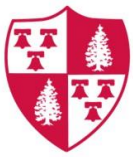


MONTCLAIR STATE UNIVERSITY

Data mining in GIS for a prototype SDSS on urban land use

Master's Project submitted to
Department of Computer Science
Montclair State University
Montclair, NJ 07043

In Partial fulfillment of the requirements for a
Master of Science degree in Computer Science

Anita Pampoore Thampi
Advisor: Dr. Aparna Varde
30th November 2012

Table of Contents

1. Abstract

As the population of the world continually increases, so do the challenges for decision makers to provide jobs, infrastructure and social security, among other essential needs. The main options available to most people are to come to urban areas and find means to live. As a result, urban areas are highly populated with much less natural green cover. All the places in a city are not the same. But, when we consider facilities, hygiene, population and infrastructure, there are good parts in a city as well as bad parts. "Urban sprawl" or suburban sprawl is a multifaceted concept centred on the expansion of auto-oriented, low-density development [13].

This project is an attempt to achieve a clearer perspective to the term "urban sprawl" by developing a prototype based on the findings from the data mining analysis which defines this condition of urban areas. This would serve as part of a bigger project that involves developing a full-fledged Spatial Decision Support System (SDSS) with several decision-making scenarios that would head towards performing the forecasting of urban land use dynamics.

Urban areas will have certain distinct features that are either directly related or hidden to sprawl. For example, population density is a feature, which directly affects sprawl. At the same time, income rates and unemployment rates are some of the features that indirectly affect sprawl. Some of these features were studied mathematically in the form of demographics, spatial data, socio-economic conditions, infrastructure usage, and accidental reports. Although the definition of the term "urban sprawl" is debatable, the term urban sprawl index is used to identify which counties in New York are prone to urban sprawl. Urban sprawl or urban spatial expansion results from three powerful sources: growing population, rising incomes, and falling commuting costs**.**

Furthermore, the data collected is correlated to sprawl to find a relation between sprawl and these factors using two of the data mining algorithms: Apriori for association rule mining and J4.8 for decision tree classification. In short, this project is a cause and effect study. These relations are saved and implemented in a prototype, which will help users with two types of decision making: first, to decide whether "urban sprawl" is occurring or there is possibility it will occur in the regions, and second, to estimate the value of a variable or variables based on the values of another variable or combination of variables. The users will enter the variables like demographics etc. and will thus be able to come to a conclusion about the probability of urban sprawl or so. These outputs can help decision makers to identify the problem and create solutions for avoiding sprawl occurrence in new rural or suburban areas and to plan accordingly, thus leading to urban sustainable development.

2. Introduction

2.1 Background and Motivation

We are living in a world which grows and urbanizes at a rapid pace. If we continue this, the destiny of mankind will be different from what we dream, and our successors will be left with no resources to sustain. Other impacts of this "growth", when the mankind spread its wings to the outskirts of the urban background, are overcrowding, pollution, unemployment, crime, poverty, disease etc. Today, expansion means cities are expanding to nearby towns and villages by converting those natural lands to impervious lands by constructing buildings, parking lots, highways etc. So there should be some method to curb urban sprawl and some limitations for urbanization in each area. This can only be achieved by taking appropriate decisions. Urban planners and engineers should take appropriate and wise decisions to protect natural land while designing any activities for new constructions. With this concept/thought, we are introducing a mini-prototype.

This project attempts to formulate an appropriate mathematical model which captures the conditions and effects of urban sprawl. There are numerous studies to find the reason for urban sprawl and for its remedies, and most of them are done by analyzing data using statistical tools. Here we are using powerful data mining algorithms to find the inherent relations between the variables collected which is related to sprawl. Although the relations and patterns derived from this project are specific to New York, these relations can be generalized to simulate other urban areas in United States. The basic structure of all cities and townships are almost the same. They all have the same infrastructure design, socio-economic conditions, transportation and facilities. It is this similarity which 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 places accordingly in order to eliminate the conditions of sprawl.

2.2 Brief Introduction of Terminologies

2.2.1 GIS (Geographic Information System)

A geographic information system (GIS) integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. GIS allows us to view, understand, question, interpret, and visualize data in many ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts [6].

