

# Mining GIS Data to Predict Urban Sprawl

Anita Pampoore-Thampi<sup>1</sup>, Aparna S. Varde<sup>1,3</sup>, Danlin Yu<sup>2,3</sup>

1. Department of Computer Science

2. Department of Earth and Environmental Studies

3. Environmental Management PhD Program

Montclair State University, Montclair, NJ, USA

(pampooretha1 | vardea | yud)@montclair.edu

## ABSTRACT

This paper addresses the interesting problem of processing and analyzing data in geographic information systems (GIS) to achieve a clear perspective on urban sprawl. The term “urban sprawl” refers to overgrowth and expansion of low-density areas with issues such as car dependency and segregation between residential versus commercial use. Sprawl has impacts on the environment and public health. In our work, spatiotemporal features related to real GIS data on urban sprawl such as population growth and demographics are mined to discover knowledge for decision support. We adapt data mining algorithms, Apriori for association rule mining and J4.8 for decision tree classification to geospatial analysis, deploying the ArcGIS tool for mapping. Knowledge discovered by mining this spatiotemporal data is used to implement a prototype spatial decision support system (SDSS). This SDSS predicts whether “urban sprawl” is likely to occur. Further, it estimates the values of pertinent variables to understand how the variables impact each other. The SDSS can help decision-makers identify problems and create solutions for avoiding future sprawl occurrence and conducting urban planning where sprawl already occurs, thus aiding sustainable development. This work falls in the broad realm of geospatial intelligence and sets the stage for designing a large scale SDSS to process big data in complex environments, which constitutes part of our future work.

## Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – data mining, scientific databases, spatial databases and GIS; I.2.1 [Artificial Intelligence]: Applications and Expert Systems – industrial automation

## General Terms

Algorithms, Performance, Experimentation, Human Factors

## Keywords

Association Rules, Classification, Geospatial Intelligence, Urban Planning, Environment, Sustainability

## 1. INTRODUCTION

We are living in a world which grows and urbanizes at a rapid pace. If we continue this trend, our successors will be left with no resources to sustain. Other impacts of this growth that occur when mankind spreads its wings to the outskirts of the urban background are: overcrowding, pollution, unemployment, crime, poverty, disease etc. Today, expansion implies that cities are growing to nearby towns and villages by converting those natural lands to impervious lands by constructing buildings, parking lots, highways etc. Thus there should be some method to curb this

urban overgrowth and impose some limitations for urbanization in each area. This can only be achieved by taking appropriate decisions. Urban planners and engineers should take appropriate decisions to protect natural land while designing any activities for new constructions. With these concepts, we introduce a prototype spatial decision support system (SDSS) by mining geospatial data on parameters relevant to urban planning. A decision support system falls in the category of expert systems designed to support or assist the users’ decisions in specific domains, thus playing the decision-making role of an expert. SDSS refers to spatial aspects of the data, such as location-specific parameters, hence the term spatial decision support system [20]. This paper aims to formulate a model which predicts the occurrence of a concept known as *urban sprawl* explained next.



Figure 1: Urban Sprawl affecting Parking Lots

Urban sprawl can be defined as a pattern of urban and metropolitan growth that reflects low density, automobile-dependent, exclusionary new development and the fringe of settled areas often surrounding a deteriorating city. Among the traits of metropolitan growth, frequently associated with sprawl are unlimited outward extension of development, low density housing and commercial development, leapfrog development, “edge cities,” and more recently “edgeless cities;” reliance on private automobiles for transportation, large fiscal disparities among municipalities, segregation of types of land use, race and class-based exclusionary housing and employment, congestion and environmental damage, and a declining sense of community among area residents [11]. The term urban sprawl generally has negative connotations due to the health, environmental and cultural issues associated with the phrase. Residents of sprawling neighborhoods tend to emit more pollution per person and suffer more traffic fatalities [2]. As a result people would start the trend

of moving to neighborhood low density areas, and to meet their requirements, more houses, parking lots (see Figure 1), roads (see Figure 2), shops etc. should be constructed which gradually leads to expansion of sprawl to those areas as well.



**Figure 2: Urban Sprawl affecting Roads**

In our work, we aim to find the inherent relations between the variables related to sprawl. Although the relations and patterns explained in this paper are specific to New York, these can be generalized to simulate other urban areas in United States. The basic structure of all cities is almost the same. They have similar infrastructure design, socio-economic conditions, transportation and facilities. It is this similarity that has encouraged us to devise a method which will help decision-makers/engineers to identify sprawl conditions in their respective areas and thus help them design their cities accordingly in order to eliminate the conditions of sprawl. In this context we use geographic information system (GIS) data for analysis on spatiotemporal parameters.

A geographic information system integrates hardware, software and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. The power of a GIS comes from its ability to relate different pieces of information in a spatial context and to reach conclusions about pertinent relationships [17]. Most of the information we have about the world contains a location reference, placing that information at some point on the globe. For example, when rainfall information is collected, it is important to know where the rainfall occurs. This is done by using a location reference system, such as longitude and latitude, and perhaps elevation. Comparing the rainfall information with other information, such as the location of marshes across the landscape, may show that certain marshes receive little rainfall [5]. This fact may indicate that these marshes are likely to dry up, and this inference can help us make the most appropriate decisions about how humans should interact with the marsh. A GIS, therefore, can reveal important new information that leads to better decision-making [14].

The focus of this paper is to understand and identify urban sprawl by mining such geospatial data coming from a GIS. The results of this mining would thus help urban planners make better decisions by providing models to predict urban sprawl that guide the planning to make the areas sustainable. In addition, we are interested in finding out how some variables pertaining to sprawl affect others, e.g., how many trucks are used for transportation within a sampled place given a certain number of housing units. To the best of our knowledge, a model to predict urban sprawl and study its parameters does not exist in the literature so far. We thus address this problem in our work.

