International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:12, 2015

Mining Multicity Urban Data for Sustainable Population Relocation

Xu Du, Aparna S. Varde

Abstract**—**In this research, we propose to conduct diagnostic and predictive analysis about the key factors and consequences of urban population relocation. To achieve this goal, urban simulation models extract the urban development trends as land use change patterns from a variety of data sources. The results are treated as part of urban big data with other information such as population change and economic conditions. Multiple data mining methods are deployed on this data to analyze nonlinear relationships between parameters. The result determines the driving force of population relocation with respect to urban sprawl and urban sustainability and their related parameters. This work sets the stage for developing a comprehensive urban simulation model for catering to specific questions by targeted users. It contributes towards achieving sustainability as a whole.

Keywords**—**Data Mining, Environmental Modeling, Sustainability, Urban Planning.

I. Introduction

INTRUSIVE unorganized land use change that happens around the boundaries of urban areas is urban decentralization. It is also called as suburbanization or urban sprawl, which leads to the relocation of population, employment, transportation, and land use types. The process and result of urban decentralization causes many negative outcomes, for example, functional open space shortage, farmland loss and habitat fragmentation, traffic congestion and accidents, air pollution and fossil fuel consumption, incline of management costs, and lack of social capital [1]. Population dynamics play a highly significant role in urban decentralization. Low population density is a major phenomenon of urban decentralization Policy implementation to reduce urban decentralization like smart growth focuses on supporting high density communities and regulating low density communities [2]. There are various factors which would influence urban population relocation. The traditional theory believes that economic factors, social amenities, health services, traffic, employment, and other variables could drive population changes [3]. The relationship between them is non-linear and changes among different cities. Recently, the population growth in urban areas has been higher compared with rural areas. From 2000 to 2010, the urban population in the U.S. grew 12.1%, while the rural population growth rate was just 0.7% (the total population growth of the U.S. was 9% during that time). This is a significant phenomenon due to the reverse direction of the population decentralization, which is the major cause of urban sprawl [2]. Identifying the key parameters of this process would be significant for urban sustainability. The relationship between population relocation, urban land use change and other conditions would be the major focus of this research. It would provide valuable information for urban management and planning agencies to promote more compact urban areas.

Data Mining involves the discovery of novel, useful, and interesting patterns and trends from huge volumes of data. It usually involves large data sets and computing. Previous research showed that some methods of data mining can be applied to the calibration of cellular automata transition rules [4]. Data mining is a very broad field, which involves statistics, machine learning, pattern recognition, numeric search, and scientific visualization [5]. Recently, there are many applications of data mining in various research fields due to three major reasons: Firstly, the amount of available data is increasing; secondly, there are more powerful computers; thirdly, there are pertinent advances in statistical and machine learning algorithms [5]. Data mining is suitable for the nonlinear relationship analysis between urban sustainability and urban conditions [6]. To process the data mining for urban sustainability research, a proper data set must be established. The empirical urban databases provide large amounts of data. However, there are no data directly related to urban development trends, usually represented as urban land use change. To replenish this data, the research herewith integrates urban simulation models and data mining. The urban simulation models would extract urban development trends from raw data in the form of indicators, matrix, or rules.

Urban simulation models are the simplified, computed form of the real urban areas. Firstly, the goal of simulation models was to determine transportation capacity needs by predicted land use trends. Then it transferred to policy objectives like reducing the air pollution. Currently, the objectives are predicting and explaining the development trends to support the urban management and planning.

In this research, the major simulation models are the land use change models. There are various land use change models: such as the cellular automata based models, statistical analysis models, Markov chain models, artificial neural network models, economic-based models, and agent-based models. Most of these models have the ability to predict the future change of land use, it also means there can extract the development trend and utilize them (land use change matrix-Markov chain, transition rules- cellular automata,

Xu Du is with the Department of Earth and Environmental Studies and with the Environmental Management PhD Program at Montclair State University, Montclair, NJ 07043, USA (e-mail: dux3@montclair.edu).

Aparna S. Varde is with the Department of Computer Science and with the Environmental Management PhD Program at Montclair State University, Montclair, NJ 07043, USA S (corresponding author; phone 973-655-4292; fax 973-655-4164; e-mail: vardea@montclair.edu).

statistical indicators- statistical analysis models and its). These extracted trends can be utilized by the data mining to discover interesting knowledge.