The power of GIS comes from the ability to relate different information in a spatial context and to reach a conclusion about this relationship. Most of the information we have about our world contains a location reference, placing that information at some point on the globe. When rainfall

information is collected, it is important to know where the rainfall is located. 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. 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 [25].

2.2.2 Urban Sprawl

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 [23].

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 [13]. As a result people would start the trend of moving to neighborhood low density areas, and to meet their requirements, more houses, parking lots (Fig.2.1), roads (Fig. 2.2), shops etc. should be constructed which gradually leads to expansion of sprawl to that areas too.

Fig.2.1: Parking lots [36]

Fig.2.2: Roads [35]

2.2.3 Urban Sustainability

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 [25].

Sustainable development means 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 [7]. Urban sprawl, which leads the land to be unsustainable, is nowadays considered as a serious issue not only in the United States but all over the world (Fig.2.3–Fig.2.6). Ecologists and environmentalists are competing to find remedies for the same.

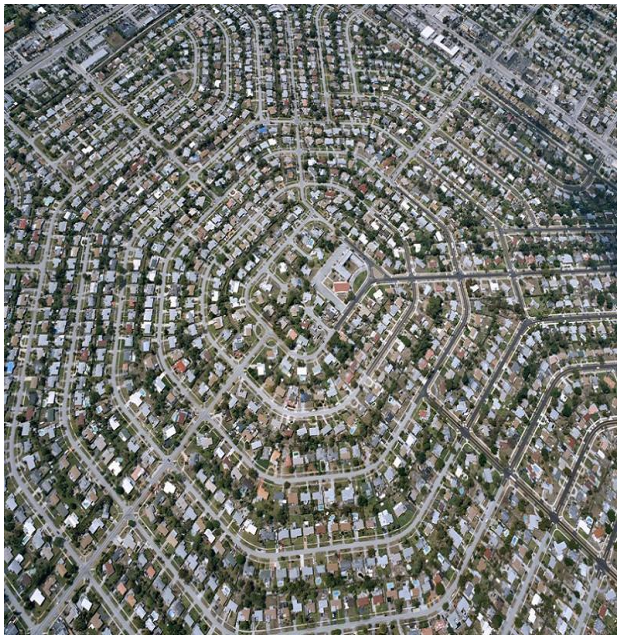


Fig.2.3: Eden Prairie, Florida [32]

Fig.2.4: Melbourne, Australia [33]

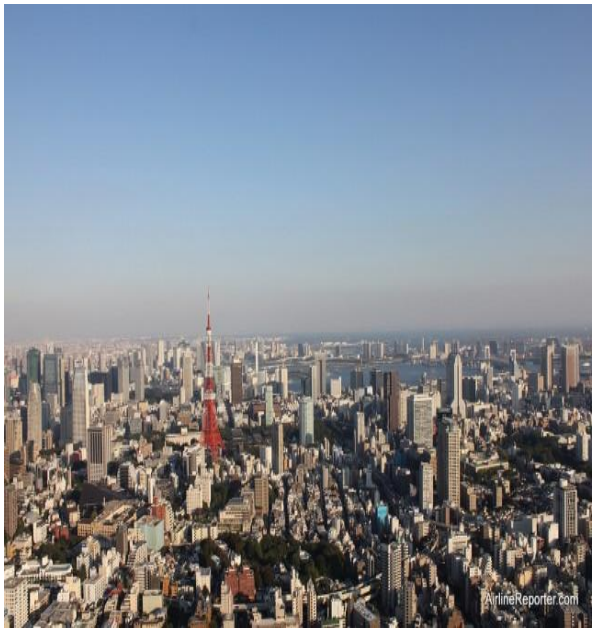


Fig.2.5: Tokyo, Japan [34]


Fig.2.6: Mumbai, India [35]

2.2.4 ArcGIS

ArcGIS is a GIS software package of Economic and Social Research Institute (ESRI), which is used for working with maps and geographic information [8]. 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.

