Optimizing Pattern-Based Calibration of Cellular Automata by Imperialistic Competitive Algorithm

Ehsan Momeni

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
OPTIMIZING PATTERN-BASED CALIBRATION OF CELLULAR AUTOMATA BY IMPERIALISTIC COMPETITIVE ALGORITHM

by

Ehsan Momeni

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

Major: Earth Sciences

The University of Memphis
May 2022
Preface

"Everything is related to everything else, but near things are more related than distant things."   First Law of Geography - Waldo R. Tobler

Due to the accuracy, efficiency, and ability of modeling complex dynamics, cellular automata (CA) have been used extensively by researchers to study natural, social or economic phenomena in numerous fields. In the city and regional planning, a two-dimensional regular-grid CA is used frequently to simulate different urban dynamics, such as urban growth. However, precise calibration of CA is a challenge due to the high degree of uncertainty and its knowledge-intensive procedure. This dissertation proposes a pattern-based calibration of CA using an Imperialistic Competitive Algorithm (ICA) in Shelby County, Tennessee, USA.

Chapter 1 of this dissertation explores socio-demographic and geographic changes that occurred in Shelby County from 1990 to 2010. Changes in the distribution and rate of population growth, as well as changes in the race/ethnicity, age, and educational structures of residents in Shelby County, were detected in this chapter. Moreover, land use and land cover changes in addition to their mixtures were analyzed at a micro-level in the same chapter. The findings of this chapter indicate a rapid urban and suburban development in Shelby County.

Chapter 2 of this dissertation introduces the Shannon relative index (SRI), as a pattern, in the Genetic Algorithm (GA), a broadly used metaheuristic, to improve the calibration of CA. This chapter also accommodates a literature review on using patterns in urban growth simulations. Results of the pattern-based calibration indicate that the Kappa coefficient (as an assessment of the model’s predictive power) achieves a higher value when SRIs are used in the
calibration, in comparison with the calibration of CA using a standard GA or a logistic regression (LR). A version of this chapter has been published in the journal of Transactions in GIS.

Chapter 3 is dedicated to the implementation of an ICA consisting of SRIs and Kappa coefficients. Reviews of literature on the ICA and also the Kappa coefficient are provided in this chapter. This chapter compares the Kappa coefficient and the total disagreement and concludes that using the Kappa coefficient in the cost function of ICA statistically significantly reaches a higher precision than the use of total disagreement in the calibration of CA using ICA. Findings also present that, in comparison with LR, using ICA and patterns results in a more realistic simulation of urban growth in Shelby County. These results indicate the superiority of ICA in the calibration of a CA.

Chapter 4 summarizes the results and the key points of all chapters as a conclusion and presents recommendations for future work.
Abstract

Land use and land cover (LULC) data analyses disclosed large land conversion in Shelby County, TN, between 1990 and 2010. LULC mixtures have significant associations with socio-demographics and travel behavior. Mathematical modeling is used to forecast urban expansion in order to promote sustainable development.

Cellular automata (CA) is a popular approach for the simulation of urban growth. Nevertheless, precise calibration of CA is challenging due to uncertainties and its knowledge-intensive process.

Shannon relative indices (SRIs) have been used in this study as an indicator of land patterns to calibrate a CA model of urban growth in Shelby County. The results of using a Genetic Algorithm (GA) indicate that including patterns in the calibration improves the simulations’ accuracy (from 93.21% to 94.84%).

Furthermore, an Imperialistic Competitive Algorithm (ICA) was implemented for the first time in the field of urban planning to calibrate the CA model of urban growth. Two alternatives including total disagreement and the Kappa coefficient have been separately implemented in the cost function of ICA. The findings indicate that the Kappa coefficient achieves higher overall accuracy in ICA, compared with the total disagreement (93.86% vs 92.37%). Moreover, adding patterns to the Kappa coefficient improves overall accuracy and increases the maximum (from 93.86% to 94.65%), the mean (from 79.76% to 81.11%), and the median (from 79.97% to 82.54%) of simulations’ accuracy. The pattern-based calibration resulted in a more realistic simulation of urban growth of over 9.49 sq. km of land in Shelby County (in comparison with the simulation without adding patterns). The results also demonstrated that ICA surpasses logistic regression (LR). While LR's overall accuracy was 92.98%, ICA's was 94.88% with
patterns added, 94.40% using the Kappa coefficient alone, and 94.03% using the total disagreement. Using ICA and including patterns resulted in a correct simulation of urban growth of over 37.56 sq. km of land in Shelby County (in comparison with when an LR is used).

Urban planners can utilize these findings to more accurately forecast urban growth while construction companies, transportation engineers, tax assessors, and utility providers will all benefit from accurately modeled urban land.
CHAPTER 1: A MICRO-LEVEL ANALYSIS OF COMMUTING AND URBAN LAND USING THE SIMPSON’S INDEX AND SOCIO-DEMOGRAPHIC FACTORS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>1</td>
</tr>
<tr>
<td>1.1 INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>1.2 METHODOLOGY</td>
<td>9</td>
</tr>
<tr>
<td>1.2.1 Data overview</td>
<td>9</td>
</tr>
<tr>
<td>1.2.1.1 Data preparation</td>
<td>11</td>
</tr>
<tr>
<td>1.2.2 Study area</td>
<td>12</td>
</tr>
<tr>
<td>1.2.3 The Simpson’s index</td>
<td>20</td>
</tr>
<tr>
<td>1.3 RESULTS</td>
<td>23</td>
</tr>
<tr>
<td>1.3.1 Land cover and land use vs commuting modes</td>
<td>23</td>
</tr>
<tr>
<td>1.3.2 Education vs commuting modes</td>
<td>27</td>
</tr>
<tr>
<td>1.3.3 Population size vs commuting modes</td>
<td>29</td>
</tr>
<tr>
<td>1.3.4 Age vs commuting modes</td>
<td>30</td>
</tr>
<tr>
<td>1.3.5 Race vs commuting modes</td>
<td>32</td>
</tr>
<tr>
<td>1.4 DISCUSSION</td>
<td>33</td>
</tr>
<tr>
<td>1.4.1 Car, van, truck</td>
<td>35</td>
</tr>
<tr>
<td>1.4.2 Public transportation</td>
<td>36</td>
</tr>
<tr>
<td>1.4.3 Motorcycle</td>
<td>38</td>
</tr>
<tr>
<td>1.4.4 Biking</td>
<td>39</td>
</tr>
<tr>
<td>1.4.5 Walking</td>
<td>40</td>
</tr>
</tbody>
</table>
1.4.6 Other transportation modes ............................................................................. 41
1.4.7 Working at home ............................................................................................... 42
1.5 CONCLUSION ....................................................................................................... 42

CHAPTER 2: PATTERN-BASED CALIBRATION OF CELLULAR AUTOMATA BY
GENETIC ALGORITHM AND SHANNON RELATIVE ENTROPY ................................. 45

ABSTRACT .................................................................................................................. 45

2.1 INTRODUCTION .................................................................................................... 46

2.2 METHODOLOGY ................................................................................................... 52

2.2.1 Study area ......................................................................................................... 57

2.2.2 Data and data preparation ................................................................................ 58

2.3 RESULTS ................................................................................................................ 60

2.4 DISCUSSION .......................................................................................................... 66

2.5 CONCLUSION ....................................................................................................... 70

CHAPTER 3: THE KAPPA COEFFICIENT IS STILL ALIVE ........................................... 73

ABSTRACT .................................................................................................................. 73

3.1 INTRODUCTION .................................................................................................... 74

3.1.1 Literature review .............................................................................................. 79

3.1.2 Cellular automata ............................................................................................. 82

3.1.3 Imperialistic competitive algorithm ................................................................. 83

3.1.4 Total disagreement .......................................................................................... 85

3.1.5 Shannon and Shannon relative indices ............................................................. 87

3.2 STUDY AREA ....................................................................................................... 88

3.2.1 Data .................................................................................................................. 91
List of Tables

TABLE 1. POPULATION CHANGE IN MEMPHIS AND SHELBY COUNTY ........................................ 14
TABLE 2. COMMUTING BY MODE (IN PERCENT) IN SHELBY COUNTY ........................................ 15
TABLE 3. CHANGES IN LAND COVER 1992-2011 (SQ. KM.) .................................................... 18
TABLE 4. CHANGES IN THE AREA OF EACH LAND USE FROM 2003 TO 2010 .......................... 19
TABLE 5. CORRELATION BETWEEN SIMPSON’S INDEX FOR LAND COVER, SIMPSON’S INDEX FOR LAND USE, AND MODES OF TRANSPORTATION ........................................ 26
TABLE 6. THE CORRELATION COEFFICIENT BETWEEN EDUCATION LEVELS AND MODES OF TRANSPORTATION .......................................................... 28
TABLE 7. THE CORRELATION COEFFICIENT BETWEEN POPULATION AND MODES OF TRANSPORTATION .......................................................... 29
TABLE 8. THE CORRELATION COEFFICIENTS BETWEEN AGE AND MODES OF TRANSPORTATION .... 31
TABLE 9. THE CORRELATION COEFFICIENT BETWEEN RACE AND COMMUTING MODES .......... 33
TABLE 10. THE OPTIMIZED PARAMETERS OF CA DERIVED FROM DIFFERENT CALIBRATION METHODS .................................................................................. 64
TABLE 11. ACCURACY RESULTS OF CA MODELING BY DIFFERENT CALIBRATION METHODS ....... 66
TABLE 12. CORRECTLY MODELED AREA IN SHELBY COUNTY USING CA WITH GA AND PATTERNS, GA WITHOUT PATTERNS, AND LR (IN SQ. KM.) ........................................ 66
TABLE 13. AN EXAMPLE OF A CONFUSION MATRIX .............................................................. 85
TABLE 14. THE OVERALL ACCURACY (%) OF URBAN GROWTH SIMULATION USING CA AND ICA WITH DIFFERENT COST FUNCTIONS ........................................... 98
TABLE 15. DESCRIPTIVE STATISTICS FOR OVERALL ACCURACIES OF CA’S CALIBRATION USING ICA WITH DIFFERENT COST FUNCTIONS .......................................................... 100

TABLE 16. COMPARISON OF ICA, GA, AND LR .................................................................. 103

TABLE 17. ELAPSED TIME FOR THE CALIBRATION OF CA BY ICA USING DIFFERENT COST FUNCTIONS ..................................................................................... 105
List of Figures

Figure 1. Memphis and Shelby County, TN, USA .......................................................... 13

Figure 2. Land cover of Shelby County for 1992, 2001, and 2011 ..................................... 17

Figure 3. Land use map of Shelby County in 2003 and 2010 ........................................ 19

Figure 4. The Simpson's index for land cover mixture in Shelby County

Figure 5. The Simpson's index for land use mixture in Shelby County
for 2003 and 2010 ........................................................................................................... 25

Figure 6. The relation between lower education and use of public transportation for commuting in 1990 .......................................................... 28

Figure 7. Flowchart of calibration of CA based on a GA with patterns (SRI) .............. 56

Figure 8. Built-up (developed) areas in Shelby County in 1992, 2001 and 2011 ........ 58

Figure 9. Examples of SRI values at different generations of GA.
   a) 10th generation, b) 50th generation, c) the reality ................................................. 62

Figure 10. The optimization process of GA with patterns, GA without patterns, and LR ................................................................................................................... 63

Figure 11. Required time for the calibration of CA using GA with patterns
based on the population size ........................................................................................ 68

Figure 12. Comparison of Kappa coefficients in 10 series of tests ................................ 69

Figure 13. Memphis Plan in 1827 (Williams, 1968) ..................................................... 89

Figure 14. The use district map for the City of Memphis in 1922
   (Use district map, 1922) .............................................................................................. 90
FIGURE 15. LAND COVER MAP OF SHELBY COUNTY 1992-2011 .................................................. 91

FIGURE 16. DISTANCE TO A) THE NEAREST MAJOR ROAD, B) A RIVER OR STREAM .................. 94

FIGURE 17. THE FLOWCHART OF THE IMPERIALISTIC COMPETITIVE ALGORITHM (ICA) ........... 95

FIGURE 18. HISTOGRAM AND THE NORMAL DISTRIBUTION (IN RED) USING DIFFERENT COST

   FUNCTIONS: A) THE KAPPA COEFFICIENT; B) THE TOTAL DISAGREEMENT; C) THE KAPPA

   COEFFICIENT AND SRI (PATTERNS) .................................................................................. 101

FIGURE 19. SIMULATION RESULTS USING ICA AND PATTERNS .............................................. 106
Chapter 1: A micro-level analysis of commuting and urban land using the Simpson’s Index and Socio-Demographic Factors

Abstract

This study explores the association between urban form, socio-demographics, and travel behavior for 1990, 2000, and 2010 in Shelby County, Tennessee, at a micro-level using U.S. Census tracts capturing active and passive transportation modes. We used bivariate correlations between land use and land cover mix (estimated separately by Simpson’s index), population, race, age, education, and commuting modes. Major findings indicate that land use mix is positively related to public transportation use while the land cover mix is negatively related; the opposite is found for both diversity measures and working from home. Greater land cover diversity discourages walking and biking and encourages car commuting; Blacks are the majority who use public transportation; older travelers are more likely to use transportation alternatives; higher-educated people tend to work from home or commute by bike. This study helps city planners in designing sustainable cities and increasing active modes use. Understanding travel patterns may help policymakers to control local/regional problems like increasing traffic congestions and emissions due to a modal shift in commuting to a private car during a COVID-19 pandemic, as well as develop strategies for encouraging active modes and public transport use in the post-COVID-19 world.

Keywords: Travel behavior, Urban growth, Land use, Land cover, Commuting, Diversity;
1.1 Introduction

Rapid urbanization and growing economic prosperity have increased the rate of motorization in developed countries. As a consequence, many metropolitan areas experience traffic-related problems: exhaust pollution and traffic congestion cause thousands of deaths annually due to cardiovascular disease and sleep disturbance; and greenhouse gas generation causes climate change (Krzyzanowski et al. 2005; McMichael et al. 2006). While road construction and decentralization of land use had led to urban sprawl and over-reliance on cars, some cities focus on public transportation. For instance, a survey in Seattle, WA, shows that in 2018 the number of people using public transportation to go to work in downtown is about double the number of people who drive alone. Seattle’s outstanding performance increased by 3% more public transit in 2017 compared to 2016 (Schmitt 2018). Additionally, encouraging people to use bikes or walking more often is part of environmentally friendly and sustainable planning for developed cities. For example, in 2015 voters approved to dedicate $2.4 billion to improve sidewalks and bike lanes in Phoenix, AZ (Russel 2015).

While many cities have relied mostly on cars and especially the single-occupant vehicles (SOVs), the provision of viable alternatives to this mode remains the goal for an efficient and effective multimodal transportation system (TDOT 2016). Recently, urban areas witnessed nominal growth in walking and biking levels due to changing demographics and shifting generational attitudes towards non-motorized travel. Importantly, in Memphis, TN, despite the overall small percent of mode share for walking and biking, the use of alternative non-motorized transportation modes has grown up. For example, riding a bike for commuting to work increased by 40.8% between 2006-2017 (McLeod 2017). However, despite the impressive growth, a small share of the working population commutes to work by biking, walking, or public transit. A desire
for a better understanding of how land use and land cover configuration may contribute towards growth in the use of a bike and other desirable travel modes for work and decrease car use accounting for socio-economic and demographic attributes drove this research. Given that land development patterns in urban areas can substantially affect the need for both car-based and non-motorized accommodations, understanding a degree of association, if there is any, between the frequently promoted land use mix and commuting modes is needed to develop policies that promote travel by other than private vehicle-based modes, and ensure many positive benefits including peak-hour modal shifts and healthier residents (TDOT 2016).

Studying and changing travel behavior will help to reach a better balance between modes of transportation, to decrease the daily travel time and travel cost as well as to achieve a more sustainable city. Modeling travel behavior allows us to form alternative scenarios and optimize predictions about the future of transportation demands (de Sá et al. 2015) as well as to define particular target groups for environmental interventions (Saris et al. 2013). Balancing transportation modes promotes physical activity which is a public health priority (Commission of the European Communities 2007). Studies from the United States and Australia have shown a positive correlation between physical activities and neighborhood walkability. Therefore, by knowing areas with a greater demand for a walk, we can place them on the top priority for walkability development (Sallis et al. 2009; Owen et al. 2007).

Commuter mode share indicates the percentage of workers who commute by various travel modes and is linked to environmental conditions, such as air pollution, reflecting the impact of infrastructure, policies, and land patterns (U.S. Department of Transportation 2019). Certain land use arrangements may be more conducive to the more desirable commuting modes providing motivation for this study.
In this study, we set out to analyze the impact of changes in both land use and land cover, and socio-demographic characteristics on travel behavior measured by various commuting modes, to capture the use of active and passive modes using three periods of 1990, 2000, and 2010. For this, we estimate the trending travel behavior, or the most frequently used commuting modes. Changes in land use and land cover development are quantified (separately) by calculating Simpson’s indices for land use and land cover (as an indicator of mix), respectively, in a large-sized urban area of the south-eastern USA. More details about the Simpson’s index are provided in Section 1.2.3.

We used the example of Shelby County, TN, as a case study. The objective of this paper is to find the relationship, if any, between various commuting modes and land use, land cover, and socio-demographic characteristics. A micro-level analysis was conducted to analyze both land use and land cover at a fine spatial resolution using Census tracts. Based on the comprehensive review of the literature on the topic, in this study, the Simpson’s index is used to measure the mixture of land use and land cover (separately), whereas socio-demographic variables are represented by population size, age, race, and education. Using the results of this paper, urban planners and decision-makers can make more reliable future transportation plans in Memphis and Shelby County, TN, USA, as well as elsewhere by identifying priority locations for building walking and biking infrastructure.

Recently, Litman (2019, page 54) argued that in order to better evaluate land use impacts, relationships between mixed-use development and travel mode selection need to be determined, while Duncan et al. (2010) recognized the effect of the geographical scale when measuring this relationship. To build and expand on these arguments, the following research questions were asked:
(1) Is there an association between various commuting modes and land use or land cover mix measured by the Simpson’s index at the fine scale of analysis such as the Census tract level?

(2) Which socio-demographical factors impacted commuting mode choice in Shelby County in 1990, 2000, and 2010?

While hypotheses are:

Hypothesis 1: land use mixture and/or land cover mixture impact commuting choice in a substantial way at the fine scale of analysis.

Hypothesis 2: commuter age, race, education, and population of neighborhoods impact commuting choice.

Both land use and land cover impact the way workers travel. Changes in the built environment, e.g. mixed-use development, are associated with changes in active commuter trips (Mackenbach et al. 2016). On the other hand, land cover conversions, such as deforestation or wetlands conversion occurring as a result of economic development are indicative of processes of human encroachment of natural areas, or urban sprawl, in large metropolitan areas in the USA, and city expansion may result in the costs of commuting (Walker 2001/1997; Antipova, Momeni & Banai 2022a).

Extensive research focuses on the influence of land use on travel behavior. Prior studies suggest that land use measures are useful in explaining commuting patterns. Residents in mixed land use areas or moderate density areas tend to take advantage of their location and commute less, while low-density residents who usually are located farther from their jobs are more likely to commute using motorized vehicles (Antipova, Wang & Wilmot 2011).

Even in places where the existing land use is mixed with little separation between workplaces and house locations and thus non-motorized trips traditionally constitute a major
portion of the total trips, the share of the non-motorized trips is decreasing primarily due to the changing socio-economic attributes and the land use pattern (Sarkar et al. 2013). In this regard, Sarkar et al. (2013) investigated the effect of land use density and land use mix measured by area index on commuting choices in Agartala, a small-size city in India for work trips, shopping trips, and all trips. Land use parameters and area index were found to be significantly influencing the utility/disutility of the motorized and non-motorized modes, respectively.

Litman (2019) studied how land use mix (measured by entropy indices and jobs-to-housing ratio) impacts travel behavior in a community including per capita vehicle travel. A more mixed development shortens travel distances and increases shares of walking and biking modes, and reduces congestion and pollution emissions, and conserves energy more.

While land use parameters improve travel behavior models, socio-economic attributes appear more significant for the mode choice. Socio-demographics at both neighborhood and individual levels shape commuting. To illustrate the role of individual attributes, minority workers tend to commute less time, and part-time workers commute less in terms of both distance and time than their full-time counterparts (Antipova et al. 2011). Among neighborhood-level consolidated socio-economic factors, both the primary (that is, variables related to socio-economic disadvantages such as poverty, unemployment, minority, and poor housing) and secondary disadvantage factors (including high ratios of part-time and immigrant workers and households with young dependents) were found to be negatively associated with car-based commuting distances (Antipova et al. 2011). Across the world, socio-demographic factors influence the market share of passengers in different modes of transportation both in developed and developing countries. These factors include sex of a traveler (e.g. Azimi et al. 2021;
Socio-demographic contexts of travel decision-making should be addressed in order to have a precise estimation of behavior changes. In 2014 Grando et al. focused on higher education students to study their travel patterns at the Federal University, Brazil. Among car, public transport, walking, and biking, the first two modes have been used most often. Quintero et al. (2016) surveyed school students and employees and collected data on total trips by modes of transportation. The analyses for 22 schools showed that the most used mode of transportation is the car followed by public transport. Meira et al. (2014) used income as a variable to study the travel behavior of 35,000 people in Brazil, finding a strong interest in the use of public transport among 60% of students and professors, while 30% prefer private cars and 10% prefer other modes of transportation.

Panter et al. (2011) studied individual commuter characteristics and physical environment factors affecting biking and walking behavior to and from work in Cambridge, UK. 53% of respondents reported biking and 30% walked to or from work. The study shows that better education and a higher disposable income are associated with biking, and not walking, while the impact of environmental determinants of physical activity was less apparent. Saris et al. (2013) explored relations between personal and environmental characteristics and active transportation in the Netherlands where younger people, females, and migrants are more likely to walk, and males are more likely to ride a bike. Bauman et al. (2012) used ethnicity and age to study the relationship between physical activities and modes of transportation. They found that Whites are positively correlated with active transportation while age is inversely correlated.
Thern et al. (2015) explored factors associated with active commuting (walking or biking to and from work/school for at least 15 minutes one-way) including gender, marital status, education, ethnicity, financial situation, and personal habits such as smoking and alcohol habits in Sweden. They found active commuters in the minority (8.3%), likely due to unsuitable outdoor temperature conditions, and the main mode of commuting as motor vehicles (63.0%).