Different cities show different development patterns, and there is no universal pattern of urban growth [7]. Various factors influence urban growth and there is no direct linear connection between the factors and responses [8]. Due to this, urban simulation models need adjustments for each specific city for proper results. Previous research works usually consider ambient variables as weighted indexes. The weighted value is determined by statistical methods in single area simulations. A proper weighted value brings more accurate results. However, when the location changes, the weight needs to be adjusted. Few urban simulations have involved multiple urban areas [9]. For this, illustrating the non-linear connection between urban conditions and urban development patterns is helpful to build a proper simulation model, which would provide more accurate information for urban environmental management. More importantly, the single urban simulation would not be able to provide enough data for the data mining process, both in terms of the quantity as well as the quality.

Population dynamics have a highly significant role in urban decentralization. Low population density is a major phenomenon in the urban decentralization process [2]. The population ratio between the urban areas and rural areas are an indicator of urban compactness [1]. Policy implementation to reduce the urban decentralization like smart growth focuses on supporting high density communities and regulating low density communities. There are various factors which would influence urban population relocation and existing theories state that economic factors, social amenities, health services, traffic, employment and other variables could drive population change. The relationship between them is non-linear and changes among different cities. To analyze this change, especially the population re-centralization in the United States between 2000-2010, the urban simulation models and data mining methods must be integrated to provide useful information, which enhances the urban sustainability.

II. Problem Definition

Decision makers and management agencies need the information and knowledge about urban population dynamics, urban land use change and urban development to frame proper policies, which could ensure sustainable growth patterns. Urban systems are very complicated and different cities have different development patterns (Schneider & Woodcock, 2008). Thus, the main problem of this research is to build scalable and flexible urban simulation models in multicity environments for conducting predictive and diagnostic analysis on relationships between spatio-temporal changes of population, land use, and urban development to enhance urban sustainability.

We define the following sub-problems in this work

1. Propose methods to analyze multicity big data
a. Generate complex urban data capturing the required spatio-temporal features and process this big data
b. Develop data mining approaches to reveal relationships between urban conditions and urban population relocation
2. Diagnose the key factors causing urban sprawl
a. What are the major parameters causing sprawl?
b. How do these parameters affect each other?
c. How does sprawl itself impact the parameters again?
3. Predict the indicators to enhance sustainability
a. What are the primary urban sustainability goals?
b. How do specific changes affect each other?
c. How exactly do sprawl and sustainability correlate?
4. Set the stage for a comprehensive urban simulation model to answer potential user questions such as
a. What is the relationship between parameters causing sprawl?
b. If size of city is a parameter, how does the model alter based on the size?
c. What is the quantitative relationship between individual factors affecting sustainability? (e.g., between population density and number of doctors).

III. Proposed Solution

In this research, the ultimate goal is to enhance urban sustainability in a multicity environment catering to various objectives such as minimizing sprawl, offering valuable information for urban management and planning agencies and improving the environmental management aspect of urban areas. In order to achieve this goal, this research aims to conduct data mining on complex urban data, which is suitable for analysis of nonlinear relationships between urban development activities and urban conditions. It proposes to perform urban simulation modeling, which helps us understand the urban development. The proposed approach involves urban land use change simulation modeling by data mining in a multicity environment and is depicted in Fig. 1.

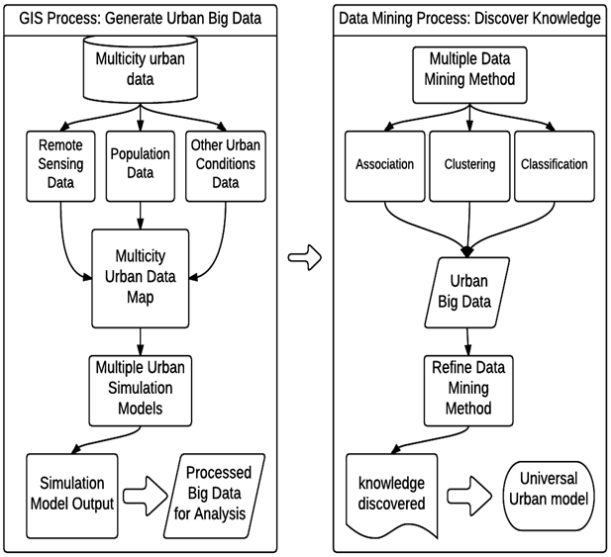

Fig. 1 Proposed Approach for Analysis

Firstly, urban data from multicity environments is gathered.

