A New Approach to Develop Large Scale Land Use Models Using Open Source Data

Ali Riahi Samani
Department of Civil Engineering
University of Memphis, Memphis, Tennessee, United States, 38152
Email: rhsamani@memphis.edu

Sabyachee Mishra
Department of Civil Engineering
University of Memphis, Memphis, Tennessee, United States, 38152
Email: smishra3@memphis.edu

David Jung-Hwi Lee
Assistant Director, Long Range Planning
Tennessee Department of Transportation
Nashville, Tennessee, 37243-0344
Email: david.Lee@tn.gov

Mihalis M. Golias
Department of Civil Engineering
University of Memphis, Memphis, Tennessee, United States, 38152
Email: mgkolias@memphis.edu

Jerry Everett
Center for Transportation Research
The University of Tennessee Knoxville, Tennessee, United State, 37996-4133
Email: jeverett@utk.edu

ABSTRACT

Developing a land-use model for large-scale cases is a topic that has received less attention in the literature while transportation engineers and urban planners continue to analyze the effect of various policies in multi-jurisdiction metropolitan areas and to some extent in statewide scale. Gravity based models when too simplistic, microsimulation models require extensive data and massive computation. This paper presents a land use model that can be applied to large-scale geographies using open source data and be able to forecast demographic and socioeconomic attributes with reasonable accuracy and acceptable computational time. The proposed model incorporates the Putman's Integrated Transportation–Land Use Package (TELUM) and Kockelman's Gravity-based Land Use Model (G-LUM) fundamentals with enhanced formulation of newly added variables and structural changes. Considering the non-convex and non-linear nature of the proposed model, we utilize an enhanced genetic algorithm for base year calibration. Further, we assess the accuracy of the model with two-fold validation including back-casting and forecasting. We utilize the state of Tennessee as the case study area and utilized all open source data available to the model application. The model results show reasonably accurate estimates of households by size, employment by industry, and land utilization by condition. As applicable the model outperforms G-LUM and TELUM by accuracy (R^2) and error measures (MAPE). The proposed land use model has the potential to be applied for medium to large scale geographies with reasonable accuracy in predicting socio-economic, demographic, and land condition estimates by using open source data.

Keywords: Integrated Land Use, Transport Model, Gravity Theory, Statewide Land Use Model, Genetic Algorithm

INTRODUCTION

The interdependence of land use and transportation led urban planners and transportation engineering to focus on each while the former has received less attention in recent years. The first generation of land use models such as aggregate spatial interaction and gravity models were introduced around 1960s. Lowry was a pioneer by introducing the model of metropolis (*1*). Later on, utility-based econometric and discrete choice models were developed. The development of advanced micro-simulation land use models and activity-based travel demand models created the need for a new generation of integrated land use-transport systems (*2*). New models such as ILUTE (*3*) and ILUMASS (*4*) were developed thereafter. Existing models such as UrbanSim (*5*), PECAS (*6*), and MUSSA (*7*) were updated to facilitate the need for advanced research in the field of integrated land use-transport modeling.

In recent years, improving the accuracy of the land use models by introducing micro-simulation garnered much attention because of behavioral interpretation, studies on their accuracy is still an evolving area of research. The problem of dealing with large-scale geographies received less attention though multiple jurisdictions are managed by metropolitan planning authorities, or state planning agencies. T the topic of accuracy versus scale became important on the regional scale, such as multi-jurisdiction metropolitan areas and in statewide applications. Despite the high accuracy of microsimulation land use models (i.e. UrbanSim (*8*)), the enormous data requirement and massive computation time, make the implementation of these models on large-scale cases challenging (*9*). The development of a large-scale land use model becomes more important when the integration of land use models with a statewide travel demand model is raised. It is important to urban planners and transportation engineers to assess and analyze the effect of policies or scenarios on broader scales. Therefore, the development of a land use model can be applied in large-scale (regional or statewide) becomes crucial.

The purpose of this research is to develop a land use model which can be applied to large-scale geographies with acceptable computational time reasonable forecasting accuracy and use of open-source data. Such large-scale models can be integrated with existing travel demand models as applicable in many areas. Few land use models are applied in large-scale geographies, such as PECAS and TELUM (*10*). PECAS is a generalized approach for simulating spatial economic systems. It operates by clearing spatial submarkets for various goods, services, and factors in a short-run equilibrium based on development event probabilities (*6*). California's statewide land use model is a statewide PECAS model, integrated with a statewide travel demand forecast model (*11*). Moreover, PECAS components applied in the development of statewide transportation land use modeling systems for Ohio and Oregon (*12*). However, the implementation of PECAS to large-scale cases has limitations on computational time and the number of zones (*13*).

TELUM is another well-known land use model. TELUM is an integrated land use and transport model that incorporates gravity theory to allocate households and employment to zones (*14*). TELUM development is based on three components, a disaggregated residential allocation model (DRAM), an employment allocation model (EMPAL), and a land consumption model (LANCON) (*10*). Presenting a user-friendly interface, GIS-base result, reasonable data requirements, and short run time have gained researchers' attention to applying this model in their projects (*15–19*). Although the implementation of TELUM is simple, this model has limitations. First, in TELUM the number of employment categories is limited to 4 to 8 employment sections and the number of household categories is limited to 5-8 categories, usually for household's income. Second, TELUM has restrictions on zone size and it is recommended that the average population in a zone lies between 3,000 and 10,000 (*10*).

Kockelman et al. tried to solve these limitations of TELUM by developing Gravity Land Use Model (G-LUM) (*20*). The G-LUM structure is based on the formulation of the ITLUP package (*21*) and includes three major sub-models for predicting changes in employment location (EMPLOC), residential location (RESLOC), and land consumption (LUDENSITY). G-LUM was used to validate the outputs of TELUM. Studies showed that, although G-LUM and TELUM use the same structure, the forecasting results were varied (*22*). The differences rooted in the method of calibration (*19*). TELUM uses a gradient search method and G-LUM uses the Nelder-Mead method with 12 different initial points. In addition to the mentioned limits, the formulation of land use consumption in TELUM and G-LUM (LANCON and

LUDENSITY) tends to generate unreasonable average land consumption values for households and jobs (as compared with base and prior year land conditions in each zone) (*23*).

The contribution of this paper is three-fold. First, improvement in the model structure of TELUM and G-LUM, so that large scale implementation is possible. Second, proposition of a new solution algorithm to enhance the accuracy of the land use models. Third, demonstration of model applicability using a case study area by only using open source data. The rest of the paper is organized as follows. The next section discusses the proposed methodology and calibration procedure. The following section presents the data used and the case study. The results section compared the performance of the proposed models with similar models in the past. The conclusion section summary of the paper, and avenues for future research.

METHODS

In this section, the proposed land use model development and specifications are discussed. First, the description and formulation of different sections of the model are provided. Then, a brief description of the travel demand model which is planning to be integrated with the proposed land use model is provided. Finally, the calibration of the proposed model is discussed. As a general description, the proposed model is based on gravity theory and incorporates the principle structure of the TELUM and G-LUM, while the details have been changed to improve the model's accuracy. The model presented in TAZ level and forecasts socioeconomic and demographic character of zones in five-year intervals.

Land Use Models

In this section, two land use models' structure are provided. The first model is called Large-Scale Land Use Model (LS-LUM) and the second model is named Large-Scale Land Use Model Without House Condition (LS-LUM-WOHC). In the following subsections, the formulation of both models is provided.

Large-Scale Land Use Model (LS-LUM)

LS-LUM contains two principal sections and two subsections. Principal sections estimate households and employments in different categories. These two models incorporate gravity theory to allocate households and employments to each TAZ. The first principal model is named *HH-AL* (Households Allocation) which is responsible for residential location choice. This model assigns households to each TAZ based on the total number of houses, vacant houses, the amount of residential land (acres), total useable land in each TAZ, and the travel cost between zones. The second principal model is called *EMP-AL* (Employments Allocation). This model allocates employments to zones based on job opportunities in the prior year (lag year), the amount of commercial, industrial, and agricultural land (acres) in each TAZ, and the travel cost between zones. Two subsections are responsible for updating house conditions and land use consumption which are the components of principal sections. Two models provided in these subsections are called *HC* (House Condition) *and LC* (Land Consumption). *HC* models the number of total and vacant houses in each TAZ and *LC* Models the amount of land (acres) in each land use class (residential, commercial, industrial, agricultural, and developable or vacant). Multiple Linear Regression (MLR) is applied in these subsections. In the following sections, the formulations of these models are provided. In general, in this paper i and j represent TAZs, $C_{i,j}$ represents the travel cost between zone i and zone j, and t represents the period of time (i.e. year 2010). Moreover, n and k stand respectively for household categories (i.e. different household size) and employments categories (i.e. NAICS sectors categories).

