22}\boldsymbol{b} = \mathbf{0} \tag{F.4}$$

3) By solving Equations (F.3) and (G.4), $\lambda = \gamma = \boldsymbol{a}'\sum_{12}\boldsymbol{b} = \rho$, then Equations (F.1) and (F.2) become

$$\sum_{12}\boldsymbol{b} - \rho\sum_{11}\boldsymbol{a} = \mathbf{0} \tag{F.5}$$

$$\sum_{21}\boldsymbol{a} - \rho\sum_{22}\boldsymbol{b} = \mathbf{0}\ \ \mathbf{0} \tag{F.6}$$

4) Multiplying Equation (F.5) by $\rho{\sum_{11}}^{-1}$ and Equation (F.6) by ${\sum_{11}}^{-1}\sum_{12}{\sum_{22}}^{-1}$.

$$\rho{\sum_{11}}^{-1}\sum_{12}\boldsymbol{b} - \rho^2\boldsymbol{I}\boldsymbol{a} = \mathbf{0} \tag{F.7}$$

$${\sum_{11}}^{-1}\sum_{12}{\sum_{22}}^{-1}\sum_{21}\boldsymbol{a} - \rho{\sum_{11}}^{-1}\sum_{12}\boldsymbol{b} = \mathbf{0} \tag{F.8}$$

Adding Equations (F.7) and (F.8), gives

$$({\sum_{11}}^{-1}\sum_{12}{\sum_{22}}^{-1}\sum_{21} - \rho^2\boldsymbol{I})\boldsymbol{a} = \mathbf{0} \tag{F.9}$$

Then, ρ^2 is the eigenvalue of ${\sum_{11}}^{-1}\sum_{12}{\sum_{22}}^{-1}\sum_{21}$ and $\boldsymbol{a}$ is the corresponding eigenvector. Similarly,

$$({\sum_{22}}^{-1}\sum_{21}{\sum_{11}}^{-1}\sum_{12} - \rho^2\boldsymbol{I})\boldsymbol{b} = \mathbf{0} \tag{F.10}$$

Then, ρ^2 is the eigenvalue of ${\sum_{22}}^{-1}\sum_{21}{\sum_{11}}^{-1}\sum_{12}$ and $\boldsymbol{b}$ is the corresponding eigenvector.

Therefore, the maximum correlation $\rho = ({\sum_{11}}^{-1}\sum_{12}{\sum_{22}}^{-1}\sum_{21})^{1/2} = ({\sum_{22}}^{-1}\sum_{21}{\sum_{11}}^{-1}\sum_{12})^{1/2}$. The largest correlation found in the first iteration of the above steps is denoted as ρ_1, which is also called the first canonical correlation coefficient of $U_1 = \boldsymbol{a}_1'\boldsymbol{X}$ and $V_1 = \boldsymbol{b}_1'\boldsymbol{Y}$. After finding U_1, V_1 and ρ_1, it is easy to get U_2, V_2 and ρ_2 by replicating the above processes. Similarly, ρ_2 is the largest correlation between U_2 and V_2. It also can be seen as the second largest correlation in the first iteration. U_2 and V_2 have to be uncorrelated with U_1 and V_1. The above processes can be continued until a threshold has met.

Solutions are usually not the same for different combinations of canonical variates. Similar to Principle Component Analysis (PCA), only pairs of canonical variates with high correlations are considered significant to the model. Interpretation of canonical variates requires attention since the original variables comprising canonical variates can be highly correlated with each other. Correlations of canonical variables with original variables can help determine which original variable contributes the most to the correlation. Furthermore, by comparing canonical variables from the same set of original variables (i.e., either response or covariates), the association between original variables can be found.

For the experiment done for exploring CCA’s ability to model LUCC, two sets of variables have been created using a modified version of the study done by Lee et al. (1999). A set of variables contains the indicator of LUCC and another set of variable contains mean values of each predictor. When CCA performs classification duties, the theory behind it is very similar to discriminate analysis (DA). The experiment was done in R with self-programmed codes. The results turned out to be unrealistically low, which are much worse than a random classification result.