For now the most important data are the land use data, the population data and other data such as economic and policy data. All these data are preprocessed into an urban data map, which is a spatial form of data derived from multiple data sources. These provide the stationary information of different time periods. To conduct further analysis, such as population relocation and land use change, other information is required. The urban simulation model gathers this information. The detailed process is explained later. The original urban data and intermediate model output are combined as the urban big data for further data mining analysis.

This research utilizes multiple data mining methods such as association rules, clustering, and classification to analyze the nonlinear relationships in the urban big data. Association rule mining addresses issues such as the driving force and consequence of population relocation by identifying suitable antecedents and consequents through relationships of the type A implies B. Cluster analysis identifies various forms of urban population relocation by determining the groups or clusters of urban population and their respective relocation. Classification is able to predict targets such as the estimated growth of the urban population over a certain period.

Once the knowledge of the relationships between land use change, population relocation and other parameters are found through the proposed approach. It sets the stage to combine them together in a comprehensive urban simulation model which has the ability to predict and explain the population relocation with universal capability, since this model is based on the knowledge form multicity urban big data. The following sections are the detailed explanation of these steps.

IV. Urban Big Data

This research identifies the causes and consequences of population relocation. To analyze the non-linear relationship among population dynamics, land use change, and other urban conditions by the data mining method, a proper data set must be generated. The data set in this research not only has a large number of urban areas, but also a large amount of attributes and indicators. It is a multi-dimensional data set of complex urban information, which constitutes urban big data.

A. Multicity System

The urban system this research aims to analyze is a system with multiple cities around the United States [10]. Different urban areas have different urban population dynamics, urban development trends, and conditions. To achieve the goal of this research, a single urban area is not suitable, since a single urban area cannot provide enough data for the nonlinear analysis, in terms of both quality and quantity. This research intends to produce a comprehensive urban simulation model, hence the knowledge form single urban area is not sufficient due to different urban have different conditions.

Due to the data available and the time limitation, it is not feasible to analyze all the urban areas in the United States. In 2014, there was research conducted on a nationwide survey of urban sprawls, they analyzed about 200 urban areas [1]. We use some of the results from this analysis for further work.

B. Data and Databases Description

The form of the urban area influences its conditions, which are measured from the empirical databases. To analyze the relationship between them, researchers point out many indicator systems to identify and analyze the dynamic of urban development.

To analyze the urban areas population dynamic and development trend, there is a large requirement of various empirical databases. These data would be obtained from the database about population, the database about employment, the database about land use, and the database about street distribution. In the USA: The population database is U.S. Census of Population and Housing. The employment databases are the Census Transportation Planning Package (CTPP) data on employment and the Local Employment Dynamics (LED) database. The land use database is the National Land Cover Database (NLCD). The street distribution database is the national dataset of street centerlines by TomTom.

C. Urban Data Map Description

All the data collected from the previous steps are combined for data mining analysis, which is presented as a multicity urban data map.

The land use data from remote sensing is based on the emission and reflection of radiation. It has only accounted the impervious surfaces percentage and constructed materials. It would not fully represent the land use types. Census data which has the spatial distribution of population, income and employment, by interaction with these two layers of data, would generate a more reasonable urban data map.

Once the land use map is generated, the other information can also be easily associated with each of the urban areas. The land use map is part of the urban big data. With this explanation, we now describe urban development activities captured by urban simulation models.

Different urban simulation models require different forms of data, the economic-base model, and statistical analysis model may just need the general report of target areas. Some other simulation models like cellular automata and agent-based models require GIS software to provide available information to process further analysis. The data table must be presented in spatial form for analysis. Fig. 2 shows an example of transferring population density data table to spatial form. The map is from socialexplorer.com.

We prefer to use the ArcGIS platform to process the raw data from different databases. ArcGIS is the most popular GIS platform and could perform analysis of the data. The urban land use simulation models require land use map of different time periods, which can be easily achieved by the normal function of ArcGIS.

