Integrated Land Use and Transportation Modeling within Data-Poor Contexts

Emad B. Dawwas (*)

(*)Assistant Professor of Urban Planning, Urban Planning Engineering Department, College of Engineering, An-Najah National University, Nablus, Palestine. P O Box: 7, dawwas@najah.edu

DOI: https://doi.org/10.33976/JERT.7.2/2020/3

Abstract

Integrated Land Use and Transportation Models (ILUTMs) are revolutionary planning support tools that have been used in the developed countries since the early 1990s. ILUTMs evolved in response to the complexity of the urban planning process, which became more communicative and collaborative process involving different stakeholders with diverse and conflicting interests. The main challenge for the ILUTMs to be used in the developing countries is the cost of rich data needed for these models to give satisfactory results. This paper discusses the technical problems facing the researchers and the urban planners in adopting ILUTMs. The research proposes an alternative modeling approach that makes ILUTMs applicable in the developing countries' context. The suggested approach is centered on the idea of functioning within data-poor context instead of the costly data-rich context. The paper concludes with the expected limitations in the new modeling approach and suggests some guidelines for the researchers in order overcome these limitations.

Keywords: Land Use Planning, Modeling Land Use and Transportation, Integrated Land Use and Transportation Modeling.

I INTRODUCTION

Planning is a process that mainly aims to produce future oriented plans used as guidelines by different levels of decision-making. This process is more complicated in the urbanized areas, in which almost 50% of the world's total population and nearly three-quarters of all Westerners live (Fragkiasô and Seto, 2007). The fast urban growth and higher densities call for sophisticated planning tools to be able to deal with the interconnected environmental, socioeconomic and geopolitical issues. Due to the environmental impacts of the urban growth and the high cost of infrastructure required to accommodate this growth, the planning tools should be able to predict where and how much growth will occur, and how strong the change will be. This motivated scientists and researchers from different disciplines to cooperate in order to build what so called Integrated Land Use and Transportation Models (ILUTMs). ILUTMs represents and advanced attempt to simulate the urban dynamics represented in the interactions among the urban development, land used changes and transportation (Dawwas, 2018).

ILUTMs are computerized models consisting of large data sets, which contain information about population, employment, commercial areas, and other socioeconomic characteristics. The data sets are stored in a central data bank at a basic period called base year (Jianquan and Ian, 2002; Dawwas, 2018). Base year data sets, which are spatially distributed, are used to run submodels, included in an ILUTM, over a period of time, five or ten years. The base year is used as a temporal reference to predict and to allocate the different changes like migration rates into or out of the study area, economic changes, and built up area changes etc.

It is very common to work on land use and transportation planning within poor data context, especially in developing countries, with no foreseeable solutions to acquiring more accurate and detailed data. This study aims to propose a conceptual model for modified ILUTMs that can function under poor data conditions in terms of the quantity and the quality of the available data sets. The proposed approach consists of developing new ideas and techniques that attempt to bridge the simplicity of the aggregate modeling approaches and the advantages of disaggregate approaches. This endeavor requires defining clear borderlines, in terms of statistical uncertainty, between lowest levels of aggregate data and the highest levels of disaggregate data for various data sets used in different submodels in an ILUTM.

II LITERATURE REVIEW

2.1 Modeling under Data-Poor Context

Urban development is a complex dynamic process because it involves high number of unrepeatable events and various actors with different patterns of behavior. Considering complexity, planners need to model the future urban development patterns in advance, which requires enormous amonts of data. Therefore, modelers working in data poor situations should abandon the traditional models requiring rich data sets in favor of simpler models that can function using the available substitute

data. An excellent example, for modeling under the scarcity of data, is a study by Fragkias and Seto about modeling the urban growth in data-spares environment (Fragkiasô and Seto, 2007). Taking into account that most of expected urban growth in the next two decades will occur in the developing countries where usually the available data are sparse (Fragkiasô and Seto, 2007), the challenge in the study was to develop an urban growth model that merely used spatially explicit data. Utilizing the available binary urban/nonurban maps, which are usually generated by satellite images, the researchers, in this study, used a discrete choice framework to evaluate the probabilities of urban growth for a baseline period by employing a spatially explicit logistic regression analysis. The model could achieve relatively high accuracy (73%-77%), and the uncertainty could be captured and reduced by an explicit policy making framework, which in turn could effectively address problems relating to the predictive bias.