Different technologies and economic features rank different modes of transportation differently in safety, services, quality, and fares. While active transportation mode refers to non-motorized and human-powered travel including walking and biking, passive transportation mode is non-active motorized travel including motorcycles, cars, and taxis. Collective passive modes refer to buses, trains, and subways (Litman 2017; Li et al. 2016). Prior research documents the importance of active transportation on health (both physical and mental). The environmental benefits of active transportation studied by Woodcock et al. (2009) and the Department for Transport (2011) conclude that people who use active modes of transportation increase their level of physical activity and are healthier than others.

Cultural issues are also an important factor in studying the mobility, e.g., as noted above, in Europe, especially in Central and Northern Europe, the bicycle and pedestrian travel modes each account for substantial mode shares, and cities (with small differences) are well adapted to that way of life. While this is not generally the case in the USA, recent impressive changes began to occur even in places where the socioeconomic structure is primarily car-dependent. According to the U.S. Census Bureau, there was a nationwide growth in bike commuting from 2000 to 2017, where the top five cities regarding the absolute counts of bike commuters are New York City (over 51,000 bike commuters), Portland (over 22,000), Chicago (over 22,000), Washington D.C. (over 18,000), and Los Angeles (over 18,000). The top five states ranked by mode share are
Washington D.C. (4.9% of commuters who bike), Oregon (2.04%), Montana (1.11%), Colorado (1.10%), and Wyoming (1.06%). In Tennessee, which is ranked the fourth-worst state in terms of the percent of commuters who bike, there was a positive change of 42% between 2006-2017 and a more modest change of 6% between 2011-2017. Within Tennessee, the city of Memphis (which is part of Shelby County) has the highest growth in the use of the bike for commuting to work with a change of 40.8% between 2006-2017 (versus 38.1% change in the larger city of Nashville, the state’s capital city) (McLeod 2017). Thus, Shelby County, which is part of Tennessee, and which is a large metropolitan area including the city of Memphis with a substantial urban population, as well as young and educated groups potentially driving the observed changes and thus justifying the choice of the study area.

The rest of this paper is organized as follows: Section 1.2 is dedicated to the methodology. This section describes the data, study area, and Simpson’s index in detail. Sections 1.3 and 1.4 are dedicated to the results and discussion, respectively. Finally, the conclusion is presented in Section 1.5.

1.2 Methodology

1.2.1 Data overview

People make trips from one place to another by choosing a certain transportation mode. Consequently, people, spatial patterns, and commuting modes are the three elements of journey-to-work travel investigated in this research. Therefore, to study commuting patterns, socio-demographic, spatial, and travel data are used; population size, age, education, and race as socio-demographic data; land use, and land cover as spatial data; different modes of transportation as travel data.
Socio-demographic data comes from the U.S. Census Bureau. To estimate trends, we compare the results for three periods including 1990, 2000, and 2010 covering two decades, which is an adequate time span for urban studies (U.S. Census Bureau 2017).

Journey-to-work travel data for 1990, 2000, and 2010 come from the National Historical Geographic Information System (NHGIS at https://www.nhgis.org/).

Remote sensing is a reliable tool to study land cover. Using an image classifier each pixel in a satellite image is assigned a label of a class (or more classes) with the maximum confidence of assignment (Momeni et al. 2018; Momeni et al. 2020). Multi-Resolution Land Characteristics Consortium, MRLC, (https://www.mrlc.gov/) provides the National Land Cover Database (NLCD) including satellite-derived land cover derived from a level-2 classification of Landsat satellite images with a spatial resolution of 30 meters (MRLC, 2021; Momeni & Antipova, 2020). At the time of this study, land cover data were only available for 1992, 2001, and 2011, which are used for further analyses of the land cover mixture.

Land use data, often provided by urban planning authorities (here, by Assessor of Property in Shelby County, https://www.assessormelvinburgess.com/welcome), are available at the parcel level in a vector format. We used land use data in 2003 and 2010. More accurate identification of the relationship between land configuration and commuting behavior is possible using both land use and land cover data, as some transportation modes are associated with land use while others are associated with land cover. These land data are available for different time periods. Therefore, to be compatible with socio-demographic attributes, we used land data that come closest regarding time (we used land use data for 2003 and 2010, and land cover data for 1992, 2001, and 2011). To avoid any error, all data (including land use and land cover) were
transformed to the same spatial unit (tract level in 2010) before calculating the Simpson’s indices and correlations (refer to Section 1.2.3).

The tract level (Geographic Terms and Concepts - Census Tract 2017) is selected as the spatial unit of study as the most commonly used scale of analysis (Klein et al. 2016; Giuliano et al. 2006; Osama et al. 2017; Graham et al. 2015). Tracts are small, relatively permanent geographic entities within counties. A tract is roughly equivalent to a neighborhood with a population between 2,500 and 8,000 residents, and boundaries that follow visible features (Census Tracts and Block Numbering Areas 2018). We used the tract level as this spatial unit of study represents the most commonly used scale of analysis (Klein et al. 2016; Giuliano et al. 2006; Osama et al. 2017; Graham et al. 2015). This study uses the most recently updated Census tract boundaries for 2010, to estimate Simpson’s indices and correlations. Census tract geography comes from the Tiger line database (Tiger/line® shapefiles and tiger/line® files 2017).

MatLab R2017(MathWorks 2017), ArcMap 10.5.1 (ArcGIS Desktop 2017), and Microsoft Excel 2016 (Office Excel 2016) are the software packages used for data preparation, statistical analysis, and map generation.

1.2.1.1 Data preparation

Initially, transportation modes in 1990 included 13 categories while 2000 and 2010 data included 7 and 20 categories, respectively. Therefore, raw data were preprocessed by combining bus, streetcar, railroad, ferryboat, and taxicab to only one public transportation category for attribute consistency with public transportation data in 2000.

To compare the results between three different years, a consistent data frame is needed, while raw data initially came in different units and frames (e.g. land cover data are at pixel level while land use data are at parcel level). Tracts are mostly homogeneous with respect to
population characteristics, economic status, and living conditions (Census Tracts and Block Numbering Areas 2018). The U.S. Census Bureau updates the boundaries of each tract over time (usually each decade). Therefore, to avoid errors in compiling and comparing data on a neighborhood (roughly a tract) over several decades, using a fixed boundary for the tract is necessary. In this study, Boundary lines of Census tracts in 2010 were selected as the most recently updated, and land cover data in 1992, 2001, and 2011, in addition to land use data in 2003 and 2010 were adjusted to Census tracts in 2010 using Data Management Tools and Spatial Analyst Tools in ArcMap.

1.2.2 Study area

Shelby County is the largest county in Tennessee, regarding population (936,130 in 2019) and geographic extent (≈ 763 sq. miles of land and 22 sq. miles of waterbodies) (Tennessee Land Area County Rank 2019; Momeni & Antipova 2022a). Memphis metropolitan area is the commercial, political, and cultural hub of the Mid-South (TN, AR, and MS) and the centerpiece of the region (Cushman & Wakefield 2012). The city of Memphis is a large-sized city in Shelby County, TN, (Figure 1) with a population of 646,889 in 2010 (Large cities ranking 2017; U.S. geographic summary data and boundary files 2017). The studies on transportation-related issues such as travel patterns and congestion reduction in the city of Memphis are lacking (Klein et al. 2016; Osama et al. 2017; Graham et al. 2015).

The locations of Shelby County in Tennessee and Memphis in Shelby County are presented in Figure 1.
The population of Shelby County increased by more than 12% from 1990 to 2010. However, population growth is not evenly distributed. While the population in the city of Memphis declined by 7.78%, the population in the suburban areas of Shelby County increased.
by more than 124% in the same time span. This population growth has changed land use, land cover, and spatial configurations. For instance, new towns and suburbs such as Barlett and Cordova grew up rapidly (Momeni & Antipova, 2020). This rapid growth of the suburban area and/or migration of population from the central area to the peripheral suburbs as a consequence of decentralization, low-density development, a poor mixture of land use, and motor vehicle-oriented development is known as urban sprawl (Wheeler & Beatley, 2014; Antipova, Momeni, & Banai, 2022a).

Within Shelby County, to understand the population dynamics, the population is studied in three different regions: the city of Memphis, suburban areas, and the entire Shelby County (which includes both Memphis and suburban areas) (Table 1).

Table 1. Population change in Memphis and Shelby County

<table>
<thead>
<tr>
<th>Year</th>
<th>Memphis</th>
<th>Shelby County (excluding Memphis)</th>
<th>Shelby County</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990-2000</td>
<td>-3.87</td>
<td>78.71</td>
<td>8.61</td>
</tr>
<tr>
<td>2000-2010</td>
<td>-4.07</td>
<td>25.82</td>
<td>3.36</td>
</tr>
<tr>
<td>1990-2010</td>
<td>-7.78</td>
<td>124.85</td>
<td>12.26</td>
</tr>
</tbody>
</table>

Despite population growth in Shelby County, the age structure did not change between 1990 to 2010 (U.S. Census Bureau 2017). However, the educational pattern has changed. The number of people with education less than a 9th degree declined from 9% in 1990 to 6.90% in 2010 (U.S. Census Bureau 2017). Also, the number of people with greater education (bachelor’s,
some college, graduate, or professional degrees) increased from 48% in 1990 to 66% in 2010 (U.S. Census Bureau 2017).

The most dominant commuting modes in Shelby County contain seven categories including car/truck/van, motorcycle, public transportation, biking, walking, other modes (e.g. lorry, boat, scooters, skating, etc.), and working at home (Table 2). Travelers in Shelby County mostly rely on private cars for journey-to-work travel as car/truck/van accounts for over 90% of travel for all three years under study (Table 2).

<table>
<thead>
<tr>
<th>Mode</th>
<th>1990</th>
<th>2000</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car, truck, van</td>
<td>91.02</td>
<td>93.33</td>
<td>93.64</td>
</tr>
<tr>
<td>Motorcycle</td>
<td>0.09</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Public transportation</td>
<td>3.29</td>
<td>2.09</td>
<td>1.49</td>
</tr>
<tr>
<td>Bike</td>
<td>0.12</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>Walked</td>
<td>3.24</td>
<td>1.46</td>
<td>1.48</td>
</tr>
<tr>
<td>Other</td>
<td>0.75</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>Worked at home</td>
<td>1.48</td>
<td>2.21</td>
<td>2.41</td>
</tr>
<tr>
<td>Sum</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The commuting share for other modes rather than private cars has remained stable during the past three decades accounting for less than 10% (Table 2). To compare, in the Netherlands, the country that is recognized as the global leader in active transportation, 30% of all trips are made by bike, 18% by walking, 5% by public transportation, 45% by car, and 2% by other modes of transportation (Bicycle Statistics 2017).

For commuting modes, we use the following: (1) car, van, or truck; (2) public transportation; (3) motorcycle; (4) bike; (5) walking; (6) working at home, and (7) other. To have
a better understanding of commuting in Shelby County, the current study analyses the relationship between the above-mentioned modes used to get to work and land use and land cover mixtures, while also looking at demographic attributes including population, age, race, and education in a bivariate correlation.

A visual comparison of land cover data (Figure 2) and a mathematical comparison of the data (Table 3) disclose the growth of developed areas in Shelby County between 1992 and 2011. This urbanization is due to the rapid growth of the population (Table 1) and the demand for residential, commercial, and industrial lands. Figure 2 illustrates the spatial distribution of land covers in Shelby County in 1992, 2001, and 2011. The land cover data consist of eight classes including developed area, barren, forest, herbaceous, planted/cultivated, shrublands, water, and wetlands (National Land Cover Database 2017). As Figure 2 shows, developed areas grew over 1992-2011 in Shelby County, while planted/cultivated areas decreased. Mostly, development occurred in rural areas outside of Memphis’s boundaries.
In addition, Table 3 shows the pattern of change in land cover between 1992 and 2011. As expected, many classes changed to the developed areas to meet the increasing demand for housing and jobs for the population. As seen in Table 3, due to sprawl and development, planted and forested areas decreased significantly while developed areas increased.
Table 3. Changes in land cover 1992-2011 (sq. km.)

<table>
<thead>
<tr>
<th></th>
<th>Barren</th>
<th>Developed</th>
<th>Forest</th>
<th>Planted</th>
<th>Water</th>
<th>Wetlands</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>0.4</td>
<td>13.4</td>
<td>0.1</td>
<td>2.2</td>
<td>0.4</td>
<td>0.5</td>
<td>17.0</td>
</tr>
<tr>
<td>Developed</td>
<td>0.4</td>
<td>535.9</td>
<td>16.4</td>
<td>6.8</td>
<td>0.8</td>
<td>1.8</td>
<td>562.1</td>
</tr>
<tr>
<td>Forest</td>
<td>1.0</td>
<td>116.3</td>
<td>166.5</td>
<td>110.8</td>
<td>3.8</td>
<td>26.4</td>
<td>424.8</td>
</tr>
<tr>
<td>Planted</td>
<td>1.7</td>
<td>218.9</td>
<td>35.6</td>
<td>437.0</td>
<td>4.1</td>
<td>17.9</td>
<td>715.1</td>
</tr>
<tr>
<td>Water</td>
<td>1.3</td>
<td>5.5</td>
<td>2.4</td>
<td>3.0</td>
<td>79.4</td>
<td>6.5</td>
<td>98.2</td>
</tr>
<tr>
<td>Wetlands</td>
<td>0.7</td>
<td>13.8</td>
<td>35.0</td>
<td>23.9</td>
<td>5.1</td>
<td>136.5</td>
<td>215.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barren</td>
<td>5.5</td>
</tr>
<tr>
<td>Developed</td>
<td>903.9</td>
</tr>
<tr>
<td>Forest</td>
<td>256.0</td>
</tr>
<tr>
<td>Planted</td>
<td>583.6</td>
</tr>
<tr>
<td>Water</td>
<td>93.6</td>
</tr>
<tr>
<td>Wetlands</td>
<td>189.6</td>
</tr>
<tr>
<td>Sum</td>
<td>2032.3</td>
</tr>
</tbody>
</table>

In addition to land cover, population growth has contributed to changes in land use in Shelby County. Land cover is defined as a physical material at the surface of the earth including grass, asphalt, trees, bare ground, water, and other land and water types (Makers of American Botany 2017), while land use is the way people use the landscape and utilize the land (NOAA 2017). Figure 3 shows the distribution of land uses in Shelby County in 2003 and 2010 consisting of seven classes including farm/agricultural, residential, commercial, industrial, mixed-use, other uses, and exempt. The majority of commercial lands (in light red color) are alongside the main roads, while industrial lands (in purple color) are around the downtown and outside Memphis’ boundaries in the southeast. This urban structure may affect the commuting mode choice.
A visual comparison of land use data (Figure 3) and a mathematical comparison of the data (Table 4) disclose the growth of industrial uses (by 26.4%) and commercial uses (by 21.3%) in Shelby County from 2003 to 2010.

Table 4. Changes in the area of each land use from 2003 to 2010

<table>
<thead>
<tr>
<th>Land use</th>
<th>Area in 2003 (Acre)</th>
<th>Area in 2010 (Acre)</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial</td>
<td>26,717.2</td>
<td>32,412.0</td>
<td>21.3</td>
</tr>
<tr>
<td>Exempt</td>
<td>86,918.4</td>
<td>95,762.3</td>
<td>10.2</td>
</tr>
<tr>
<td>Farm/Agriculture</td>
<td>152,762.1</td>
<td>140,439.5</td>
<td>-8.1</td>
</tr>
<tr>
<td>Industrial</td>
<td>15,885.7</td>
<td>20,077.2</td>
<td>26.4</td>
</tr>
<tr>
<td>Mixed used</td>
<td>7,884.6</td>
<td>63.6</td>
<td>-99.2</td>
</tr>
<tr>
<td>Residential</td>
<td>156,727.7</td>
<td>158,459.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Other</td>
<td>1,432.0</td>
<td>6.9</td>
<td>-99.5</td>
</tr>
</tbody>
</table>
The land use and/or land cover pattern and their mixtures affect travel behavior such as traffic volumes and travel modes (Nasri & Zhang 2019; Kuzmyak 2012; Sarkar & Mallikarjuna 2013; Dargay & Hanly 2003). The greater use of private vehicles is expected in developing areas with less density whereas the greater use of public transport is expected in densely populated areas as a more practical and economic alternative. A higher density and a better mixture of both land use and land cover are closely related to shorter and more frequent trips and provide a higher chance for biking and walking (Dargay & Hanly 2003). For instance, Mäki-Opas et al. (2016) found that Finns who were physically active during their leisure time and lived in an area with a proportion of green space to home from 25% to 49.9% were more likely to be physically active while commuting. Also, de Vries et al. (2010) found that walking and biking to schools in the Netherlands are strongly associated with the presence of green spaces in neighborhoods.

1.2.3 The Simpson’s index

The Simpson's index is one of three closely related indices. The general Simpson's Index (D) ranges between 0 (infinite diversity) and 1 (no diversity). To overcome the inherent problem with an increase in values of D meaning the lower diversity, D is often subtracted from 1. The second index is the Simpson's Diversity Index that also ranges between 0 and 1 with greater values indicating greater diversity. The Simpson's Reciprocal Index has the lowest value of 1 representing an area containing only one land use/land cover (LULC) class, while higher values mean greater diversity. The maximum value is the number of LULC classes in the study area. We use the second index described above to understand the impacts, if any, a land mixture has been calculated using Simpson's index (Rocchini et al. 2013).

The Simpson’s index can be defined as the probability that two randomly selected locations within an area are parts of the same group (e.g. land uses or land covers) (Antipova,
The Simpson’s index is sensitive to the distribution of the size of land use and land cover (separately) in an area. The index is more sensitive to the larger classes and less sensitive to the smaller classes (Koomen & van Eck 2007).

This diversity index should be distinguished from a LULC fragmentation index useful in studies of landscape processes related to habitat divisions into smaller and more isolated patches often caused by ever-increasing urbanization and development of areas observable in all parts of the world. Due to the rapid increase in the built-up areas, the mean patch size decreases, and the number of patches increases resulting in landscape diversity decline. The described situation of low diversity combined with high fragmentation threatens people- and environment-friendly mobility, also known as “soft mobility” (Cramer 2009), while increased landscape fragmentation is considered a health risk as it facilitates disease transmission, increases human exposure to the disease, and may increase the risk to the emergence of some diseases (Wu & Smithwick 2016; Antipova, Momeni & Banai, 2022a).

We use Simpson’s index to summarize a large data set of polytomous variables consisting of multiple categories (e.g. eight different land covers or seven different land uses) into only one single value (one for the mixture of land use and one for the mixture of land cover, separately). The Simpson’s index is among entropy indices that estimate a mixture of land use and a mixture of land cover within an area, and it has been widely used in prior research (Rocchini et al. 2013; Banai, Antipova & Momeni 2021). This index presents a balance between different categories within an area. The Simpson’s index is defined using the following:

\[
D=1-\sum_{i=1}^{N} \left( \frac{A_i}{N_A} \right)^2
\]  

[1]
where \( N \) is the number of classes in the area, \( A_i \) is the area of \( i^{th} \) class and \( T_A \) is the total area. The range of this Simpson’s index \( D \) is between 0 and 1. A lower value indicates less mixture of classes in the area (seen as less desirable for a “good urban form”), while higher values mean there are more classes and a more mixture of classes in the area (more desirable) (Rocchini et al. 2013; van Eck & Koomen 2008). Promoting land use mixture (e.g. higher \( D \) values) in new developments often is associated with smart growth and transit-oriented development initiatives. Mixed-used development (e.g. higher values of \( D \)) contributes positively to the creation of enlivening urban districts by meeting the everyday community needs. For example, mixed-use highly-dense built environments decrease the travel distance between complementary land uses, and support public transit, biking, and walking as alternatives to private vehicles (KYOVA Interstate Planning Commission 2013).

In the next step, we use a bivariate correlation between diversity and mobility to determine if a mutual relationship between the two variables exists. Accordingly, the correlation analyses between Simpson’s index of land use and land cover (calculated separately) with different modes of transportation were implemented. Socio-demographic and commuting data were available for 1990, 2000, and 2010. However, land cover data were available for 1992, 2001, and 2011 (see Section 1.2.1 for more details). Thus, to estimate the Simpson’s index of land cover data in 1990, 2000, and 2010, land cover data in 1992, 2001, and 2011 were used, respectively. The same procedure was used for land use data: to estimate the Simpson’s index of land use data in 2000 and 2010, land use data in 2003 and 2010 were used, respectively.

It should be noted that in contrast to a cause-and-effect relationship between cause and effect, where one event actually causes the other, we are just looking to explore the relationship between these two variables with no implication that one causes the other. LULC change and its
influence on (soft) mobility and other human aspects is a key issue for future smart and sustainable cities (Cramer, 2009). Knowing that correlation exists justifies further examination of the situation to determine if causation can be established between the variables (subject to future studies).

1.3 Results

1.3.1 Land cover and land use vs commuting modes

To calculate the Simpson’s index, a MatLab script was programmed. In order to compare indices in different years, the same borders (2010 tract borders) were used, and Simpson’s indices were calculated using the consistent geography of the Census tract. The Simpson’s index is an indicator of the mixture of classes within an area. To illustrate, in an area with a low mixture (where the Simpson’s index is close to zero), such as an agricultural field, the use of a private car, van, or especially trucks is expected to be greater while using public transportation is not expected. Conversely, in the area with a higher mixture of classes and shorter distances between different land uses (or land covers) (where the Simpson’s index is closer to one), more walking and biking are expected as people may satisfy their needs by walking or biking between places.