1. *HH-AL:*

$$N_{i,t}^{n} = \eta^{n} \sum_{j} a_{n}\, E_{j,t}^{T} \frac{W_{i,t-1}^{n}\, C_{i,j,t-1}^{\alpha^{n}}\ \exp\left(\beta^{n} C_{i,j,t-1}\right)}{\sum_{i} W_{i,t-1}^{n}\, C_{i,j,t-1}^{\alpha^{n}}\ \exp\left(\beta^{n} C_{i,j,t-1}\right)} + (1-\eta^{n}) N_{i,t-1}^{n} \qquad (1)$$

Where

$$W_{i,t-1}^{n} = (H_{i,t-1}^{T})^{o^n} (H_{i,t-1}^{V})^{p^n} (1 + \frac{L_{i,t-1}^{Res}}{L_{i,t-1}^{T}})^{q^n} \tag{2}$$

In **Equation 1**, $N_{i,t}^{n}$ is the number of households in category n in zone i in time t, a_n is the proportion of the population to employment in zone i, $E_{j,t}^{T}$ is the total number of employments, $W_{i,t-1}^{n}$ is the attractiveness function of zone i to which attract employment in zone j to live in zone i in year $t-1$. $W_{i,t-1}^{n}$ is a weighted multiplication of different components in a zone. In **Equation 2**, $H_{i,t-1}^{T}$ is the total number of houses, $H_{i,t-1}^{V}$is the number of vacant houses, $LP_{i,t-1}$ is the residential land value, $L_{i,t-1}^{Res}$ is the amount of residential land are in zone i in year $t-1$. Finally, η, α, β, o, p, and q are parameters estimated in the calibration procedure.

2. *EMP-AL:*

$$E_{j,t}^{k} = \lambda^{k} \sum_{i} N_{i,t-1}^{T} \frac{M_{j,t-1}^{k}\, C_{i,j,t-1}^{\alpha^k}\ \exp(\beta^k C_{i,j,t-1})}{\sum_i M_{j,t-1}^{n}\, C_{i,j,t-1}^{\alpha^k}\ \exp(\beta^k C_{i,j,t-1})} + (1-\lambda^n) E_{j,t-1}^{k} \tag{3}$$

Where,

$$M_{j,t-1}^{k} = (E_{j,t-1}^{k})^{g^k} (L_{j,t-1}^{Com} + L_{j,t-1}^{Ind} + L_{j,t-1}^{Agr})^{h^k} \tag{4}$$

In **Equation 3**, $E_{j,t}^{k}$ is the number of employments in category k, $N_{i,t-1}^{T}$ is the total number of households, and $M_{j,t-1}^{k}$ is the attractiveness function shows how much zone j is attractive for peopled living in zone i to find a job. $M_{j,t-1}^{k}$ is calculated based on, job opportunities in year $t-1$ ($E_{j,t-1}^{k}$), amount of commercial ($L_{j,t-1}^{Com}$), industrial ($L_{j,t-1}^{Ind}$), and agricultural ($L_{j,t-1}^{Agr}$) land in zone j in year $t-1$. λ, α, β, g, and h are parameters estimated in the calibration procedure.

3. *HC:*

In this subsection, the total number of houses and the number of vacant houses in each TAZ are updated. First, the total number of houses in each TAZ is calculated by applying a Multiple Linear Regression. As **Equation 5** shows, the total number of houses in zone i and in year t ($H_{i,t}^{T}$) is the dependent variable; while, the number of total houses in the previous year ($t-1$), the amount of vacant land ($L_{i,t-1}^{Vac}$), and the total number of households are the independent variables.

$$H_{i,t}^{T} = \theta_0 + \theta_1(H_{i,t-1}^{T}) + \theta_2\left(H_{i,t-1}^{V}\right) + \theta_3\left(L_{i,t-1}^{Vac}\right) + \theta_4(N_{i,t}^{T}) + \varepsilon \tag{5}$$

In **Equation 5**, $L_{i,t-1}^{Vac}$ is the amount of vacant or developable land in zone i and ε is the error associated in regression. In this equation, θ_0 is the intercept and θ_1 to θ_4 are coefficient estimated in calibration.

After calculating the total number of houses, the number of vacant houses in each TAZ can be estimated as follow:

$$H_{i,t}^{V} = H_{i,t}^{T} - \sum_{n} N_{i,t}^{n} \tag{6}$$

4. *LC:*

Finally, in *LC,* the amount of land in different land use classes is updated to feed the two principal models (*HH-AL* and *EMP-AL*) in order to forecast future years' demographic and socio-economic conditions.

$$L_{i,t}^{Res} = R_0 + R_1 L_{i,t-1}^{Vac} + R_2(L_{i,t-1}^{Res}) + R_3(N_{i,t-1}^{T}) + R_4(N_{i,t}^{T}) + \varepsilon \tag{7}$$

$$L_{i,t}^{Com} = C_0 + C_2(L_{i,t-1}^{Vac}) + C_2(L_{i,t-1}^{Com}) + C_3(E_{i,t-1}^{Com}) + C_4(E_{i,t}^{Com}) + \varepsilon \tag{8}$$

$$L_{i,t}^{Ind} = I_0 + I_1(L_{i,t-1}^{Vac}) + I_2(L_{i,t-1}^{Ind}) + I_3(E_{i,t-1}^{Ind}) + I_4(E_{i,t}^{Ind}) + \varepsilon \tag{9}$$

$$L_{i,t}^{Agr} = A_0 + A_1\left(L_{i,t-1}^{Vac}\right) + A_2\left(L_{i,t-1}^{Agr}\right) + A_3\left(E_{i,t-1}^{Agr}\right) + A_4(E_{i,t}^{Agr}) + \varepsilon \tag{10}$$

$$L_{i,t}^{Vac} = L_{i,t-1}^{Vac} - \left(L_{i,t-1}^{Res} - L_{i,t}^{Res}\right) - \left(L_{i,t-1}^{Com} - L_{i,t}^{Com}\right) - \left(L_{i,t-1}^{Ind} - L_{i,t}^{Ind}\right) - \left(L_{i,t-1}^{Agr} - L_{i,t}^{Agr}\right) \tag{11}$$

In **Equations 9** to **10**, E^{Arg} refers to the number of employments in NAICS sector 11 (agriculture, forestry, fishing, and hunting), E^{Com} is the number of employments in NAICS sectors 44, 45, 51, 52, 53, and 72 (retail trade, finance and insurance, real estate and rental and leasing, accommodation and food services), and E^{Ind} is the number of employments in NAICS sectors 21, 31, 33, and 42 (mining, quarrying, oil and gas extraction, manufacturing, and wholesale trade).

Large-Scale Land Use Model Without House Condition (LS-LUM-WOHC)

LS-LUM-WOHC is the second model developed in this paper. The formulation of this model is very similar to LS-LUM and the only difference is that in this land use mode, the house condition subsection (*HC*) and its components have been removed from LS-LUM. The purpose of developing this land use model is to evaluate the effect of adding *HC* to land use modeling. In other words, developing LS-LUM-WOHC provides the opportunity to compare the presence and absence of *HC*. Therefore, LS-LUM-WOHC is consist of two principal models and a subsection model. The principal models are responsible for allocating households and employment. LS-LUM-WOHC incorporates the same model for allocating employment. *EMP-Al* is applied in this model too. However, the household allocation model is different in comparison with LS-LUM. Household allocation model in LS-LUM-WOHC is called *HH-AL2*, where the formulation is as follow:

$$N_{i,t}^{n} = \eta^{n} \sum_j a_n\, E_{j,t}^{T} \frac{W'^{n}_{i,t-1}\, C_{i,j,t-1}^{\alpha^n}\ \exp\left(\beta^n c_{i,j,t-1}\right)}{\sum_i W'^{n}_{i,t-1}\, C_{i,j,t-1}^{\alpha^n}\ \exp\left(\beta^n c_{i,j,t-1}\right)} + (1-\eta^n) N_{i,t-1}^{n} \tag{12}$$

Where,

$$W'^{n}_{i,t-1} = \left(1 + \frac{L_{i,t-1}^{Res}}{L_{i,t-1}^{T}}\right)^{q'^n} \tag{13}$$

In the formulation of *HH-AL2*, the attractiveness of each zone ($W'^{n}_{i,t-1}$) is calculated only by using the amount of residential ($L_{i,t-1}^{Res}$) and the total land ($L_{i,t-1}^{T}$) in zone i and in the prior year. In addition, the formulation for the subsection model is similar to LS-LUM; where, *LC* is applied to forecast the amount of land in five different land use classes.

Model Explanation

This integrated modeling framework starts with forecasting employment in different categories and for each zone (see **Figure 1**). This section of the model gets employment, the amount of agricultural, commercial, industrial lands, and travel cost in each zone and for the prior year. The output of this section is the forecasted employment (by different categories) in each TAZ. The output of the *EMP-AL* would serve as input for the *HH-AL*. The *HH-AL* incorporates the current total employment (from *EMP-AL*), the total number of houses, the number of vacant houses, and the proportion of residential to total land in each zone for the prior year. The output of this section is the number of households (by different categories, e.g. income). Then *HC* computation is processed, by forecasting how many houses will be built in each TAZ. This section needs total and vacant number of houses, the amount of vacant land in the prior year, and the forecasted total number of households (from *HH-AL* section). Considering the total number of forecasted houses and total households in each TAZ, *HC* forecasts the number of vacant houses in each zone by subtracting the total number of households from the total number of houses in each zone. The output of *HC* feeds the *HH-AL* by providing the number of total and vacant houses. Lastly, *LC* forecasts the amount of residential, commercial, industrial, agricultural, and vacant land (developable) in each zone and for each forecasting year. The output of *LC* directly affects other models' sections. By connecting *LC* to other sections, capturing the effect of land use changes on the socio-economic character of each TAZ would be possible and more accurate results can be obtained. The amount of commercial, industrial, and agricultural land modeled in this section are added to employment section. The amount of residential land is added to the *HH-AL*. Finally, the amount of vacant land is one of the components involved in forecasting the total number of houses in a zone. Moreover, the amount of vacant land in each zone works as a development restriction. Because in the model, if all the vacant land had allocated to other land use classes, no more development will happen, and the model will stop adding a new area to other land use classes (residential, commercial, industrial, and agricultural).