2.2.5 Data mining

Data mining is about solving problems by analyzing data already present in databases and thus creating mining model. To create a model, an algorithm first analyzes a set of data, looking for specific patterns and trends. The process must be automatic or semi-automatic. The patterns discovered must be meaningful in that and they lead to some advantage, usually an economic advantage. Useful patterns allow us to make non trivial predictions on new data. The algorithm then uses the results of this analysis to define the parameters of the mining model [28].

The mining model that an algorithm creates can take various forms [25], including:

- A set of rules that describe how products are grouped together in a transaction.

- A decision tree that predicts whether a particular customer will buy a product.
- A mathematical model that forecasts sales.
- A set of clusters that describe how the cases in a dataset are related.

2.2.6 Apriori

One of the most popular data mining approaches is to find frequent itemsets from a dataset and derives association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation Agarwal el at., 1993. It is characterized as a level-wise complete search algorithm using anti- monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order Wu el at., 2006.

2.2.7 J4.8 (C4.5) Decision Tree

J48 is an open source Java implementation of the C4.5 algorithm**.** A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes. The branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset [30].

In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained. For the other cases, we then look for another attribute that gives us the highest information gain. Hence we continue in this manner until we either get a clear decision of what combination of attributes gives us a particular target value, or we run out of attributes. In the event of, we run out of attributes, or if we cannot get an unambiguous result from the available information, we assign this branch a target value that the majority of the items under this branch possess [26].

2.2.8 WEKA

WEKA (Waikato Environment for Knowledge Analysis) is a user-friendly data mining tool which contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionalities. It includes a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from our own Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [10, 11].

2.2.9 SDSS- Spatial Decision Support System

Spatial decision support system (SDSS) is an 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. A system which models decisions could be used to help identify the most effective decision path, Sprague el at., 1982.

An SDSS is sometimes referred to as a policy support system, and comprises a decision support system (DSS) and a geographic information system (GIS). This entails use of a database management system (DBMS), which holds and handles the geographical 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 [9].

2.3 Problem Definition

The trends used in urban areas lead to "urban sprawl" in most regions. Suburban areas are expanding day by day. The natural covers left in these regions are left to almost zero value. As stated in the abstract a proper definition to urban sprawl is complicated and so arises the problem of mathematically defining it. After taking a closer look into the definition and some references in environmental studies, we have taken into consideration certain variables and factors which would help define sprawl in the relevant form [2, 14], but there are of course limitations to the variables and its availability in this short time period.

Given the brief description of the various terminologies involved in the domain, we can outline the central goal of this project. The focus of this study is to understand and identify urban sprawl and thereby help decision makers to make better decisions by providing them some models which can predict urban sprawl or help them to plan accordingly in a way that the new suburban areas are sustainable enough and thus maintain a balance between nature and development. In addition the tool being developed is looking to find the value of some variables with respect to others by finding the hidden patterns from the dataset. For example, how many trucks are used for transportation, within a sampled place if the number of housing units is given for that place?

A tool to identify sprawl or sprawl like conditions based on these kinds of data is not available for any particular area in general. We would like to address this problem.

2.4 Proposed Solution

A prototype SDSS is what we propose. SDSS stands for Spatial Decision Support System. This prototype is based on the cause and effect theory and involves reverse engineering. The theory is simple, we study the effects of the interesting variables and thus with the help of effects we try to define the cause of urban sprawl. To study the patterns and trends among these variables, two data mining algorithms are applied on to the dataset.

Decision tree algorithms are used since the primary goal of this project was prediction of the discrete target attribute 'sprawl' based on other attributes in the dataset [2]. The secondary goal of this project was to find the correlation between these variables and to find the patterns hidden in them. This is very similar to transaction database analysis where the best algorithm to apply is association rule mining [20]. Therefore, Apriori [26, 29] for association rule mining and J4.8 [26, 29] for decision tree classifications are used here to explore the New York county dataset. In this case, variables are the causes for urban sprawl. Models were derived based on the relations, which we got from these data mining techniques. Based on those models a mini user interactive prototype was made which can help urban planners, city dwellers and various other users in making decisions pertaining to urban land, for example buying houses, building apartment complexes, investing in massive land projects, building utilities in developed regions and so forth.