Generally, a lower value for the Simpson’s index is interpreted as a less mixture of classes in the area and is considered less desirable rather than being a “good urban form”. However, in the case of land cover in the urbanized area, commonly the conversion to developed/built-up area is observed from the previous land cover categories as evidenced by data in Table 3. Accordingly, inside the city of Memphis, the index of land cover mix has decreased between 1992 and 2011 due to a conversion of some planted/cultivated, forest, and barren lands to developed land in 2010 (Figure 2) making areas there more homogeneous and developed.
Figure 4 visualizes this transition and presents Simpson’s indices computed for land cover in Shelby County for different years. Simultaneously, the value of the index has increased beyond the boundaries of the city of Memphis in the east and northeast of the study area between 1992 and 2011. Simpson’s indices increased in those neighborhoods because of urban sprawl (Figure 2) which brought more development to rural areas usually covered only by forest or planted/cultivated land.

Simpson’s indices increased in those neighborhoods because of urban sprawl (Figure 2) which brought more development to rural areas usually covered only by forest or planted/cultivated land.

Figure 4. The Simpson's index for land cover mixture in Shelby County for 1992, 2001, and 2011

Simpson’s indices computed for land use in Shelby County are presented in Figure 5. A lower value for the Simpson’s index measuring land use mix is interpreted as a less mixture of
land use in the area and is considered less desirable rather than being a “good urban form”. From 2003 to 2010 Simpson’s indices for land use mix have decreased in many neighborhoods within the city (seen as less desirable), that is, land use became more homogeneous resulting in a lower mix of land use indicating a less desirable urban form. This situation is more conducive to the use of a car since the different types of land use are farther away from each other. The findings indicate that the values of the land cover mix are almost the mirror image of those of the land use mix, that is, tracts with higher values of land use mix (mostly located in the central part of the study area) have lower values of the land cover mix, while suburban areas have higher-mixed land cover areas.

![Simpson's index of land use](image.png)

Figure 5. The Simpson's index for land use mixture in Shelby County for 2003 and 2010

To understand the degree of an association between land cover mix, land use mix, and various commuting modes, we conducted a bivariate correlation analysis. Table 5 shows the
correlation coefficient, as an indicator of the strength of a relationship between Simpson’s indices computed for land cover mix, land use mix, and different modes of transportation.

Table 5. Correlation between Simpson's index for land cover, Simpson’s index for land use, and modes of transportation

<table>
<thead>
<tr>
<th>Year</th>
<th>Car, van, or truck</th>
<th>Public transportation</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Walked</th>
<th>Other</th>
<th>Worked at home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Land cover  
1990 | 0.16*              | -0.34***              | -0.04      | -0.15* | 0.08   | 0.02  | 0.15*         |
| 2000 | 0.06               | -0.29***              | 0.04       | -0.1  | -0.15* | 0.01  | 0.17*         |
| 2010 | 0.19**             | -0.24***              | 0.07       | -0.01 | 0.01   | 0.13* | 0.21***       |
| Land use  
1990 | N/A                | N/A                   | N/A        | N/A  | N/A    | N/A   | N/A           |
| 2000 | -0.12              | 0.14*                 | -0.03      | 0.02  | -0.06  | -0.07 | -0.14*        |
| 2010 | -0.11              | 0.29***               | -0.18**    | 0.05  | -0.02  | 0.07  | -0.07         |

* P value < 0.05  
** P value < 0.01  
*** P value < 0.001

Based on the results, the correlation between land use mixture and the use of public transportation is significant for all time periods investigated, which means in the area of higher mixed-uses (the value of Simpson’s index is closer to “1”) people tend to use public transportation more often. The opposite is true for land cover diversity and public transportation for 1990, 2000, and 2010: the greater the mix in land covers, the less the use of public transportation. This is the expected finding since the greater values of land cover diversity can be found in the outskirts and in the suburbs where developed (built-up) lands tend to intermix with agricultural and other land cover types (refer to Figure 4), while in the central areas developed land is dominant. The opposite relationship is found for both diversity measures and working
from home: a statistically significant positive relationship is found between land cover mix and working from home for all years studied, and a negative association is found for a land use mix. Greater land cover diversity of the suburbs discourages walking and biking, and contributes to greater car use, while no links have been found between active and passive commuting and land use mix.

1.3.2 Education vs commuting modes

Similarly, we tested the strength of a relationship between land cover and land use (separately) and various socio-demographic indicators such as education, population size, age, and race. Changes in education between 1990 and 2010 (already mentioned in Section 1.2.2) might affect travel patterns. Data analysis shows a strong correlation between education and certain modes of transportation (Table 6). Commuters whose education is less than 9th degrees rely on public transportation more than other groups in 1990 and 2000 (Figure 6). As mentioned in Section 1.2.2, in 2010 fewer people had less than a 9th degree of education, therefore making the correlation between less education and use of public transportation not significant in 2010.

Also, based on the analysis in 2000, people who are working at home are more likely to have a graduate, or professional degree and less likely to have a high school degree or some diploma (Table 6).

In addition, in 2010, in the area with more people with graduate and professional degrees, biking is more likely to occur (Table 6).
Table 6. The correlation coefficient between education levels and modes of transportation

<table>
<thead>
<tr>
<th>Year</th>
<th>Car, van, or truck</th>
<th>Public transportation</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Walked</th>
<th>Other</th>
<th>Worked at home</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>-0.04</td>
<td>0.72***</td>
<td>-0.1</td>
<td>-0.06</td>
<td>-0.06</td>
<td>0.23**</td>
<td>-0.32***</td>
</tr>
<tr>
<td>2000</td>
<td>0.02</td>
<td>0.63***</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.23**</td>
<td>0.19**</td>
<td>-0.26***</td>
</tr>
<tr>
<td>2010</td>
<td>0.08</td>
<td>0.03</td>
<td>-0.07</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.02</td>
<td>-0.16*</td>
</tr>
<tr>
<td>1990</td>
<td>0.69***</td>
<td>0.26***</td>
<td>0.18*</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.35***</td>
<td>0.22**</td>
</tr>
<tr>
<td>2000</td>
<td>0.53***</td>
<td>0.37***</td>
<td>0.18**</td>
<td>-0.06</td>
<td>0.03</td>
<td>0.37***</td>
<td>0.03</td>
</tr>
<tr>
<td>2010</td>
<td>0.08</td>
<td>0.36***</td>
<td>-0.1</td>
<td>-0.09</td>
<td>0.09</td>
<td>0.26***</td>
<td>-0.27***</td>
</tr>
<tr>
<td>1990</td>
<td>0.93***</td>
<td>-0.33</td>
<td>0.19*</td>
<td>0.09</td>
<td>-0.01</td>
<td>0.23**</td>
<td>0.78***</td>
</tr>
<tr>
<td>2000</td>
<td>0.95***</td>
<td>-0.29***</td>
<td>0.26***</td>
<td>0</td>
<td>-0.06</td>
<td>0.25***</td>
<td>0.78***</td>
</tr>
<tr>
<td>2010</td>
<td>0.86***</td>
<td>-0.09</td>
<td>0.21***</td>
<td>0.13*</td>
<td>0.16*</td>
<td>0.08</td>
<td>0.49***</td>
</tr>
<tr>
<td>1990</td>
<td>0.63***</td>
<td>-0.36***</td>
<td>0.02</td>
<td>0.07</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.73***</td>
</tr>
<tr>
<td>2000</td>
<td>0.64***</td>
<td>-0.32***</td>
<td>0.17*</td>
<td>0.02</td>
<td>0.04</td>
<td>0.1</td>
<td>0.79***</td>
</tr>
<tr>
<td>2010</td>
<td>0.66***</td>
<td>-0.23***</td>
<td>0.11</td>
<td>0.17**</td>
<td>0.12</td>
<td>-0.05</td>
<td>0.65***</td>
</tr>
</tbody>
</table>

* P value < 0.05
** P value < 0.01
*** P value < 0.001

Figure 6. The relation between lower education and use of public transportation for commuting in 1990
1.3.3 Population size vs commuting modes

We also examined the relationship between population size and modes of transportation (Table 7). As expected, using a car, van, or truck has a strong correlation with the population size as these modes are dominant. Also, for all three years, there is a significant correlation between population size and working at home. It means that in areas with more population more people are working remotely.

Results of the bivariate correlation show a significant correlation between the population size and commuters who used a motorcycle between 1990 and 2010. Meanwhile, no correlation was found between the population size and the use of public transportation.

The correlation between population size and use of other modes of transportation (such as lorry, boat, scooter, and skating) was significant in 1990 and 2000 (no longer significant in 2010). Investments in developing E-scooter networks (such as Bird, or Lime) may help people to use them more often as a commuting mode.

Also, no correlation was found between population size and the number of commuters who walked or biked to work (Table 7).

Table 7. The correlation coefficient between population and modes of transportation

<table>
<thead>
<tr>
<th>Year</th>
<th>Car, van, or truck</th>
<th>Public transportation</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Walked</th>
<th>Other</th>
<th>Worked at home</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.92***</td>
<td>0.06</td>
<td>0.23**</td>
<td>0.11</td>
<td>0.14</td>
<td>0.44***</td>
<td>0.54***</td>
</tr>
<tr>
<td>2000</td>
<td>0.93***</td>
<td>0.05</td>
<td>0.28***</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.4***</td>
<td>0.54***</td>
</tr>
<tr>
<td>2010</td>
<td>0.93***</td>
<td>-0.09</td>
<td>0.23***</td>
<td>0.09</td>
<td>0.01</td>
<td>0.11</td>
<td>0.53***</td>
</tr>
</tbody>
</table>

* P value < 0.05  
** P value < 0.01  
*** P value < 0.001
1.3.4 Age vs commuting modes

Table 8 summarizes the correlation coefficients between age and commuting modes in different years. Using a car, van, or truck has a strong correlation with all age groups because more than 90% of people are using this category for commuting (Table 2). Some significant correlations between age and certain modes of transportation were found including between biking and young people (25 to 34 years old) in 2010 Table 8). That correlation was significant in 1990 for people aged 18-24.

In 1990 and 2000, for almost all age groups (except people above 74) there was a significant correlation between age and use of other commuting modes (in 2010, it was significant for young commuters under the age of 24). The correlation between age and walking is significant for people at age 18-24 in all three years. This age group includes students who choose walking to go to school or their jobs (probably on campus or in the university district).

In addition, in all three years, the number of people in all age groups has a significant correlation with the number of people who worked at home (Table 8).
Table 8. The correlation coefficients between age and modes of transportation

<table>
<thead>
<tr>
<th>Year</th>
<th>Car, van, or truck</th>
<th>Public transportation</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Walked</th>
<th>Other</th>
<th>Worked at home</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>0.78***</td>
<td>0.21**</td>
<td>0.19**</td>
<td>0.04</td>
<td>0.05</td>
<td>0.42***</td>
<td>0.39***</td>
</tr>
<tr>
<td>Under 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.81***</td>
<td>0.13*</td>
<td>0.27***</td>
<td>-0.1</td>
<td>-0.05</td>
<td>0.42***</td>
<td>0.39***</td>
</tr>
<tr>
<td>2010</td>
<td>0.85***</td>
<td>-0.04</td>
<td>0.19**</td>
<td>0.02</td>
<td>-0.1</td>
<td>0.14*</td>
<td>0.39***</td>
</tr>
<tr>
<td>1990</td>
<td>0.48***</td>
<td>0.06</td>
<td>0.32***</td>
<td>0.25***</td>
<td>0.77***</td>
<td>0.53***</td>
<td>0.2**</td>
</tr>
<tr>
<td>18-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.63***</td>
<td>0.16*</td>
<td>0.22**</td>
<td>0.08</td>
<td>0.37***</td>
<td>0.35***</td>
<td>0.19**</td>
</tr>
<tr>
<td>2010</td>
<td>0.56***</td>
<td>0.03</td>
<td>0.1</td>
<td>0.12</td>
<td>0.44***</td>
<td>0.15*</td>
<td>0.29***</td>
</tr>
<tr>
<td>1990</td>
<td>0.87***</td>
<td>-0.04</td>
<td>0.24**</td>
<td>0.16*</td>
<td>0.11</td>
<td>0.3***</td>
<td>0.44***</td>
</tr>
<tr>
<td>25-34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.83***</td>
<td>-0.03</td>
<td>0.2**</td>
<td>0.08</td>
<td>0.05</td>
<td>0.29***</td>
<td>0.4***</td>
</tr>
<tr>
<td>2010</td>
<td>0.78***</td>
<td>-0.07</td>
<td>0.19**</td>
<td>0.18**</td>
<td>0.02</td>
<td>0.04</td>
<td>0.28***</td>
</tr>
<tr>
<td>1990</td>
<td>0.93***</td>
<td>-0.13</td>
<td>0.19*</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.34***</td>
<td>0.67***</td>
</tr>
<tr>
<td>35-44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.92***</td>
<td>-0.08</td>
<td>0.29***</td>
<td>-0.05</td>
<td>-0.08</td>
<td>0.33***</td>
<td>0.61***</td>
</tr>
<tr>
<td>2010</td>
<td>0.89***</td>
<td>-0.16*</td>
<td>0.23***</td>
<td>0.11</td>
<td>-0.07</td>
<td>0.04</td>
<td>0.50***</td>
</tr>
<tr>
<td>1990</td>
<td>0.87***</td>
<td>-0.06</td>
<td>0.1</td>
<td>0.02</td>
<td>-0.09</td>
<td>0.35***</td>
<td>0.62***</td>
</tr>
<tr>
<td>45-54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.90***</td>
<td>-0.08</td>
<td>0.3***</td>
<td>-0.06</td>
<td>-0.07</td>
<td>0.36***</td>
<td>0.69***</td>
</tr>
<tr>
<td>2010</td>
<td>0.87***</td>
<td>-0.12</td>
<td>0.21**</td>
<td>0.07</td>
<td>-0.04</td>
<td>0.09</td>
<td>0.6***</td>
</tr>
<tr>
<td>1990</td>
<td>0.87***</td>
<td>-0.03</td>
<td>0.1</td>
<td>0.01</td>
<td>-0.1</td>
<td>0.35***</td>
<td>0.61***</td>
</tr>
<tr>
<td>55-64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.78***</td>
<td>0.02</td>
<td>0.27***</td>
<td>-0.08</td>
<td>-0.06</td>
<td>0.32***</td>
<td>0.60***</td>
</tr>
<tr>
<td>2010</td>
<td>0.82***</td>
<td>-0.1</td>
<td>0.21**</td>
<td>0.04</td>
<td>0.02</td>
<td>0.11</td>
<td>0.63***</td>
</tr>
<tr>
<td>1990</td>
<td>0.87***</td>
<td>-0.06</td>
<td>0.1</td>
<td>0.02</td>
<td>-0.09</td>
<td>0.35***</td>
<td>0.62***</td>
</tr>
<tr>
<td>65-74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.56***</td>
<td>0.14*</td>
<td>0.13</td>
<td>-0.07</td>
<td>0.05</td>
<td>0.26***</td>
<td>0.43***</td>
</tr>
<tr>
<td>2010</td>
<td>0.64***</td>
<td>-0.1</td>
<td>0.22***</td>
<td>0</td>
<td>0.02</td>
<td>0.11</td>
<td>0.54***</td>
</tr>
<tr>
<td>1990</td>
<td>0.24**</td>
<td>0.14</td>
<td>0.01</td>
<td>0.06</td>
<td>-0.08</td>
<td>0.07</td>
<td>0.19*</td>
</tr>
<tr>
<td>Above 74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>0.43***</td>
<td>0.08</td>
<td>0.06</td>
<td>0.02</td>
<td>0.24***</td>
<td>0.14*</td>
<td>0.3***</td>
</tr>
<tr>
<td>2010</td>
<td>0.39***</td>
<td>-0.08</td>
<td>0.22***</td>
<td>0.02</td>
<td>0.07</td>
<td>0.04</td>
<td>0.37***</td>
</tr>
</tbody>
</table>

* P value < 0.05  ** P value < 0.01  *** P value < 0.001
1.3.5 Race vs commuting modes

In some cities, there is a relationship between race/ethnic groups and a certain mode of transportation with race/ethnic minorities and immigrants more often relying on public transport and using cars as passengers (Transport for London 2012; Syam et al. 2012). Accordingly, we examined a relationship between race groups and transportation modes in Shelby County for all three years (Table 9). As expected, there is a strong significant relationship between all the races and using a car, van, or truck as a dominant transportation mode. Others, including Asian and Hispanic groups, show a significant, however less strong, correlation with biking. That is, in the area with more people of the race other than Whites or Blacks, people tend to use biking more often. For all three years, the correlation between Blacks and public transportation is stronger than for Whites.

All race groups are significantly associated with working at home in all three years (Table 9), with Whites having a positive and stronger correlation while for Blacks it is negative and weaker. Whites have a significant correlation with using the motorcycle in all three years. The correlation between Whites and Other commuting modes is not significant, while for Blacks is significant. In 1990, the correlation between Other races and walking was significant. However, the correlation is not significant for any race after 1990.

Table 9 summarizes the correlation coefficients between race and commuting modes in different years.
Table 9. The correlation coefficient between race and commuting modes

<table>
<thead>
<tr>
<th>Year</th>
<th>Car, van, or truck</th>
<th>Public transportation</th>
<th>Motorcycle</th>
<th>Bike</th>
<th>Walked</th>
<th>Other</th>
<th>Worked at home</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>-0.5***</td>
<td>0.29***</td>
<td>0.15*</td>
<td>0.13</td>
<td>0.17*</td>
<td>0.81***</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>-0.43***</td>
<td>0.21**</td>
<td>0.01</td>
<td>-0.06</td>
<td>0.08</td>
<td>0.83***</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>-0.39***</td>
<td>0.28***</td>
<td>0.11</td>
<td>0.04</td>
<td>-0.12</td>
<td>0.69***</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>0.09</td>
<td>-0.09</td>
<td>-0.07</td>
<td>0</td>
<td>0.34***</td>
<td>-0.36***</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.2**</td>
<td>0.1</td>
<td>-0.07</td>
<td>0.09</td>
<td>0.38***</td>
<td>-0.31***</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td>0.19**</td>
<td>-0.09</td>
<td>-0.09</td>
<td>0.01</td>
<td>0.3***</td>
<td>-0.22***</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>0.49***</td>
<td>-0.25***</td>
<td>0.31***</td>
<td>0.51***</td>
<td>0.54***</td>
<td>0.32***</td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>0.56***</td>
<td>-0.17*</td>
<td>0.1</td>
<td>0.19**</td>
<td>0.1</td>
<td>0.32***</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.16*</td>
<td>0.08</td>
<td>0.22***</td>
<td>0.06</td>
<td>-0.12</td>
<td>0.3***</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P value < 0.05
** P value < 0.01
***P value < 0.001

1.4 Discussion

In this section, significant relationships between modes of transportation and other factors are further discussed.

Since both land cover and land use are used to study diversity or fragmentation and are even sometimes used interchangeably, they represent nevertheless different concepts. Land cover represents the physical cover of an area and land use represents the land utilization. Thus, mapping of urban areas with both land cover and land use systems is one of the key research topics in human-environment monitoring (Sertel et al., 2018) and is essential for a sustainable urban planning.
We show that using both types of measures in combination is beneficial for the effective assessment of land cover /land use maps by decision-makers. The findings indicate that the values of land cover mix are almost the opposite of those of land use mix, that is, tracts with higher values of land use mix indicating a better and more desirable urban form (mostly located in the central part of the study area) have lower values of the land cover mix due to a transition to mostly a developed land from other types of land cover compared with suburban areas which have higher mixed land cover areas and lower land use mix values. Accordingly, measures based on land cover and land use including land cover diversity and land use diversity are different as evidenced by our figures (compare land cover mixture shown in Figure 4-C and land use mixture in Figure 5-B, respectively). Further, these measures have various relationships with different transportation modes.

Our results (Section 1.3) indicate statistically significant positive relationships between the land cover mixture and car/van/truck as well as working at home, and a negative relationship with public transit for 1990, 2000, and 2010: the greater the mix in land covers, the less the use of public transportation. As mentioned above, the greater land cover diversity is typical for the suburbs where commuting by car is a preferred option while public transportation is much less developed in the US context. Both land diversity measures have the opposite associations with working from home: while the land cover mix is statistically significantly positively related for all years studied, land use mix has a negative association. Additionally, higher diversity in land cover is related to greater car use and less commuting by walking and biking, while no links have been found between active and passive commuting and land use mix.

These different associations provide additional evidence of the benefit of the simultaneous use of both land measures in studying travel behavior. Thus, one contribution of
this study is that our research provides a more comprehensive view on commuting patterns using both land cover and land use data to better reveal the complex commuting patterns.

More discussion about each commuting mode is provided in Section 1.4.1 to Section 1.4.7, separately.

1.4.1 Car, van, truck

The analysis of data and the results of correlation presented in section 1.3, showed that there is a relationship between the land cover mixture and the use of car, van, or truck in 1990 and 2010 (insignificant for 2000, see Table 5). This relationship can be interpreted as follows: where there is a greater land cover mixture, people tend to use a car, van, or truck more. In addition, the analysis indicated that the use of a car, van, or truck is not linked to the land use mixture (Table 5). This contrasts findings by Sarkar et al. (2013) where any change in the existing land use pattern in Agartala (an Indian small-sized city) shifts choices towards the use of private modes. In Shelby County, the correlation between Simpson’s index of land use and the use of public transportation is significant. Therefore, a change in land use patterns may affect the use of public transportation. Our finding is in agreement with Chatman (2003) who found that higher employment density is associated with a lower likelihood of automobile commuting, and residential density is correlated with the choice of public transit and walking for commuting in the United States in 1995. People who prefer to use alternative modes may choose to live in a dense mixed-use neighborhood which is often more walkable and have better access to public transit (Chatman 2003).

In all three years, using a car, van, or truck is correlated significantly with the number of people with a bachelor’s/some college, graduate, or professional degree (Table 6). Also, that mode of transportation has a strong and significant correlation with all age groups (Table 8). The
population is meaningfully and significantly correlated with the use of a car, van, or truck (Table 7).

Blacks use private vehicles less than other ethnicities (Table 9) which is in agreement with the American Community Survey: the rate of driving alone to work is less for Blacks (71% in 2006) in comparison with Whites (80% in 2006) in the United States (McKenzie 2015).