There is another study by Jianquan and Ian (2002), in which the researchers worked under poor data conditions. They tried to answer a fundamental question about what should be modeled in spatial patterns of urban growth by modeling the urban growth pattern at three levels. The first level was the macro level, which was defined as the probability or the possibility of land use changing from nonurban area to urban use. The second level was the meso level defined as 'the density' or the possibility of land-use change agglomerated in any pixel. The third level was the micro level defined as 'the intensity' or the possibility of high-density land-use change intensified in any pixel. The results of this study were incomplete because of two reasons. First reason was that the third level reflecting more spatial behavior was excluded from the model due to the highly detailed data required in the spatial dimension. This micro level of details required more disaggregated data at parcel, census block, or building level, and from which information like number of floors, ownership, and land value can be extracted. The second reason was that the number of the independent variables used in the macro and meso level was limited because of the data limitation. The results showed that the hierarchical system used in the study was constrained by data limitation and it could partially provide a conceptual and logical framework for the spatial analysis and spatial patterns of urban growth.

2.2 Aggregated Models vs. Disaggregated

ILUTMs are disaggregated models that are mainly based on the discrete choice models, which are in turn based on the choice behavior. Gensch and Ghose (1997) attempted to compare aggregated with disaggregated models by studying one of the discrete choice models' property namely the Independence of Irrelevant Alternative (IIA) at the two levels. According to this property, the ratio between two choices is assumed constant when more choices are introduced into the choices set where the two alternatives exist (Koppelman and Bhat, 2006).

When Gensch and Ghose tested the IIA violation at the individual level and at the aggregate level, they found that even when the IIA assumption is valid for each individual, IIA is always violated at the aggregate level. The only exception occurs when there was no heterogeneity among the individuals' choices pattern. This implicitly means that all individuals have identical choice patterns. Therefore, heterogeneity across the individuals could be the reason behind the violation of the IIA at the aggregate level rather than the violations at the individual level. These significant findings make it essential to look at the IIA property from a full choice set (the aggregated level) rather than a single pair perspective (the disaggregated level). Consequently, the authors recommended that instead of developing sophisticated and complex choice models that require enormous data at the individual level, it is possible, in some cases, to develop more aggregated choice models that do not require highly detailed data and in the same time segment the study area in order to reduce the heterogeneity. This study will build on this pivotal finding to adapt ILUTMS from rich-data modeling approach into poor-data modeling approach.

Recently, the competition between aggregated and disaggregated modeling approaches has risen in travel demand modeling. The aggregated approach is represented by the traditional "4-step" travel demand models that relies on aggregate demographic data at a traffic analysis zone (TAZ) level. The disaggregated approach, on the other hand, is represented in the activity-based microsimulation methods that employs robust behavioral theory while focusing on individuals and households. One of the few studies have compared the two approaches is the one by McWethy and Kockelman (2007) who compared the microscopic activity-based and traditional models of travel demand. Using identical sets of data, they tried to search for the tradeoffs between these two methodologies. They calibrated and then applied a based activity and traditional aggregate model on the same study area. The results of the analyses showed several differences regarding the performance and accuracy. Activity-based models required more calibration and application effort in order to ensure synthetic populations matched key criteria and that activity schedules matched surveyed behaviors. At the same time, the modeling process is accomplished while being realistic and consistent across household members. On the other hand, activity-based models were found to be more sensitive to the changes in model inputs such as the capacity expansion and employment location tests (McWethy and Kockelman, 2007). This is an additional support to the notion about the aggregate models that they ignore behavioral distinction across the population.