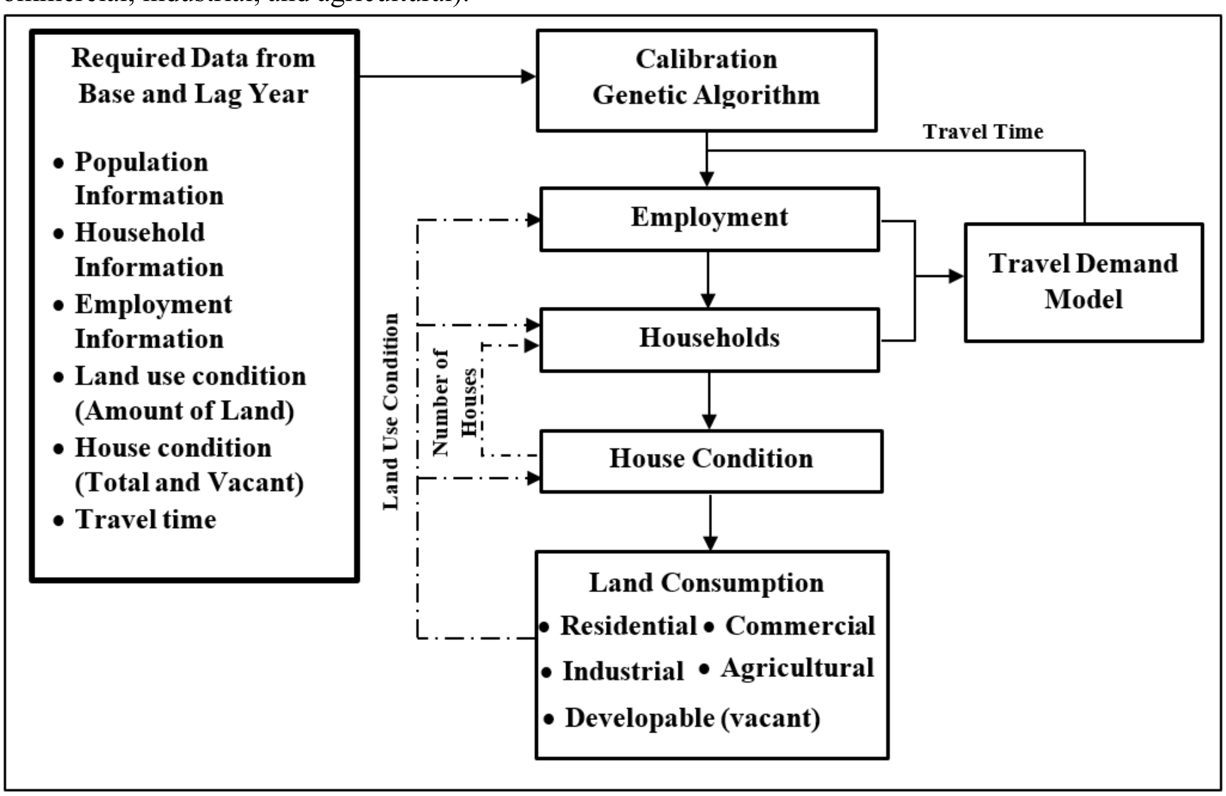

Figure 1 Integrated land use transport model's flowchart (dashed lines represent one period ($t-1$) lagged feedback of information; each period is 5 years).

Travel Demand Model

The Travel Demand Model (TDM), which integrated with the land use model and the travel time derived form, is the Tennessee Statewide Travel Model (TSTM) version 3 (*24*). This version of TSTM is a traditional four-step, TDM consisting of three different components, short distance passenger model (trips less than 50 miles), long-distance passenger model, and freight model. The underlying geographic area of operation is at the TAZ level. The total number of TAZs in TSTM is 3,687. Zonal attributes include the number of households, categorized by income, size, worker, presence of student, presence of seniors, and the number of vehicles; and the number of employments categorized by 20 sectors of NAICS codes. The TSTM3 can be understood at a high level as comprised of input network and socioeconomic data together with some component demand models and a highway assignment model. The demand components can be gathered in three broad groups related to short-distance passenger demand, long-distance passenger demand, and freight and truck demand. The TSTM3 uses TransCAD's implementation of the tri-conjugate Frank-Wolfe algorithm for multi-class user equilibrium traffic assignment (*25*). The accessibility matrices which serve as input for the land use model are obtained from TSTM's assigned networks using shortest path method.

Calibration

The parameters of four models (*HH-AL*, *EMP-AL*, *HC*, and *LC*) need to be estimated through a calibration process. The calibration of the proposed models is categorized into two sections. The first section is dedicated to the estimation of the parameters of *HC* and *LC*. These two models are Multiple Linear Regression and the intercept and coefficients are estimated using least square method. The objective of the second section is to estimate the parameters of *HH-AL* and *EMP-AL*. The calibration is conducted through maximum likelihood approach where the two following objective functions are defined. First, for *HH-AL* the objective function is as below:

$$Z_1 = Min \ \sum_i \sum_n \frac{\left(N_{i,t_{Obs}}^n - N_{i,t_{Est}}^n\right)^2}{\left(\sigma_{P_{i,t_{Obs}}^n}\right)^2} \tag{14}$$

Where, $N_{i,t_{Est}}^n$ is defined in equation (1), (2), and is illustrated here for convenience.

$$N_{i,t_{Est}}^n = \eta^n \sum_j a_n \, E_{j,t}^T \frac{W_{i,t-1}^n \, C_{i,j,t-1}^{\alpha^n} \ \exp\left(\beta^n C_{i,j,t-1}\right)}{\sum_i W_{i,t-1}^n \, C_{i,j,t-1}^{\alpha^n} \ \exp\left(\beta^n C_{i,j,t-1}\right)} + (1-\eta^n) N_{i,t-1}^n \tag{1}$$

$$W_{i,t-1}^n = (H_{i,t-1}^T)^{o^n} (H_{i,t-1}^V)^{p^n} \left(1 + \frac{L_{i,t-1}^{Res}}{L_{i,t-1}^T}\right)^{q^n} \tag{2}$$

In the objective function Z_1, $N_{i,t_{Obs}}^n$ and $N_{i,t_{Est}}^n$ are respectively, the number of observed and estimated households in category n and zone i and $\sigma_{P_{i,t_{Obs}}^n}$ is the standard deviation of observations. Where, the decision variables are η, α, β, o, p, and q (the calibration parameter mentioned in **Equations 1** and **2**). Moreover, a similar objective function is defined for *EMP-AL* as follows:

$$Z_2 = Min \ \sum_j \sum_k \frac{\left(E_{j,t_{Obs}}^k - E_{j,t_{Est}}^k\right)^2}{\left(\sigma_{E_{j,t_{Obs}}^k}\right)^2} \tag{15}$$

Where, $E_{j,t_{Est}}^k$ is defined in equation (1), (2), and is illustrated here for convenience.

$$E_{j,t}^{k} = \lambda^{k} \sum_{i} N_{i,t-1}^{T} \frac{M_{j,t-1}^{k}\, C_{i,j,t-1}^{\alpha^{k}}\ \exp\left(\beta^{k} C_{i,j,t-1}\right)}{\sum_{i} M_{j,t-1}^{n}\, C_{i,j,t-1}^{\alpha^{k}}\ \exp\left(\beta^{k} C_{i,j,t-1}\right)} + \left(1 - \lambda^{n}\right) E_{j,t-1}^{k} \tag{3}$$

$$M_{j,t-1}^{k} = \left(E_{j,t-1}^{k}\right)^{g^{k}} \left(L_{j,t-1}^{Com} + L_{j,t-1}^{Ind} + L_{j,t-1}^{Agr}\right)^{h^{k}} \tag{4}$$

Similarly, in objective function Z_2, $E_{j,t_{Obs}}^{k}$ and $E_{j,t_{Est}}^{k}$are, the number of observed and estimated employments in category k and zone j and $\sigma_{E_{j,t_{Obs}}^{k}}$ is the standard deviation of observations respectively. Where, the decision variables are α, β, g, and h, defined in **Equations 3** and **4**.

Both objective functions Z_1 and Z_2 are non-linear, non-convex, and are not subjected to any constraints. In the previous land use models (TELUM and G-LUM), a gradient search method and the Nelder-Mead method with 12 different initial points applied. Previous approaches add strict limitations to the solution approach. First, due to the non-convexity of objective functions, using gradient search method would increase the chance of trapping in local optimum solution. Second, accuracy and final solution of the Nelder-Mead method with initial points is highly sensitive to selection of initial points (*23*). Therefore, in order to eliminate these limitations, in this paper, an evolutionary algorithm is applied to solve the above-mentioned optimization problem. The following section discusses the proposed solution approach.