Two approaches are used for performing data mining in this study.

a) Implementing the Apriori algorithm in Java
b) Running J4.8 algorithm using the data mining tool WEKA

The knowledge discovered by mining is then used to implement the functionality of the decision support system in Java, in order to allow the user to enter data on certain parameters for the system to predict whether sprawl occurs.

3. Data Collection and formatting

Based on the findings of the pilot studies and literature reviews used, there are hundreds of variables which by itself or by some combinations can directly and/or indirectly affect urban sprawl (can be the reason of sprawl). Among those variables, the ones that had the most impact on urban sprawl and which was available in this limited time were selected for this study [23, 19].

The data collected for this project are all real data which are based on New York State. The reason why the counties were selected was to differentiate which areas of New York was affected by urban sprawl. Even though New York is famous for its urbanized metropolitan cities (Fig.3.1 and Fig.3.2), this state still have many counties which have strong rural character (Fig.3.3 and 3.4) [14].

Fig.3.1: New york City [38]

Fig.3.2: New york City [38]

Fig.3.3: Allegany [31]

Fig.3.4: Washington [37]

The dataset consisting of 27 variables, excluding shape file, are all continuous 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. For some variables, interpolation had to be done since those data were not available for the years 2000 and 2010. Seeing as there weren't any existing studies prior, there was no data-set available, so we had to collect each variable from many government sources such as http://www.census.gov/, a range of research papers, and also government publications. The data collected were in many various formats such as word, pdf, excels csv, notepad, .shp [11] files etc.

Among these 27 variables, the last one is the target attribute which defines the presence of sprawl occurrence. That is whether sprawl's values are either yes or no. Among 124 instances 31 have target attribute positive which shows presence of sprawl and rest 93 have target attribute negative which shows absence of sprawl. This target attribute is finalized based on a project about "Measuring the health effects of sprawl" done by smart growth America [19]. Once after the data was collected preprocessing was carried out and combined all together into an excel sheet (Fig .3.5 and Fig.3.6).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	County_Name	Latitude	Longitude	Regions	Land_Area	Time	Population	Population_Density,	Percentage of foreign born	White people	African p	Asians	TotalPersonaIncome	Employed	Unemploy	FarmLand	Ttlh
2	Albany	42.6525793	73.7562317	7	524	2000	294565	562.1469466	6.5	83.19	11.08	2.75	9787234	50.85	1.77	84.8	1
3	Allegany	42.0900647	78.4941887	1	1030	2000	49927	48.47281553	1.8	97.03	0.72	0.72	947140	43.46	2.2	149.2	
4	Bronx	40.85 N	73.866667	11	42.1	2000	1332650	31654.3943	29	29.87	35.64	3.01	25929212	33.9	2.61	0	4
5	Broome	17.9512214	122.244327	6	707	2000	200536	283.6435644	5.3	91.33	3.28	0.19	5055744	47.27	1.8	102.4	
6	Cattaraugus	42.2318132	78.7476207	1	1310	2000	83955	64.08778626	1.4	94.63	1.06	0.46	1743948	46.81	2.26	114	
7	Cayuga	39.9486488	87.4597385	3	693	2000	81963	118.2727273	2.3	93.34	3.99	0.42	1843616	47.46	1.95	121.8	
8	Chautauqua	42.209774	79.4668444	1	1062	2000	139750	131.5913371	1.9	94.04	2.18	0.36	2955871	46.8	1.93	82.8	
9	Chemung	42.007381	76.625473	3	408	2000	91070	223.2107843	2.2	90.96	5.82	0.23	2214176	45.02	1.98	88	
10	Chenango	42.4972314	75.6208087	6	893.55	2000	51401	57.524481	1.7	97.65	0.82	0.28	1092543	45.72	1.95	131.2	
11	Clinton	38.7651145	76.8983058	5	1039	2000	79894	76.89509143	4.5	93.33	3.58	0.67	1798974	46.06	2.25	147	
12	Columbia	4.570868 N	74.2973325	9	636	2000	63094	99.20440252	4.4	92.09	4.52	0.8	1783783	48.34	1.74	110.6	
13	Cortland	42.6011813	76.1804842	3	500	2000	48599	97.198	2.2	96.95	0.86	0.41	1071770	47.53	2.06	135.2	
14	Delaware	38.9108325	75.5276698	8	1446	2000	48055	33.23305671	3.4	96.44	1.18	0.53	1037264	44.32	1.87	171.2	
15	Dutchess	41.7784372	73.7477857	9	802	2000	280150	349.3142145	8.4	83.66	9.32	2.52	8816205	48.15	1.57	56.8	1
16	Erie	42.1292240	80.085059	2	1045	2000	950265	909.3444976	4.5	82.18	13	1.46	26290382	47.24	2.06	56.6	4