More specifically, the problem addressed in this paper is defined with two specific goals as follows:

1. *Predict the occurrence of urban sprawl based on given parameters*
2. *Analyze the impact of the parameters on each other and on the urban sprawl*

The broader impact of this work addresses urban sustainability as explained below. Urban sustainability involves a reexamination of urban development, including environmental, social and economic policies, politics and practices, and an acknowledgement of the role of cities in global environmental change [15]. Sustainable development implies improving the quality of life of a population within the capacity of Earth's finite resources. The needs of the present generation must be met, particularly those of the poor, without compromising the ability of future generations to meet their own needs. This is a dynamic process whereby the decision-makers involved in any area plan, implement and then re-examine their ideas and policies over time. In cities the goal of sustainability has been increasingly highlighted over the past few decades as problems and issues arise from unsustainable practices and developments [18]. Thus a spatial decision support system (SDSS) that predicts the occurrence of urban sprawl by mining data on geographic information systems, and helps understand the relationships between parameters affecting urban sprawl, would have a positive impact on urban sustainability. It would help urban planners, city dwellers and other related personnel make better decisions that help to promote sustainable growth and development.

The rest of this paper is organized as follows. Section 2 describes our proposed approach to address the problem of urban sprawl. Section 3 describes the implementation and evaluation of the SDSS developed to predict urban sprawl. Section 4 outlines related work. Section 5 gives the conclusions and future work.

## 2. PROPOSED APPROACH

A spatial decision support system (SDSS) is proposed herewith as an approach to address the problem of predicting urban sprawl and analyzing the impact of its parameters. A SDSS is an intelligent, interactive computer-based system designed to assist in decision-making while solving a semi-structured spatial problem. It is designed to assist the spatial planner with guidance in making land use decisions. This entails use of spatiotemporal databases to store and process the geospatial data, a library of potential models that can be used to forecast the possible outcomes of decisions, and an interface to aid the user's interaction with the computer system and to assist in analysis of outcomes [20]. In this paper, we study the effects of the interesting variables and predict the occurrence of urban sprawl. We also study the impacts of the concerned variables on each other. In order to discover the patterns and trends among these variables, two classical data mining algorithms are deployed: Apriori for association rule mining [1] and J4.8 for decision tree classification [6]. These are adapted with specific reference to geospatial data.

The justification for decision tree classification in this work is that the primary goal of this work is the prediction of the discrete target attribute 'sprawl' based on other relevant attributes. Thus decision trees would be well-suited based on their paths that represent the concerned attributes along with their values, and the leaves that represent the decisions, i.e., predicting the occurrence of sprawl.

The justification for association rule mining is that the secondary goal here is to understand the correlation between the

attributes pertaining to sprawl. Discovering patterns in the form of association rules of the type  $A \Rightarrow B$  would be very useful here to indicate how one attribute impacts another.

We now describe the steps of our proposed approach involved in building the SDSS. There are three main steps as listed below.

- Data Preprocessing Step
- Association Rule Mining Step
- Decision Tree Classification Step.

In order to explain the concepts in these respective steps, we use a running example based on the state of New York. Real GIS data from New York counties is used in our work. Counties in NY are studied to distinguish areas affected by urban sprawl. It is to be noted that although NY is famous for its metros, it still has counties with a strong rural character. We thus use it as an exemplary data set.

## 2.1 Data Preprocessing Step

In order to discover knowledge from the GIS data, we first map it by superimposing a shape file on the urban land use file and then extract relevant information for conversion to suitable formats as appropriate for the concerned mining algorithms. For this purpose we use the ArcGIS software briefly described below.

ArcGIS is a GIS software package of Economic and Social Research Institute (ESRI), which is used for working with maps and geographic information [19]. It is used for:

- Creating and using maps
- Compiling geographic data
- Analyzing mapped information
- Sharing and discovering geographic information
- Using maps and geographic information in a range of applications
- Managing geographic information in a database.

Using ArcGIS data preprocessing is conducted as follows, referring to the example New York data set. This NY dataset consisting of 27 variables, excluding shape file, are all continuous data. A partial snapshot of this data appears in Figure 3.

	P	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD						
1	Employee	Unemployee	FarmLand	Ac	HousingUnit	Mean	time	Total	Accident	Public	Supply	No. of house	or Poverty	Births	per Death	Rat	Education	Gasoline	Truck	train	Transit	an	Education	Target
2	50.85	1.77	84.8	129972	19.7	5429	292	15780	10.6	11.2	10.1	33.3123	19124	89132	19563	313713	Y							
3	43.46	2.2	149.2	24505	21.2	988	25.24	1433	15.5	10.9	9.8	17.1389	4489	6784	4739	35914	N							
4	33.9	2.61	0	490639	41.7	13662	1332.65	48312	30.7	17.1	7.5	14.6458	14372	52333	81953	409965	Y							
5	47.27	1.8	102.4	88817	18.2	2465	168.8	7265	12.8	11	11.1	22.6593	14481	36568	8469	15021	N							
6	46.81	2.26	114	39839	21.4	958	51.39	3327	13.7	12.1	10	14.8724	6431	14941	3413	33291	N							
7	47.46	1.95	121.8	35477	20.9	984	54.1	2686	11.1	11.6	8.9	15.5373	7106	52963	4685	10570	N							
8	46.8	1.93	82.8	64900	17.6	1665	111.28	5670	13.8	11	11	16.9415	10058	34543	2541	4963	N							
9	45.02	1.98	88	37745	19.6	853	87.36	2792	13	11.9	10.5	18.5818	4073	13936	3454	15943	N							
10	45.72	1.95	131.2	23890	21.9	711	21.28	2206	10.7	11.2	10.3	14.4487	2702	10667	2873	943	N							
11	46.06	2.25	147	33091	18.8	973	48.58	9548	13.9	9.8	7.8	17.803	9380	22050	1506	3271	N							
12	48.34	1.74	110.6	30207	25.1	941	29.94	3661	9	10.5	11.8	22.5938	5802	9013	3473	5809	N							
13	47.53	2.06	135.2	20116	20.8	679	33.86	1874	15.5	11.6	8.8	18.8494	4534	5534	1755	2599	N							
14	44.32	1.87	171.2	28952	21.3	631	22.78	1702	12.9	9.6	11.8	16.6314	5166	8238	1944	1598	N							
15	48.15	1.57	56.8	100610	29.9	4524	204.55	11895	7.5	11.9	7.8	27.6286	14445	25140	17451	237458	N							