The ArcGIS platform preprocesses the raw data in this research. The raw data may not always contain all the information we need. For example, some of the raw land use maps may just contain remote sensing data with the information of impervious rate and land cover to identify the related land use type. No population data are contained in this kind of map. Thus if we need to process any urban simulation which requires

population, ArcGIS is useful to connect all the population information from other databases to the raw map. This process can generate more relevant land use maps. ArcGIS can easily achieve this goal by the joint function in the attribute table.

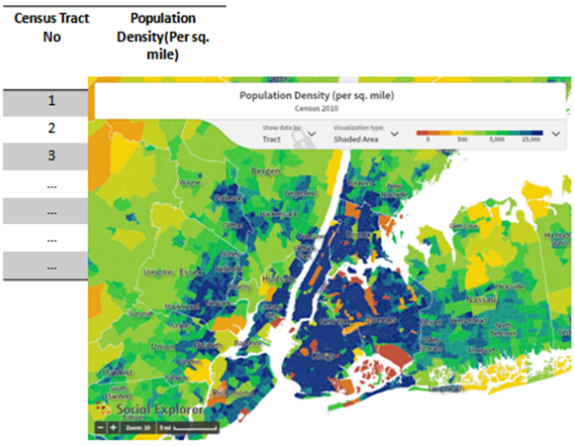

Fig. 2 From the data table to the spatial map

The ArcGIS platform not only serves the purpose of processing data for the simulation, but in addition the visualization function could also be utilized for result checking and presentation.

D. Urban Data Map to Urban Big Data

The urban data map just contains the static information of urban areas. However, the urban development trends that are usually represented as the land use changes play a significant role in the population dynamics. These can only be captured by applying urban simulation models. Different urban simulations measure different indicators and utilize them to explain current urban development and predict future trends. These indicators have different forms and would be valuable for data mining analysis since they contain information about the urban development trends. The simulation model involved with artificial neural networks is not utilized much here due to the fact that it is a black box process. On the other hand, we find that the simulation model based on empirical assessment provides valuable knowledge. Examples include the land use transition matrix in Markov-chain model and the transition rules of cellular automata model. They contain the information about land use change trends. The combination of urban data maps and development trends would be treated as urban big data. The data mining methods would be applied on these. This following example of development trends extraction is using cellular automata method. In 2004, Xia Li, Anthony Gar-On Yeh [4] applied the decision tree learning model on the calibration of historical observation data to generate transition rules. This method has its own limitation due to being highly adapted to the sample areas. On the other hand, it means that the transition rules obtained by this method are highly associated with the sample areas and contain the information about local development trends.

The basic principle is to compare the difference between two observational land use maps, and to generate the transition rules through the reverse direction of decision trees. The rules should reduce the entropy of data each time the divide the data into categories.

Based on some rule, the total data S has been divided into j categories. For all cells on the map, their state change situation and neighborhood condition becomes the data set S. They change or remain stable under certain conditions, hence each category has C1, C2….Cj, the entropy of the data set now is

$$\text{Entropy} = -\sum_{j=1}^{R} \frac{C_j}{S} * log_2 \frac{C_j}{S}$$

If the division is efficient, it will get a smaller entropy value than the previous one. The efficient decision tree learning would ensure "the gain ratio is maximized at each node of the tree" [4]. This principle would also prevent generating too many transition rules.

We now describe the analysis conducted on this data after generating the urban big data through data maps and existing models including transition rules.

V. Non-Linear Relationship Analysis

The urban big data from previous step is utilized to discover knowledge about the relationships between urban population relocation, urban land use change and other urban conditions. These parameters are considered as constraints in the urban simulation and data mining. For example, the urban simulation model could generate a data set with population data as a constraint, e.g., population above a certain number, the land use type changes etc. Furthermore, the data mining methods could discover knowledge about the relationships between urban land use change trend and population dynamics.

This analysis helps to answer questions such as: "What are the reasons for the differences between urban development trends of different cities?" We could find quantifying information, which is valuable for a comprehensive urban simulation model.

We find that association rule mining is suitable for analysis of nonlinear relationships between urban growth and urban conditions. Association rule mining is the technique of detecting rules among data sets, the rules are typically of the type: A => B where A is the antecedent and B is the consequent [11]. This means that if a trend A occurs, then B is likely to occur. These rules have interestingness measures called Confidence and Support. The Confidence C of a rule A=>B is the probability of B given A [i.e., C = P (B|A)], while the Support S is the probability of A and B occurring together in the entire data set [i.e., S = P (A ^ B)]. This is standard terminology is association rule mining.