1.4.2 Public transportation

As Table 5 reports, a meaningful negative correlation has been detected between the use of public transportation and land cover mixture: in places with more mixed land cover, people tend to use public transportation less. Conversely, in places with less land cover mixture, people tend to use public transportation more. This negative correlation occurs in the developed area (“Developed” as a land cover class, see Figure 2), where there is not a good mixture of other classes due to the single dominance of developed areas, while people are using public transportation more in the developed area. Obviously, in the non-developed area (where there is a greater mixture of land covers including barren, forest, herbaceous, shrubland, planted, water, and wetlands), there is little to no public transportation compared to developed and urbanized areas.

Also, public transportation has a significant positive correlation with land use mixture (Table 5): where there is more mixture of land uses, people tend to use public transportations more. These findings are in line with those by Schmitt (2018) in Seattle, WA, where the number of people using public transportation to go to work in downtown (a place with more mixture of land uses) is more than the number of people who drive alone. The use of public transportation correlates with educational level, too (Table 6). Having less education (high school diploma or less than 9th degree) is positively correlated with the use of public transportation, while better
education (some college degrees or higher education), is negatively correlated. It means reliance on public transportation is lower in areas with more educated people, while in an area with less educated people public transportation use is higher. These findings are in contrast with those by Meira et al. (2014) in Brazil where university students and professors use public transportation (60%) more than private cars (30%) and other modes (10%).

We also detected some associations between age and travel modes. Younger people (under 24) tend to use public transport more often evidenced by the positive correlations (Table 8). The younger population prefers to drive less. These findings are in agreement with those by Bauman et al. (2012) that age is inversely correlated with the use of active transportation. Table 8 reports a correlation between the use of public transportation and the number of people aged 65-74 in 2000.

The population size is not correlated with the number of people who use public transportation (Table 7). However, there is a negative correlation between the use of public transportation and Whites, and a positive correlation for Blacks (Table 9). This finding is in agreement with well-documented examples of Black communities in the UK, relying more heavily on transit compared with Whites (Transport for London, 2012). In an area with more Whites, people tend to use public transportation less, while in an area with more Blacks, people use public transportation more.

The negative relationship between White ethnicity and public transportation gets weaker captured by a decreasing magnitude of the correlation coefficient (-0.50, -0.43, and -0.39 in 1990, 2000, and 2010, respectively) (Table 9) indicating the likelihood that more Whites may use public transportation in 2010 than in 1990. This finding is in line with the findings of Bauman et al. (2012) that White ethnicity is correlated with active transportation. Similar trends have been
observed for Blacks: the correlation coefficient is 0.73, 0.56, and 0.4 for 1990, 2000, and 2010, respectively (Table 9). Based on Table 9, the correlation between Blacks and public transportation is stronger than for Whites in all three years. This finding is in agreement with that by Battelle (2000) that ethnicity and gender play an important role: African-Americans and Asians used public transit more often than Hispanics and significantly more often than Whites from 1983 to 1995 in the United States, while African-American women depend heavily on public transit as they are nine times as likely to use public transit as Whites.

The correlation results between car, van, or truck and ethnicity confirm the above conclusion. From 1990 to 2010, the correlation between the number of Whites and the use of a car, van, or truck has declined. So, fewer Whites tend to use private vehicles in 2010 and tend to use public transportation or other modes more. From 1990 to 2010, the correlation coefficient between Black ethnicity and use of a car, van, or truck increased from 0.09 to almost 0.2 meaning more Blacks are using private cars, while less relying on public transportation.

1.4.3 Motorcycle

No significant correlation between commuting by motorcycle and land cover mixture has been detected (Table 5). This travel behavior in Shelby County is in contrast to that in China, and Albania, where the proportion of people using a motorcycle is much smaller in cities with higher urbanization rate (high urbanization rate yields a lower Simpson’s index due to a less mixture of land cover; see Section 1.2.3) (Zhu et al. 2017; Instat 2014). For land use, there was a negative correlation between the mixture and the number of people who use motorcycles (Table 5): in the area with less land use mixture, people use motorcycles more. In all three years, there was a correlation between the number of people who have Bachelor's or some college degrees with commuting by motorcycle (Table 6).
In 2010 (Table 8), there is a significant correlation between people aged above 65 and using a motorcycle as a commuting mode, while in 2000 and 1990 there was no significant relationship. The strongest correlation between the use of motorcycles and age groups is found for commuters aged 18-24 in 1990, but in 2000 and 2010 it was the most significant for those aged 45-54, and 35-44, respectively (Table 8). In comparison, in 2013, Vicki & McLaughlin (2013) used a survey of 424 motorcyclists and found that the largest group of motorcyclists in Virginia who ride more annual miles (on average 7,942) than others for work as well as for pleasure are mostly males (93% males, 7% females) aged 40-59 years old (averaging 50 years old).

The population size also has a correlation with the number of people who are using motorcycles (Table 7). While Whites in all three years use motorcycles for commuting, Blacks did not use them more often. In 1990, other ethnicities such as Asians or Hispanics tended to use a motorcycle as a mode of commuting, but in both 2000 and 2010, they used it less (Table 9).

1.4.4 Biking

In a commuting study, Cervero (1996) concluded that except for walking and biking modes, residential densities exhibit stronger influences on mode choices than land use mixture levels. However, our findings do not present any correlation between biking and neither land cover nor land use mixture (except for land cover mixture in 1990) (Table 5).

Table 6 reports a correlation between the number of people who use bikes for commuting and the number of people who have higher education in 2010 (absent in the 1990s and 2000s). It is probably due to the local governments’ and social efforts that encouraged people to use bikes more often. The strongest correlation is for people with graduate or professional degrees. This finding in Shelby County is compatible with that by Panter et al. (2011) in the UK: people with a
degree-level education commute by bike more often in Cambridge, UK. However, it is in contrast with the finding by Meira et al. (2014) in Brazil that for students and professors public transportation is their first priority for commuting.

In terms of age (Table 8), the strongest correlation is for the 18-24 age category in 1990, and the age of 25-34 in 2010. This finding in Shelby County agrees with the American Community Survey in 2014 that younger people biked to work more than other age groups between 2008 and 2012 in the United States (McKenzie 2014). These age groups are probably represented by students, as we found a correlation between higher education and the use of bikes.

Also, in terms of race (Table 9), other ethnicities than White or Black use bikes more often as a mode of commuting.

1.4.5 Walking

Walking is strongly associated with land use diversity (defined as the number of different land uses in a given area and the degree to which they are represented in land area, floor area, or employment) (Ewing & Cervero 2010). Walking frequency is linked to land use mix (Duncan et al. 2010). Our findings in Shelby County (Table 5) do not show any significant correlation between walking and the Simpson’s index of neither land cover nor land use (except for a land cover in 2000) and agree with a study by Lu et al. (2018) on a marginal effect by a land diversity on walking.

Age is associated with non-motorized commuting. A significant correlation was found between the number of people whose age is 18-24 and walking to work in all three years (Table 8) which is compatible with that by the American Community Survey for the United States (McKenzie 2014) and with a study by Saris et al. (2013) who had similar results for the Netherlands. However, in Shelby County, a correlation between the number of people older than
74 and the number of people who walk is found in 2000. As mentioned in Section 1.4.3, older people in Shelby County were found to use motorcycles more than walking in 2010.

White ethnicity is positively correlated with the use of active transportation (walking or biking) in low-income and middle-income countries (Bauman et al., 2012). We did not find a significant correlation between neither Whites nor Blacks and walking or biking (except for White and biking in 1990) (Table 9). However, the United States is classified as a high-income country (The World Bank 2019), making these findings unsuitable for comparison.

1.4.6 Other transportation modes

Other commuting modes such as lorry, boat, scooters, skating, etc. (Vehicles and Transportation Vocabulary Word List 2017), are also correlated with a mixture of land cover in 2010 (Table 5). Except for people with graduate and professional degrees, there is a significant correlation between the number of people who are using other modes of transportation for commuting and educational level (Table 6). As already discussed, people with graduate and professional degrees prefer to use a bike, public transportation, or private vehicles as a mode of transportation rather than other modes.

In addition, Blacks prefer to use other modes more often than other ethnicities in all three years (Table 9). This finding from 1990 to 2010 is in agreement with that by Battelle (2000) from 1983 to 1995: African-Americans (2.66%) used other modes of transportation more often than Whites (1.45%), Hispanics (2.54%), and Asians (1.41%) in the United States from 1983 to 1995 (Battelle 2000).

As Table 7 presents, the correlation between population size and the number of people who used other commuting modes was significant in 1990 and 2000 while in 2010 the correlation is weak and insignificant. Investments in developing E-scooter and E-skateboard
networks (such as Bird, or Lime) may bring the motivation back to people to use them more often for commuting.

1.4.7 Working at home

We found a significant correlation between people who were working at home and a mixture of land cover (Table 5): where there is a better mixture of land cover, more people are working at home. Also, even though there was a correlation between almost all educational degrees and the number of people who were working at home, this correlation is much stronger for higher educated people (Table 6). Higher educated people are more likely to work at home than people of other education levels. For people who have education less than a 9th degree, the correlation was negative. Less educated people are less likely to work at home. The elderly population in the United States had shorter commute times between 1983 to 1995 due to the much greater tendency to work at home (Battelle 2000). We found all age groups correlate significantly with working at home in all three years (Table 8), with Whites having the strongest association (Table 9).

1.5 Conclusion

The premise of the paper is that the relationship between land use mixture, land cover mixture, and commuting choice matters, and these factors continue to influence each other. To examine this relationship, we examined the association between commuting modes and both land use mixture and land cover mixture (measured by the Simpson’s index) as well as socio-demographics in Shelby County in 1990, 2000, and 2010. Studying seven modes of commuting indicated that both land use and land cover mixture, as characteristics of a built environment, were associated especially significantly with the use of public transportation and working at home suggesting that city and regional planners should focus on reducing land cover segregation,
and also land use segregation, to improve the share of public transportation and working at home.

No significant association was found between neither land use mixture nor land cover mixture and walking (except for land cover mixture in 2000) in Shelby County at the Census tract level. As Duncan et al. (2010) mentioned, the geographical scale of the measure does matter for this relationship. Therefore, to evaluate the influence of mixture on walking, future studies are needed using different geographical units.

Collecting primary data, by using a survey or an interview, provides the most updated data for transport-related studies. However, primary data collection is very time-consuming and expensive for studying big groups of people. In this research, the use of secondary data provided by the Census was due to the extensive population (936,130 in 2019) and large spatial coverage (763.17 sq. miles) over the entire extent of Shelby County. Also, to compare the results and finding the trends, data from different years (1990, 2000, and 2010) were needed. However, as the mode shares of public transit (1.49% in 2010) and bike (0.13% in 2010) were very low in the study area, the interpretation of results for modes other than private vehicles (with a share of 93.64% in 2010) needs caution due to the sample’s small size. Future studies could use multivariate models for a better understanding of these complex relationships and consider the influence of income, gender, and household life cycle. Moreover, future studies are needed to determine whether causation can be established between the variables with significant association.

We add to the well-documented impact of socio-demographic attributes on commuting and show that land use mixture, land cover mixture, commuter’s age, race, education, and population of neighborhoods, influence commuting mode choice. City planners responsible for
the development of a city can use these results to make neighborhoods more conducive to an active lifestyle. The comparison of the results in different years enables estimation of future changes in commuting mode choice and helps to identify priorities for the development of a specific mode of transportation. Moreover, policies can be more practical to prevent or control local/regional problems like urban sprawl, traffic congestion, and emissions which may increase substantially since shift in a preferable commuting mode may occur from public to private transportation by car during a pandemic such as COVID-19 (Das et al. 2021). Understanding travel behaviors and the need of travelers in an area may help policymakers to develop strategies for encouraging active modes and public transport use in the post-COVID world.
Chapter 2: Pattern-based Calibration of Cellular Automata by Genetic Algorithm and Shannon Relative Entropy

Abstract

While cellular automata (CA) are considered an effective algorithm to model urban growth, their precise calibration can be challenging. Shannon relative index (SRI) is an indicator of urban sprawl accounting for dispersion or concentration of built-up/non-built-up areas. This study uses SRIs directly in the calibration of CA as patterns, applying a genetic algorithm (GA). Moreover, the Kappa coefficient is used in the calibration process. CA was calibrated using data for 2001 and 2006 and validated using 2011 data to model urban growth in Shelby County, TN, USA.

Results indicate that the Kappa coefficient achieves the highest value using the proposed method (89.48%) compared with a GA without patterns (86.15%, that underestimates 32.22 sq. km) or logistic regression (85.83%, that underestimates 36.76 sq. km). A more precise calibration of urban growth using the proposed method helps city planners to provide more realistic models for the future of the region.

Keywords: Urban growth; sprawl; cellular automata; calibration; Shannon relative index; Kappa coefficient;
2.1 Introduction

Urban development and land-use patterns have been modeled early on by Weber (1909), Burgess (1925), Hoyt (1939), and Harris and Ullman (1945). These early static overly simplistic models were criticized for the lack of practical assumptions, being non-operational and failing to capture existing patterns (Liu, 2009). The systems approach in urban modeling considers urban systems consisting of population, land, employment, services, and transport, where these elements interact with each other and the environment and thus shape the evolution and change of the system. A new understanding of cities as evolutionary and complex self-organizing systems (Allen, 1997; Batty, 2009; Batty, 1997; Batty et al., 1997) encouraged new urban growth simulation modeling techniques including those based on the automata to study the behavior of a self-organizing system, the patterns, and processes of urban development shaped by human behavior in space and various social processes (Liu, 2009).

Cellular automata (CA), a collection of fixed cells regularly arranged in a cellular space, is the simplest model and a special type of automata. A two-dimensional regular grid is the most common CA (Liu, 2009), however, other arrangements such as one-dimensional CA (Di Gregorio & Festa, 1981), or CA with irregular cells (Barreira-Gonzalez et al., 2019) have been developed to represent objects and elements with different shapes and sizes. Each of the regular spatial cells that constitute a CA can only have one of a finite number of possible values or states (e.g. urban or rural land) determined by locally defined transition rules which in turn depend only on the states of the nearest neighbors (Liu, 2009). The simple rules when combined generate complex structures (Wolfram, 1984) and serve as the algorithms to simulate and realistically represent the complex behavior of the system (Liu, 2009) including spatial patterns (Soares-Filho & Coutinho, 2002). The spatial pattern is an arrangement of objects on the Earth and also...
includes the space in between of those objects, thus, various patterns may form due to different arrangements, density, quantity, clustering or relationship between objects (Keys-Mathews, 2003).

CA modeling is widely used due to the ability to accurately and efficiently replicate observed complex dynamics using a simple rule system, and the incorporation of socioeconomic and biophysical forces driving land use change (Newland et al., 2015). Since cities can be conceived of as consisting of a two-dimensional regular grid of $n \times n$ cells (Liu, 2009), with simple transition rules generating complex urban development patterns due to the CA’s self-organization and self-reproduction ability, CA have been adopted in modeling various natural, social and economic phenomena (Sidiropoulos & Fotakis, 2011) including urban modeling to simulate the urban growth (Barreira-Gonzalez et al., 2019; Xia et al., 2018; Liu et al., 2017; Rienow, 2016), understand processes of land use changes (Roodposhti et al., 2020) and in various optimization problems (Strange et al., 2001; Heinonen & Pukkala, 2007; Mathey et al., 2008; Afshar & Shahidi, 2009; Guo et al., 2007). Since the early 2000s, urban growth and land use change were simulated by the SLEUTH, a modified cellular automata (CA) model which is an acronym for input data layers of the slope, landuse, exclusion, urban extent, transportation, and hillshades (Clarke, 2008a and 2008b; Clarke et al, 2007). The SLEUTH’s predefined growth rules are fixed, but based on behavioral parameters vary from perfect to zero impact at each time step (Clarke-Lauer & Clarke, 2011; Clarke, 2008a; Silva & Clarke, 2002). A reader is referred for a detailed review of applications of the SLEUTH model to Chaudhuri and Clarke (2013).

CA can be represented by a set of mathematical equations used to model self-organization through homogenous states (Aithal et al., 2017; Li et al., 2013). Self-organization is the tendency of a system to naturally develop ordered patterns (Torrens & O’Sullivan, 2001), or the process
when internal organization increases in complexity with no outside guidance, such as when a randomly chosen “disordered” initial state can generate some structure (Liu, 2009).

While CA is a common algorithm to model urban growth, its precise calibration is knowledge-intensive and time-consuming due to a high degree of uncertainty (Roodposhti et al., 2020) and the process can be challenging, resulting in less precise urban growth modeling. Since some parts of the real-world systems are not reproduced by the model, validating the results of the model is essential (Bharath et al., 2018; Liu, 2009). Calibration of CA is a process of estimating the best combination of CA values such that the modeled urban growth can better match the real process (Shan et al., 2008). During calibration, the model’s outcomes are verified to test model performance. Accuracy is assessed with the degree of conformity to test if the simulated forecast conforms to the observed data achieved with statistical techniques including factor analysis and regression (Liu, 2009). A model is often calibrated to estimate its parameters that fit an observed data the best, that is, the “goodness-of-fit” (Bell et al., 2000). The accuracy of the model’s predictions determines the validity of a simulation model (Kilbridge et al., 1970).

Regardless of different techniques of CA calibration, a cell-to-cell calibration of CA is predominantly performed; e.g. at each iteration, the future state of all cells is simulated and compared with the real state of those cells, while trying to minimize a cost function (also called “a fitness function”, see Newland et al., 2015) which is a measure of how close a simulation is to the reality. Cost functions such as a Kappa coefficient or overall accuracy are based on the cell-to-cell comparisons. The calibration of CA which uses these cost functions fails to capture spatial patterns, such as a diversity of urban land cover, in each neighborhood. Spatial patterns such as shapes, densities, and diversities are mostly used only in the model validation (Naghibi et al., 2016a and 2016b; Bharath et al., 2018). To improve the precision of calibration, a cost function
that realistically estimates spatial patterns is required to model complex environmental issues including urban growth (Soares-Filho & Coutinho, 2002). As information derived from spatial patterns can describe changes in urban structures, including a diversity of urban land cover, the calibration of CA can better describe neighborhood characteristics and hence, achieve higher calibration precision (Newland et al., 2015).

A limited number of studies used landscape metrics in urban growth modeling. A pioneering study by Soares-Filho & Coutinho (2002) used landscape structure measures including fractal dimension, contagion index, and the number of patches for each land use and land cover class to study land cover dynamics and to evaluate the performance of the manual calibration of CA. Other studies used the Shannon Entropy index in addition to landscape metrics to examine changes in urban land use patterns and quantities (Araya & Cabral, 2010).

Studies of urban growth use various spatial metrics to verify the calibration of CA. The performance of a calibration process depends on the choice of spatial metrics; thus, it is important that the selection process be more fully informed. The most common ones include the total built-up and non-built-up areas, a number of patches, edge length, the largest patch index, patch area, contagion, the distance to the nearest patch of the same type, patch density, shape index, clumpiness index, percentage of land adjacencies and cohesion index (Mustafa et al., 2018; Aithal et al., 2017; Araya & Cabral, 2010). Other studies rely on different spatial landscape metrics including patch size, edge density, shape index, fractal dimension, interspersion and juxtaposition of edges, Shannon’s diversity index, and Simpson’s diversity index (these ecology-based metrics quantify landscape structure) (Newland et al., 2015).

While the above-mentioned studies used spatial landscape metrics just for model validation, they can also be used in the calibration process of CA (Feng & Tong, 2018; Li et al.,
2013). For instance, a cost function of the calibration defined as a linear combination of a percentage of landscape, largest patch index and landscape division in addition to the overall accuracy, yields a better calibration (7.2% error) in comparison with the regular calibration of CA using a genetic algorithm (GA) (33.6% error) and logistic regression (27.7% error) in urban growth modeling (Li et al., 2013). Using differential evolution (DE) into CA to simulate land use patterns and evaluating with a cell-by-cell-based statistical comparison yields the same performance as CA calibrated using GA reaching 92.4% in overall accuracy (Feng & Tong, 2018).

The individual and overall accuracies are measured by the percentage of correctly categorized cells and exclude the omission and commission errors, and thus, represent a simple measure of agreement discounting chance matching (Liu, 2009). The Kappa coefficient is often used to assess the predictive power of the model, for this, the results of the simulation are compared to a real map (a reference) for the same year computing the Kappa coefficient (Araya & Cabral, 2010). The Kappa analysis produces a Kappa coefficient computed as the difference between the observed agreement and the chance agreement. When the value of the Kappa coefficient approaches 100%, it indicates the greater agreement between the simulated and the reference land use/land cover maps. The Kappa analysis is commonly integrated into virtually every accuracy assessment and became a required component of most image analysis software packages that include accuracy assessment procedures (Congalton & Green, 2009). However, the Kappa coefficient is criticized for first, being a single summary statistic of the agreement while having the two components of disagreement, quantity and allocation disagreements, allegedly better helps to understand and interpret results (Pontius & Millones, 2011), and second, due to a potential underestimation of the classification agreement and overestimation of the degree of a
chance agreement (Foody, 1992). Despite the criticisms, the Kappa coefficient has very powerful properties including its ability to test for significant differences between two independent coefficients, and thus, it represents a vital accuracy assessment measure (Congalton & Green, 2009; Sicre et al., 2020; Momeni et al., 2020; Mohammed et al., 2019; Farg et al., 2019; Elmaizi et al., 2019).