Figure 3: Partial Snapshot of NY State Urban Land Use Data

These variables consist of demographics, socio-economic conditions, infrastructure usage, and accidental reports. The dataset consists of data for the 62 counties for the years 2000 and 2010. Among these 27 variables, the last one is the target attribute which defines the presence of sprawl occurrence. This target attribute is finalized based on a project about “Measuring the health effects of sprawl” done by smart growth America [10]. The collected dataset for the years 2000 and 2010, with a county based shape file for New York, as shown in Figure 4, is plotted as two

separate maps using the Arc GIS software to show the urban sprawl in both the years. The map representation gives a clear visual depiction of the spatial data. The steps followed to plot the map are listed below with reference to the given example.

- Arc GIS 10.1 is opened to add the shapefile of NY.
- Under properties, counties are named using “Label”, selecting ‘Name’, from the attribute table.
- Using the Join option the data file is joined with the existing attribute table of New York state.
- Two separate maps displaying Urban Sprawl in year 2000 and year 2010 are prepared.

FD	Shape	STATEFP	COUNTYFP	COUNTYNS	CITYOFF	NAME	NAMESAD	LSAD	CLASSFP	INTFC	CSAPP	CSAPP	NETDVP	FUNCT	ALAND	AWATER	INTPTLAT	INTPTL
1	Polygon	36	003	0037448	16103	Suffolk	Suffolk County	06	H1	04020	400	55020	55004	A	2382010230	3704300106	+42.9435539	-472.8602
2	Polygon	36	003	0037410	16003	Albany	Albany County	06	H1	04020	400	55020	55004	A	3685964575	131537575	+42.3419710	-478.020
3	Polygon	36	009	0097428	16059	Nassau	Nassau County	06	H1	04020	400	55020	55004	A	737304529	158450375	+40.7266070	-478.590
4	Polygon	36	013	0097415	16013	Chautauque	Chautauque County	06	H1	04020	27460	55020	55004	A	2745973843	1139480992	+42.3042159	-478.4075
5	Polygon	36	011	0097414	16011	Cayuga	Cayuga County	06	H1	04020	532	12100	55020	A	179189918	145666465	+43.0085455	-478.5740
6	Polygon	36	015	0097416	16015	Chemung	Chemung County	06	H1	04020	21300	55020	55004	A	105235004	8877181	+42.1581760	-478.7450
7	Polygon	36	001	0097409	16001	Albany	Albany County	06	H1	04020	104	10500	55020	A	135462509	27164206	+42.9482712	-478.9740
8	Polygon	36	005	0097411	16005	Bronx	Bronx County	06	H6	04020	400	55020	55004	C	198028919	39830244	+40.8487110	-473.8520
9	Polygon	36	007	0097412	16007	Dutchess	Dutchess County	06	H1	04020	400	55020	55004	A	268697485	19688302	+41.7551147	-478.7390
10	Polygon	36	009	0097410	16010	Columbia	Columbia County	06	H1	04020	104	24100	55020	A	268002519	36635472	+44.7527102	-478.6250
11	Polygon	36	005	0097416	16005	Fulton	Fulton County	06	H1	04020	104	24100	55020	A	120143236	9688333	+43.1596962	-478.4230
12	Polygon	36	011	0097414	16011	Orange	Orange County	06	H1	04020	400	55020	55004	A	210037006	71717891	+41.4324696	-474.3365
13	Polygon	36	007	0097412	16007	Chemung	Chemung County	06	H1	04020	400	55020	55004	A	231427920	1161207	+42.4897320	-478.8940
14	Polygon	36	047	0097422	16047	Kings	Kings County	06	H6	04020	400	55020	55004	C	181444846	67610124	+40.6351132	-478.6840
15	Polygon	36	063	0097430	16063	Nassau	Nassau County	06	H1	04020	180	15300	55020	A	132366881	158601586	+40.5487318	-478.7920
16	Polygon	36	067	0097432	16067	Columbia	Columbia County	06	H1	04020	532	45900	55020	A	201602073	78251380	+43.0685268	-478.1960
17	Polygon	36	063	0097430	16063	Schenectady	Schenectady County	06	H1	04020	104	10500	55020	A	53841388	1285622	+43.2175420	-478.9430
18	Polygon	36	021	0097418	16021	Columbia	Columbia County	06	H1	04020	104	26400	55020	A	164087380	59121459	+42.2477238	-478.6200
19	Polygon	36	023	0097410	16023	Columbia	Columbia County	06	H1	04020	286	16800	55020	A	129170269	7144280	+42.5930237	-478.0760
20	Polygon	36	021	0097418	16021	Essex	Essex County	06	H1	04020	400	55020	55004	A	464708310	115600026	+44.1088711	-473.7770
21	Polygon	36	079	0097438	16079	Putnam	Putnam County	06	H1	04020	400	55020	55004	A	588591433	41202700	+41.4278035	-473.7430
22	Polygon	36	057	0097427	16057	Montgomery	Montgomery County	06	H1	04020	104	11200	55020	A	1044165243	1864888	+42.8916991	-474.4380
23	Polygon	36	099	0097447	16099	Saratoga	Saratoga County	06	H1	04020	464	42900	55020	A	83033338	17235889	+42.7322943	-478.8270
24	Polygon	36	099	0097433	16099	Ontario	Ontario County	06	H1	04020	464	40300	55020	A	168766583	4784138	+42.6868640	-478.3010
25	Polygon	36	049	0097423	16049	Lewis	Lewis County	06	H1	04020	400	55020	55004	A	330138884	3623099	+43.7883865	-478.4420
26	Polygon	36	113	0097454	16113	Warren	Warren County	06	H1	04020	104	24200	55020	A	2245394142	167448317	+43.5551052	-478.0300

Figure 4: Shape File for New York State

These maps prepared using ArcGIS are shown in Figures 5 and 6 respectively. In these maps, the counties shown in red indicate those with sprawl and the ones in green represent the absence of sprawl or low-intensity sprawl. By comparing these figures we can see that 5 more counties were affected with urban sprawl in 2010 than in 2000. In the map for 2000 the counties Putnam, Orange, Ditches are not in the list of sprawl-affected ones. But the expansion of the nearby highly dense counties affects these counties over those 10 years.