Now consider this with reference to our work. The urban big data as a data set is:

$$D = \{I_{p,q}\}$$

Here, the variable *I* relates to a certain parameter or trend, while

p marks cities, and q marks certain attributes. Thus, in our context we define C as the confidence between parameters x and y which is given as:

$$C = \frac{\{I_{1,x}, I_{3,x}, \dots I_{n,x} \cap I_{1,y}, I_{3,y}, \dots I_{n,y}\}}{\{\{I_{1,x}, I_{3,x}, \dots I_{m,x}\} \cap D\}}$$

This is the ratio between number of urban areas in which both x and y occur and the number of urban areas in which only x occurs. When we measure the confidence as defined herewith, it could be utilized for prediction, thus the more similar the urban condition is to the association rule, the more likely it is to occur.

Cluster analysis is helpful in urban simulation as follows. Clustering is a data mining technique that divides the entire data set of different objects into groups based on their similarity [12]. The urban big data contains information that could be used as indicators of similarity. For example, a transition rule may be described as: in city A, non-urban lands have B% chance to change into the urban lands when it has C distance to the center of the cities and D% of the neighborhood is urban land. The A, B, C and D are the indicators, and by using cluster analysis, a lot of significant knowledge would be revealed. For example: in City A, X….Z, non-urban lands have the similar chance to change into urban lands under similar neighborhood conditions. Thus, when we build the comprehensive urban model, in the cities similar to City A, X…Z, this transition rule is applicable.

The classification analysis intends to produce rules which discover the relationships between urban land use change, population relocation and other indicators, and help to predict a target. For example, it could predict that under certain land use change and economic conditions, the amount of population change would be within a certain range. There are various of classification analysis methods. A common one is J4.8 decision tree learning [13]. Decision tree learning follows an inductive approach to learn from an existing data set and build a stem and leaf structure such that root represents a starting point, the intermediate nodes represent certain parameters and leaf nodes represent the final outcomes, e.g., in our case this could be urban sprawl.

With this description of the approach used in our work, we now proceed with a summary of our experimental evaluation.

VI. Experimentation

In 2014, Hamidi and Ewing conducted a research to measure the urban decentralization around the USA from 2000 to 2010 [14]. They analyzed 163 census urbanized areas. We use their results as one source of input. The data sets are publicly accessible [15]. They contain four main types of numeric indicators as follows [14].

- *Density factor:* The density factor refers to population, employment and build-up land density.
- *Mix Factor:* The mix factor or mix use factor pertains to the condition of population and employment land use type mixture.
- *Center Factor:* The center factor or centering factor determines the condition of how the population and employment concentrates in the urban center.
- *Street Factor:* The street factor is the condition of accessibility of the urban area.

In addition, there is also a *Numeric Composite Factor* which relates to the urban sprawl. It contains information about population and employment (density factor), urban land use/urban form (mix and centering factor), and accessibility (street factor). They utilize the statistical models to output indicators of population dynamics and urban development, which is suitable as a form of urban big data. Thus, we use these factors in our analysis with data mining methods.

Based on these inputs and the urban big data that we have generated, we run association rule mining, clustering and classification as described next.

The association rule mining finds several rules from which we can infer some interesting facts as follows. We find that the greater the mix factor, the lower is the tendency of the urban sprawl occurrence. This is due to the composite index being "compact". This implies that as there is a better mix between population and employment land use, the city tends to be more compact and less sprawl-prone. Likewise, we discover other interesting trends. Examples of association rules discovered from this analysis are shown in Fig. 3.

1. mix factor10=high composite index10=compact 63 ==> mix factor00=high 63 conf:(1)
2. mix factor10=high composite index10=compact 63 ==> mix factor00=high 63 conf:(1)
3. mix factor10=high composite index10=compact composite index00=compact 61 ==> mix factor00=high 61 conf:(1)
4. mix factor10=high street factor10=high 58 ==> mix factor00=high 58 conf:(1)
5. mix factor10=high 85 ==> mix factor00=high 84 conf:(0.99)
6. density factor10=high 64 ==> density factor00=high 62 conf:(0.97)
7. mix factor10=high composite index00=compact 63 ==> composite index10=compact 61 conf:(0.97)
8. mix factor10=high composite index10=compact 63 ==> composite index00=compact 61 conf:(0.97)
9. mix factor10=high mix factor00=high composite index00=compact 63 ==> composite index10=compact 61 conf:(0.97)
10. mix factor10=high composite index10=compact mix factor00=high 63 ==> composite index00=compact 61 conf:(0.97)