2.3 Missing Data and Imputation

The presence of missing values is an important issue facing modelers and planners because these missing values make data analysis and usage problematic. This problem is more challenging in poor data environment because the missing data are usually duplicate. Analyses from some of the highway agencies show that up to 50% permanent traffic counts have missing values (Zhong et. al., 2002). In this case, it will be difficult to eliminate such a significant portion of data from traffic analysis. Therefore, these missing data must be substituted through a process called *data imputation* that includes different methods with different levels of accuracy associated with these methods.

There are many studies concentrating on different methodologies for analyzing missing data, including basic concepts and applications of multiple imputation techniques and for analyzing results from multiply imputed data sets (Yang, 2002). Bradley (1994) discussed three main topics related to missing data: (1) bootstrap methods for missing data, (2) the relationship of bootstrap methods to the theory of multiple imputation, and (3) computationally efficient ways of executing them. The results showed that the simplest form of nonparametric bootstrap confidence interval turns out to give convenient and accurate answers. In addition, there were interesting practical and theoretical differences between bootstrap methods and the multiple imputation approach, as well as some useful similarities. In another study, a fully conditional-specification for multiple imputation of discrete and continuous data was used (Buuren, 2007). In this paper, two approaches for imputing multivariate data were presented and their results were compared: joint modeling (JM) which is based on parametric statistical theory and fully conditional-specification (FCS), which is a semi-parametric and flexible alternative. JM and FCS were applied to a data set containing 3801 observations with missing data. Imputations for these data sets were created under two models: a multivariate normal model with rounding and a conditionally specified discrete model. The JM approach introduced biases in the reference curves, whereas FCS did not. The paper concluded that FCS was a useful and easily applied flexible alternative to JM when no convenient and realistic joint distribution can be specified.

Regarding ILUTMs, using a proper imputation method can help maintain data integrity and improve the output accuracy of the models. This practically means improving the model capabilities to predict the future change in land use. Based on a pattern matching technique, Zhong et. al. (2006) used a new method for estimating data imputation for data from an automatic traffic recorder (ATR) in Alberta, Canada. According to their results, the new method improved the model outputs and its level of performance over the traditional models. In another study, genetically designed neural network and regression models, factor models, and autoregressive integrated moving average (ARIMA) models were developed. It was found that genetically designed regression models based on data from before and after the failure had the most accurate results. Average errors for refined models were lower than 1% and with stable patterns, and for counts with relatively unstable patterns, average errors were lower than 3% in most cases (Zhong et. al. 2004).

This study will take advantage of these results in order to propose an alternative approach to ILUTMs to effectively function within data-poor context.

III POOR-DATA CONTEXT FRAMEWORK

Modeling in data-poor context means using lower levels of detailed data, which is widely available and can be easily obtained, as input to the ILUTMs. One of the data-poor contexts is to use aggregated data (course resolution) instead of disaggregated (fine resolution). Data aggregation can be classified into two main types, spatial and temporal aggregation. The spatial aggregation means increasing the size of the smallest spatial unit in which people choices, activities, and behavior are assumed to be homogenous. For example, instead of using data at parcel level, we use neighborhood boundaries within a city. The temporal aggregation means aggregating data over longer period of time like months and could reach a year instead of days and daily time intervals. Data-poor context could also lead to dealing with inconsistent data that come from different resources, collected by different techniques, and from different periods. This usually leads to misunderstanding, and misreporting about what these data sets mean and how the data should be interpreted. Because of the data aggregation and the data inconsistency, more missing data are expected to appear, which increases the uncertainty that already exists in any model.