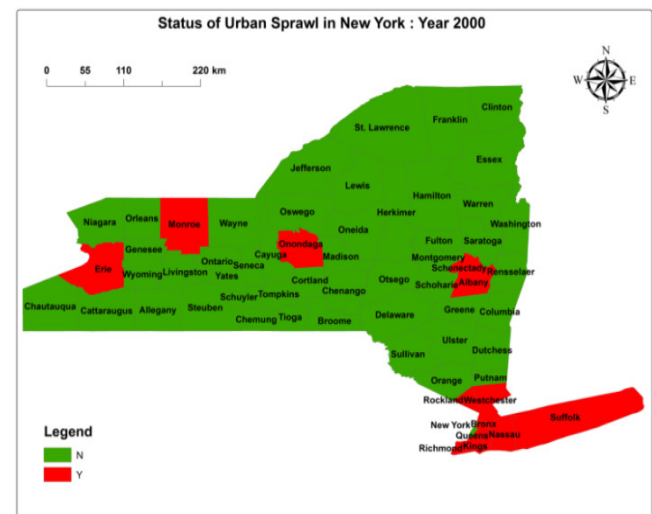
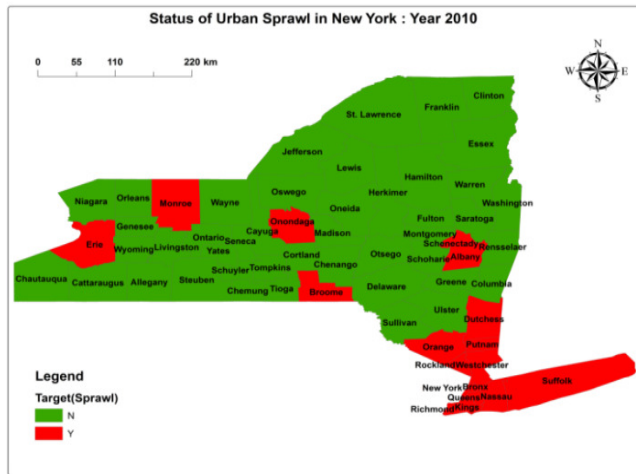


Figure 5: NY Map with Presence of Sprawl for Year 2000





**Figure 6: NY Map with Presence of Sprawl for Year 2010**

The data preprocessing using ArcGIS sets the stage for discovering knowledge from the GIS data through the data mining techniques of association rules and decision trees as described in the next two subsections.

## 2.2 Association Rule Mining Step

The goal of the association rule mining step is to understand the causal relationships between pertinent variables in the geospatial data. We adapt the classical Apriori algorithm [1] for association rule mining in order to discover relationships among the various spatial attributes affecting urban sprawl.

Although this is the secondary goal of our work, we conduct this prior to decision tree classification. This is in order to gain an understanding of individual parameter impacts on each other before predicting how they all impact urban sprawl. Since the GIS data is spatial, we write a Java program based on the Apriori algorithm adapted to geospatial data. We carry out further processing within the algorithm such that continuous data are mapped to binary attributes, implementing relevant filters as needed. After running our implementation of Apriori for spatial data over the NY dataset, we get some association rules, examples which are shown in Figure 7.

```
[BirthRate_Range2 -> GasolineStations_Range2 Target_Sprawl,
Income_Range3 -> HousingUnits_Range3 Target_Sprawl,
Asians_Range3 -> HousingUnits_Range3 Target_Sprawl,
HousingUnits_Range3 -> BirthRate_Range2 Target_Sprawl,
Target_Sprawl -> BirthRate_Range2 GasolineStations_Range2,
GasolineStations_Range2 -> BirthRate_Range2 Target_Sprawl,
BirthRate_Range2 Target_Sprawl -> Income_Range3
HousingUnits_Range3, Income_Range3 Target_Sprawl ->
BirthRate_Range2 HousingUnits_Range3, Income_Range3
BirthRate_Range2 -> HousingUnits_Range3 Target_Sprawl,
Asians_Range3 Target_Sprawl -> BirthRate_Range2
HousingUnits_Range3, Asians_Range3 BirthRate_Range2 ->
HousingUnits_Range3 Target_Sprawl, Asians_Range3
Income_Range3 -> HousingUnits_Range3 Target_Sprawl,
HousingUnits_Range3 Target_Sprawl -> Income_Range3
BirthRate_Range2, Income_Range3 HousingUnits_Range3 ->
BirthRate_Range2 Target_Sprawl, Asians_Range3
HousingUnits_Range3 -> BirthRate_Range2 Target_Sprawl,
BirthRate_Range2 HousingUnits_Range3 -> Income_Range3
Target_Sprawl]
```

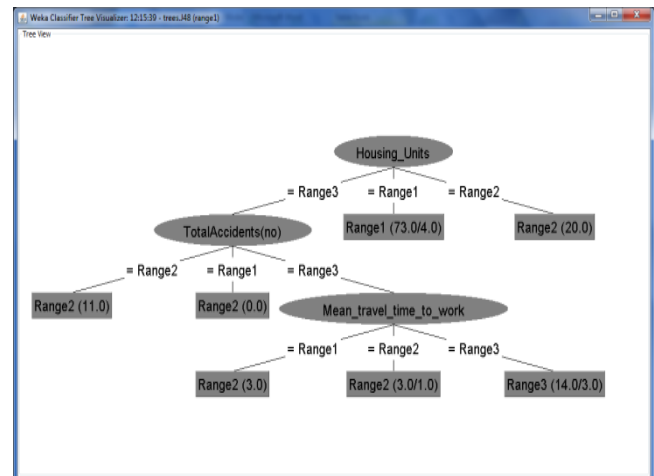
**Figure 7: Examples of Association Rules over GIS Data**

The ranges shown in this figure are explained as follows. In order to apply data mining algorithms efficiently, continuous variables are converted into discrete variables by grouping them into various ranges. Depending on the values of the variables they are grouped into ranges numbered 2, 3 and 4 respectively. For example, in Figure 7, an association rule states that if the income rate is in Range 3 of income (greater than 10 per 1000 population) then the housing units in that area would be in Range 3 of housing (greater than 100000). If these conditions prevail, it would lead to urban sprawl.