The research question of this study is whether including the Shannon relative index (SRI) as an indicator of land use/land cover patterns in the calibration process of CA can improve the calibration precision in urban growth modeling. In the current study, we use a pattern-based calibration of CA by using a GA and the SRI of each neighborhood directly in the calibration process. Also, in comparison with other studies that used overall accuracy to analyze the goodness of fit in the cost function, the current research uses the Kappa coefficient to consider commission and omission errors in the cost function of calibration (Rossiter, 2004; ArcGIS Pro, 2018). Using patterns in the cost function of the calibration makes the results of the optimal parameters more realistic (Newland et al., 2015). Therefore, a special cost function is designed in the calibration process to use the Kappa coefficient and SRIs during the multiple iterations of GA. In addition, compared with other studies, more layers of information (see Section 2.2) are used as driving forces to achieve more realistic calibration results. CA will be calibrated using 2001 and 2006 data and will be validated using data for 2011 in Shelby County, TN. The results of the proposed method will be compared with the results of the conventional methods such as a GA and logistic regression (LR). Using the proposed method, city planners may potentially calibrate CA with greater precision, and therefore produce a more reliable model for the future of cities. Ultimately, decision-makers will be able to direct urban development to smart growth and avoid urban sprawl.
This paper is organized as follows: Section 2.2 discusses the methodology. The concepts of the Kappa coefficient and the SRI are also introduced briefly in this section. In addition, the study area and the required data are discussed in the same section. Section 2.3 presents and compares the results of the calibration. Section 2.4 and Section 2.5 are dedicated to the discussion and conclusion, respectively.

2.2 Methodology

Similar to previous studies that used binary states to model urban land-use changes using only urban and non-urban land (Roodposhti et al., 2020) or simulate the process of non-urban to urban conversion (Li et al., 2000), we defined the cells as discrete land use states such as either non-built-up or built-up. In this urban growth applications of CA, each cell represents the urban landscape and can have only one of two possible development states (from here onwards, termed also “a state”). Verburg et al. (2004) define a cell’s probability of conversion (e.g., from a non-built-up to a built-up area) as:

\[
\log \left( \frac{P_i}{1-P_i} \right) = \beta_0 + \sum \beta_{N,l} F_{i,N,l} \tag{2}
\]

where \(F_{i,N,l}\) is an independent variable for cell \(i\), neighborhood \(N\) and land use/land cover \(l\). Also, \(\beta_{N,l}\) is a coefficient to be estimated for the neighborhood \(N\) and land use/land cover \(l\). For the cell \(i\), \(P_i\) indicates the probability of conversion (Verburg et al., 2004). The optimized values in Equ. 2 (such as \(\beta_{N,l}\) and \(\beta_0\)) can be estimated by GA. Also, GA can simultaneously determine an appropriate threshold for the probabilities of conversion to categorize them. For example, if a cell has a development state as a non-built-up area, and if the probability of conversion is less than the threshold, the state of the cell will remain as non-built-up in the next iteration, otherwise
its state will change to built-up. The cost function of the GA is defined according to the problem domain (Kramer, 2017; Mathworks, 2018). In this study, we used the GA with a cost function based on the Kappa coefficient and SRI.

The Kappa coefficient is a statistic that measures an inter-rater agreement of qualitative (categorical) items. It is generally thought a more robust measure than simple percent agreements such as overall accuracy which excludes the omission and commission errors (Pontius & Millones, 2011). The Kappa coefficient subtracts the estimated effect of chance assignment of cells to categories and thus adjusts the percentage of correctly categorized cells (Liu, 2009). The Kappa coefficient ranges from 0 to 1. While a value of 0 means no agreement between the results and the reference data, the value of the Kappa coefficient of 1 means that the results and reference data are identical (Liu, 2009).

Equ. 3 shows the Kappa coefficient based on a confusion matrix. A confusion matrix (also known as an “error matrix”) describes the performance of a supervised-learning algorithm. A confusion, or error matrix, is commonly used to express the accuracy of land cover classification for remotely sensed data where the actual reference data representing the ground truth is compared with the classified image or simulated data (the model-generated), on a cell-by-cell basis (Liu, 2009). In the confusion matrix, each row counts the instances of a predicted state while each column counts the instances of an actual state (or vice versa) (Powers, 2011). A confusion matrix produces values of overall accuracy, omission error, and commission error.

\[
K = \frac{N \sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r} (X_{i+}X_{+i})}{N^2 - \sum_{i=1}^{r} (X_{i+}X_{+i})} \tag{3}
\]
where \( r \) is the number of rows (or columns) in the confusion matrix, \( N \) is the total number of observations, \( X_{ii} \) is the observation in row \( i \) and column \( i \) of the confusion matrix, \( X_{i+} \) is a marginal total of row \( i \), and \( X_{+i} \) is a marginal total of column \( i \) (Cohen, 1960).

The Shannon index as a diversity index is a robust indicator of urban sprawl which can account for the dispersion or concentration of built-up or non-built-up areas at state, county, tract or block group level (Chong, 2017; Herries, 2019). The Shannon index is defined mathematically as:

\[
H_n = -\sum_{i=1}^{S} p_i \log (p_i) \tag{4}
\]

where \( S \) is the total number of species (e.g. land cover classes), and \( p_i \) is the probability of occurrence of species \( i \) (e.g. land cover \( i \)) in the zone \( n \) (Roghair & Dolloff, 2013). In this study, a zone, the spatial unit of computation, is a block group (the total number of block groups = 653). A block group is the next level above census blocks (the smallest geographic area for which the Bureau of the Census collects decennial census data) in the geographic hierarchy. A block group is the smallest geographic entity for which the decennial census tabulates and publishes sample data (U. S. Census Bureau, 2018).

The value of the Shannon index, \( H_n \), ranges from 0 to \( \log_e(n) \). To limit the index values between 0 to 1, in order to compare them, the Shannon relative index (SRI) is calculated for each block group using the following equation:

\[
H'_n = H_n / \log_e(n) \tag{5}
\]
where $H_n$ is the value of the Shannon index and $n$ is the total number of zones (Deka et al., 2012). A higher value of SRI indicates the dispersion of built-up areas and an occurrence of sprawl in the zone, while a lower value of SRI indicates a high concentration of a dominant class (either built-up or non-built-up) in the zone, therefore further conversion is less likely to occur (Chong, 2017; Yeh & Xia, 2001). Since the SRI values computed for each of the block groups indicate a certain pattern within a county (which includes several block groups, here $N=653$ in total), an integration of SRI into the CA calibration process is referred to in the study as “the GA with patterns”, while the process of calibration of CA using just a GA with no SRI calculation, is referred to in the study as “the GA with no patterns”.

In the proposed calibration method, at each iteration of GA, which is also known as a “generation” (see Deb, 2004), the CA values are determined and the states of all cells (e.g. built-up or non-built-up) are simulated based on these values. The simulated state of cells is compared with the actual state of cells, cell by cell, and based on that a confusion matrix will be generated for the simulation process. Then based on Equ. 3, the Kappa coefficient is calculated for the whole simulation. In addition, the SRI for each block group is calculated (using Equ. 5) and compared with the SRI of block groups in reality. The Kappa coefficient integrated with the maximum absolute error for simulation of SRIs was used as the cost function of the GA (Equ. 6).

$$\text{Cost}_\text{function} = \alpha \times (1-K) + (1-\alpha) \left( \max \{|\text{errors of simulation for SRIs}|\right) \quad [6]$$

where $\alpha$ is a constant defining the weights, $K$ is the Kappa coefficient and $|...|$ denotes the absolute value. Based on Equ. 6, at each iteration, the values of SRI (for each block group) are calculated and compared with the real values of SRI to determine errors in the simulation of
SRIs. Then the maximum of absolute errors and the error for the Kappa coefficient are combined proportionally (based on $\alpha$) to calculate the cost of simulation at that iteration.

Figure 7 shows the procedure of the CA calibration based on the GA with patterns. As Figure 7 shows, the SRI is calculated at each generation (for each block group) and fed into the cost function to be compared with the actual SRI using Equ. 6.

Figure 7. Flowchart of calibration of CA based on a GA with patterns (SRI)
The results of the calibration of CA based on the GA with patterns will be compared with the results of the CA calibration based on the GA without patterns. Moreover, the results will be compared with the results of the CA calibration based on the conventional logistic regression (LR).

2.2.1 Study area

Shelby County, the state’s largest county both in terms of population (927,644 in 2010) and geographic extent (1976.60 sq. km), is located in southwest Tennessee, USA. Its county seat, Memphis is the largest city in Shelby County with a population of 646,889 in 2010 which is also located southwest of Shelby County (U.S. Census Bureau, 2019).

The population of Shelby County increased by more than 12% from 1990 to 2010. Meanwhile, land cover in Shelby County has changed to provide facilities, housing, and jobs for people; while developed areas expanded by 60%, planted/cultivated areas shrunk by more than 18% and forested areas diminished by almost 40% (Momeni & Antipova, 2022b). Figure 8 shows the built-up (developed) areas in Memphis and Shelby County in 1992, 2001 and 2011 (MRLC, 2019). As Figure 8 presents, over the past two decades, developed areas have grown up in Shelby County, while the majority of development has occurred in rural areas out of Memphis’s boundaries, especially in the east and south-east of the city.
Figure 8. Built-up (developed) areas in Shelby County in 1992, 2001 and 2011

As population changes, Memphis sprawls eastward, where suburbs and towns such as Cordova and Bartlett grow up rapidly (Ciscel, 2000; Shelby County Government, 2019). Therefore, to know the possible extent of Shelby County’s developed areas in the future, a simulation of growth is required. This study is attempting to provide a precise calibration of CA, as a base for growth modeling for the nearest future in Shelby County.

2.2.2 Data and data preparation

Based on the previous simulation studies, we used a retrospective prediction approach by predicting the known past and the present using historical data (Liu, 2009). Remote sensing is a reliable tool to study urban features and land covers. Using an image classifier each pixel in a satellite image is assigned a label of a class (or more classes) with the maximum confidence of assignment to produce a land cover map (Momeni et al., 2020; Momeni et al., 2018). Land cover maps of Shelby County are obtained from the National Land Cover Database (MRLC, 2019) for
2001 and 2006. These maps are used to obtain empirical information for the calibration of CA. Also, the land cover map of 2011 is used for validation of the simulation (as reference data). These maps are derived from Landsat satellite imagery at a spatial resolution of 30m×30m by using a modified level 2 classification system (Yang et al., 2018). That is, each cell on the model represents an area of 30 × 30 sq. meters on the ground. Urban development is the result of various forces including geographical (such as physical constraints including water bodies and topographic features that may limit or accelerate urban development) and socio-economic conditions (such as accessibility to employment and administrative nodes, and facilities such as urban infrastructure, transportation network, proximity to existing urban lands, etc., that may impact urban development). The commonly used factors driving urban development include proximity to the city center, distance to primary/ secondary/ main roads, distance to the existing urban settlements, population density, land policy change/ spatial constraints for land development (such as water bodies and other conservation regions), and distance to commercial/ district centers, with land use policy, slope, and distance to roads found to be the most important drivers of urban expansion (Mustafa et al., 2018). While distance to the secondary roads impacting the formation of the urban growth pattern the most, the densification process is mainly determined by zoning, slope, distance to different roads and richness index (Feng & Tong, 2018; Mustafa et al., 2018; Bharath et al., 2018; Al-Darwish, 2017). Accordingly, in addition to land cover maps, other data including airports, cities, Central Business Districts (CBDs), highways, major roads/ streets, land uses, rivers and streams, railroads, DEM and borderlines of block groups for Shelby County and surrounding counties, were collected to calculate the driving forces of the growth in Shelby County (IMPUS, 2018; U.S. Census Bureau, 2019; Earth Explorer, 2018; Assessor of property, 2018; TNGIS, 2018; DIVA-GIS, 2018).
In the data preparation step, land cover maps were imported to ArcMap 10.5.1 and reclassified as developed and non-developed areas. In addition, using spatial analysis in GIS, Shelby County was divided into 30m×30m geo-referenced cells (2,258,471 cells in total) according to the geo-referenced pixels of land cover maps, and 14 driving forces of growth were calculated for each cell. Based on the review of the relevant literature, driving forces for each cell include the distance to the 1) nearest airport, 2) nearest CBD, 3) nearest city, 4) nearest developed area, 5) nearest highway, 6) nearest major road, 7) nearest railroad, 8) nearest residential area, 9) nearest river/stream, 10) nearest waterbody (lake, bay, wetlands) as well as 11) easting coordinate of the cell, 12) northing coordinate of the cell, 13) the elevation of the cell, and 14) the number of developed cells in a 7-cell×7-cell neighborhood of the cell (Li et al., 2013; Liao et al., 2016; Aburas et al., 2017; Mohammady & Delavar, 2016; Votsis, 2017). These driving forces of urban sprawl represent the likelihood of each cell converting to a developed area. Driving forces are context-based and can vary from region to region and country to country. Also, some of them might not be independent, as some factors are interrelated and interact with each other (Li et al., 2013).

Besides, to calculate the SRI (using Equ. 5) each cell of CA was assigned to a block group using a proximity analysis and a select-by-location tool in ArcMap. A cell is assigned to a block group when the center of the pixel falls within the block group’s boundaries.

2.3 Results

To calibrate CA using the proposed method which uses SRI in the calibration process to indicate patterns of urban development at the block group level (with the higher values indicative of dispersion of built-up areas and sprawl in the zone, and the lower values of SRI indicating a high concentration of either a built-up or a non-built-up area in the zone), a GA was programmed
in Matlab R2018. Also, to reduce the required time for the code to run, the vectorization technique was considered in the programming to avoid loop-based, and scalar-oriented processes, when it was possible (for more details refer to Xia et al., 2018; and MathWorks, 2020). In order to avoid singularity in calculations, all 14 driving forces of growth in Shelby County were normalized before importing to the program. While driving forces function as variables in Equ. 2 (e.g. $F_{LN1}$), the optimized coefficients (e.g. $\beta_{N1}$) and constants (e.g. $\beta_0$) for the probability of conversion of cells were estimated (using data for 2001 and 2006) over different generations of the GA with the proposed cost function (Equ. 6). In this case, all the parameters of CA are considered as the chromosomes in the GA (Li et al., 2013).

To implement the GA, framework parameters such as population size, crossover and mutation rates, and the number of generations should be assigned. While crossover partially exchanges chromosomes of parents (selected from the population) to generate offspring, mutation changes elements in an individual. These parameters are usually determined based on the domain of the problem according to the user’s experience (Li et al., 2013; Deb, 2004). Technical details about the GA and its implementation are provided in the “Introduction to genetic algorithms for engineering optimization” by Deb, 2004. Model calibration includes the manual adjustment of model parameters (Newland et al., 2018). In this study, population size is assigned as 1000, crossover and mutation rates both as 0.1 and the number of generations as 100 (the values are assigned through the trial and error process to reach a convergent optimization). By using patterns in the calibration process ($\alpha = 0.95$, where alpha is defined through a sensitivity analysis to reach a more-realistic result), parameters of CA were optimized and the Kappa coefficient (Equ. 3) of 87.11% was reached in the calibration. This means that the simulation achieved an accuracy that is 87.11% better than what would be expected from the
chance assignment of cells to categories, a high value means that the simulation results are effective. Figure 9 illustrates an example of SRIs at different generations of the GA (here: (a) the 10th iteration, (b) the 50th iteration, and (c) the “reality” state). At the beginning of the calibration (e.g. at the 10th iteration; Figure 9-a) the calculated SRI values are very different from the reality (Figure 9-c). Therefore, the GA tries to minimize the differences and match the calculated SRIs to those in reality. Consequently, the calculated SRI values at the 50th iteration (Figure 9-b) are more similar to the reality (Figure 9-c) with less difference. Therefore, taking advantage of the patterns in the cost function, a GA will optimize the calibrated parameters more precisely.

Figure 9. Examples of SRI values at different generations of GA. a) 10th generation, b) 50th generation, c) the reality

To compare the calibration results, CA was calibrated by using the GA without computing the SRIs (that is, without patterns) into the cost function (e.g. $\alpha = 1$ in Equ. 6), using the same data and framework parameters. In this case, parameters of CA were optimized and the Kappa coefficient of 85.13% was reached (this compares with the Kappa coefficient of 87.11% reached when CA was calibrated using the GA and SRIs (using patterns)).
In addition, the results of calibration based on the proposed method were compared with the results of the statistical-based calibration of CA using Logistic Regression (LR). LR is a commonly used method of calibration for CA. In this method, a logistic regression fits the predictions to known categorical data (e.g. built-up vs non-built-up). Using the same data and the Logistic Regression package provided by the Earth Science Learner (2020) the coefficients of CA were calculated and the Kappa coefficient of 85.44% was achieved from the calibration (in comparison with 87.11% derived from the GA with patterns, and 85.13% derived from the GA without patterns).

Figure 10 demonstrates the optimization process of the GA with patterns versus the optimization process of the GA without patterns and LR.

Figure 10. The optimization process of GA with patterns, GA without patterns, and LR
As Figure 10 shows, throughout the optimization process in all generations, the GA with patterns reached the higher Kappa coefficient in comparison with the GA without patterns and LR. Also, the GA with patterns reaches a convergent solution almost after 40 generations while the GA without patterns converged after 80 generations.

Table 10 presents the optimized parameters of CA derived from the GA with patterns, the GA without patterns, and LR. The $\beta$s (parameters of CA in Equ. 2) and their description are also provided in the same table.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>GA (with patterns)</th>
<th>GA (without patterns)</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>Constant</td>
<td>-12.22</td>
<td>-25.60</td>
<td>-0.41</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>Dist. to the nearest airport</td>
<td>-4.29</td>
<td>12.65</td>
<td>-0.35</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>Dist. to the nearest CBD</td>
<td>3.60</td>
<td>-6.31</td>
<td>-0.23</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>Dist. to the nearest city</td>
<td>10.59</td>
<td>17.54</td>
<td>-0.16</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>Dist. to the nearest developed area</td>
<td>25.74</td>
<td>6.87</td>
<td>-0.28</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>Dist. to the nearest highway</td>
<td>11.18</td>
<td>-2.14</td>
<td>-0.19</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>Dist. to the nearest major road</td>
<td>-3.99</td>
<td>-13.46</td>
<td>-0.19</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>Dist. to the nearest railroad</td>
<td>0.78</td>
<td>25.96</td>
<td>-0.21</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>Dist. to the nearest residential area</td>
<td>5.31</td>
<td>56.15</td>
<td>-0.25</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>Dist. to the nearest river/ stream</td>
<td>-2.85</td>
<td>7.98</td>
<td>0.04</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>Easting coordinate</td>
<td>-10.29</td>
<td>20.18</td>
<td>-0.07</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>Northing coordinate</td>
<td>2.46</td>
<td>11.33</td>
<td>0.14</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>Elevation</td>
<td>13.90</td>
<td>6.04</td>
<td>0.14</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>Dist. to the nearest waterbody</td>
<td>14.06</td>
<td>15.81</td>
<td>-0.38</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>No. developed cells in the neighborhood</td>
<td>-74.92</td>
<td>-51.27</td>
<td>1.44</td>
</tr>
<tr>
<td>THR</td>
<td>Threshold</td>
<td>-12.25</td>
<td>-25.58</td>
<td>0.50</td>
</tr>
</tbody>
</table>
In Table 10, the same variable may have opposite signs due to the nature of metaheuristic algorithms, as the algorithms start by using a random initial population and due to the different thresholds. The presence of the opposite signs was likewise observed in prior studies (e.g., Li et al., 2013).

In order to evaluate the performance of the proposed method and compare the difference in performance between the proposed method and a conventional measure, the Kappa coefficient (Equ. 3) and the overall accuracy (OA) of the simulated models were calculated using the land cover map in 2011. The overall accuracy indicates the proportion of data that was modeled correctly. It can be calculated as

\[ OA = \frac{\sum_{i=1}^{r} X_{ii}}{N} \times 100 \]  

where \( r \) is the number of rows (or columns) in the confusion matrix, \( N \) is the total number of observations, \( X_{ii} \) is the observation in row \( i \) and column \( i \) of the confusion matrix (Congalton & Green, 2009). Although all methods resulting in an overall accuracy greater than 85% are recommended for effective and reliable analysis and modeling of land cover change (Araya & Cabral, 2010), the overall accuracy is greater when the CA calibration is based on the proposed method of the GA using patterns, reaching 94.84% (with the Kappa coefficient 89.48%), while the overall accuracy of CA calibration using the GA without patterns and LR achieved 93.21% (with the Kappa coefficient 86.15%) and 92.98% (with the Kappa coefficient 85.83%), respectively. Table 11 summarizes the overall accuracies and the Kappa coefficients achieved from the calibration (using data for 2001 and 2006) and validation (using data for 2011) of simulated models.
Table 11. Accuracy results of CA modeling by different calibration methods

<table>
<thead>
<tr>
<th></th>
<th>GA (with patterns)</th>
<th>GA (without patterns)</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Acc.</td>
<td>0.9373</td>
<td>0.9279</td>
<td>0.9289</td>
</tr>
<tr>
<td>Kappa coef.</td>
<td>0.8711</td>
<td>0.8513</td>
<td>0.8544</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Acc.</td>
<td>0.9484</td>
<td>0.9321</td>
<td>0.9298</td>
</tr>
<tr>
<td>Kappa coef.</td>
<td>0.8948</td>
<td>0.8615</td>
<td>0.8583</td>
</tr>
</tbody>
</table>

Even though overall accuracies reached by different calibration methods (Table 11) have similar values making one think the differences are negligible, nevertheless they constitute a difference in terms of the amount of correctly-modeled area. For example, applied to Shelby County, in absolute terms, the proposed method is more accurate versus a standard GA (the GA without patterns) by over 32.22 sq. kilometers correctly modelled ((0.9484 - 0.9321) × 1976.60 = 32.22). Similarly, in the context of Shelby County, the method performs better than a conventional LR resulting in a larger total correctly-modeled area of 36.76 sq. kilometers ((0.9484 – 0.9298) × 1976.60 = 36.76). Table 12 summarizes the total correctly-modeled area in Shelby County using different methods. This correctly-modeled area contains the correctly-modeled developed land and correctly-modeled non-developed land.

Table 12. Correctly-modeled area in Shelby County using CA with GA and patterns, GA without patterns, and LR (in sq. km.)