Generally, uncertainty is a major problem that enters into all aspects of the model development at two phases. First phase is the development of a conceptual model, which is a qualitative representation of the relationships between different parts of the urban system being modeled. Uncertainty at this phase does not increase due to the data aggregation or data inconsistency because no quantitative data is needed. Therefore, in this phase, we have almost the same level of uncertainty in both contexts: the data-rich context and in the data-poor context. The second phase is the development of a quantitative model, in which variables representing the relationships developed in the conceptual model are identified, and parameters of these variables are generated. This is a critical phase because all data are entered to the model, as well as the model outputs are obtained in this phase. The difference in the uncertainty between data-rich and data-poor contexts is expected to appear here, and it should be tested here as well.

According to the flowchart in Figure (1), the modeling process is conducted through three main steps discussed as follows:

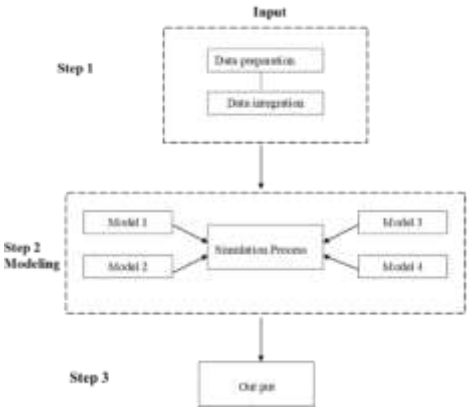

Fig. 1. Proposed Modeling Framework

Step 1) Input Data

This step is the main step in which the data will be prepared to be input in the following steps. Major changes should take place to move from the full-data modeling approach to data-poor modeling approach.

1. Data Preparation

The data sets are prepared on two stages as follows:

a. Data Aggregation

Defining the limits of aggregation and the relevant aggregate population is the first issue to resolve in making aggregate forecast. This study will exploit the available procedures of aggregating data. Here are the methods that will be used for aggregating the available data in order to be used as input to the aggregate model (Ben-Akiva and Lerman, 1985):

1. Average individual: in this method, an average individual will be constructed for the population and this average will be used as an approximation for the weight of individual;
2. Classification: the population in this method is divided into a number of nearly homogenous subgroups with different sizes, and the choice probability of the average individual within each subgroup is used;
3. Statistical differentials: this method is a "technique for approximating the expected value of a function of random variables from information about the moments of their joint distribution";
4. Explicit integration: in this method, the distribution of the attributes in the population is represented with an analytically continuous distribution. The main assumption in this method is that the population is defined in a way that all individuals in it have the same choice set. If this condition is violated then the population must be divided, and separate aggregate forecasts must be made for each subgroup;
5. Sample enumeration: this method uses a random sample of the population to represent the entire population.

b. Missing Data

When treating large amount of data from different sources with different resolutions, missing data problem is inevitable. Common methods for solving this problem usually introduce substantial bias and yield in most cases lower standard error (David, 2002). However, there are good methods that do not look so bad. Three of these methods will be used, in this study, which are the Listwise Deletion (LD), Maximum Likelihood (ML), and Multiple Imputation Method (MI). Selecting one of these methods depends mainly on the assumption of the missingness and on the model we are estimating, as shown in Figure (2). Following is a brief description of these methods (David, 2002).

Listwise Deletion

Listwise Deletion (LD) is simply accomplished by deleting any observation with missing data on any variable in the model of interest and then doing the required analysis for the complete data set. Taking into account that the missing data are MAR and the amount of missing data are tolerable, LD has two obvious advantages. The first is its ability to be used for any kind of statistical analysis, and the second is that no special computational methods are required. Furthermore, the obtained standard errors and test statistics with the LD data set are as appropriate as the full data.