Having understood the impact of these parameters on urban sprawl, through association rule mining, we now proceed with decision tree classification to predict the occurrence of sprawl.

## 2.3 Decision Tree Classification Step

The purpose of the decision tree classification step is to achieve the most important goal in our work, namely, prediction of whether urban sprawl occurs given certain conditions. We deploy the popular J4.8 algorithm for decision tree classification [6] over the spatial GIS data. Adaption is performed similar to the Apriori algorithm. Figure 8 shows a partial snapshot of the decision trees learned from the given GIS data. Following the example of the NY data set, consider the 27 attributes. Since population density and percentage of Asians have most impact on sprawl and had comparatively very high impact when compared with other variables, we show those patterns in the J4.8 results in this figure. The ranges in the figure are similar to those in association rule mining, depicting the respective categories of the parameters.



**Figure 8: Partial Snapshot of Output from J4.8 over GIS Data**

Additionally, in this step, bagging and boosting [7] are also carried out. Bagging, which generates repeated bootstrap samples of the data, is useful in order to learn a more generic hypothesis, in this case, the knowledge discovered from the GIS data to predict the target attribute “sprawl” based on ranges. Boosting, which adjusts the weights of the training instances is useful to incorporate their relative importance to enhance the discovered knowledge. Thus, bagging and boosting are conducted in the development of our SDSS in order to improve the prediction rules with the intention of enhancing performance in the SDSS for prediction. Results after bagging and boosting appear in Figures 9 and 10 respectively.

REPTree

=====

TotalPersonalIncome < 11713160

| Employed(of tot pop) < 18.88 : Y (2/0) [1/0]

| Employed(of tot pop) >= 18.88 : N (61/1) [30/0]

TotalPersonalIncome >= 11713160 : Y (19/0) [11/0]

Size of the tree : 5

REPTree

=====

TotalPersonalIncome < 11286795 : N (65/3) [33/1]

TotalPersonalIncome >= 11286795 : Y (17/1) [9/0]

Size of the tree : 3

REPTree

=====

White people(of tot pop) < 82.31 : Y (20/2) [10/1]

White people(of tot pop) >= 82.31 : N (62/0) [32/1]

Size of the tree : 3

REPTree

=====

Percentage of foreign born < 5.25 : N (56/0) [27/1]

Percentage of foreign born >= 5.25

| FarmLand(Acres) < 44.7

| | TotalPersonalIncome < 373943200 : Y (15/0) [5/0]

| | TotalPersonalIncome >= 373943200 : N (2/0) [1/0]

| FarmLand(Acres) >= 44.7 : N (9/2) [9/3]

Size of the tree : 7

REPTree

=====

TotalPersonalIncome < 10641129

| FarmLand(Acres) < 71.6

| | Percentage of foreign born < 4.7 : N (6/0) [2/0]

| | Percentage of foreign born >= 4.7 : Y (6/2) [5/2]

| FarmLand(Acres) >= 71.6 : N (51/0) [25/0]

TotalPersonalIncome >= 10641129 : Y (19/0) [10/1]

**Figure 9: Output after Bagging**

Class 1 (Target(Sprawl)=Y)

Decision Stump

Classifications

Mean travel time to work (minutes) <= 37.7 : -0.417205238913497

Mean travel time to work (minutes) > 37.7 : 1.5486319364513992

Mean travel time to work (minutes) is missing : -0.09551223079474705

Class 2 (Target(Sprawl)=N)

Decision Stump

Classifications

Mean travel time to work (minutes) <= 37.7 : 0.4172052389134981

Mean travel time to work (minutes) > 37.7 : -1.5486319364514127

Mean travel time to work (minutes) is missing : 0.09551223079474797

Class 1 (Target(Sprawl)=Y)

Decision Stump

Classifications

Gasoline stations <= 16471.0 : -0.5648782405415602

Gasoline stations > 16471.0 : 0.8344461567041037

Gasoline stations is missing : 0.07978121875324665

Class 2 (Target(Sprawl)=N)

Decision Stump

Classifications

Gasoline stations <= 16471.0 : 0.5648782405415594

Gasoline stations > 16471.0 : -0.8344461567040931

Gasoline stations is missing : -0.07978121875324777

**Figure 10: Output after Boosting**

### 3. IMPLEMENTATION & EVALUATION

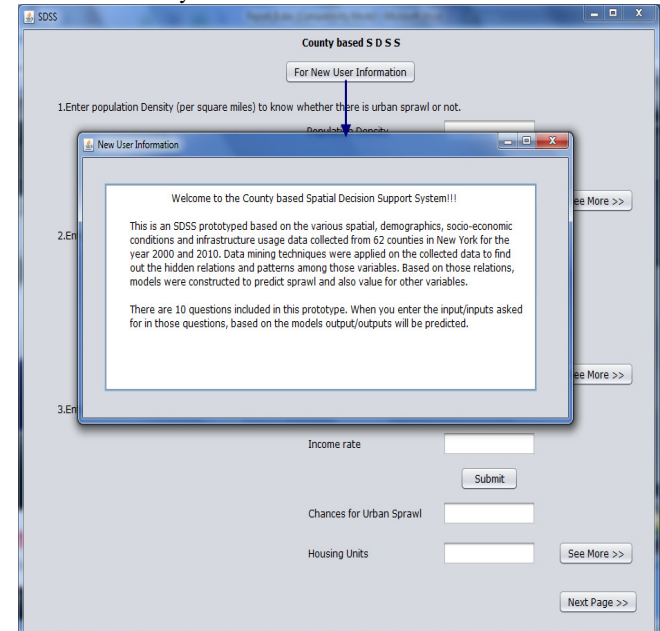
After discovering knowledge from the spatial GIS data, the results of the association rule mining and decision tree classification are used to implement the SDSS. The most important parameters affecting urban sprawl as obtained from the knowledge discovery process are critical in the SDSS development and are used to design questions that targeted users could pose to the system.