Solution Approach

In this paper, a Modified Genetic Algorithm (GA) is applied to solve these optimization problems. GA applies to solve non-convex and non-linear optimization problems because of its superiority in evolutionary search computation over other search techniques which are limited by the continuity, differentiability, and unimodality of the evaluated functions (*26*). GA operates by maintaining and modifying the characteristics of a set of trial solutions (population) over iterations (generations). Each of the GA steps is further illustrated below:

- **Encoding:** the initial step in operating the genetic algorithms is forming an initial population (initial trial solution set). Each individual solution in the population is represented by a binary string which is called chromosome. Each chromosome contains model parameters that are encoded in the form of binary codes (called genes). In the initial set, the values of the model parameters are randomly assigned. In this research, the model generates 1,000 chromosomes in the initial population.
- **Reproduction:** the initial population will not provide an optimal solution. A genetic algorithm works by trying to reproduce other chromosomes that are better solutions. The reproduction process is simply a selection process where those chromosomes that have a better objective function will have a higher chance of reproduction. Through this process, the overall quality of the population will gradually be improved.
- **Crossover:** in the crossover, genetic materials (genes) between chromosomes are exchanged to generate a new chromosome. Various crossover methods, such as single-point crossover, multiple-point crossover, and uniform crossover, may be used.
- **Mutation:** in order to avoid becoming trapped in a local optimal solution, a mutation process is used. In this process, some genes in the chromosomes are selected randomly and their values changed. Mutation is generally damaging rather than beneficial to the optimization process. However, it reduces the probability of trapping in a local optimum point.
- **Evaluation:** the purpose of this step is to evaluate the goodness of each chromosome. The evaluation is done by finding the objective function value (in this paper, the value of Z_1 Z_2. The less Z1 and Z2 the better the chromosome is.
- **Stopping criterion:** There are several strategies for stopping the evolution process of GA. Usually, two procedures are adopted as convergence criterion: (1) the iteration stops when the variation in the fitness level among generations is within a user-defined range; and (2) when the number of generations

reaches to a predetermined level. Both approaches are applied in this research; while the procedure stop when each stopping criterion pleased first. In this research the predefined range for variation in the fitness level is set 0.001 and the number of generations is set as 100*the number of decision variables (Z_1 has 6 and Z_2 has 5 decision variables).

In comparison with TELUM and G-LUM, applying the Genetic Algorithms would increase the flexibility of the model, by eliminating the need of testing different initial points. In addition, adding more variables in the model would increase the chance of trapping in a local optimum solution and makes finding the optimum solution harder. In problems with the high local optimum solutions, Genetic Algorithms reduce the chance of trapping in local optimum solution (*27*).

DATA REQUIREMENT AND DATA PREPARATION

The proposed land use model needs six sets of input data. Households, employment, house conditions (total and vacant houses), amount of land in five land use classes, and travel time. These data sets are needed for two periods of time with a time interval of five years. The household data, along with categories (total population, total households, household income, household size, household worker, household seniors, household students, quarter group) collected from census data. This data set is available for every 10 years. The employments data containing 20 categories of NAICS codes are available through Longitudinal Employment and Household Dynamics (LEHD). The house condition (the number of total and vacant houses) is collected through census data. The land use condition in five land use classes collected from parcel data (*28*); Each parcel has a year of the built attribute allowing extending data for previous years. By using this information, land use conditions generated from the year 2000 to 2020 every five years. Lastly, travel time data is obtained from TSTM.

RESULTS

This section discusses the result of model validation and accuracy. In order to illustrate the model applicability and validity, LS-LUM and LS-LUMS-WHC are implemented in the state of Tennessee, United States. The state of Tennessee has 95 counties and 3,293 TAZs. Due to data collection limitation (especially parcel data, even though available but need to be requested), in this paper, the model is applied in 39 counties (see **Figure 2**). The selected study area has 1,451 TAZs with a population of 2,881,195 and total employment of 1,755,491 in 2010. To test the model performance, households are modeled in 9 categories (total population, total households, households with 1 to 6 persons, and households with 7 or more persons) and employments modeled by 21 categories (total employment and 20 NAICS employment categories). The three-step validation process is discussed below.

First, models are developed for the base year (2010), then backcasting and forecasting accuracies are presented. At each step, the goodness-of-fit measure $R^2$and the error, Mean Absolute Percentage Error (*MAPE*) are provided. The proposed model results are compared with G-LUM. Generally, a model with higher R^2 and smaller *MAPE* is a better model. In addition, based on Chin (*29*) study which proposed a rule of thumb for acceptable R^2, where R^2 greater than 0.66 is substantial, between 0.33 and 0.66 is moderate, and less than 0.33 is week, in this paper, R^2 greater than 0.66 is considered as acceptable. Two points should be mentioned. First, since G-LUM does not model total houses, agricultural, and vacant land, the R^2 and *MAPE* for these variables are provided only for LS-LUM. Second, because the only difference between LS-LUM and LS-LUM-WHOHC is in modeling households, the results for LS-LUM-WHOHC are provided just for households.

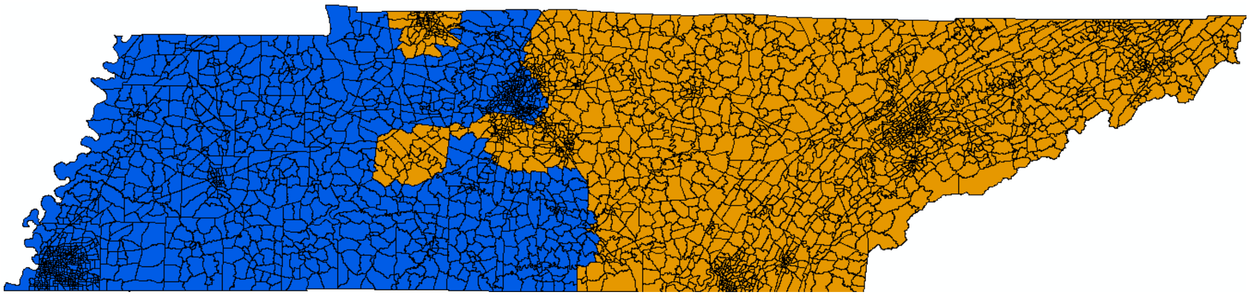

Figure 2 The State of Tennessee with 95 Counties and 3293 TAZs; The Model Implemented on the Blue Part (39 Counties with 1451 TAZs)

Developing the Model for The Base Year 2010

In the first step, the model developed for the year 2010. and 2005 is considered as the lag year. As **Figures 3** to **8** present, both LS-LUM and G-LUM are fitted very well in all the categories, except for employment NAICS sector 51. The goodness of fit in the households and land use conditions (**Figures 3** and **5**) is better in LS-LUM in comparison with G-LUM. The differences in the land use section are more significant. In addition, **Figure 3** shows that removing the *HC's* variables from the model is reduced the R^2 of households, specifically in households with 7 or more person group. Also, the *MAPE* is increased significantly in comparison with both LS-LUM and G-LUM.

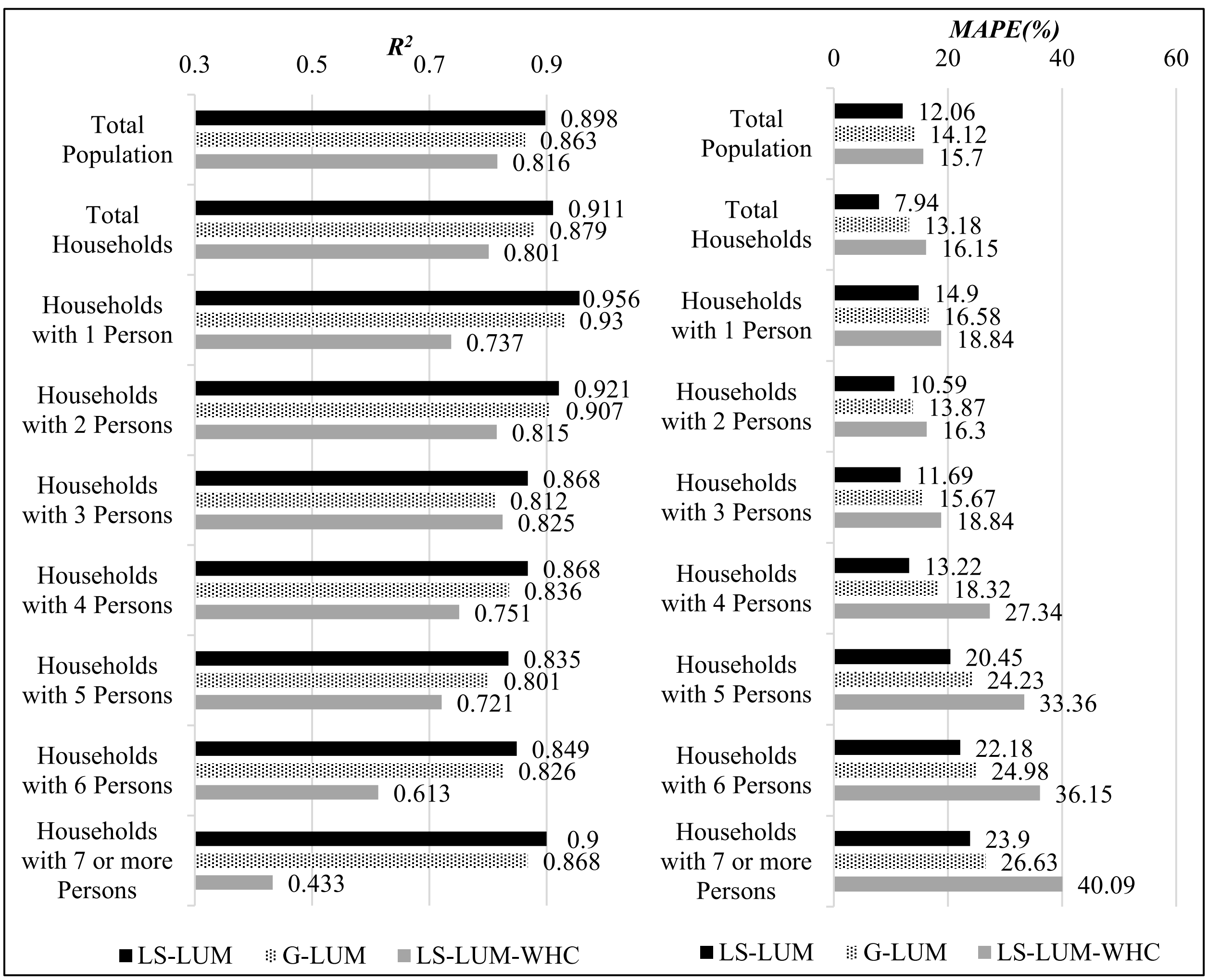

Figure 3 The R^2 and *MAPE* of three models for households for the base year 2010