Fig. 3 Examples of Association Rules

Cluster analysis is then performed on the data set. Snapshots of the results appear in Figs. 4 and 5. In these figures, Cluster 0 and 3 are compact clusters, Cluster 1 and 2 are sprawl clusters, Cluster 4 is the sprawl+ cluster. Based on these figures, it is observed that urban areas with high centering factor are more compact, and those with low centering factor tend to be sprawl. Thus, urban centering, i.e., concentration of employment and population in the urban center would reduce the urban sprawl. Also, we notice that when the street factor improves, sprawl would reduce. This can be interpreted as follows. Proper design of streets would limit edge development by stimulating the growth of developed urban areas, thereby decreasing the urban sprawl.

Attribute	Full Data (162)	0 (46)	1 (35)	2 (30)	3 (34)	4 (17)
density factor10	mid	high	mid	mid	high	mid
mix factor10	high	high	mid	high	high	low
centering factor10	high	mid	high	mid	high	mid
street factor10	high	high	mid	mid	high	low
composite index10	compact	compact	sprawl	sprawl	compact	sprawl+
density factor00	high	high	mid	mid	high	mid
mix factor00	high	high	high	high	high	low
centering factor00	high	mid	high	mid	high	mid
street factor00	mid	high	low	mid	high	low
composite index00	compact	compact	sprawl	sprawl	compact	sprawl+
density factor change	reduce	reduce	reduce	reduce	stable	stable
mix factor change	reduce	reduce	reduce	reduce	reduce	reduce+
centering factor change	stable	stable	stable	reduce	increase	stable
street factor change	increase	increase	increase+	increase	increase	increase+
composite index change	reduce	reduce	increase	reduce	increase	increase

Fig. 4 Clustering Result Example

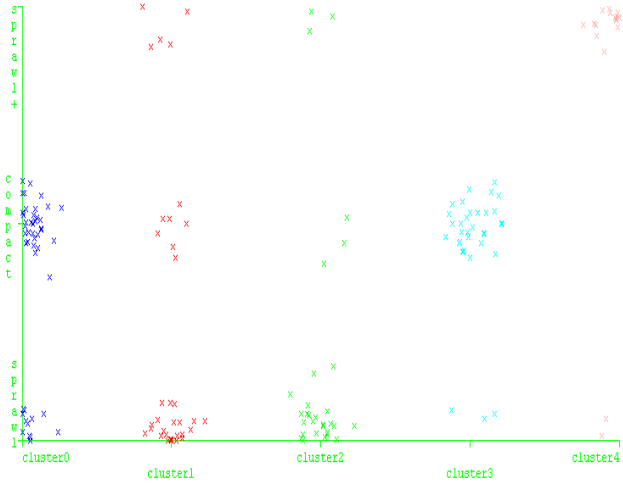

Fig. 5 Visualization of Clusters

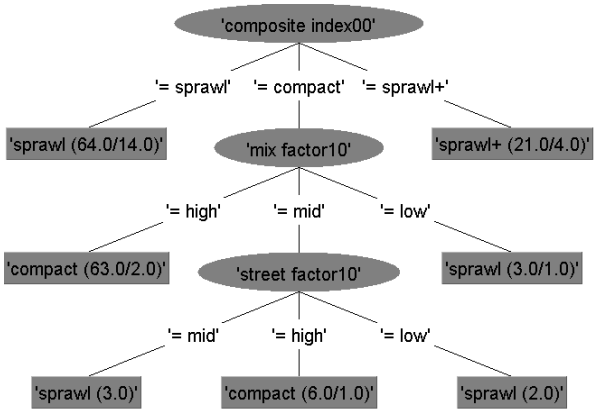

Fig. 6 Visualization of Decision Tree

Classification analysis is performed with J4.8 decision tree learning for the same data set. A partial snapshot of example results appears in Fig. 6. This example gives a result with 82.716% correctly classified instances. This result shows that most of the urban areas maintain their development conditions in these 10 years and the mix factor and street factor showed significant influence on the composite index. Low mix factor urban areas have the tendency to become sprawl. The low street factor with medium mix factor would also lead to sprawl. This result also follows the previous result as the mix factor and street factor have strong influence of the urban development. However, it does not include any influence of the changing trends of the factors. This along with other issues is being addressed in ongoing work.