For example, in the NY state data, we find that population density is the major cause of urban sprawl. Moreover, due to the increase in population other related factors also increase. We also find interesting results such as how the number of housing units, truck transportations and gas stations relate to each other in sprawl-affected areas. Based on this learning, the anticipated questions in the SDSS are designed which guide user interaction. These cater to various user queries, a few examples of which are listed below.

- If the unemployment rate is greater than a certain value, would sprawl be likely to occur?
- If personal income is in a given range & percentage of Asians is above a certain value, would sprawl occur?
- What is the relationship between birth rate and sprawl occurrence?
- What is the relationship between housing units and gas stations?

Note that the first query predicts the likelihood of sprawl based on a single parameter while the second one predicts its likelihood based on multiple parameters. The third one is to understand the impact of a given parameter on sprawl whereas the fourth one is to understand the impact of parameters on each other. Such queries are designed during the SDSS implementation based on the GIS data mining and serve as the basis for creating the user interface for interaction with the system.

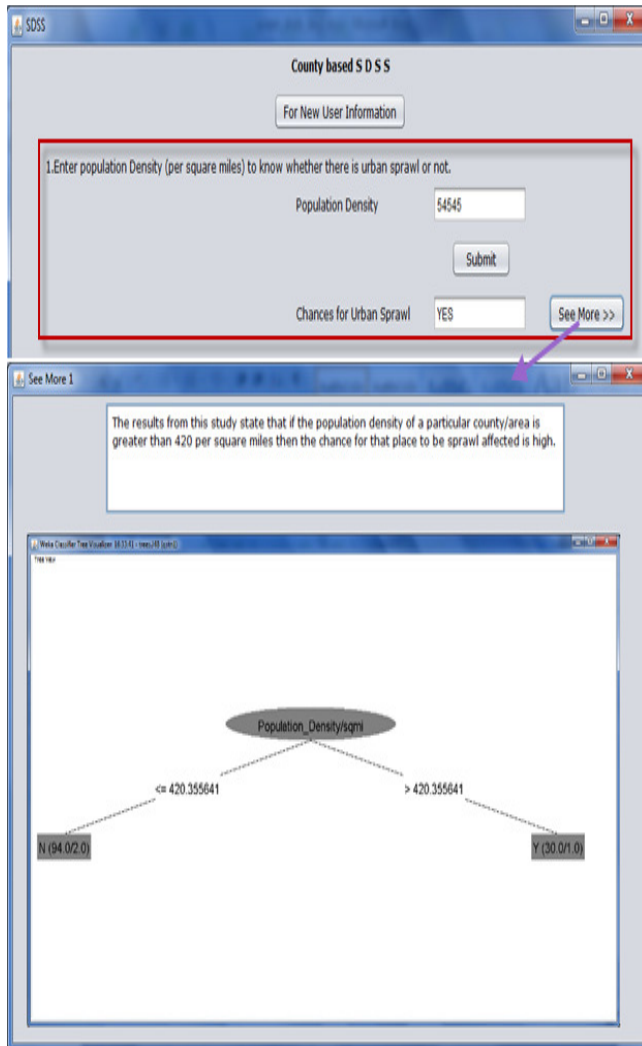
This SDSS is implemented in Java using the knowledge discovered by mining the GIS data. A screen-dump of the SDSS is shown in Figure 11. This system offers an interface through which the user can pose queries in the form of several decision-making scenarios to the system.



**Figure 11: Prototype SDSS for Urban Sprawl**

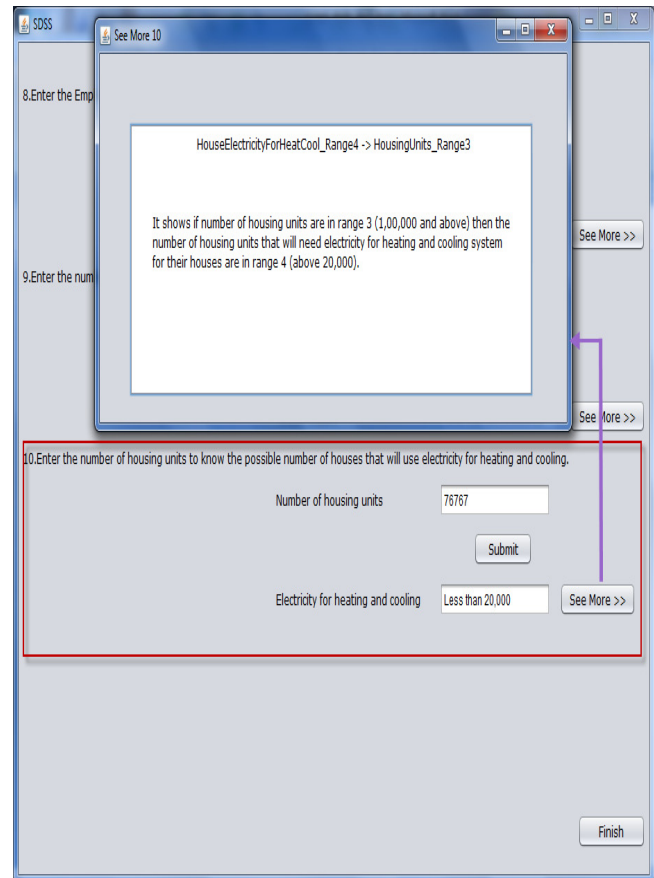
We conduct a thorough evaluation of this SDSS by running various user queries. Two such sample queries are shown in the scenarios below, along with the responses of the SDSS to the respective scenarios.