<table>
<thead>
<tr>
<th></th>
<th>GA (with patterns)</th>
<th>GA (without patterns)</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1874.60</td>
<td>1842.38</td>
<td>1837.84</td>
<td></td>
</tr>
</tbody>
</table>

2.4 Discussion

The experience of calibrating CA in Shelby County indicates that with crossover and mutation rates of 0.1, the GA converges prematurely when the population size is less than 80. In that case, solutions trap in a local optimum and the precision of the model will not improve over
generations because of homogeneity in initial solutions. Moreover, increasing the population size will lead to a better solution and higher precision. However, in that case, the GA consumes more time to optimize parameters. As an example, with crossover and mutation rates of 0.1, the number of generations equal to 100, and population size equal to 100, the GA with patterns required almost 5 hours to reach 83.35% as the Kappa coefficient in the calibration process, while for the same data and framework parameters but population size as 1000, the Kappa coefficient reached 87.11%. However, the elapsed time was almost 27 hours in the latter case.

One concern of getting a better fit is an increased computational run time for running the GA. Despite an increase in the computational cost (the run time), that occurs as a result of the improvement of the simulation and due to an increase in the population size, reaching a higher precision might be substantial for some stakeholders including private companies or governments. For instance, the professional stakeholders such as government agencies planning to build a road and evaluating the effects of the new road on urban growth, need precise results even when it means spending more hours using sophisticated methods such as a genetic algorithm or fuzzy methods. Improving prediction comes at a cost and the decision on a trade-off between computational costs versus accuracy is both purpose- and place-specific.

Figure 11 shows the sensitivity analysis of calibration time based on the population size of the GA with crossover and mutation rates of 0.1 and the number of generations 100, using a computer with a processor of an Intel Core i7 with CPU @ 2.20 GHz.
As Figure 11 shows, increasing the population size will increase the elapsed time. Also, based on the best fit line, the relationship between the population size and the time is a linear function (elapsed_time = 1.6434 \times \text{population}_\text{size}).

Due to the random sampling methods for creating initial solutions, metaheuristics such as the GA may result in different solutions over different tests using the same data and the same framework parameters (Deb, 2004). To evaluate the performance of the proposed method over different tests, CA were calibrated in 10 independent tests using the same data and framework parameters (population size = 100; crossover and mutation rates = 0.1, generations = 100) once using patterns and once without patterns. The population size in these 10 tests is set to 100 to reduce the run time of the program (in comparison with the population size of 1000; see Figure 11 for run time). The solutions that were trapped on a local optimum were removed from the
tests, and the results were sorted in descending order to compare them. Figure 12 shows the Kappa coefficient values of modeling following 10 series of tests.

![Comparison of Kappa coefficients in 10 series of tests](image)

Figure 12. Comparison of Kappa coefficients in 10 series of tests

As Figure 12 illustrates, in comparison with the GA without patterns the Kappa coefficient reaches higher values when the GA with patterns is used (that is, when SRIs are included in the GA). In some tests, such as test 1 and test 10, the Kappa coefficients resulting from both methods are close together while in some tests, such as the tests 4 and 7, the Kappa coefficient is much greater when the GA is used with patterns (more than 10%). However, the GA with patterns reached the greater Kappa coefficient in all 10 tests. The average of the Kappa coefficients in those 10 tests is 74.36% and 65.57% for the GA with patterns and the GA without patterns, respectively. We conclude that including the SRI as an indicator of land use/land cover
patterns in the calibration process of CA can improve the calibration precision in urban growth modeling.

2.5 Conclusion

CA models are able to handle the complexity of urban systems. In an effort for a more precise quantitative modeling of urban growth in Shelby County, this research introduced a new method for calibration of CA with a genetic algorithm (GA) which uses a land use /land cover indicator as patterns in the cost function. The goal was to test whether using the Shannon Relative Index (SRI) as an indicator of land use/land cover patterns might improve a calibration measured by the Kappa coefficient with a higher value indicating a better precision. Data in 2001 and 2006 were used to calibrate a CA model of the urban growth in Shelby County, TN. The Kappa coefficient of 87.11% derived from a calibration based on the proposed method proves the importance of using the GA with patterns, such as the SRI, in the calibration of CA (it compares with the Kappa coefficient of 85.13% derived from the GA without patterns and the Kappa coefficient of 85.44% derived from an LR). Moreover, 2011 data was used to validate the CA model of urban growth. The Kappa coefficient of 89.48% was reached from the CA when patterns are used in the calibration. The Kappa coefficients of 86.15% and 85.83% were derived from the traditional CA (without patterns) and LR, respectively. With the proposed method, the GA was able to correctly simulate a larger area (32.22 sq. kilometers than the GA without patterns). In light of the outcomes of the proposed method, a more precise future extent of Memphis and Shelby County may be simulated, and a better management plan can be selected to avoid sprawl in the area.

In this study, a linear combination of the Kappa coefficient and the maximum absolute error of SRIs (Equ. 6) were used in the cost function of a GA. While the weight of the Kappa
The coefficient was assigned as 95%, the optimum weight for each factor in the cost function could be determined by a sensitivity analysis. Moreover, non-linear combinations may lead to higher precision. However, using a non-linear combination and sensitivity analysis of weights may increase the computational cost of the calibration based on their mathematical complexity (e.g. it increases the elapsed time), which needs further investigation.

Although the proposed method can achieve appropriate simulation results (overall accuracy of 94.84% at validation using data for 2011), it is subject to some limitations under a GA. For instance, the GA is sensitive to the diversity of initial solutions and may converge prematurely when a solution traps on local optima. Therefore, to yield higher precisions for CA modeling of urban growth in Shelby County, the SRI may be used as patterns in more advanced optimization algorithms for future studies. Moreover, due to the assumptions and limitations of a standard CA (e.g. predefined growth rules), modified CA models such as the SLEUTH may be used in future studies.

Despite using the Kappa coefficient as an agreement metric has become part of the culture in remote sensing, there is some criticism about using it. Using disagreement metrics, such as quantity and allocation disagreements, in the calibration process may decrease the computational cost (e.g. it decreases the elapsed time) as they are simpler metrics (Pontius et al., 2011).

Also, including socio-demographic data (such as population, income, age, etc.) and spatial data (such as the distance to the closest medical center) in addition to the 14 driving forces used in the current study (see Section 2.2), may result in a more comprehensive simulation which can predict the urban growth in different scenarios. For instance, a model that precisely predicts the urban growth at the time of an economic recession due to a mortgage crisis (e.g. the
financial crisis of 2007–2008) or a pandemic (e.g. coronavirus/COVID-19), should incorporate various financial or health-related driving forces. Identifying possible scenarios and related driving forces can be the subject of future studies.
Chapter 3: The Kappa coefficient is still alive

Abstract

We investigate the potential of total disagreement as a suggested alternative to the Kappa coefficient, in urban growth simulation using Cellular Automata (CA). To overcome some limitations of metaheuristic approaches such as trapping in local optima and those by logistic regression such as multicollinearity, an Imperialistic Competitive Algorithm (ICA), an optimization algorithm used in various fields, was used to calibrate CA in a study of urban growth. Total disagreement and the Kappa coefficient were separately implemented in the cost function of ICA. Moreover, this study integrates the Kappa coefficient and Shannon relative indices (SRIs) into the cost function of ICA to provide a more precise approach for the calibration of CA. Data for 2001 and 2006 were used for calibration, and data for 2011 was used for validation of CA simulations in Shelby County, Tennessee, USA. The findings indicate that utilizing the Kappa coefficient achieves statistically significantly higher precisions than using the total disagreement (overall accuracy of 93.86% vs 92.37%). Moreover, integrating SRIs, as patterns, and the Kappa coefficient improves the simulation, although not significantly (overall accuracy of 94.65% vs 93.86%). The results of this paper can assist city developers, construction companies, transportation engineers, tax assessors, and utility providers to make more accurate decisions.

Keywords: Urban growth; sprawl; cellular automata; calibration; diversity index; measure of agreement;
3.1 Introduction

The Kappa coefficient is a measure of an agreement to compare two different sets of classified remote sensing data, such as comparing land cover categories in a simulated map with the reference data for the accuracy assessment purpose (Pontius & Millones, 2011).

The introduction of the Kappa coefficient as a measure of inter-judge agreement for categorical data is often credited to Cohen (1960), who used it in psychology. Congalton (1981) introduced the Kappa coefficient into the field of remote sensing by assessing the accuracy of a Landsat image classification. This accuracy assessment measure, Kappa, is often reported along with a proportion of observations classified correctly. While most of the previous accuracy assessment measures were based on parametric statistical techniques (with the standard assumptions of continuous data and normal distribution), the Kappa coefficient is used in a discrete multivariate analysis, which is more appropriate for assessing categorized data where the data either fall into a particular category or they do not (Cohen, 1960; Congalton, 1981; Congalton, Oderwald, & Mead, 1983). Over the succeeding decade, the Kappa coefficient became popularized, implemented, and developed in more studies such as Congalton et al. (1983), Hudson & Ramm (1987), Fung & LeDrew (1988), and many more.

The Kappa coefficient is defined as the proportion of agreement after the chance agreement is removed from the consideration (Equ. 8):

\[
\text{Kappa}_\text{coef.} = \frac{P_o - P_c}{1 - P_c}
\]

where \(P_o\) is the proportion of units in which the judges agreed, and \(P_c\) is the proportion of units in which agreement is expected by chance (Cohen, 1960). While Cohen (1960) gave
examples of the units as “psychological test protocols” or “small groups,” other units can be considered in remote sensing including pixels, clusters, or polygons (Anand, 2012). $P_o$ and $P_c$ can be calculated based on a confusion matrix (also known as an “error matrix” or a “contingency table”) using equations 9 and 10, respectively:

$$P_o = \frac{1}{N} \sum_{i=1}^{r} X_{ii}$$

[9]

$$P_c = \frac{1}{N^2} \sum_{i=1}^{r} (X_{i+}X_{+i})$$

[10]

where $r$ is the number of rows (or columns) in the confusion matrix, $N$ is the total number of observations, $X_{ii}$ is the observation in row $i$ and column $i$ of the confusion matrix, $X_{i+}$ is the marginal total of row $i$, and $X_{+i}$ is the marginal total of column $i$ (Liu, 2009).

The Kappa coefficient ranges from $[-P_c / (1 - P_c)]$ to +1 (Landis & Koch, 1977). The Kappa coefficient is 0 when the obtained agreement ($P_o$) equals the chance agreement ($P_c$) and is a positive value when the obtained agreement ($P_o$) is greater than the chance agreement ($P_c$). The upper limit of the Kappa coefficient (one) occurs when (and only when) there is a perfect agreement between the judges (Cohen, 1960).

The Kappa coefficient is commonly used in accuracy assessment. Most image analysis software packages offering accuracy assessments report the Kappa coefficient (Congalton & Green, 2009).

Since the early 1980s, there have been strong criticisms about the use of the Kappa coefficient. For instance, Foody (1992) claimed that due to overestimating the chance agreement ($P_c$), the Kappa coefficient underestimates the classification agreement (Foody, 1992). To overcome the limitations of the Kappa coefficient researchers developed some modifications of
the standard Kappa coefficient such as Kno (Kappa for no ability), Klocation (Kappa for location), Kquantity (Kappa for quantity), conditional Kappa, and weighted Kappa (Pontius, 2000; Rossiter, 2014). However, Pontius & Millones (2011) claimed that all “Kappa indices are useless, misleading, and/or flawed for the practical applications in remote sensing that we have seen” and recommended that “the profession abandoned the use of Kappa indices for purposes of accuracy assessment and map comparison” (Pontius & Millones, 2011). Extending the argument, quantity disagreement and allocation disagreement were proposed as alternative measures. The former, also known as a “composition”, was defined as the disagreement because of a “less than perfect match in the proportions of the categories” (Pontius & Millones, 2011, p. 4409). The latter measure, also known as a “configuration”, was defined as the disagreement because of “the less than optimal match in the spatial allocation of the categories, given the proportions of the categories in the reference and comparison maps” (Pontius & Millones, 2011, p. 4409). In addition, that study suggested the total disagreement as the sum of the overall quantity disagreement and overall allocation disagreement (Pontius & Millones, 2011).

Should the Kappa coefficient be archived due to the claimed superiority of the suggested alternatives? This paper is aimed to statistically examine 1) whether a total disagreement can be used as an alternative to the standard Kappa coefficient, and 2) whether including the Shannon relative indices (SRIs), as a pattern, in the cost function of a calibration can increase the simulation’s accuracy. To answer the first question, we calibrate a cellular automata (CA) model of urban growth by the Imperialistic Competitive algorithm (ICA) using two different cost functions: the Kappa coefficient, and the total disagreement, and the results will be compared together statistically. For the second purpose, we calibrate the CA model of urban growth by the ICA using a cost function containing combined Kappa coefficient and SRIs, and the results will
be statistically compared with the results of the calibration by ICA using a cost function containing only the Kappa coefficient. In all the above-mentioned methods, a CA model of urban growth in Shelby County, Tennessee, will be calibrated using 2001 and 2006 data and validated with 2011 data. Moreover, the simulation results based on the described measures will be compared with a traditional simulation based on a standard Logistic Regression (RL).

The study set out to investigate whether (1) the total disagreement is a statistically superior alternative to the standard Kappa coefficient in the cost function of the calibration process of urban growth simulation using the CA model, and (2) whether including the SRIs, as a pattern, into the cost function of calibration can increase the simulation’s accuracy.

Accordingly, the following hypotheses will be tested:

(1) $H_0: \mu_{ICA,Kc} = \mu_{ICA,td}$

$H_1: \mu_{ICA,Kc} > \mu_{ICA,td}$

(2) $H_0: \mu_{ICA,Kc} = \mu_{ICA,K+SRI}$

$H_1: \mu_{ICA,Kc} \neq \mu_{ICA,K+SRI}$

where $\mu_{ICA,Kc}$ denotes the population mean of the results using ICA and the Kappa coefficient, $\mu_{ICA,td}$ denotes the population mean of the results using ICA the total disagreement, and $\mu_{ICA,K+SRI}$ denotes the population mean of the results using the Kappa coefficient and SRIs.

To use CA for simulating urban growth, the best combination of CA values (see Section 3.1.2) must be determined through a process called calibration. During the calibration, a cost function (also known as “a fitness function”, see Newland et al., 2015) is used to measure the goodness of fit between the simulated data and reality. Depending on the nature of the problem and purposes, the cost function can be a linear or non-linear combination of different landscape metrics (see Section 2.1). To achieve higher accuracies in the calibration, and consequently in the
simulation, metaheuristics, such as GA, are used to optimize the cost function (e.g. to achieve the maximum value of the goodness-of-fit and thus the minimum difference between the simulated data and the reality) through an iterative process.

This paper attempts to provide a more accurate approach for the calibration of CA in the studies of land use/land cover dynamics and urban growth modeling. This study is a pioneer in the implementation of ICA for optimizing the cost function of a CA calibration to simulate urban growth, as the technique has not yet been used in that field despite its application in other areas for optimization purposes (see Section 3.1.1). Moreover, this study explores whether the Kappa coefficient is usable in the cost function of the calibration (which will be optimized by ICA) or it should be disused. A study by Momeni and Antipova (2020) demonstrated that adding SRIs (as indicators of LULC patterns) to the cost function of a CA calibration increases the simulation accuracies when the cost function is optimized by GA. However, as this is the first execution of ICA for optimizing the cost function of CA calibration in urban growth modeling, the potential influence of adding SRIs to the cost function of the calibration which will be optimized by ICA is discussed as well.

The results of this study can help city planners to produce a more reliable model of future development. Moreover, transportation engineers, tax assessors, utility providers, and other stakeholders may directly benefit from the results of this paper. Also, by anticipating the future extension of a city, local governments can promote smart development while avoiding undesirable effects of urbanization including sprawl and its negative environmental outcomes (Antipova, Momeni, & Banai, 2022a).

The rest of this paper is organized as follows: the literature on the topic is summarized in Section 3.1.1. In Sections 3.1.2 to 3.1.5, the concepts of cellular automata (CA), imperialistic
competitive algorithm, total disagreement, and the Shannon relative index (SRI) are introduced, respectively. The study area and the required data for urban growth simulation using CA are discussed in Section 3.2. The methodology of this study is explained in Section 3.3 in detail. Section 3.4 presents the results of different simulations. Discussions about results are presented in Section 3.5. Finally, Section 3.6 summarizes the study and compares our findings with the findings of other researchers as a conclusion.

3.1.1 Literature review

In many studies, CA is considered a powerful tool for modeling urban growth (Guan & Rowe, 2016; Roodposhti et al., 2019; Siddiqui et al., 2018). CA contains a large finite number of cells that can change their states based on their initial states, the neighbors, and some pre-defined transition rules (Roodposhti et al., 2019).

Among the alternative approaches for simulation of urban growth, including Lowry and agent-based models, a two-dimensional CA is considered the most proper approach as it models complex spatial situations in a simplified scientific way (Guan & Rowe, 2016; Maithani, 2010; Roodposhti et al., 2019). However, precise calibration of CA is still a challenge for scholars. Calibration of CA is a process to determine the best parameters of CA, in a way that the simulated urban growth can better represent the real world by reaching the highest “goodness-of-fit” criterion (Bell, Dean, & Blake, 2000; Li et al., 2013; Shan, Alkeder, & Wang, 2008). Calibration of CA is a complex and challenging process as many coefficients interact with each other (due to inter-related variables) and do not essentially provide a unique solution (Verburg et al., 2004).

In 1997, White & Engelen used CA to study land use patterns and provide an insight into the socio-economic consequences of global climate change in the Caribbean island of St. Lucia.
In that study, CA was not calibrated. However, one scenario was developed to demonstrate the behavior of the model. The study concluded that the model can only support “what-if” experiments which allow a planner to illustrate different types of development in the future (Maithani, 2010; White & Engelen, 1997). In another research by White et al. (1997), CA was used to model land use patterns in Cincinnati, Ohio, by a trial-and-error calibration process. The results were relatively accurate and proved the potential of CA in planning contexts (Maithani, 2010; White, Engelen, & Uljee, 1997). In a one-day symposium about the state of the art in CA modeling of urban systems held in 2001 in the Centre for Advanced Spatial Analysis at the University College London, researchers were encouraged to investigate advanced approaches for calibration of CA as a manual tuning of transition values by the trial-and-error process was very time consuming and no longer recommended by other scholars (Torrens & O’Sullivan, 2001; Verburg et al., 2004).

Li et al. (2001) proposed calibration of CA based on Artificial Neural Networks (ANN) to model built-up/non-built-up areas in Dongguan, China. In that study, seven spatial variables including distance to the major urban areas, distance to sub-urban areas, distance to the closest road, distance to the closest expressway, distance to the closest railway, agricultural suitability, and the amount of development in the neighborhood were used in the model. The overall accuracy of that model reached 79% (Li & Yeh, 2001). Verburg et al. (2004) used statistics-based calibration of CA using Logistic Regression (LR) to study land use patterns in the Netherlands. In that study, neighborhood land use characteristics were analyzed and land use conversions were explained by the occurrence of land uses in each neighborhood. However, Verburg argued that other factors such as accessibility and environmental suitability also affect the pattern of land use (Verburg et al., 2004). Even though LR is considered the commonly used
method for calibration of CA, it has its limitations. For example, LR may encounter multicollinearity if some variables are highly correlated. It also has difficulties with non-linear functions. Moreover, LR does not include pattern-based calibrations as it is a cell-based regression that cannot be used in iterative processes (Li et al., 2013; Li & Yeh, 2001; Momeni & Antipova, 2020).

To overcome the above-mentioned limitations in the calibration of CA, metaheuristic optimization algorithms have been used to calibrate CA. For instance, Feng et al. (2011) implemented particle swarm optimization (PSO) for calibration of CA to model urban growth in the Fengxian District of Shanghai Municipality, China (Feng et al., 2011). Li et al. (2013) used a genetic algorithm (GA) to calibrate the CA model of growth in Guangdong, China (Li et al., 2013). Momeni & Antipova (2020) also used GA to calibrate the CA model of urban growth in Shelby County, Tennessee, USA. In that study, SRIs of neighborhoods were imported into the cost function of GA as patterns to improve the accuracy of the calibration (Momeni & Antipova, 2020). The artificial bee colony (ABC) optimization of CA was implemented in 2016 by Naghibi et al., to model the urban growth in Urmia, Iran (Naghibi & Delavar, 2016a). Moreover, Bharath et al. (2017) studied the growth of five Indian megacities (Delhi, Mumbai, Pune, Chennai, and Coimbatore) by calibration of CA using fuzzy logic (Bharath et al., 2018).

ICA is a comparatively new algorithm used in various optimization problems in different fields such as product-service systems (Yin & Gao, 2019), power systems (Aghdam & Hagh, 2019), rock mechanics (Armaghani et al., 2019), machine learning (Ahmadi & Chen, 2019), and even in medical sciences (Reisi et al., 2019). However, the ICA has not been used in the field of urban growth simulation yet. Specifically, in this study, the ICA is used for the calibration of CA to compare the potential of ICA with the traditional methods, such as LR, and overcome some
limitations of metaheuristics, such as trapping in local optima. The details about the ICA are provided in Section 3.1.3.

3.1.2 Cellular automata

The concept of CA was introduced by Neumann and Ulam in the early 1950s (Neumann & Burks, 1966). The concept remained in its formative stage until the 1970s when Conway disseminated CA in the Game of Life, a cellular automata zero-player game where the evolution of each cell in the game is only determined by its initial state (Schiff, 2006). Nowadays, CA is widely used for modeling dynamic patterns such as noise filtration (Jeelani & Qadir, 2018), land use/land cover changes (Roodposhti et al., 2019), modeling pedestrian dynamics (Padovani, Neto & Cereda, 2018), urban growth (Guan & Rowe, 2016) and other applications.

In urban growth studies, CA consists of five major components: 1) Lattice: a uniform array or geometrical grid with discrete variables at each cell; 2) Cell: the smallest unit of a lattice which has only one state at any instant of time; 3) State: a variable which takes different values (e.g. non-built-up or built-up; 0 or 1); 4) Neighbors: cells in the lattice which are physically close to a specific cell, based on a predefined distance, and can affect the state of the cell in the future. Further, the neighborhood of a cell includes the cell itself; 5) Transition rules: a set of conditions that can change the future state of a cell based on its current state and its neighbors (e.g. proximity to a built-up area) (Maithani, 2010).