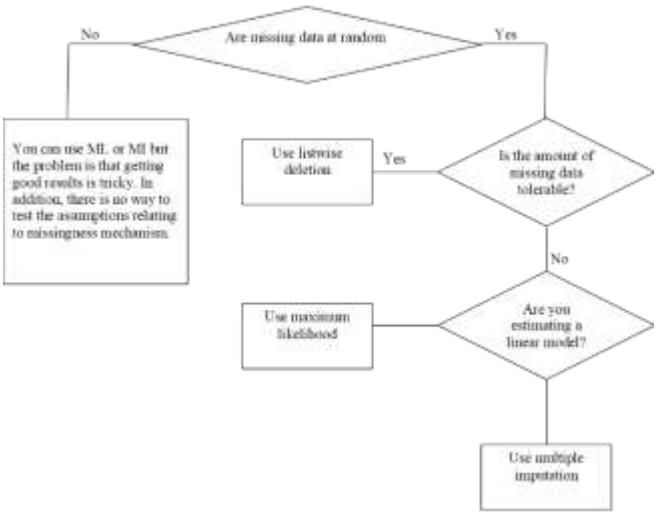

Fig. 2. Missing Data Treatment

Maximum likelihood

Maximum likelihood (ML) can be used if the missing data are MAR but the amount of missing data are intolerable, and, in the same time, if we are dealing with linear or log linear models. One of the major limitations of this method is that it is limited to linear and log linear models which will create the need to find a suitable alternative to ML.

Multiple Imputation

Multiple Imputation (MI) is an alternative approach with the same optimal properties as ML, but it can deal with any kind of data and any kind of models. Furthermore, MI usually produces consistent estimates when the data are MAR.

When the data are nor MAR, the ML and MI can be used but the problem is to obtain high quality results, because these methods are very sensitive to the assumptions relating to the missing mechanism. In addition, there is no way to test these assumptions.

2. Data Integration

ILUTMs require extensive amounts of data which makes the acquisition, maintenance, and calibration of these data the largest time consuming part of the modeling process. ILUTMs data requirements can be mainly classified into:

a. Socioeconomic data including data about the population and household characteristics;
b. Land use data including available land use and land supply, land use plans, and density of development;
c. Economic data including businesses and employment;

d. General measures of accessibility in urban areas;
e. Data about the environmental constrains.

The data integration requires dealing with all these data coming from different sources and at different levels of details ranging from parcel level to growth boundary level. After preparing the data, a database containing these data will be built. In addition, a series of interconnected data tables will be connected to the study area, which will be converted to grid cells. Therefore, each grid cell will contain these data and the related policies specifying the development rules, according to which the database will be updated overtime during the modeling process.

Step 2) Modeling (Adjusting Models to New Data and Creating Scenarios)

The main difference between the full model and the aggregate model is the input data, so existing models will be modified in order to be able to deal with the aggregated data prepared in step 1. Therefore, there will be differences in the number of variables in both models due to the differences in the input data. However, the variables used for the aggregate model will be a subset of those used in the disaggregate model.
To examine the sensitivity of each model, policies and scenarios should be applied to the models. Based on the scenarios, different results of the models can be compared and evaluated in step 3. Some changes will be required in the existing models in order to accommodate the distinct changes in the input data.

Step 3) Modeling Output

The output of modeling process will be sets of different maps and tables representing the results of different scenarios based on different sets of policies and regulations. More specifically, the results may include:

- Acreage by land use
- Housing units by housing type
- Square feet of nonresidential space by type
- Property values
- Businesses and employment by sector
- Households by type (income, age, presence of children, household size)
- Accessibility measures to employment by type and population by type

IV CONCLUSION & RECOMMENDATIONS

The output of the proposed modeling approach will certainly have lower levels of details when compared to the data-rich modeling approach. The results, however, will enable the decision makers and urban planners from predicting the future trends and patterns of the land uses and the corresponding transportation demands at higher levels of accuracy.
Urban planners who will adopt the proposed modeling approach and researchers who will do further research should keep in mind that there are some limitations in this modeling approach. Firstly, there will be uncertainty about the source of errors whether they come from the data aggregation and the input variation or from the change in the aggregation model parameters. The uncertainty in the model parameters estimates may be a significant source of uncertainty in some model outputs. Secondly, the validity of all results will be constrained by the limitations of the used model as integrated land use-transportation modeling software. Finally, the results cannot be generalized because incomplete and missing data are unlimited to one or two studies. Several studies should be conducted before we reach acceptable levels of generalizability where the results of one study can be applicable to other cases.