*Scenario 1:* The user queries whether sprawl occurs by entering a certain population density (54545 as in Figure 12). The system predicts that sprawl occurs in this case (YES). In order to understand why system gives this response, the user can select the “See More” option of the SDSS. It shows that based on decision tree learning over the given data set, if population density per square miles is more than 420, then sprawl is likely to occur. This example predicts the occurrence of sprawl shows the impact of a given parameter on the target attribute “sprawl”.



**Figure 12: How Population Density affects Sprawl**

*Scenario 2:* The user needs to know how many houses would need electricity for heating and cooling, given the total number of housing units at that area (76767 as in Figure 13). The SDSS responds “less than 20,000” as shown. The “See More” option indicates how this was learned by association rules. This example shows the impact of sprawl-related parameters on each other.



**Figure 13: Relation between Housing Units & Possible Number of Houses with Electricity for Heating, Cooling**

Likewise, users can pose many queries to the SDSS to predict sprawl occurrence based on given conditions and to understand the impact of sprawl-causing factors on each other. This prototype SDSS is useful to urban planners and geoscientists and sets the stage for the development of a large scale SDSS for predictive analysis in urban planning by mining over big data with the potential use of cloud technology and other advances. More algorithms and solutions would be proposed in the further development of the SDSS in the future.

## 4. RELATED WORK

In recent years geospatial technology is developing rapidly. New techniques in the GIS field and large resources in the form of digital imagery, statistical data are encouraging more and more researchers to put in efforts into topics such as urban sprawl and sustainability, which is now a big concern among geologists, environmentalists and city planners. We cite some of the literature herewith.

In Jiang et al. [3], 2007 they take Beijing as a case study and put forward that urban sprawl can be measured from spatial configuration, urban growth efficiency and external impacts, and then develops a geo-spatial indices system for measuring sprawl, using a total of 13 indicators. Thus, the authors try to measure the sprawl or rate the sprawl after it occurs. However, in our work, we predict the occurrence of sprawl and also find the patterns that caused the sprawl.

Luo et al. [4], strive to model urban expansion with the use of geographic information systems by remote sensing. The focus of

their work is more on building a model through remote sensing technologies. Our work assumes that the GIS data has already been collected and works further to analyze that data to predict the occurrence of sprawl in the future and study the impact of the parameters causing the sprawl.

In the study conducted by Sudhira et al. [12], the authors study sprawl using spatial data along with other attributes. The study area is India, and their study attempts to identify sprawl, quantify it by defining new metrics, understand the dynamic process and subsequently model the same. The authors in this study have used statistical analysis. However, we consider data mining algorithms over GIS data to find the sprawl causing patterns.

The authors of Sun et al. [13], implement land use analysis for the City of Calgary, Alberta, Canada, using an object-oriented approach. They aim to simulate the land use pattern using Markov Chain analysis and Cellular Automata analysis based on the interactions between the land use and the transportation network. However, they do not develop tools to predict sprawl occurrence to assist urban planning. In our work we analyze GIS data with the adaption of data mining techniques over the geospatial attributes. Also, we predict the manner in which land use is likely to develop in the future. Moreover, we develop a prototype SDSS to assist decision-making in urban planning by predicting urban sprawl and its parameter impact.

## 5. CONCLUSIONS

We have addressed the issue of predicting urban sprawl by mining GIS data. We have collected data for variables directly or indirectly related to urban sprawl and adapted the data mining algorithms of Apriori for association rule mining and J4.8 for decision tree classification to this spatiotemporal data. Based on the results of knowledge discovery over this data, we have implemented a prototype SDSS that can assist in decision-making on urban sprawl.

This can be used by urban planners, city dwellers and any other users interested in finding out the likelihood of sprawl occurrence and understanding how sprawl-causing variables affect each other. This work on the whole contributes to the computer science community of spatiotemporal data mining and the geoscience community of urban sustainable development. It spans the realm of geospatial intelligence has an impact on the environment and public health.

Our future work includes deploying this prototype as the basis for research and development of a large scale SDSS in urban planning by knowledge discovery over big data. This could potentially entail the proposal of advanced techniques in cloud computing, and / or new paradigms in areas overlapping GIS and big data to solve the concerned problems. Its broader impact would be to enhance urban sustainability by helping to make decisions for promoting the development of sustainable cities and surrounding areas.