Verburg et al. (2004) defined the probability of a conversion of a cell’s state as:

\[
 \log \left( \frac{P_i}{1-P_i} \right) = \beta_0 + \sum \beta_{N,i} F_{i,N,i}
\]

[11]
where \( F_{i,N,l} \) is an independent variable for cell \( i \), neighborhood \( N \), and land use/land cover \( l \). \( \beta_{N,l} \) is a coefficient to be estimated for the neighborhood \( N \) and land use/land cover \( l \). For the cell \( i \), \( P_i \) is the probability of conversion, e.g. from non-built-up to built-up (Verburg et al., 2004). To go in-depth with the details of CA and its applications, the “Introduction to Cellular Automata” by J. L. Schiff (2006) is recommended to a reader.

Calibration of CA, or finding \( \beta_{N,l} \) in Equ. 11, is challenging as many variables may interact with each other (due to inter-relatedness of variables). Therefore, calibration of CA does not essentially reach a unique solution (Verburg et al., 2004). In this study, the imperialistic competitive algorithm is used for the calibration of CA.

### 3.1.3 Imperialistic competitive algorithm

The imperialistic competitive algorithm (ICA) is a relatively new algorithm introduced by Atashpaz in 2007, mainly used to solve multi-objective optimization problems. The purpose of optimization is to find an optimal solution in terms of the variables of a problem. The initial idea of ICA came from using the analogy of the imperialistic competitions between empires in history where imperialists try to extend the power and the rule of their government beyond its boundaries (Atashpaz-Gargari & Lucas, 2007; Yin & Gao, 2019).

Similar to other evolutionary algorithms, such as GA, the ICA begins with a population of random initial solutions called countries. Each country is equivalent to a chromosome in GA and represents a basic configuration rule. Power is assigned to each member of the population based on a cost function (also referred to as a “fitness function”). A precise solution results in a lower cost and therefore, a higher power. Some members of the population with the highest power are considered imperialists and the remainder of the population is considered colonies. These imperialists are equivalent to elites in GA. Each imperialist, based on its power, takes possession
of some colonies to shape an empire. The total power of each empire is a mathematical function of its imperialist’s power and its colonies’ power (Equ. 12).

\[ \text{power}(\text{empire}_i) = [\alpha \times \text{power}(\text{imperialist}_i)] + 
\quad [(1 - \alpha) \times \text{average}(\text{power(all colonies of empire}_i))] \]  

where \( \text{power}(\text{empire}_i) \) is the total power of the \( i^{th} \) empire, and \( \alpha \) is a weight ranging from 0 to 1.

Over different generations of ICA, each empire takes possession of colonies of the rival empires based on its power (which is called an inter-empire competition). Consequently, powerful empires will progressively grow (getting more colonies, and thus, more power) during the inter-empire competition, and weak empires will be gradually eliminated by losing their colonies (and their power). Meanwhile, empires have to increase their colonies’ power by assimilation (equivalent to a crossover in GA) or revolution (equivalent to a mutation in GA). If a country gets more power (i.e., a lower cost function value) than its imperialist, it will take the control of the empire and behave as the imperialist for that empire. Therefore, over generations, colonies will become closer to the imperialist of the most powerful empire as a kind of convergence. Eventually, the ICA will stop further generations when some pre-defined conditions (e.g. a single empire with colonies that have become very close to their imperialist) are reached. In this situation, the most powerful (the most precise) imperialist is considered the best (the optimum) solution (Atashpaz-Gargari & Lucas, 2007; Reisi et al., 2019; Yin & Gao, 2019). More details about the history of the ICA and its development process can be found in the original work by Atashpaz in “Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialistic Competition”, 2007.
3.1.4 Total disagreement

Different measures have been introduced to measure the goodness of fit for calibration and validation of CA. Using a confusion matrix (Table 13), Pontius & Millones (2011), defined equation 13 as the category-level quantity disagreement, $q_g$, for a category $g$. The quantity disagreement is defined as the amount of the difference between simulated data and the reference data due to the less than perfect match in the proportions of the categories (Pontius & Millones, 2011).

Table 13. an example of a confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>j = 1</th>
<th>j = 2</th>
<th>...</th>
<th>j = J</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>i = 1</td>
<td></td>
<td>$p_{11}$</td>
<td>$p_{12}$</td>
<td>...</td>
<td>$p_{1j}$</td>
<td>$\sum_{j=1}^{J} p_{1j}$</td>
</tr>
<tr>
<td>i = 2</td>
<td></td>
<td>$p_{21}$</td>
<td>$p_{22}$</td>
<td>...</td>
<td>$p_{2j}$</td>
<td>$\sum_{j=1}^{J} p_{2j}$</td>
</tr>
<tr>
<td>i = J</td>
<td></td>
<td>$p_{J1}$</td>
<td>$p_{J2}$</td>
<td>...</td>
<td>$p_{Jj}$</td>
<td>$\sum_{j=1}^{J} p_{jj}$</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>$\sum_{i=1}^{I} p_{i1}$</td>
<td>$\sum_{i=1}^{I} p_{i2}$</td>
<td>...</td>
<td>$\sum_{i=1}^{I} p_{ij}$</td>
<td>1</td>
</tr>
</tbody>
</table>

$$q_g = \left| \left( \sum_{i=1}^{I} p_{ig} \right) - \left( \sum_{j=1}^{J} p_{gj} \right) \right| \quad [13]$$

Where in Equ 13 and Table 13, $p_{ij}$ is the proportion of the study area that is simulated as a category $i$ while belonging to a category $j$ in the real world (reference data). $J$ also denotes the number of categories in the study area.
The overall quantity disagreement, $Q$, for all $J$ categories can be calculated by dividing the summation of the category-level quantity disagreements by two (Equ. 14), as an underestimation in one category goes along with an overestimation in another category (Pontius & Millones, 2011)

\[ Q = \frac{\sum_{g=1}^{J} q_g}{2} \tag{14} \]

In addition to a category-level quantity disagreement, a category-level allocation disagreement, $a_g$, for a category $g$ can be calculated using Equ. 15. Given the proportions of the categories in the simulated data and the reference data, allocation disagreement is defined as the amount of difference between the simulated data and the reference data due to the less than optimal match in the spatial allocation of the categories (Pontius & Millones, 2011).

\[ a_g = 2 \times \min\left\{ (\sum_{i=1}^{J} p_{ig}) - p_{gg}, \ (\sum_{j=1}^{J} p_{jg}) - p_{gg} \right\} \tag{15} \]

Moreover, the overall allocation disagreement, $A$, for all $J$ categories can be calculated by dividing the summation of the category-level allocation disagreements by two (Equ. 16) as an underestimation in one category goes along with an overestimation in another category (similar to the overall quantity disagreement).

\[ A = \frac{\sum_{g=1}^{J} a_g}{2} \tag{16} \]
Eventually, the total disagreement, $D$, can be calculated by summation of the overall quantity disagreement (Equ. 14) and the overall allocation disagreement (Equ. 16) as follows (Pontius & Millones, 2011):

$$D = Q + A$$  \[17\]

### 3.1.5 Shannon and Shannon relative indices

Urban studies use a diversity index as an indicator of the simultaneous relations among all different types (e.g. species, land uses, land covers, etc.) in a community (Tucker et al., 2017). Shannon index is a robust diversity index used as an indicator of urban sprawl to estimate the dispersion or concentration of developed and non-developed areas (Banai, Antipova & Momeni, 2021; Momeni & Antipova, 2020).

Using the total number of types, $S$, and the probability of type $i^{th}$ incident, $p_i$, in zone $n$, Roghair et al. (2013) mathematically defined the Shannon index as:

$$H_n = -\sum_{i=1}^{S} p_i \log (p_i)$$ \[18\]

To limit the range of Shannon index between 0 and 1, Shannon relative index (SRI) can be calculated as:

$$H_n' = H_n / \log (n)$$ \[19\]

Where $H_n$ denotes Shannon index, $n$ denotes the total number of zones, and $H_n'$ denotes SRI (Deka, Tripathi, & Khan, 2012).
While lower values of SRI indicate a high concentration of a dominant type in the community, a higher value indicates scattering of developed land reflecting more likelihood of urban sprawl (Antipova, Momeni, & Banai, 2022b).

### 3.2 Study area

The first firm foothold of Memphis, Tennessee, in its climb to success as the metropolitan center of the Mid-South goes back to the nineteenth-century economy, politics, real estate promotion, and early land speculation. At that time roads were practically non-existent and locations on navigable rivers, such as the Mississippi river, were considered prime town sites. Therefore, the population moved into Tennessee, Kentucky, and Northwest Territory en-mass to find spots for future development (Williams, 1968). Meanwhile, West Tennessee lands were generously granted by North Carolina to support urbanization in Tennessee. In October 1818, two U. S. commissioners, Andrew Jackson and Isaac Shelby, successfully negotiated a treaty with the Chickasaw Indians to purchase some of their lands and limit the tribe to an area south of the present Tennessee-Mississippi state line (Williams, 1968). Shelby County was purchased by the United States for a total of $300,000 along with the rest of West Tennessee and was drawn on Tennessee maps in 1819 (“A Brief History of Shelby County,” 2020). Shelby County court’s first action, at a cost of $142.50, was a survey of the county. The surveyors reported that the county contained 625 sq. miles (nowadays ≈ 763 sq. miles of land and 22 sq. miles of waterbodies) with a population between 250 and 350 people (nowadays ≈ 936,130 in 2019) (“A Brief History of Shelby County,” 2020; U.S. Census Bureau, 2022).

While the exact location is unknown, historians estimate that the first Shelby County road headed somewhere between present-day Dyersburg and Jackson, Tennessee. Currently, Shelby County has built and maintains 1,400 miles of roads (“A Brief History of Shelby County,” 2020).
Figure 13 shows the Memphis plan in 1827. A locally adaptive threshold in MatLab ("Imbinarize," 2020) was applied to the original image to enhance it and make it more readable.

In 1833 several schemes have been proposed by railroad promoters for building a line from the cotton-producing interior of West Tennessee to one of the Mississippi river towns- Fulton, Randolph, Memphis, or Ft. Pickering. Two years later, the Memphis Railroad Company was chartered by an act of the Tennessee Legislature (Williams, 1968). As a result, Memphis turned into a commercial center while South Memphis became a fashionable site for fine homes.

Figure 14 shows the tentative use district map for the City of Memphis that was first published on May 7, 1922, in the Commercial Appeal Sunday morning local newspaper and used for the 1922 Memphis zoning code; the first zoning code not only in Memphis but in the entire state of Tennessee ("Historic Zoning Codes and Maps," 2020; "Use district map," 1922).
Figure 14. The use district map for the City of Memphis in 1922 (Use district map, 1922)

With the growth of population, Memphis continued to expand and some rural areas within Shelby County got developed. For instance, an analysis of Census data and satellite images from 1990 to 2010 shows a population growth of 12.26% in Shelby County and 124.85% in Shelby County excluding Memphis, which caused the growth of urbanized areas (developed areas) by 60% from 1992 to 2011 in Shelby County (Momeni & Antipova, 2022b). Figure 15 illustrates developed areas in 1992 and 2011 in Shelby County with barren land, forests, herbaceous, planted/cultivated, shrublands, water, and wetlands. The majority of developments in Shelby County arose in the east and southeast of Memphis’ boundaries (Figure 15).
In this study, to investigate the capability of a total disagreement to be used as a suggested alternative to the Kappa coefficient, urban growth in Shelby County is simulated using CA. ICA is implemented for the calibration of the CA model using total disagreements and using the Kappa coefficient, separately.

3.2.1 Data

Boundary lines of Shelby County, Memphis, and Census block groups were extracted from the Tiger line database ("TIGER/Line Shapefiles," 2019) to define the geographic entities.
The Multi-Resolution Land Characteristics Consortium (https://www.mrlc.gov/) provides the national-wide land cover maps of the USA based on a classification of Landsat satellite images at a spatial resolution of 30 meters (“Land cover data,” 2020; Yang et al., 2018). In the classification process, each pixel in the satellite image is assigned to a certain class (e.g. developed, barren, forests, herbaceous, planted/cultivated, shrublands, water, wetlands, etc.) with the maximum confidence of assignment to produce a land cover map (Momeni et al. 2020). At the time of this study, land cover data were available only for 1992, 2001, and 2011 which are used for further calculation of urban growth driving forces such as the distance to the nearest developed area and the percentage of development in a neighborhood. A driving force of urban growth increases the likelihood of a change in the state of an area (e.g., a transition from undeveloped land to developed land) (Antipova, Momeni, & Banai, 2022a).

Land use data have been extracted from the Assessor of Property (https://www.assessor.shelby.tn.us/) to calculate driving forces such as the distance to the nearest residential area (Assessor of Property, 2020). Land cover and land use data are enormously used by researchers interchangeably; Nevertheless, each concept has a very specific meaning. Land cover data define the physical material on the surface of a land including Developed, Barren, Forest, Planted, Water, etc (Makers of American Botany 2017). However, land use data define the way we utilize that land (NOAA 2017). For example, an area that is classified as Developed in land cover data, may be used for industrial, commercial, or residential purposes.

The digital elevation model (DEM) provided by the USGS (https://earthexplorer.usgs.gov/) is used to estimate elevations in Shelby County (“Earth Explorer,” 2020). Moreover, data for roads, railroads, and rivers /streams /waterbodies in Shelby County are also collected for further
calculation of related driving forces (explained in Section 3.3) (“DIVA-GIS,” 2020; Map Cruzin, 2020; TN GIS, 2020).

MatLab R2018 (Matlab, 2020), ArcMap 10.5.1 (ArcMap, 2020), and Microsoft Excel 2016 (Microsoft Excel, 2020) are the software packages used for data preparation, statistical analysis, and map generation.

3.3 Methodology

In this study, ICA is used to optimize the calibration of CA. Calibration of CA is the process of determining the best combination of CA values ($\beta_{N,l}$ in Equ. 11) in a way that the simulated growth using CA can better match the real world (Momeni & Antipova, 2020).

To define a lattice for CA, Shelby County is divided into 2,258,471 geo-referenced cells (each cell measuring 30m×30m) matching the geo-referenced pixels of land cover maps. Fourteen commonly-used driving forces of urban growth ($F$ in Equ. 11) are computed for each cell using proximity analysis in ArcMap including a distance to the nearest 1) airport, 2) CBD, 3) city, 4) developed area, 5) highway, 6) major road, 7) railroad, 8) residential area, 9) river/stream, 10) waterbody (lake, bay, wetlands) as well as 11) easting coordinate of the cell, 12) northing coordinate of the cell, 13) the elevation of the cell, and 14) the number of developed cells in a 7-cell×7-cell neighborhood of the cell (Aburas et al., 2017; Li et al., 2013; Liao et al., 2016; Mohammady & Delavar, 2016; Momeni & Antipova, 2020; Votsis, 2017). All driving forces are normalized before importing into the algorithm, to avoid singularity in further calculations.

Figure 16 shows the heatmaps for distances between each cell to the nearest major road (driving force #6) and the nearest river/stream (driving force #9).
Figure 16. Distance to a) the nearest major road, b) a river or stream
Driving forces of urban growth define the probability of a conversion of a cell’s state (e.g. from non-built-up to built-up). Driving forces are subject-based and can vary from region to region. Moreover, they may not be totally independent as some factors are interrelated (Li et al., 2013; Momeni & Antipova, 2020).

The flowchart of ICA that is used in this study is illustrated in Figure 17.

![Figure 17. The flowchart of the Imperialistic Competitive Algorithm (ICA)](image)

ICA was programmed in MatLab, and to make the program run faster, the vectorization technique is implemented to circumvent time-consuming loop-based and scalar-oriented processes, when it is possible (for more details about vectorization refer to Xia. et al., 2018 and MathWorks, 2020).
The framework parameter of ICA such as the number of initial countries, the number of initial empires, assimilation rate, revolution rate, and the number of generations, should be defined before running the program. These parameters are usually defined according to the user’s experience, based on the domain of the problem or previous studies in the area (Deb, 2004; Li et al., 2013; Momeni & Antipova, 2020). To compare the results of this study with the results of a previous study in Shelby County by Momeni & Antipova (2020), that used GA for the calibration purpose, the number of initial countries is defined as 100, assimilation and revolution rate as 0.1, and the number of generations as 100.

Any change in the framework parameters of a metaheuristic optimization algorithm, including ICA, may result in different outputs. On the other hand, using the same data and the same framework parameters, the results of a metaheuristic optimization algorithm, including ICA, might be different over different tests as they begin with a population of random initial solutions (Deb, 2004; Momeni & Antipova, 2020). Therefore, to compare the results, urban growth in Shelby County is simulated 30 times, independently. Over these 30 times of simulations, the same data and framework parameters are used for calibration of CA using ICA with three different cost functions: 1) the Kappa coefficient (Equ. 8), 2) the total disagreement (Equ. 17), and 3) a combination of the Kappa coefficient and SRIs, as patterns, defined below:

\[
\text{Cost function} = \alpha \times (1-K) + (1-\alpha) \times \max |\text{errors of simulation for SRIs}| \quad [20]
\]

where \(\alpha\) is a constant defining the weights, \(K\) represents the Kappa coefficient, and \(|...|\) denotes the absolute value. Based on Equ. 20, at each iteration, the values of SRI are calculated and compared with the real values of SRI at the Census block group level, to determine errors of
SRIs at the simulation. Then the maximum of absolute errors and the error for the Kappa coefficient are combined proportionally (based on $\alpha$) to calculate the cost of simulation at that iteration. This cost function (Equ. 20) is referred to as the cost function with patterns in the rest of this paper, for the sake of simplicity.

In all simulations, CA is calibrated using data for 2001 and 2006 and validated using 2011 data in Shelby County, TN, USA. For statistical comparison of the results, a t-test will be applied to the output of simulations after testing the normality of outputs.

3.4 Results

To calibrate CA, ICA was programmed in MatLab using three different cost functions: the Kappa coefficient (Equ. 8), the total disagreement (Equ. 17), and a combination of the Kappa coefficient and SRIs (Equ. 20). Fourteen driving forces of urban growth (listed in Section 3.3) were used as independent variables ($F_{i,N,1}$) to optimize the coefficients ($\beta_{N,1}$) and the constant ($\beta_0$) in Equ. 11, in order to estimate the probability of conversion for each cell. Finally, the proportion of all cells that were modeled correctly was calculated and reported as the overall accuracy.

Table 14 presents the overall accuracy for urban growth simulation using CA and ICA in Shelby County over 30 times of simulations.
Table 14. The overall accuracy (%) of urban growth simulation using CA and ICA with different cost functions

<table>
<thead>
<tr>
<th>Test #</th>
<th>Kappa coef.</th>
<th>Total disagreement</th>
<th>Kappa coef. and SRIs (patterns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>83.82</td>
<td>87.13</td>
<td>78.53</td>
</tr>
<tr>
<td>2</td>
<td>80.14</td>
<td>58.62</td>
<td>79.89</td>
</tr>
<tr>
<td>3</td>
<td>78.39</td>
<td>91.40</td>
<td>92.82</td>
</tr>
<tr>
<td>4</td>
<td>86.59</td>
<td>67.46</td>
<td>72.52</td>
</tr>
<tr>
<td>5</td>
<td>77.31</td>
<td>55.87</td>
<td>68.62</td>
</tr>
<tr>
<td>6</td>
<td>85.33</td>
<td>79.41</td>
<td>94.09</td>
</tr>
<tr>
<td>7</td>
<td>89.16</td>
<td>64.18</td>
<td>89.70</td>
</tr>
<tr>
<td>8</td>
<td>63.22</td>
<td>59.85</td>
<td>62.22</td>
</tr>
<tr>
<td>9</td>
<td>89.43</td>
<td>88.90</td>
<td>92.02</td>
</tr>
<tr>
<td>10</td>
<td>63.40</td>
<td>74.63</td>
<td>83.70</td>
</tr>
<tr>
<td>11</td>
<td>79.80</td>
<td>63.18</td>
<td>91.67</td>
</tr>
<tr>
<td>12</td>
<td>78.84</td>
<td>86.32</td>
<td>77.97</td>
</tr>
<tr>
<td>13</td>
<td>69.01</td>
<td>70.17</td>
<td>85.19</td>
</tr>
<tr>
<td>14</td>
<td>72.62</td>
<td>83.31</td>
<td>65.25</td>
</tr>
<tr>
<td>15</td>
<td>69.86</td>
<td>69.98</td>
<td>70.65</td>
</tr>
<tr>
<td>16</td>
<td>89.07</td>
<td>78.02</td>
<td>74.46</td>
</tr>
<tr>
<td>17</td>
<td>87.51</td>
<td>73.00</td>
<td>76.24</td>
</tr>
<tr>
<td>18</td>
<td>73.03</td>
<td>78.34</td>
<td>94.65</td>
</tr>
<tr>
<td>19</td>
<td>86.28</td>
<td>66.49</td>
<td>80.86</td>
</tr>
<tr>
<td>20</td>
<td>86.96</td>
<td>65.15</td>
<td>83.61</td>
</tr>
<tr>
<td>21</td>
<td>82.14</td>
<td>62.28</td>
<td>82.77</td>
</tr>
<tr>
<td>22</td>
<td>70.66</td>
<td>55.70</td>
<td>81.59</td>
</tr>
<tr>
<td>23</td>
<td>93.86</td>
<td>74.08</td>
<td>93.93</td>
</tr>
<tr>
<td>24</td>
<td>79.38</td>
<td>79.64</td>
<td>86.95</td>
</tr>
<tr>
<td>25</td>
<td>59.77</td>
<td>78.05</td>
<td>88.00</td>
</tr>
<tr>
<td>26</td>
<td>90.93</td>
<td>87.95</td>
<td>90.57</td>
</tr>
<tr>
<td>27</td>
<td>92.86</td>
<td>71.17</td>
<td>82.31</td>
</tr>
<tr>
<td>28</td>
<td>83.56</td>
<td>92.37</td>
<td>58.60</td>
</tr>
<tr>
<td>29</td>
<td>70.08</td>
<td>67.52</td>
<td>93.25</td>
</tr>
<tr>
<td>30</td>
<td>79.72</td>
<td>84.01</td>
<td>60.74</td>
</tr>
</tbody>
</table>

From Table 14, it is obvious that overall accuracies reached by different cost functions are different. Whether these differences are negligible or statistically significant is discussed in the next section.
3.5 Discussions

The results of urban growth simulation using CA and ICA are nonidentical in the different tests (Table 14). For instance, the overall accuracy of the simulation is 83.82% in test #1 for the validation of simulation, when the Kappa coefficient was used as the cost function. However, the overall accuracy is 86.96% in test #20 for the validation of simulation with the same data, the same framework parameters, and the same cost function (the Kappa coefficient). Similarly, using the total disagreement or patterns as the cost function results in dissimilar accuracies in different tests. For example, the overall accuracies of the simulations are 87.13% and 78.53% in test #1 for the validation of simulation, when the total disagreement and patterns were used as the cost function, respectively. However, the overall accuracies are 65.15% and 83.61% in test #20 for the validation of simulation with the same data, the same framework parameters, using the total disagreement and patterns, respectively.