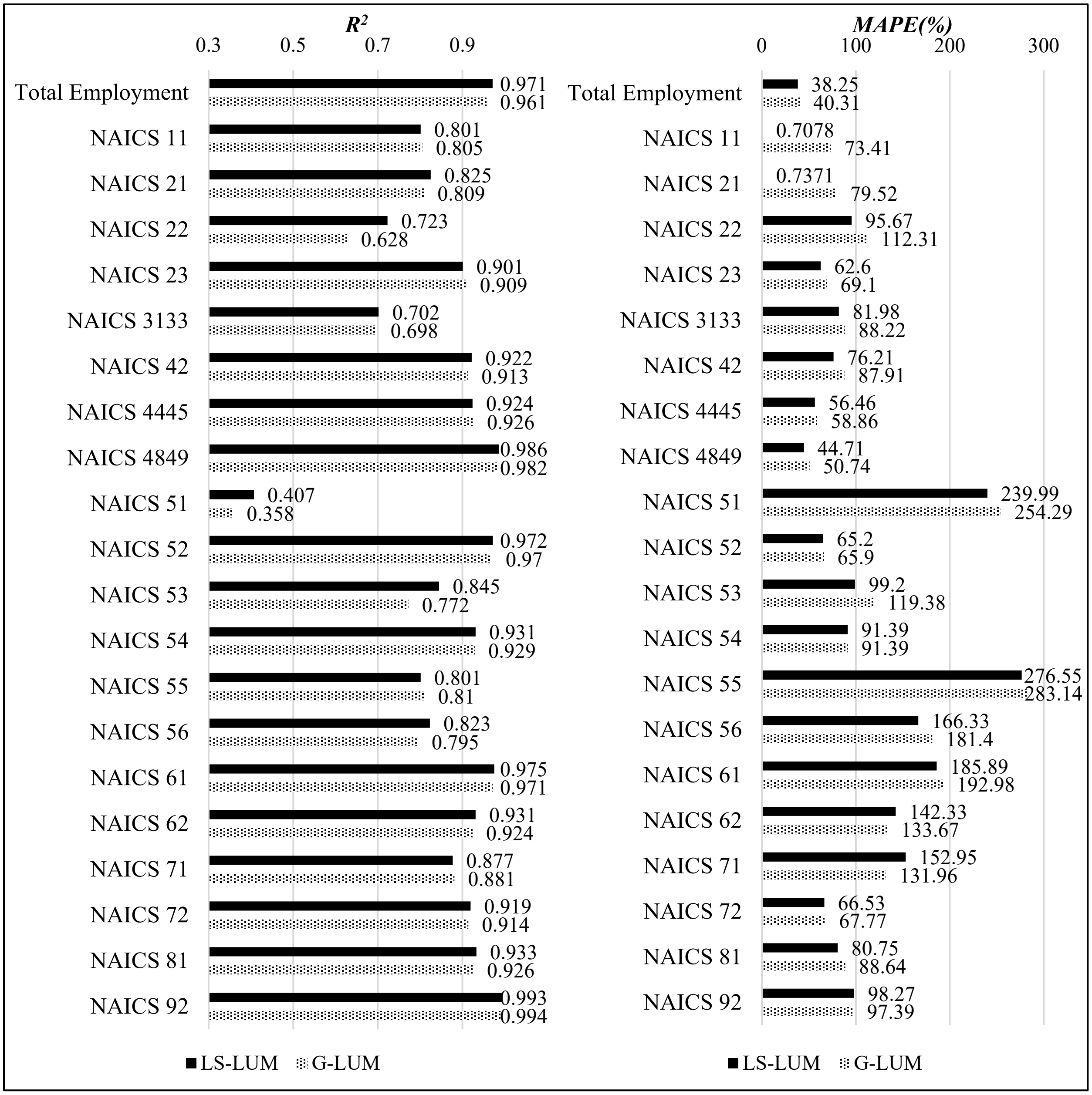

Figure 4 The R^2 and *MAPE* of employments for the base year 2010

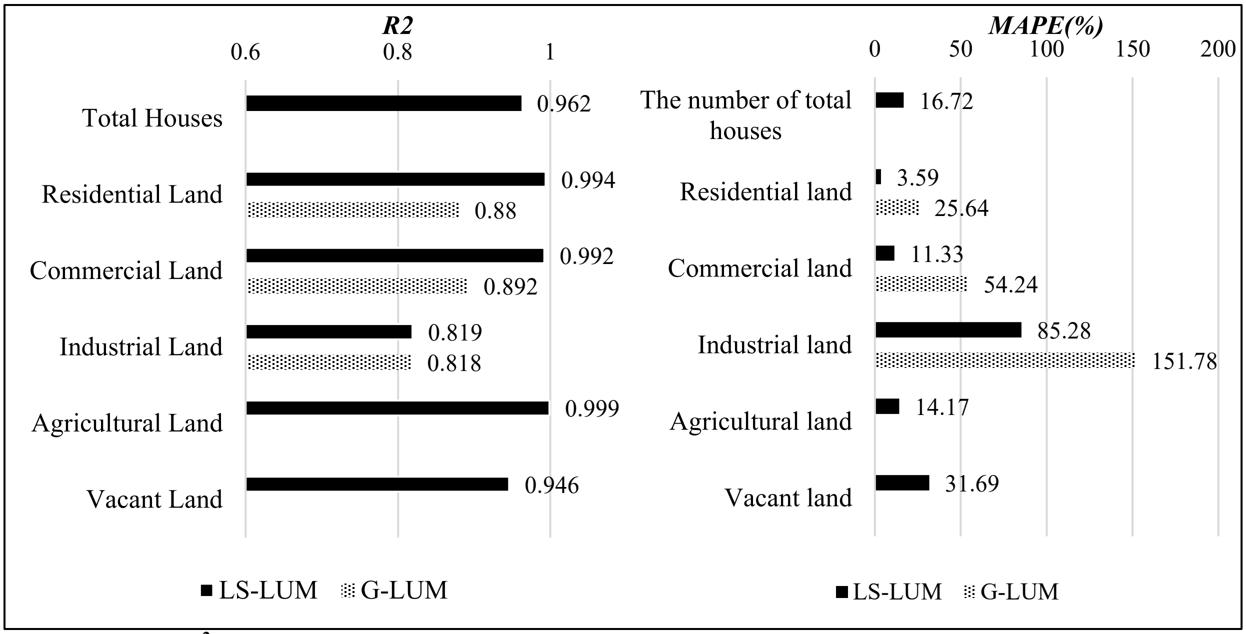

Figure 5 The R^2 *and MAPE* of total houses and land use conditions for the base year 2010

Model Backcasting Validation

After developing the model for the base year 2010, the model backcast the households, employment, total houses, and land use condition in the year 2005. **Figures 6** to **9** show backcasting validation results illustrating LS-LUM could predict all categories with acceptable accuracy. However, the performance of LS-LUM in household and land use condition prediction is better than the employment section. The R^2 in all categories is greater than 0.85. In addition, the difference between LS-LUM and G-LUM is more significant in these sections. In the employment section, both models show similar goodness of fit (**Figure 7**). However, the value of *MAPE* in LS-LUM is lower in all employments' categories in comparison with G-LUM (**Figure 7**).