VII. Related Work

There is interest in the field of urban sustainability today from several perspectives. Urban decentralization has negative outcomes, e.g., traffic congestion, air pollution, lack of social capital [1]. Low population density is feature of urban decentralization. Policy implementation to reduce decentralization, e.g., smart growth supports high density communities, regulates low density communities [2]. Theories claim that economic factors, social amenities, health services etc. drive population changes [3]. Recently, there is applied data mining research in many fields since amount of available data is increasing, there are more powerful computers and there are advances in statistics & machine learning [5]. Prior research showed that data mining can be applied to calibration of cellular automata transition rules [4]. Data mining is suitable for the nonlinear relationship analysis between urban sustainability and urban conditions [6].

We address issues that have not been predominant in earlier works, e.g., factors affecting sprawl and sustainability and the relationships between them. Also, existing research typically has single-city environments while we consider a multicity global context. Our work also entails analysis of complex urban big data with the generation of the data itself involving multiple procedures including using GIS, remote sensing, and data from existing simulation models.

Our earlier work on Mining GIS Data to Predict Urban Sprawl [16] appeared in ACM KDD 2014. We analyzed data on *urban sprawl* (overgrowth & expansion of low-density areas with issues like car dependency and segregation of residential & commercial use). Spatiotemporal features on real GIS data e.g., population growth & demographics were mined using Apriori for association rules [12] and J4.8 for decision tree classification [13], adapted to geospatial analysis, with ArcGIS for mapping. Knowledge discovered was used to build a spatial decision support system (SDSS) to predict whether "urban sprawl" was likely to occur with reasons. In our current work, we delve deeper into specific aspects of urban sprawl and sustainability and head towards generating a comprehensive urban simulation model to cater to various interesting user questions. Our proposed research activity would contribute to the state-of-the-art by discovering knowledge useful to environmental scientists, urban planners and other interested users.

Recently, there is much interest in the development of Smart Cities [17]-[19]. These entail several characteristics, among which our work would potentially make contributions to Smart Governance and Smart Environment. The Smart Governance aspect includes transparent governance and participation in

decision-making, where our work on providing useful information pertaining to sprawl and sustainability could play a role. The Smart Environment aspect includes features such as greenness and energy efficiency [20], conserving natural resources and living sustainably [21]. Thus, our work has the effect of contributing in that avenue due to the analysis of sustainability parameters and goals of sustainable population relocation as a whole. Hence, this work has a broader impact in the context of Smart Cities [17]-19].

VIII. Conclusions and Ongoing Work

In this research, we address the issue of multicity urban simulation. We propose to integrate urban simulation and data mining to conduct predictive and diagnostic analysis about the relationship between population dynamics, land use change, and urban development. Our experimentation reveals that data mining methods have the ability to discover knowledge from the national level urban data sets that contain urban development trends and urban conditions. The following are some interesting findings from this work.

- Greater the mix factor, lower the tendency of urban sprawl occurrence
- Urban areas with high centering factor are more compact and those with low centering factor tend to cause sprawl
- Mix factor and street factor combined have a significant influence on sustainable urban development
- Proper design of streets is an important indicator of sustainability

The outcomes obtained from some experiments could be even further improved by future work in this research. With respect to the techniques, we could potentially consider other methods such as: an ensemble of classifiers constituting a mixture of experts scenario for prediction in the real world; discovering associations and using them to build classifiers; clustering followed by classification and more.

With respect to the data, we could enhance the data set itself so that it includes text and image data in addition to the sources already considered. We could also mine opinions from social media data. This would be very useful given that the public satisfaction is very important in aspects such as urban development and population relocation. Public opinions are often expressed over social media and hence it would be useful to capture them in the mining process. The data on social media itself could consist of textual, numeric and image data. This would need more advanced techniques for mining. Thus, we could conduct further analysis with enhanced data sets and use that to generate a comprehensive urban simulation model. Mining over such data could potentially yield even more interesting results.