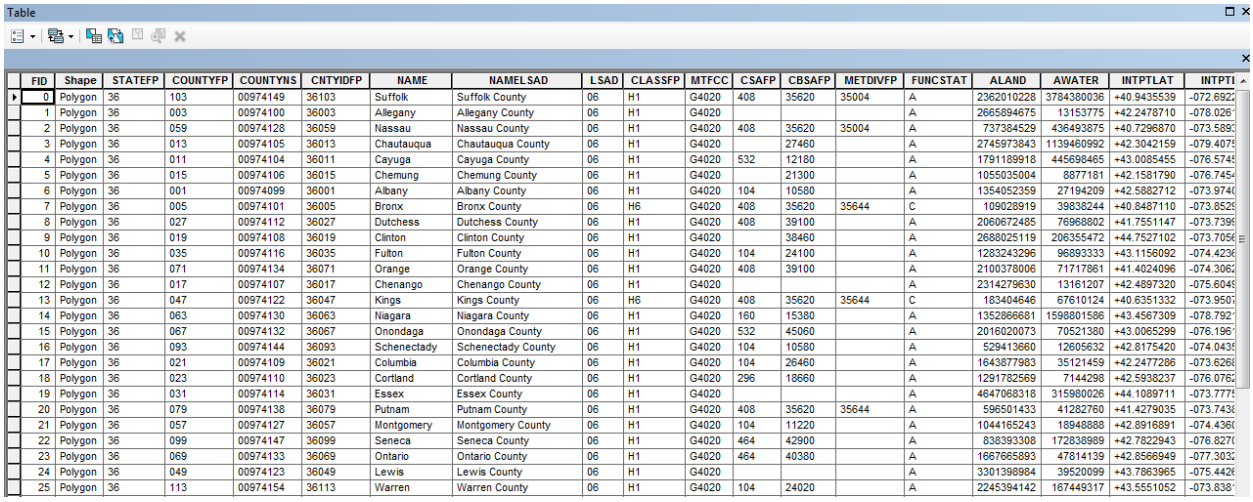
Fig.3.5: Whole data in excel

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	Employed(	Unemploye	FarmLand(Ac	TtlhousingUnit(no)	Mean travel time	TotalAccidents	Public supply,	No of house ur	Poverty_p	Births per	Death Rat	Education(pe	Gasoline s	Truck tran	Transit an	Education	Target(
2	50.85	1.77	84.8	129972	19.7	5429	292	15788	10.6	11.2	10.1	33.3123	19124	89132	19563	313713	Y
3	43.46	2.2	149.2	24505	21.2	588	25.24	1433	15.5	10.9	9.8	17.1809	4499	6734	4739	35914	N
4	33.9	2.61	0	490659	41.7	13662	1332.65	48312	30.7	17.1	7.5	14.6458	14372	52333	81953	469965	Y
5	47.27	1.8	102.4	88817	18.2	2465	168.8	7265	12.8	11	11.1	22.6503	14481	36568	8469	15031	N
6	46.81	2.26	114	39839	21.4	958	51.39	3327	13.7	12.1	10	14.8724	6431	14341	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	34343	2541	4963	N
9	45.02	1.98	88	37745	19.6	853	87.16	2792	13	11.9	10.5	18.5818	4073	13936	3454	19343	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	9348	13.9	9.8	7.8	17.803	9380	21050	1506	3271	N
12	48.34	1.74	110.6	30207	25.1	941	29.84	3661	9	10.5	11.8	22.5938	5802	9013	2473	5809	N
13	47.53	2.06	135.2	20116	20.8	679	33.86	1874	15.5	11.6	8.8	18.8404	4554	5524	1755	2599	N
14	44.32	1.87	171.2	28952	21.3	631	22.78	1781	12.9	9.6	11.8	16.6314	5166	8238	2344	1398	N
15	48.15	1.57	56.8	106103	29.9	4524	204.55	11695	7.5	11.9	7.8	27.6266	14445	25140	17450	237458	N