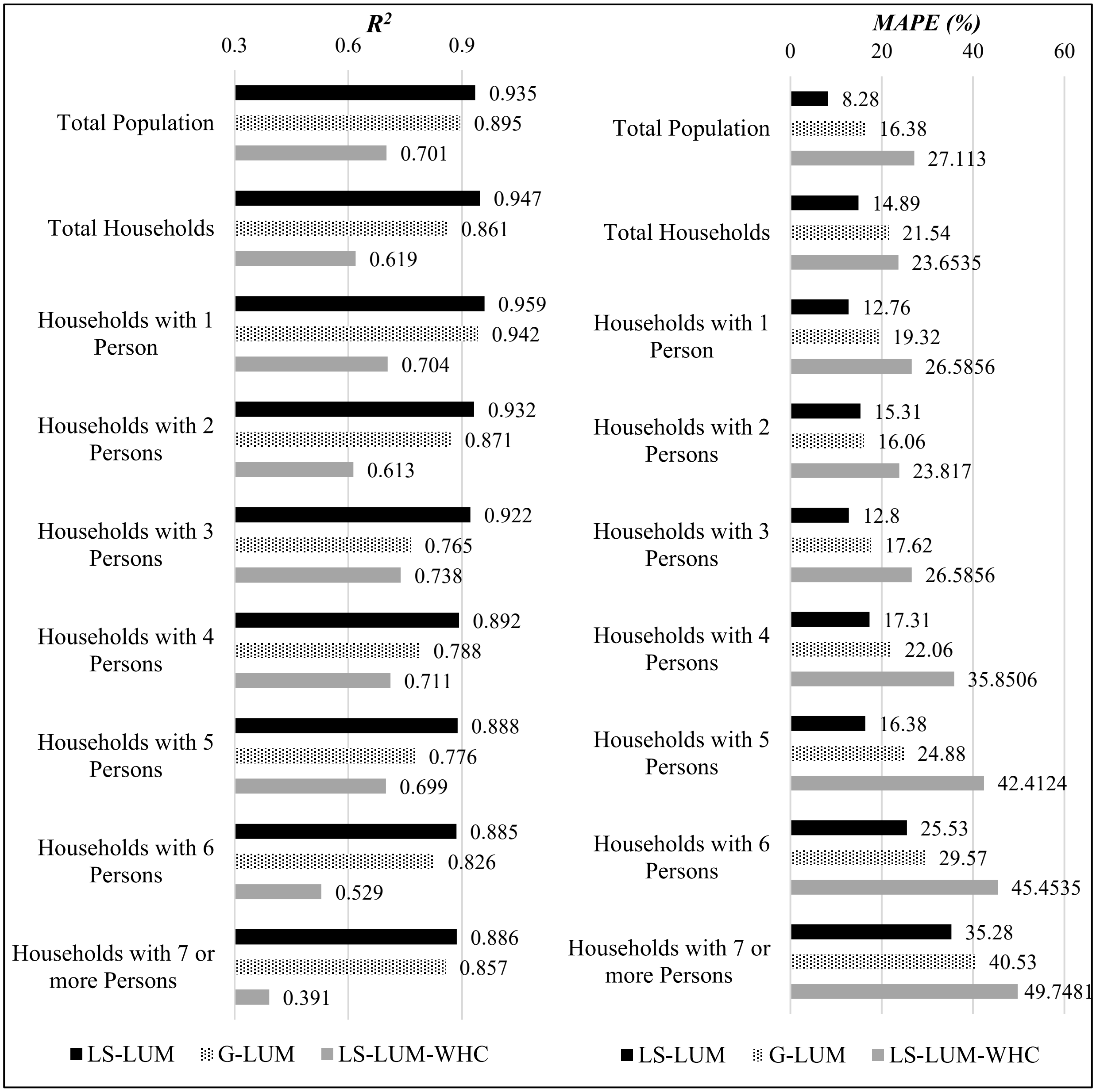

Figure 6 The R^2 of backcasting the households for the year 2005

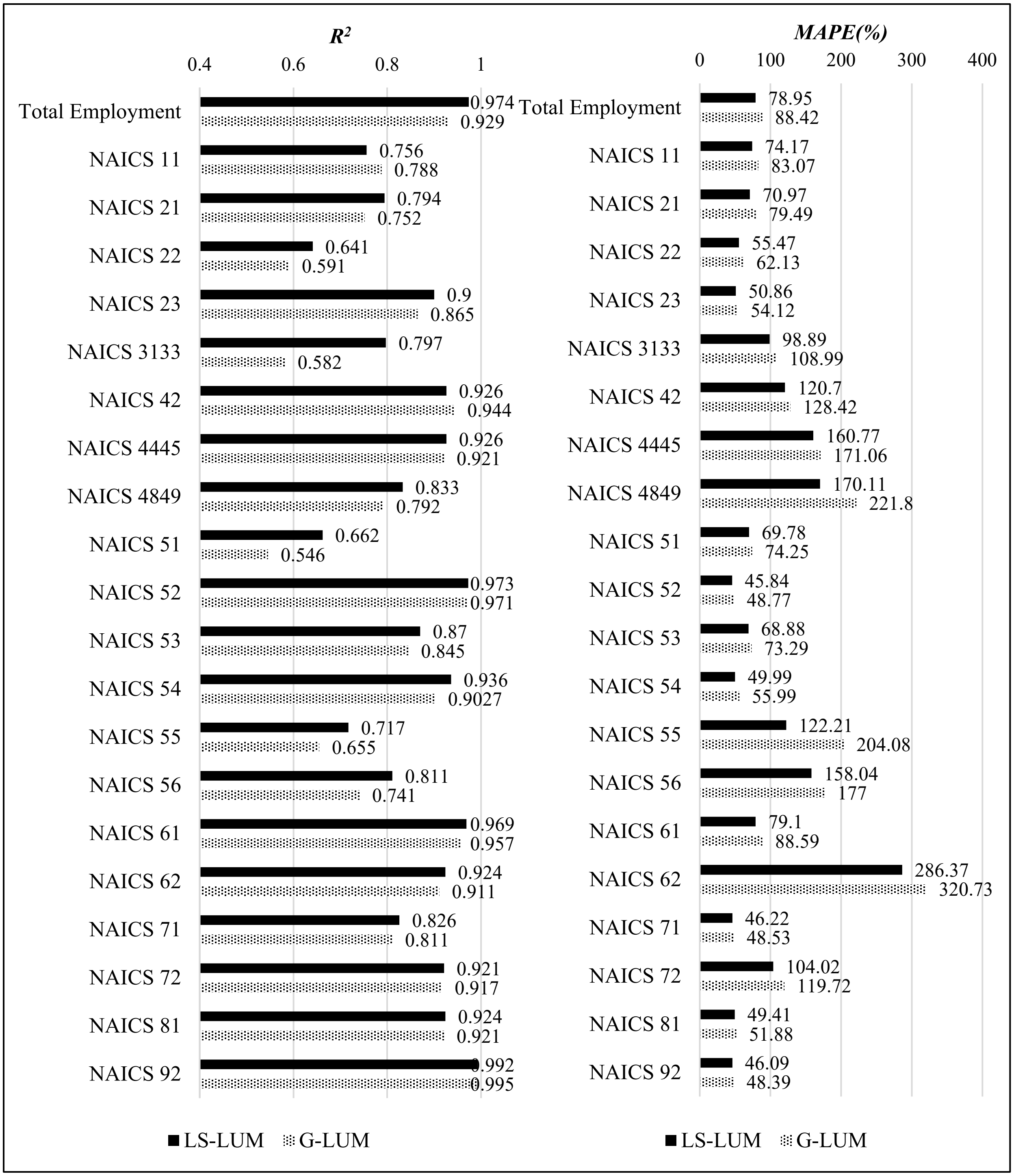

Figure 7 The R^2 and *MAPE* of backcasting employments for the year 2005

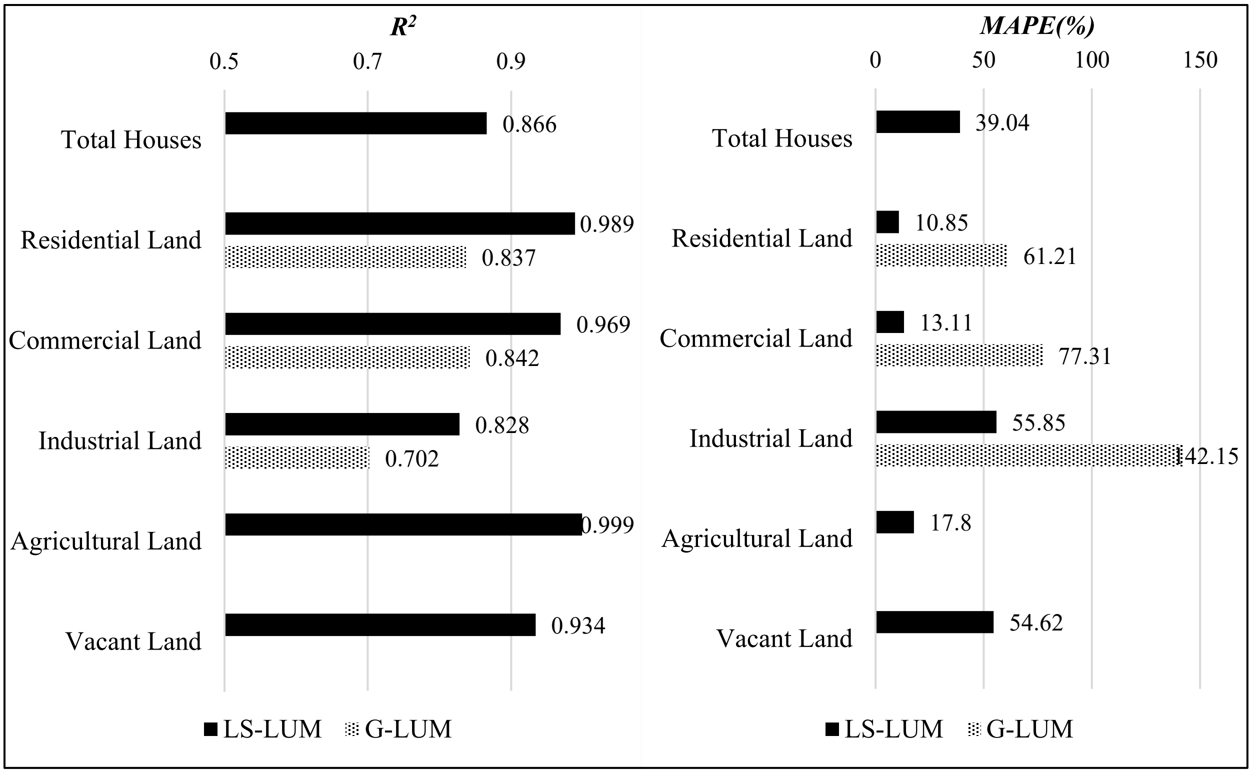

Figure 8 The R^2 and *MAPE* of backcasting land use conditions for the year 2005

Model Forecasting Validation

Forecasting validation had limits. Since the latest available household data was for the year 2010, the model forecasting accuracy is provided for employments and the land use condition. **Figure 10** shows the R^2 and *MAPE* of the estimated employment in year 2015. Except NAICS sector 53 and 48-49, LS-LUM forecast the employment better compared to G-LUM. Both models have deficits in forecasting NACIS 54 and 71. However, the improvement of LS-LUM in NAICS 11, 21,51, and 56 is significant.