## 6. REFERENCES

- [1] Agrawal, R., Imielinski, T., and Swami, A., Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD*, pp. 207-216, 1993.
- [2] Brueckner Jan K., Urban Sprawl: Diagnosis and Remedies. *International Regional Science Review* 23, 2: 160-171, 2000.
- [3] Jiang F., Liu S., Yuan H., Zhang Q. Measuring urban sprawl in Beijing with geo-spatial indices. *Journal of Geographical Sciences*. pp. 469-478, 2007.
- [4] Luo, J., Yu, D.L., Miao, X., Modeling Urban Growth Using GIS and Remote Sensing, *GIS Science & Remote Sensing*. 45(4): 426-442, 2008.
- [5] O'Sullivan, D., Unwin, D.J., *Geographic Information Analysis*. Wiley, NJ, 2002.
- [6] Quinlan, J.R., Induction of Decision Trees. *Machine Learning* 1(1):81-106, 1986.
- [7] Quinlan, J.R., Bagging, Boosting and C4.5. *AAAI*, pp. 725-730, 1996.
- [8] Ly, Z.Q., Dai F.Q., Sun C., Evaluation of urban sprawl and urban landscape pattern in a rapidly developing region. *Environmental Monitoring and Assessment*. pp 6437-6448, 2012.
- [9] McCann B. A., Reid, E.. Measuring the Health Effects of SPRAWL. *Smart Growth America Surface Transportation Policy Project* 2003.
- [10] Sprague, R. H., and E. D. Carlson. *Building effective Decision Support Systems*. Prentice-Hall, 1982.
- [11] Squires G. D. *Urban Sprawl: Causes, Consequences, & Policy Responses*. The Urban Institute, 2002.
- [12] Sudhira H.S., Ramchandra, T.V., Jagadish, K.H., Urban Sprawl: Metrics, Dynamics and Modeling using GIS, *International Journal of Applied Earth Observation and Geoinformation*, 5(1):29-39, 2004.
- [13] Sun H., Forsythe W. and Waters N., Modeling Urban Land Use Change and Urban Sprawl, *Networks and Spatial Economics*. pp 353-376, 2007.
- [14] Tang H., McDonald S. Integrating GIS and Spatial Data mining Technique for Target marketing of University Courses. *Symposium on Geospatial Theory, Processing and Applications*, 2002.
- [15] Warner M., Hinrichs C., Schneyer J., and Joyce L., Sustaining the Rural Landscape by Building Community Social Capital, 5(2), CarDI, Cornell University, 1997.
- [16] Wu X., Kumar V., Quinlan J. R., Ghosh J., Yang Q., Motoda H., McLachlan G. J., Ng A., Liu B., Yu P., Zhou Z., Steinbach M., Hand D. J. and Steinberg D. *Top 10 algorithms in data mining*. *ICDM*, 2006.
- [17] [http://egsc.usgs.gov/isb/pubs/gis\\_poster/](http://egsc.usgs.gov/isb/pubs/gis_poster/)
- [18] [http://www.makingthemodernworld.org.uk/learning\\_module\\_s/geography/04.TU.01/](http://www.makingthemodernworld.org.uk/learning_module_s/geography/04.TU.01/)
- [19] [http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What\\_is\\_ArcGIS\\_Desktop/](http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What_is_ArcGIS_Desktop/)
- [20] [http://en.wikipedia.org/wiki/Spatial\\_decision\\_support\\_system](http://en.wikipedia.org/wiki/Spatial_decision_support_system)
- [21] <http://www.cs.waikato.ac.nz/ml/weka/>
- [22] <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>
- [23] <http://maya.cs.depaul.edu/~Classes/Ect584/Weka/preprocess.html>

## 7. APPENDIX FOR DATASETS

Shape Files:

<http://www2.census.gov/cgi-bin/shapefiles2009/state-files?state=36>

Vital Statistics of New York State:

- a) [http://www.health.ny.gov/statistics/vital\\_statistics/](http://www.health.ny.gov/statistics/vital_statistics/)
- b) [http://www.health.ny.gov/statistics/vital\\_statistics/2010/](http://www.health.ny.gov/statistics/vital_statistics/2010/)



- c) [http://www.health.ny.gov/statistics/vital\\_statistics/2000/toc.htm](http://www.health.ny.gov/statistics/vital_statistics/2000/toc.htm)

Population Details:

- a) [http://www.health.ny.gov/statistics/vital\\_statistics/2010/table01.htm](http://www.health.ny.gov/statistics/vital_statistics/2010/table01.htm)  
b) <http://pad.human.cornell.edu/counties/projections.cfm>  
c) <http://pad.human.cornell.edu/counties/projections.cfm>

Death and Birth Rates:

- a) [http://www.health.ny.gov/statistics/vital\\_statistics/2000/toc.htm](http://www.health.ny.gov/statistics/vital_statistics/2000/toc.htm)  
b) [http://www.health.ny.gov/statistics/vital\\_statistics/2010/](http://www.health.ny.gov/statistics/vital_statistics/2010/)

Land Area Details

- a) [http://www.health.ny.gov/statistics/vital\\_statistics/2000/table02.htm](http://www.health.ny.gov/statistics/vital_statistics/2000/table02.htm)  
b) [http://www.health.ny.gov/statistics/vital\\_statistics/2010/table02.htm](http://www.health.ny.gov/statistics/vital_statistics/2010/table02.htm)

Census:

- a) <http://esd.ny.gov/NYSDataCenter/Census2010.html>  
b) <http://esd.ny.gov/NYSDataCenter/Census2000.html>

Total Personal Income:

<http://esd.ny.gov/NYSDataCenter/PersonalIncomeData.html>

Per Capita Personal Income:

<http://esd.ny.gov/NYSDataCenter/PersonalIncomeData.html>

Population Density:

- a) <http://esd.ny.gov/NYSDataCenter/Census2010.html>  
b) [http://www.health.ny.gov/statistics/vital\\_statistics/2006/table02.htm](http://www.health.ny.gov/statistics/vital_statistics/2006/table02.htm)

Total Housing Unit:

[http://esd.ny.gov/NYSDataCenter/Population\\_Housing\\_Data.html](http://esd.ny.gov/NYSDataCenter/Population_Housing_Data.html)

Foreign Born Percentage:

- a) <http://quickfacts.census.gov/qfd/states/36/36001.html>

Mean travel time to work:

- a) <http://quickfacts.census.gov/qfd/states/36/36001.html>  
b) <https://www.dot.ny.gov/divisions/policy-and-strategy/darb/dai-unit/ttss/jtw>

Population by Race and Hispanic or Latino Origin:

- a) <http://esd.ny.gov/NYSDataCenter/Census2010.html>  
b) <http://www.labor.ny.gov/stats/nys/statewide-population-data.shtm>  
c) [http://www.nyc.gov/html/dcp/html/census/demo\\_tables.shtml](http://www.nyc.gov/html/dcp/html/census/demo_tables.shtml)

Accident Data:

- a) <https://www.dot.ny.gov/divisions/operating/oss/highway/accident-rates>  
b) <http://www.dmv.ny.gov/stats.htm>  
c) <http://www.dmv.ny.gov/stats-arc.htm>

Employment and Unemployment Details:

- a) <http://www.labor.ny.gov/home/>  
b) <http://www.labor.ny.gov/stats/lslaus.shtm>  
c) <http://www.cdrpc.org/Employment/EMPTable.html>

Poverty Rate:

<http://quickfacts.census.gov/qfd/states/36000.html>

Education:

<http://www.p12.nysed.gov/irs/statistics/public/>

Transit Data:

<https://www.dot.ny.gov/divisions/policy-and-strategy/darb/dai-unit/ttss>