Fig.3.6: Whole data in excel

4. Mapping data using ArcGIS 10.1

Collected dataset for the years 2000 and 2010, along with county based shape file for New York (Fig.4.1) is plotted as two separate maps using Arc GIS software to show the urban sprawl in both the years. Map representation gives a clearer visual representation of the data.



Table

FID	Shape	STATEFP	COUNTYFP	COUNTYNS	CNTYIDFP	NAME	NAMELSAD	LSAD	CLASSFP	MTFCC	CSAFP	CBSAFP	METDIVFP	FUNCSTAT	ALAND	AWATER	INTPTLAT	INTPTI
0	Polygon	36	103	00974149	36103	Suffolk	Suffolk County	06	H1	G4020	408	35620	35004	A	2362010228	3784380036	+40.9435539	-072.692
1	Polygon	36	003	00974100	36003	Allegany	Allegany County	06	H1	G4020				A	2665894675	13153775	+42.2478710	-078.026
2	Polygon	36	059	00974128	36059	Nassau	Nassau County	06	H1	G4020	408	35620	35004	A	737384529	436493875	+40.7296870	-073.589
3	Polygon	36	013	00974105	36013	Chautauqua	Chautauqua County	06	H1	G4020		27460		A	2745973843	1139460992	+42.3042159	-079.407
4	Polygon	36	011	00974104	36011	Cayuga	Cayuga County	06	H1	G4020	532	12180		A	1791189918	445698465	+43.0085455	-076.574
5	Polygon	36	015	00974106	36015	Chemung	Chemung County	06	H1	G4020		21300		A	1055035004	8877181	+42.1581790	-076.745
6	Polygon	36	001	00974099	36001	Albany	Albany County	06	H1	G4020	104	10580		A	1354052359	27194209	+42.5882712	-073.974
7	Polygon	36	005	00974101	36005	Bronx	Bronx County	06	H6	G4020	408	35620	35644	C	109028919	39838244	+40.8487110	-073.852
8	Polygon	36	027	00974112	36027	Dutchess	Dutchess County	06	H1	G4020	408	39100		A	2060672485	76968802	+41.7551147	-073.739
9	Polygon	36	019	00974108	36019	Clinton	Clinton County	06	H1	G4020		38460		A	2688025119	206355472	+44.7527102	-073.705
10	Polygon	36	035	00974116	36035	Fulton	Fulton County	06	H1	G4020	104	24100		A	1283243296	96893333	+43.1156092	-074.423
11	Polygon	36	071	00974134	36071	Orange	Orange County	06	H1	G4020	408	39100		A	2100378006	71717861	+41.4024096	-074.306
12	Polygon	36	017	00974107	36017	Chenango	Chenango County	06	H1	G4020				A	2314279630	13161207	+42.4897320	-075.604
13	Polygon	36	047	00974122	36047	Kings	Kings County	06	H6	G4020	408	35620	35644	C	183404646	67610124	+40.6351332	-073.950
14	Polygon	36	063	00974130	36063	Niagara	Niagara County	06	H1	G4020	160	15380		A	1352866681	1598801586	+43.4567309	-078.792
15	Polygon	36	067	00974132	36067	Onondaga	Onondaga County	06	H1	G4020	532	45060		A	2016020073	70521380	+43.0065299	-076.196
16	Polygon	36	093	00974144	36093	Schenectady	Schenectady County	06	H1	G4020	104	10580		A	529413660	12605632	+42.8175420	-074.043
17	Polygon	36	021	00974109	36021	Columbia	Columbia County	06	H1	G4020	104	26460		A	1643877983	35121459	+42.2477286	-073.626
18	Polygon	36	023	00974110	36023	Cortland	Cortland County	06	H1	G4020	296	18660		A	1291782569	7144298	+42.5938237	-076.076
19	Polygon	36	031	00974114	36031	Essex	Essex County	06	H1	G4020				A	4647068318	315980026	+44.1089711	-073.777
20	Polygon	36	079	00974138	36079	Putnam	Putnam County	06	H1	G4020	408	35620	35644	A	596501433	41282760	+41.4279035	-073.743
21	Polygon	36	057	00974127	36057	Montgomery	Montgomery County	06	H1	G4020	104	11220		A	1044165243	18948888	+42.8916891	-074.436
22	Polygon	36	099	00974147	36099	Seneca	Seneca County	06	H1	G4020	464	42900		A	838393308	172838989	+42.7822943	-076.827
23	Polygon	36	069	00974133	36069	Ontario	Ontario County	06	H1	G4020	464	40380		A	1667665893	47814139	+42.8566949	-077.303
24	Polygon	36	049	00974123	36049	Lewis	Lewis County	06	H1	G4020				A	3301398984	39520099	+43.7863965	-075.442
25	Polygon	36	113	00974154	36113	Warren	Warren County	06	H1	G4020	104	24020		A	2245394142	167449317	+43.5551052	-073.838