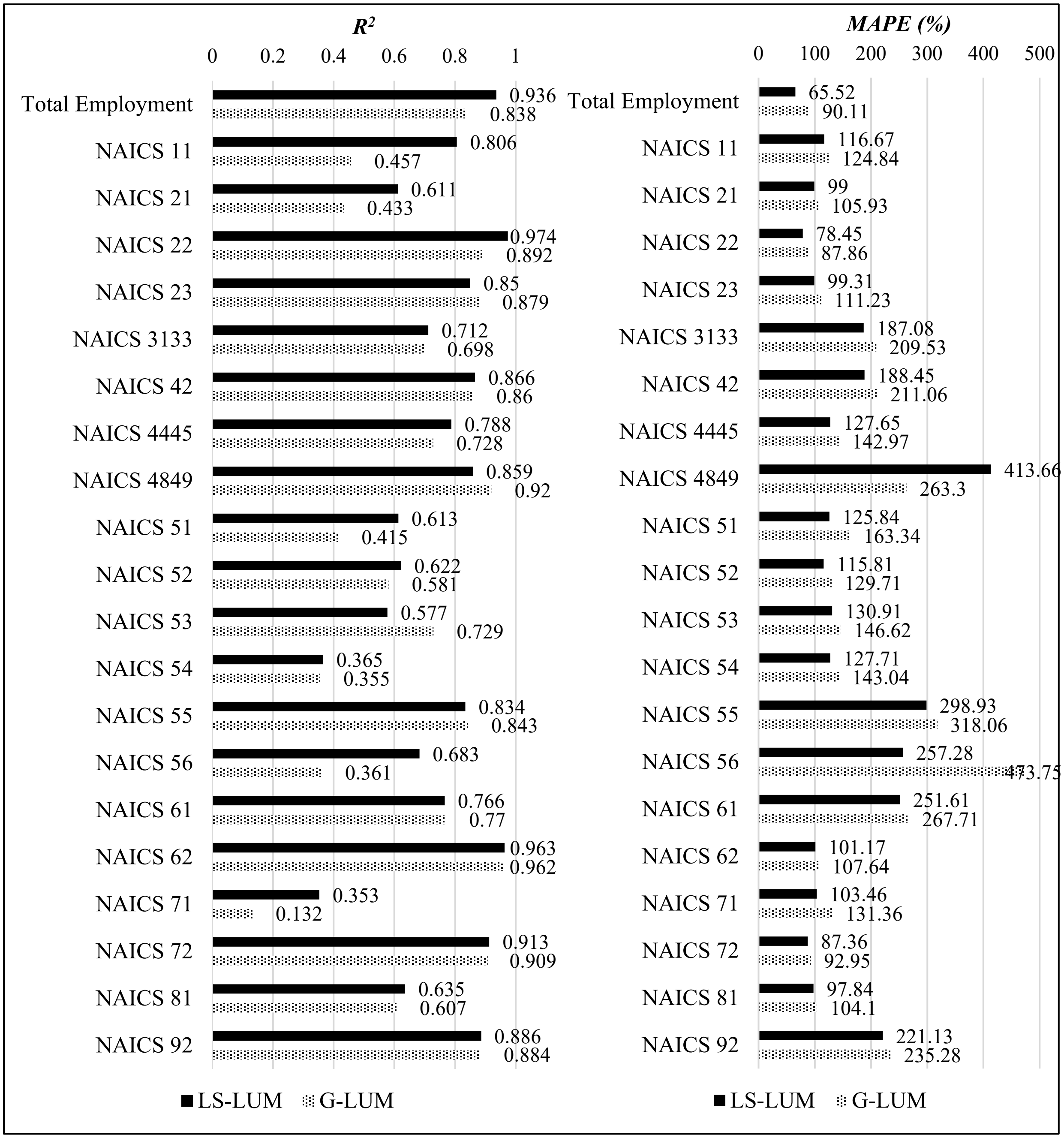

Figure 10 The R^2 and *MAPE* of forecasting employments for the year 2015

Figure 11 presents the models' accuracy in forecasting the land use condition in year 2015. In addition, **Figure 12** shows the accuracy of forecasting for the year 2020. Since the parcel data was available for the year 2020, it was possible to calculate the goodness of fit of the land use condition in 2020. Generally, similar to backcasting the accuracy of LS-LUM in predicting land use condition is much better than G-LIM. Since the components of *LC* directly affect the prediction of households and employments and due to the cumulative nature of errors, the accuracy of prediction in this section becomes more important. Moreover, as **Figure 12** shows, by increasing the year of prediction, the difference between the goodness of fit in LS-LUM and G-LUM increases. These results show that LS-LUM would work better in a long-range forecasting year.