Also, Table 14 indicates that the overall accuracy is different when different cost functions are used. For example, the overall accuracies are 83.82%, 87.13%, and 78.53% in test #1 using cost functions based on the Kappa coefficient, total disagreement, and patterns, respectively. To examine whether the same results are obtained by using the total disagreement and the Kappa coefficient (the first aim of this study), and also to examine whether including patterns into the cost function increases the accuracy of calibration (the second aim of this study) statistical tests were applied to the results.

Table 15 summarizes all the results by showing the descriptive statistics.
Table 15. Descriptive statistics for overall accuracies of CA’s calibration using ICA with different cost functions

<table>
<thead>
<tr>
<th>Cost function</th>
<th>Kappa coef.</th>
<th>Total disagreement</th>
<th>Kappa coef. and SRI(patterns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. tests</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Min</td>
<td>59.77</td>
<td>55.70</td>
<td>58.60</td>
</tr>
<tr>
<td>Max</td>
<td>93.86</td>
<td>92.37</td>
<td>94.65</td>
</tr>
<tr>
<td>Mean</td>
<td>79.76</td>
<td>73.81</td>
<td>81.11</td>
</tr>
<tr>
<td>StdD</td>
<td>09.23</td>
<td>10.81</td>
<td>10.61</td>
</tr>
<tr>
<td>Median</td>
<td>79.97</td>
<td>73.54</td>
<td>82.54</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.50</td>
<td>0.05</td>
<td>-0.66</td>
</tr>
<tr>
<td>z_Skewness</td>
<td>-1.18</td>
<td>0.12</td>
<td>-1.55</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.59</td>
<td>-1.05</td>
<td>-0.30</td>
</tr>
<tr>
<td>z_Kurtosis</td>
<td>-0.71</td>
<td>-1.26</td>
<td>-0.36</td>
</tr>
<tr>
<td>Normality</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

As Table 15 shows, in the simulation of urban growth in Shelby County, the minimum of overall accuracies is higher when the Kappa coefficient was used as the cost function (59.77%) rather than patterns (58.60%) or the total disagreement (55.70). However, the maximum of overall accuracies is higher when patterns were used as the cost function (94.65%) rather than the Kappa coefficient (93.86%), or total disagreement (92.37%). In terms of the means, the mean of overall accuracies is higher when patterns were used as the cost function (81.11%) rather than the Kappa coefficient (79.76%) or the total disagreement (73.81%). Also, the medians of overall accuracies are visually different indicating a better performance of patterns (82.54%) in comparison with the Kappa coefficient (79.97%) or the total disagreement (73.54%).

Based on Table 15, calculated z-scores of skewness and kurtosis indicate a normal distribution for overall accuracies in all three methods (using patterns, the Kappa coefficient, and the total disagreement). The distribution is considered normal when the z-scores fall between -1.96 and +1.96 for a conventional significance level of 0.05 (Foreman & Corder, 2014). This
condition has been satisfied for the results of all three different cost functions indicating normality of the results. Figure 18 illustrates histograms of overall accuracies from Table 14.

Figure 18. Histogram and the normal distribution (in red) using different cost functions: a) the Kappa coefficient; b) the total disagreement; c) the Kappa coefficient and SRI (patterns)

Particularly noticeable from the histograms are the differences in the shape of the distribution for each of the three methods. There is a skewness (-0.50) in the results when the Kappa coefficient was used as the cost function (Figure 18.a). Skewness indicating the symmetry of the data is +0.05 and -0.66 using the total disagreement and patterns, respectively (Figure 18.b and Figure 18.c). The kurtosis of results measuring how flat or peaked the distribution are -0.30, -0.59, and -1.05 for use of patterns, the Kappa coefficient, and the total disagreement,
respectively (Table 15). As z-scores of both skewness and kurtosis fall between -1.96 and +1.96, the results of all methods (Table 14) are not statistically significantly different from a normal distribution (Ali Farzan, 2018; Foreman & Corder, 2014).

Descriptive statistics of the results (Table 15) and histograms (Figure 18) show visual differences in those three methods (using patterns, the Kappa coefficient, and the total disagreement as the cost function). However, considering the normality of the results, a t-test was applied to the results to determine whether there is a significant difference between them.

To examine the first hypothesis of this study, that the total disagreement is a superior alternative to the standard Kappa coefficient, an alternative hypothesis of the t-test was defined as $H_1: \mu_{ICA,kc} > \mu_{ICA,td}$, where $\mu_{ICA,kc}$ denotes the population mean of the results using ICA and the Kappa coefficient, and $\mu_{ICA,td}$ denotes the population mean of the results using ICA the total disagreement. The output of the t-test ($t(58) = 2.2935$, $p = 0.0127$) at a conventional significance level of 0.05 rejected the null hypothesis ($H_0: \mu_{ICA,kc} = \mu_{ICA,td}$), which means that using the Kappa coefficient in ICA (Mean = 79.76%, St. D. = 9.23) statistically results in a significantly higher precision (a more realistic simulation) than using the total disagreement (Mean = 73.81%, St. D. = 10.81) at the simulation of urban growth. This finding shows that the Kappa coefficient is a useful measure for practical applications and the profession should not abandon the Kappa indices for purposes of accuracy assessment and map comparison.

To examine the second hypothesis of this study, that including SRIs as a pattern in the cost function of calibration (Equ. 20) can increase the simulation’s accuracy, a t-test was applied to the results with an alternative hypothesis as $H_1: \mu_{ICA,kc} \neq \mu_{ICA,K+SRI}$, where $\mu_{ICA,kc}$ denotes the population mean of the results using the Kappa coefficient, and $\mu_{ICA,K+SRI}$ denotes the population mean of the results using the Kappa coefficient and SRIs. The output of the t-test ($t(58)$...
=0.5277, \( p = 0.5997 \) at a conventional significance level of 0.05 failed to reject the null hypothesis \( H_0: \mu_{ICA,kc} = \mu_{ICA,K+SR1} \), which means that adding patterns (SRIs) to the Kappa coefficient in the cost function of the calibration (Mean = 81.11\%, St. D. = 10.61) will not significantly improve the accuracy of simulation, in comparison with a cost function containing only the Kappa coefficient (Mean = 79.76\%, St. D. = 9.23). Even though using patterns reached a higher accuracy (94.65\% vs 93.86\%), a higher mean (81.11\% vs 79.76\%) and a higher median (82.54\% vs 79.97\%) than using the Kappa coefficient alone.

Population size is an important parameter in metaheuristics. Due to the random sampling methods for creating initial solutions (e.g. chromosomes in GA, or countries in ICA), an increase in the population size may improve the performance. However, it escalates the elapsed time drastically (Momeni & Antipova, 2020). The current research provides the first application of ICA in the urban planning field. To compare the overall performance of ICA with GA, the CA model of urban growth was also calibrated using the population size of 1000. Table 16 presents the Kappa coefficients and overall accuracies of calibration and validation using ICA, GA, and LR.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cost function</th>
<th>Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Kappa coef. (%)</td>
<td>Overall accuracy (%)</td>
</tr>
<tr>
<td>ICA</td>
<td>Kappa coef. and patterns</td>
<td>87.87</td>
<td>94.10</td>
</tr>
<tr>
<td></td>
<td>Kappa coef.</td>
<td>87.20</td>
<td>93.77</td>
</tr>
<tr>
<td></td>
<td>total disagreement</td>
<td>86.10</td>
<td>93.25</td>
</tr>
<tr>
<td>GA*</td>
<td>Kappa coef. and patterns</td>
<td>87.11</td>
<td>93.73</td>
</tr>
<tr>
<td></td>
<td>Kappa coef.</td>
<td>85.13</td>
<td>92.79</td>
</tr>
<tr>
<td>LR*</td>
<td>-</td>
<td>85.44</td>
<td>92.89</td>
</tr>
</tbody>
</table>

* extracted from Momeni & Antipova, 2020
As Table 14 and Table 16 show, increasing the population size from 100 to 1000 slightly increases the overall accuracy of validation from 94.65\% to 94.88\% using patterns, from 93.86\% to 94.40\% using the kappa coefficient, and from 92.37\% to 94.03\% using the total disagreement. Moreover, with a population size of 1000, the Kappa coefficient and the overall accuracy in all methods are higher than LR’s (92.98\%) which is a commonly used method of calibration for CA (Table 16).

It is noteworthy to mention that only one percent increase in the overall accuracy results in the correct simulation of 19.77 sq. km of land in Shelby County. For example, a simulation of CA by ICA using patterns (with an overall accuracy of 94.88\%) is more realistic than the simulation of CA by LR (with an overall accuracy of 92.98\%) over 37.56 sq. km of land in Shelby County ([94.88-92.98] \times 19.77 = 37.56). Moreover, the simulation of CA by ICA using patterns (with an overall accuracy of 94.88\%) is more realistic than the simulation of CA by ICA using the Kappa coefficient (with an overall accuracy of 94.40\%) over 9.49 sq. km of land in Shelby County ([94.88-94.40] \times 19.77 = 9.49).

The analyses of elapsed time show that using a computer with a processor of an Intel Core i7 with CPU @ 2.20 GHz, the calibration of CA using ICA with patterns and the population size of 100, takes 138 minutes, on average, to reach a convergent result. Calibration using the Kappa coefficient and the total disagreement take 64 and 68 minutes, on average. When the population size rises to 1000, the elapsed time rises almost to 23 hours for calibration of ICA using patterns, 13 hours each for calibration of ICA using the Kappa coefficient, and calibration of ICA using the total disagreement. Table 17 summarizes the elapsed time based on population size and the cost functions.
Table 17. elapsed time for the calibration of CA by ICA using different cost functions

<table>
<thead>
<tr>
<th>Population size</th>
<th>Kappa coef.</th>
<th>Total disagreement</th>
<th>Kappa coef. and SRIs (patterns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>64 minutes</td>
<td>68 minutes</td>
<td>138 minutes</td>
</tr>
<tr>
<td>1000</td>
<td>13 hours</td>
<td>13 hours</td>
<td>23 hours</td>
</tr>
</tbody>
</table>

Figure 19 shows the results of the urban growth simulation in Shelby County in 2011. Correctly-simulated developed lands, correctly simulated non-developed lands, omission errors, and commission errors are illustrated there.

Errors of omission are false negative errors, where pixels of a known class are classified as something other than that class (ArcMap, 2022). E.g. a developed area in reality that was classified as a non-developed area in the simulation. The highest simulation accuracy that obtained by ICA and patterns resulted in 3.4% error as omission. A visual inspection of errors indicates that most of the omissions occurred around the scattered urban areas in Shelby County’s rural area. Adding secondary roads as a driving force into the simulation may decrease this error as most of the secondary roads were omitted in the simulation. Moreover, some commercial/industrial lands (such as Figure 19-b) were omitted as it was surrounded by non-developed areas, was close to river, streams, and wetlands, and was far away from residential areas. Including land use data, such as commercial and residential data, in the simulation may improve the performance of CA in those areas, which can be a topic of research for further studies.
Figure 19. Simulation results using ICA and patterns
In comparison with omission errors, errors of commission are false positive errors, where pixels are incorrectly classified as a known class when they should have been classified as something else. (ArcMap, 2022). E.g. a non-developed area in reality that was classified as a developed area in the simulation. The highest simulation accuracy that obtained by ICA and patterns resulted in 1.6% error as commission. A visual inspection of errors indicates that most of the commissions occurred around the edges of developed patches (e.g. Figure 19-c). Using higher-resolution land cover data, which provides sharper segregation of developed land from its surrounding non-developed land, may decrease this type of error. However, due to the finer resolution of land cover data and the smaller size of CA cells, the computational cost of simulation may increase.

As mentioned in Section 3.2.1, land cover data provided by MRLC are available only at a spatial resolution of 30 meters, based on the classification of Landsat satellite images. Classification of higher-resolution satellite images such as Sentinel-2A (with the resolution of 10 m), RapidEye (with the resolution of 5 m), and IKONOS (with the resolution of 0.8 m) (Satimagingcorp, 2022) produce finer land cover data. However, to avoid adding extra errors to the simulation, the same satellite images, the same radiometric and geometric correction techniques, and the same classification method should be used to produce land cover data needed for calibration and validation of CA.

Finally, this study provides a basis for future implications to predict the future of urban growth in Shelby County. E.g. to predict developed areas in Shelby County in 2030. These predictions can assist city planners to optimize their plans to control city-related issues such as traffic congestion and air pollution and promote healthy urban growth. To have more realistic
predictions, it will be important to include economical factors (e.g. land price, income, job-tohousing ratio, etc.), demographic factors (e.g. population density), environmental factors (e.g. natural disaster data, soil type, etc.), and policies (e.g. national conservation lands, tax rates, zoning plans, etc.) to the simulations.

3.6 Conclusion

This study successfully applied ICA, for the first time, in the field of urban planning to calibrate a CA model of urban growth. We used the case study set in Shelby County, Tennessee that is known for its significantly higher per capita land consumption (978 sq. meters/person) compared to Washington D.C. (480 sq. meters/person) and Portland, OR (120 sq. meters/person) based on 2001 data (Banai et al., 2021; Calthorpe & Fulton, 2001).

The first aim of this study was to examine whether the total disagreement is statistically an alternative for the Kappa coefficient in the calibration process of urban growth simulation using the CA model. Results of urban growth simulation in Shelby County indicated that using the Kappa coefficient in the cost function of ICA reaches statistically significantly higher precision (Mean = 79.76%, St. D. = 9.23) than the use of total disagreement in the calibration of CA (Mean = 73.81%, St. D. = 10.81), results in a more realistic model. The results support that the Kappa coefficient is usable for applications including accuracy assessment and map comparison, and should not be abandoned by scholars.

The second aim of this study was to examine whether including the SRIs, as a pattern, into the cost function of calibration can increase the simulation’s accuracy. The results indicate that adding patterns to the Kappa coefficient in the cost function of the calibration increases the overall accuracy (from 93.86% to 94.65%), the mean (from 79.76% to 81.11%), and also the median (from 79.97% to 82.54%) of simulations; though not significantly. The pattern-based
calibration resulted in a more realistic simulation of urban growth over 9.49 sq. km of land in Shelby County (in comparison with simulation without patterns).

The results also show that ICA, regardless of the cost function, can reach a higher accuracy than LR as the commonly used method for calibration of CA. While the overall accuracy of the simulation was 92.98% using LR, ICA reached the highest overall accuracy of 94.88% using patterns, 94.40% using the Kappa coefficient, and 94.03% using the total disagreement. Using ICA and patterns resulted in a more realistic simulation of urban growth over 37.56 sq. km of land in Shelby County (in comparison with LR). These results indicate the superiority of ICA in the calibration of a CA.

In contrast with ANN, and similar to other evolutionary optimization algorithms, ICA does not need the gradient of the function in its optimization process (Abdi et al., 2011). Also, ICA, similar to GA, is sensitive to the initial random solutions. However, in contrast to GA, ICA attempts to reduce the chance of trapping in local optima by defining different empires and inter-empire competitions. Our study shows that in practical urban simulation problems, ICA can still fall into local optima and provide pre-mature optimizations. Increasing the number of initial empires or increasing the number of generations may mitigate this issue, which can be a topic for further studies.

Due to the random selection of initial solutions in evolutionary optimization algorithms such as ICA and GA, these algorithms return different results under the same conditions using the same data. Therefore, to enable statistical analysis, we should run different methods adequately large enough number (n) of times independently (in this study n=30), to collect sufficient samples for the probability distribution for each algorithm. However, in practice, we cannot be assured to know how many tests (n) are required to reach reliable results. A rule of thumb is to
have at least \( n = 30 \) for each algorithm. This study includes the required amount of \( n \), however, more tests may lead the t-test to a more robust decision. This also can be a topic for further studies.

The results of this paper can help city developers, construction companies, transportation engineers, tax assessors, and utility providers to make more realistic decisions in Shelby County.
Chapter 4: Conclusion

This dissertation included three parts. The first part was dedicated to the inspection of the socio-demographics and LCLU trends in Shelby County, Tennessee. The findings indicate that from 1990 to 2010 the population in Shelby County increased by 12.26%. While the population in Memphis decreased by 7.78%, the population in Shelby County excluding Memphis increased by 124.85%, which means a disproportional population growth over the county. The decentralization of the population resulted in rapid urban development in the suburbs of Memphis, causing car-oriented, leap-frog sprawled development in Shelby County. The findings also indicate several significant associations between socio-demographic attributes and commuting. LCLU mixtures, travel behaviors, commuters’ age, race, education, and population of neighborhoods, influence each other and drive urban growth in Shelby County. The results of this section can help urban planners better understand the driving forces of urban growth. Moreover, understanding travel patterns help policymakers to control local/regional problems such as traffic congestion and emissions due to a shift in choices towards the use of private modes, as well as develop strategies for encouraging active modes and public transport use in the post-COVID-19 world.

Various models are applied to estimate urban growth. Cellular Automata (CA) is the commonly used simulation of urban growth yet its calibration remains challenging. Accordingly, the feasibility of using patterns in the calibration of a CA model was investigated in the second part. In that study, a linear combination of the Kappa coefficient and the maximum absolute errors of SRIs was directly used in the cost function of a genetic algorithm (GA). The proposed calibration method resulted in a more realistic simulation of urban expansion in Shelby County.
Utilizing a non-linear combination and a sensitivity analysis of combination weights (e.g. \( \alpha \) in Equ. 6) can be investigated by future studies. A version of this study has been published in the journal of Transactions in GIS.

My dissertation proposes an improvement in the existing modeling approach. In the third part, an Imperialistic Competitive Algorithm (ICA) was implemented for the calibration of the CA in order to further improve the simulation of urban growth. Different cost functions, including the pattern-based cost function, were used in the calibration process. The most accurate simulation, with an overall accuracy of 94.88\%, was obtained when ICA was used with the patterns included. In comparison with the commonly used conventional method of CA calibration, logistic regression (LR) reached an overall accuracy of 92.98\% which underestimates 37.56 sq. km of urban growth. Sensitivity analyses of the number of initial empires and number of generations can be explored next. Moreover, regardless of the algorithm used for the optimization (e.g. GA or ICA), simulation of urban growth using irregular-tessellation CA can be another research topic. In that case, urban development will be simulated at a neighborhood level using Census block group boundaries.

By using the findings of this dissertation local governments can direct urbanization to a smart and sustainable development while avoiding the undesirable effects of sprawl and its negative environmental outcomes.
References


Ali Farzan, I. (2018). *The Interplay Of Culture, Motivation, And Self: An Investigation Of Math Achievement Gap Between Middle School Students In The Us And Those In The Top-Performing Countries In East Asia* (The University of Memphis). Retrieved from [https://search.proquest.com/docview/2049687261](https://search.proquest.com/docview/2049687261)


Bell, M., Dean, C., & Blake, M. (2000). Forecasting the pattern of urban growth with PUP: a


122
to simulate present and future land use. *Transactions in GIS*, 22(2), 582-601. DOI: 10.1111/tgis.12331


Kuzmyak, J. R. (2012). Land use and traffic congestion. Arizona Department of Transportation Research Center. Retrieved from
https://repository.asu.edu/attachments/108918/content/Land%20Use%20and%20Traffic%20Congestion.pdf

http://www.kyovaipc.org/2040%20MTP%20Chapter%208%20Land%20Use%20Considerations.pdf

Land cover data. (2020). MRLC. Retrieved from
https://www.mrlc.gov/data?f%5B0%5D=category%3Aland


Momeni E., Antipova, A. (2022a). The Kappa coefficient is still alive.


MRLC. (2019). *Land cover Data*. Retrieved from
https://www.mrlc.gov/data?f%5B0%5D=category%3Aland%20cover


objective optimisation framework for calibration of cellular automata land-use models.

*Environmental Modelling & Software*, 100 (2018), 175-200. DOI:

10.1016/j.envsoft.2017.11.012

NOAA. (2017). What is the difference between land cover and land use?. Retrieved from

http://oceanservice.noaa.gov/facts/lclu.html


10.1016/j.amepre.2007.07.025


https://doi.org/10.1016/j.procs.2018.04.165


Roghair, C. N., & Dolloff, C. A. (2013). *Application of Diversity Indices and an Index of Biotic Integrity to a Basin Area Stream Survey (BASS ) Fish Dataset Collected on the Ouachita*

Retrieved from


http://www.itc.nl/personal/rossiter/teach/R/R_ac.pdf

Russel, M. (2015). *Phoenix voters approve a plan to raise money for transportation; vastly expand the city’s light rail and bus networks*. Retrieved from


134


growth dynamics of an Indian metropolitan using CA Markov and Logistic Regression.


https://doi.org/10.1016/j.ejrs.2017.11.006


https://doi.org/10.1016/S0198-9715(01)00014-X


https://doi.org/10.1111/j.1939-7445.2001.tb00073.x


https://www.tn.gov/content/dam/tn/tdot/documents/Mobility_022316.pdf


U.S. Census Bureau. (2022). Retrieved from https://www2.census.gov/geo/docs/maps-data/data/gazetteer/counties_list_47.txt


https://doi.org/10.1016/J.COMPENVURBSYS.2017.04.005


https://doi.org/10.1068/b240235

https://ideas.repec.org/a/pio/envirb/v24y1997i3p323-343.html