REFERENCES

[1] Ben-Akiva, M. and Lerman S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand (pp. 253-275). London: The MIT Press, Cambridge, Massachusetts.
[2] Bradley, E. (1994). Missing Data, Imputation, and the Bootstrap. *American Statistical Association, Journal of the American Statistical Association*, June 1994, Vol. 89, No. 426.
[3] Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, Vol. 16, No. 3, 219-242 (2007)
[4] David, P. (2002). Missing Data (pp.4-12). California: Sage Publications Inc.Dawwas B. Emad (2018). Towards a Land Use –Transportation Interactive Modeling: a Conceptual Model for Collaborative Planning. Journal of Engineering and Architecture, Vol. 6, No 1, 91-100 (2018).
[5] Fragkiasô, M. and Seto, K. (2007). Modeling urban growth in data-sparse environments: a new approach. Environment and Planning B: Planning and Design 2007, volume 34, pages 858-883
[6] Gens, Ch. and Ghose, S. (1997). Differences in Independence of Irrelevant Alternatives at Individual vs Aggregate Levels, and at Single Pair vs Full Choice Set. *Omega, Int. J. Mgmt Sci. Vol. 25, No. 2, pp. 201-214, 1997 © 1997 Elsevier Science Ltd.*
[7] Jianquan, C. and Ian, M. (2002). Modelling urban growth patterns: a multiscale perspective. Environment and Planning A 2003, volume 35, pages 679 – 704
[8] Koppelman, F. and Bhat, C. (2006). A Self Instructing Course in Mode Choice Modeling: Multinomial and Nested Logit Models. (Accessed on 02/29/2008 at www.civil.northwestern.edu/people/koppelman/PDFs/LM_Draft_060131Final-060630.pdf
[9] McWethy, L. and Kara, M. (2007). Comparing Microscopic Activity-Based and Traditional Models of Travel Demand: An Austin Area Case Study. Center for Transportation Research University of Texas at Austin.

[10] Parsons Brinckerhoff Quade & Douglas, Inc. (1998). Land Use Impacts of Transportation: A Guidebook. Transportation Research Board and National Research Council.(Accessed on 02/19/2008 at http://nepa.fhwa.dot.gov/ReNepa/ReNepa.nsf/)
[11] Springfield MPO website. (http://www.best-places.net/city/Springfield_OR-54169600000.aspx
[12] Waddell, P. (2002). UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. (Accessed on 02/15/2008 at the UrbanSim Website: http://www.urban-sim.org/Papers/)
[13] Yang, C. (2002).Multiple Imputation for Missing Data: Concepts and New Development. (Accessed on 02/29/2008 at www.sas.com/rnd/app/papers/multiple-imputation.pdf)
[14] Zhong, M., Sharma, S.C. and Lingras, P. (2006). Matching Patterns for Updating Missing Values of Traffic Counts. Journal of Transportation Planning and Technology, Vol. 29.
[15] Zhong, M. and Lingras, P. , and Sharma S. (2002). Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. (Accessed on 02/29/2008 at www.unb.ca/civil/MingZhong.htm)
[16] Zhong, M., Lingras, P. and Sharma, S.C. (2004). Estimation of Missing Traffic Counts Using Factor, Genetic, Neural, and Regression Techniques. Transportation Research Part C: Emerging Technologies , No. 12, pp. 139-166.

Emad B. Dawwas
An assistant professor in Urban Planning Engineering at An-Najah National University (ANU) – College of Engineering. I got my PhD in 2011 from University of Washington – Seattle, USA. I have been working with municipalities, village councils and Palestinian ministries since I got my master degree in 2002. I prepared many master physical plans and strategic investment plans for Palestinian communities in the last five years. My research interests are focused on land use and transportation planning and employing the new technologies in imporving the planning process. I mainly interested in developing planning support systems and urban analytical models that suits developing countries where less data and budgets are available.