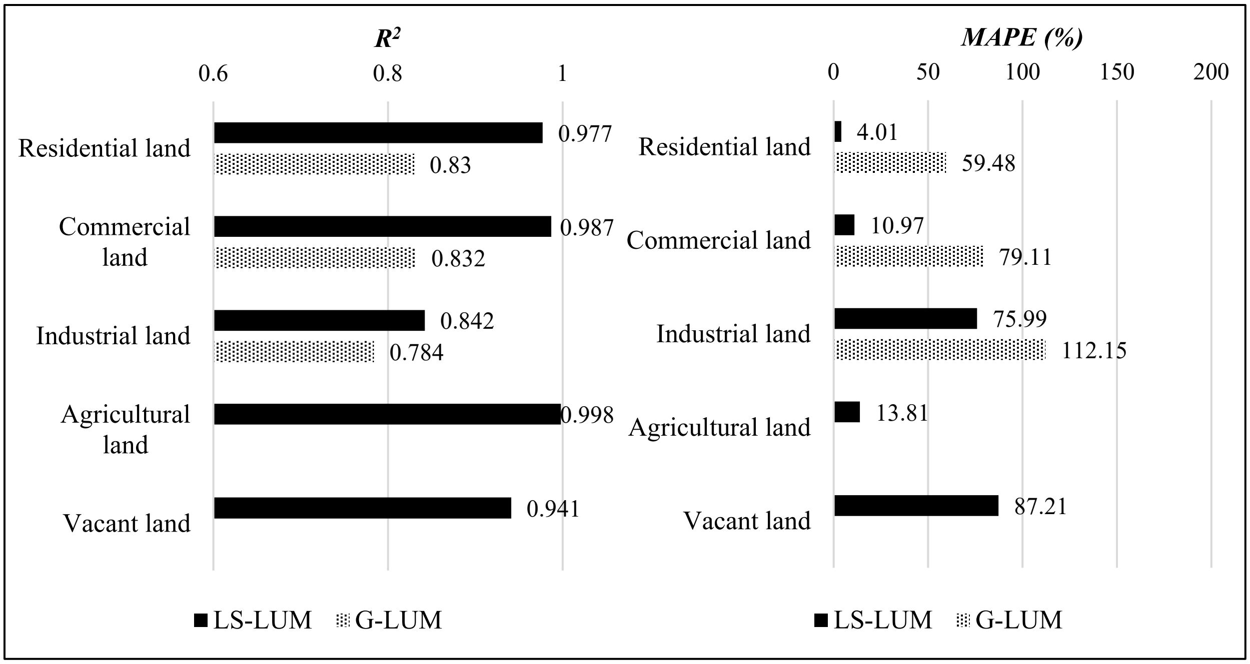

Figure 11 The R^2 and *MAPE* of forecasting land Use condition for the year 2015

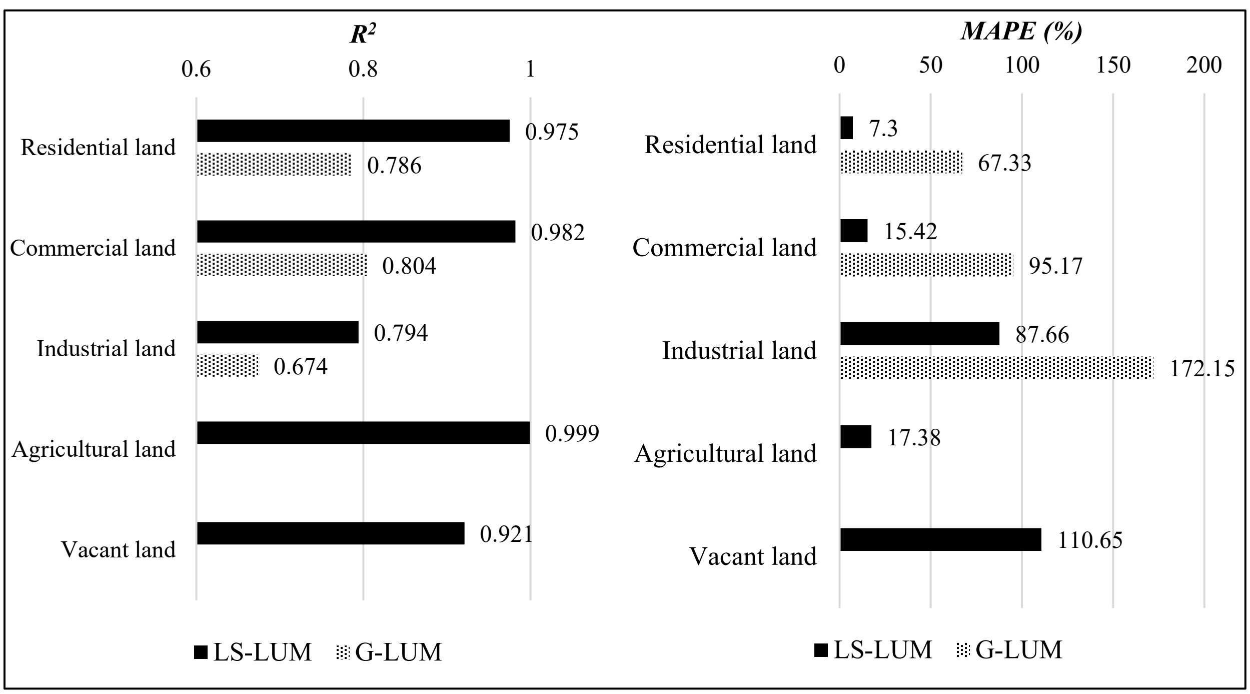

Figure 12 The R^2 and *MAPE* of forecasting land use condition for the year 2020

DISCUSSION

The overall results showed that, LS-LUM provides better accuracy compared to the similar land use model (G-LUM). Several approaches were tested to improve accuracy. The first action was involving land use conditions in the Employment section directly. In the previous models, the result of the land consumption section (LANCON in TELUM and LUDENSITY in G-LUM) does not directly affect the employment allocation section. In LS-LUM the amount of commercial, industrial, and agricultural land, were added to the *Employment* section. Although significant improvement was not observed, minor improvement in some categories is not deniable. Comparing the results of LS-LUM and LS-LUM-WOHC showed that when House Condition section (*HC*) is added to the model, the model accuracy and stability increased significantly. LS-LUM showed higher R^2 lower and *MAPE* in comparison to LS-LUM-WOHC both in developing and backcasting sections.

The new formulation for modeling land consumption shows improved accuracy. LC compared to LANCON in G-LUM, provides higher R^2 and lower MAPE in addition to predicting agricultural and vacant land. The improvements in land consumption section are important from a different point of view; due to the LC results affecting the other sections directly, the model can retain its accuracy for additional years of forecasting.

CONCLUSIONS

The purpose of this paper was to develop a land use model that can be applied to large-scale geographies with reasonable computational time and acceptable accuracy using only open source data. The proposed model, large scale land use model (LS-LUM) incorporates the underlying concepts TELUM and G-LUM with improved model formulation, and enhanced solution algorithm. The improved model formulation consists of new variables addition in the form of total and vacant houses. LS-LUM involves the amount of commercial, industrial, and agricultural land in predicting the number of employments. A new evolutionary computation-based solution approach is presented to enhance accuracy and optimality. Although the model shows acceptable results in the form of improved R^2 and lower MAPE for household and land type, additional research is needed to enhance the accuracy of the *EMP-AL* and socioeconomic conditions of large-sale cases.

Future studies can consider the effect of other components, similar to land price and salary. Conducting policy and scenario analysis is another avenue of future research. Finally, improving the calibration accuracy and runtime is another direction for future studies by exploring additional heuristic and other evolutionary algorithms.

ACKNOWLEDGMENTS

This research was funded by Tennessee Department of Transportation (TDOT). The authors would like to thank TDOT for providing data and support during the research period. In addition, computational facilities at University of Memphis is greatly acknowledged. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the above-mentioned agencies

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: AR, SM; data collection: AR, SM, JE; analysis and interpretation of results: AR, SM; draft manuscript preparation: AR, SM. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Lowry, I. S. *A Model of Metropolis*. Rand Corp Santa Monica Calif, 1964.
2. Acheampong, R. A., and E. A. Silva. Land Use–Transport Interaction Modeling: A Review of the Literature and Future Research Directions. *Journal of Transport and Land use*, Vol. 8, No. 3, 2015, pp. 11–38.
3. Salvini, P., and E. J. Miller. ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems. *Networks and spatial economics*, Vol. 5, No. 2, 2005, pp. 217–234.
4. Moeckel, R., C. Schürmann, and M. Wegener. Microsimulation of Urban Land Use. 2002.
5. Waddell, P. UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning. *Journal of the American planning association*, Vol. 68, No. 3, 2002, pp. 297–314.
6. Hunt, J. D., J. E. Abraham, and D. D. Silva. PECAS—for Spatial Economic Modeling. *Calgary, Alberta: HBA Specto Incorporated*, 2009.
7. Martinez, F. MUSSA: Land Use Model for Santiago City. *Transportation Research Record*, Vol. 1552, No. 1, 1996, pp. 126–134.
8. Waddell, P. Integrated Land Use and Transportation Planning and Modelling: Addressing Challenges in Research and Practice. *Transport Reviews*, Vol. 31, No. 2, 2011, pp. 209–229.
9. Moeckel, R., C. L. Garcia, A. T. M. Chou, and M. B. Okrah. Trends in Integrated Land-Use/Transport Modeling. *Journal of Transport and Land Use*, Vol. 11, No. 1, 2018, pp. 463–476.
10. Manual, U. TELUM (Transportation Economic and Land Use Model) Version 5.0.
11. Gao, S., E. Lehmer, Y. Wang, M. McCoy, R. A. Johnston, J. Abraham, and J. D. Hunt. Developing California Integrated Land Use/Transportation Model. 2009.
12. Abraham, J. E., and J. D. Hunt. Design and Application of the PECAS Land Use Modeling System. *University of California, Davis*, 2003.
13. Hunt, J. D. 1 Desing and Application of the PECAS Land Use Modeling System. 2003.
14. Spasovic, L. N. TELUM-Interactive Software for Integrated Land Use and Transportation Modeling. New Jersey Institute of Technology. 2013.
15. Merlin, L. A., J. Levine, and J. Grengs. Accessibility Analysis for Transportation Projects and Plans. *Transport Policy*, Vol. 69, No. C, 2018, pp. 35–48.
16. Dimitrijevic, B. A Method for Assessing Transportation Impacts of New Land Developments Using Integrated Land Use and Transportation Network Modeling.
17. Wang, C.-H., N. Chen, and S.-L. Chan. A Gravity Model Integrating High-Speed Rail and Seismic-Hazard Mitigation through Land-Use Planning: Application to California Development. *Habitat international*, Vol. 62, 2017, pp. 51–61.
18. Pozoukidou, G. Forecasting Land Use Changes: An Empirical Approach for East Thessaloniki. 2017.
19. Morton, B. J. Land Use Forecasting Models for Small Areas in North Carolina. 2013.
20. Paul, V. V., and B. Zhou. Documentation for Use of the Gravity-Based Land Use Model (G-LUM): With and Without In-Built Travel Demand Model Assumption. 2009.
21. Putman, S. H. Integrated Urban Models: Policy Analysis of Transportation and Land Use. Pion Limited. *London, UK*, 1983.
22. Valsaraj, V., K. Kockelman, J. Duthie, and B. Zhou. Forecasting Employment & Population in Texas: An Investigation on TELUM Requirements, Assumptions, and Results, Including a Study of Zone Size Effects, for Austin and Waco Regions.
23. Zhou, B., K. M. Kockelman, and J. D. Lemp. Applications of Integrated Transport and Gravity-Based Land Use Models for Policy Analysis. *Transportation research record*, Vol. 2133, No. 1, 2009, pp. 123–132.
24. RSG. *Tennessee Statewide Travel Model (Version 3) Develpment and Validation Report*. 2014.

25. Bernardin Jr, V. L., N. Ferdous, H. Sadrsadat, S. Trevino, and C.-C. Chen. Integration of National Long-Distance Passenger Travel Demand Model with Tennessee Statewide Model and Calibration to Big Data. *Transportation Research Record*, Vol. 2653, No. 1, 2017, pp. 75–81.
26. Mathew, T. V, S. Khasnabis, and S. Mishra. Optimal Resource Allocation among Transit Agencies for Fleet Management. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 6, 2010, pp. 418–432.
27. Durand, N., and J.-M. Alliot. A Combined Nelder-Mead Simplex and Genetic Algorithm. 1999.
28. Tennessee Comptroller of the Threasury. https://comptroller.tn.gov/office-functions/pa/gisredistricting/land-use-data.html.
29. Chin, W. W. The Partial Least Squares Approach to Structural Equation Modeling. *Modern methods for business research*, Vol. 295, No. 2, 1998, pp. 295–336.