In order to address this, we need to solve various sub-tasks in this research, e.g., defining precise interestingness measures for association rules, selecting appropriate classifiers in an ensemble, pre-processing the urban big data to extract relevant information for mining, extracting and interpreting important social media data etc. All of this constitutes our ongoing research. We claim that this would discover even more interesting knowledge that would be of greater value to urban planners and other users.

Acknowledgment

This research is supported by a Doctoral Assistantship from the Environmental Management Program at Montclair State University. The authors thank the former program director Dr. Dibyendu Sarkar and the current program director Dr Stefanie Brachfeld for the research funding.

References

[1] Ewing, R., & Hamidi, S. (2014). Measuring Sprawl 2014. Retrieved from http://www.smartgrowthamerica.org/documents/measuring-sprawl-2014.pdf

[2] Smartgrowthamerica.org, 'What is "smart growth?" | Smart Growth America', 2015. (Online). Available: http://www.smartgrowthamerica.org/what-is-smart-growth.

[3] Nagy, R., & Lockaby, B. (2010). Urbanization in the Southeastern United States: Socioeconomic forces and ecological responses along an urban-rural gradient. Urban Ecosystems, 14(1), 71-86. doi:10.1007/s11252-010-0143-6

[4] Li, X., & Gar-On Yeh, A. (2004). Data mining of cellular automata's transition rules. International Journal Of Geographical Information Science, 18(8), 723-744. doi:10.1080/13658810410001705325

[5] Miller, H., & Han, J. (2001). Geographic data mining and knowledge discovery. London: Taylor & Francis.

[6] Rajasekar, U., & Weng, Q. (2009). Application of Association Rule Mining for Exploring the Relationship between Urban Land Surface Temperature and Biophysical/Social Parameters. Photogrammetric Engineering & Remote Sensing, 75(4), 385-396. doi:10.14358/pers.75.4.385

[7] Schneider, A., & Woodcock, C. (2008). Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. Urban Studies, 45(3), 659-692. doi:10.1177/0042098007087340

[8] Göktuğ, M. (2012). Urban Sprawl and Public Policy: A Complexity Theory Perspective. Emergence: Complexity & Organization, 14(4), 1-16.

[9] Santé, I., García, A., Miranda, D., & Crecente, R. (2010). Cellular automata models for the simulation of real-world urban processes: A review and analysis. Landscape And Urban Planning, 96(2), 108-122. doi:10.1016/j.landurbplan.2010.03.001

[10] Branch, G. (2015). 2010 Urban Area Facts - Geography - U.S. Census Bureau. Census.gov. Retrieved 18 February 2015, https://www.census.gov/geo/reference/ua/uafacts.html

[11] Agrawal R., Imieliński T. and Swami A., 'Mining association rules between sets of items in large databases', ACM SIGMOD Record, vol. 22, no. 2, pp. 207-216, 1993.

[12] MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281—297.

[13] Quinlan J.R., C4.5. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[14] Hamidi, S., & Ewing, R. (2014). A longitudinal study of changes in urban sprawl between 2000 and 2010 in the United States. Landscape And Urban Planning, 128, 72-82. doi:10.1016/j.landurbplan.2014.04.021

[15] Gis.cancer.gov,. (2015). County Level Urban Sprawl Indices - Geographic Information Systems & Science. Retrieved 25 April 2015, from http://gis.cancer.gov/tools/urban-sprawl/

[16] Pampoore-Thampi, A. & Varde, A. (2014). Mining GIS Data to Predict Urban Sprawl, ACM conference on Knowledge Discovery and Data Mining (KDD Bloomberg Track), New York City, NY, pp 118-125.

[17] IEEE Smart Cities, http://smartcities.ieee.org/

[18] Vienna University of Technology et al., European Smart Cities, www.smart-cities.eu

[19] Smartcitiescouncil.com, "Smart Cities Council | Definitions and overviews", 2015. (Online). Available: http://smartcitiescouncil.com/smart-cities-information-center/definitions-and-overviews.

[20] Pawlish, M., Varde, A., Robila, S. and Ranganathan, A. (2014). A Call for Energy Efficiency in Data Centers, Journal of ACM's Special Interest Group on Management of Data Record (SIGMOD Record), 2014, Vol. 43, No. 1, pp. 45-51.
[21] Varde A., and Du X., Multicity Simulation with Data Mining for Urban Sustainability, Presentation at *Bloomberg Data Science Labs*, March 2015.