Fig.4.1: Shape file for New York

These are the steps followed to plot the map.

- Open the Arc GIS 10.1 and add the shapefile of New York

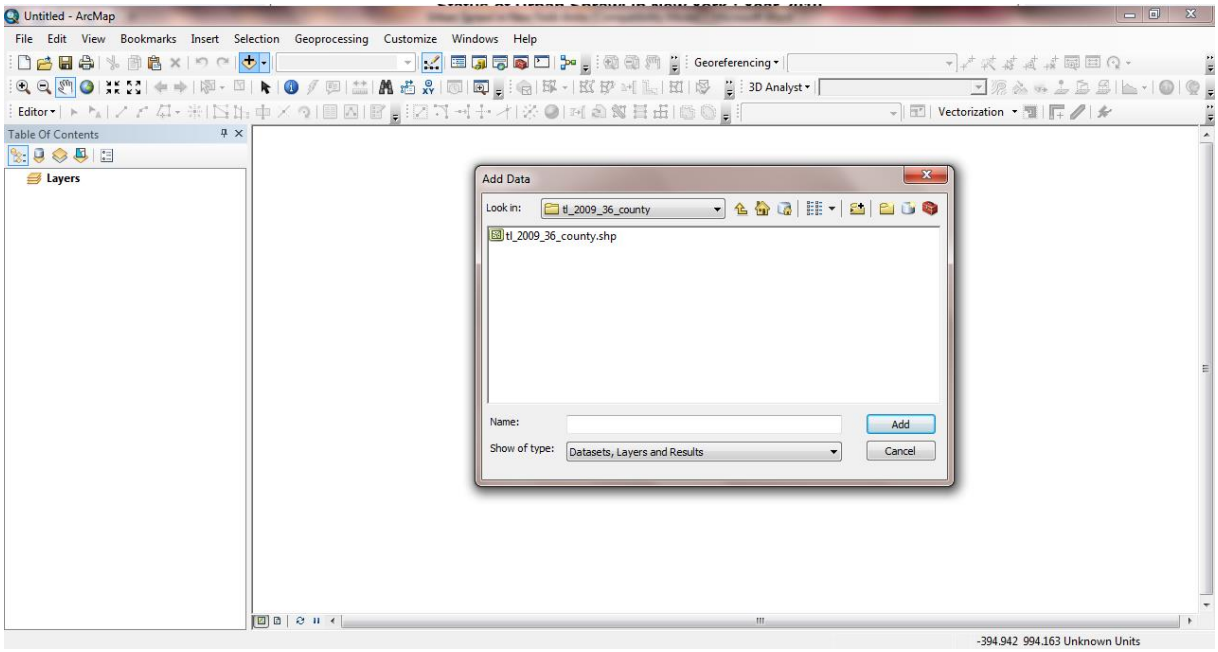


Fig.4.2: Opening shape file in Arc GIS desktop application

- This represents the blank county based map for New York state.

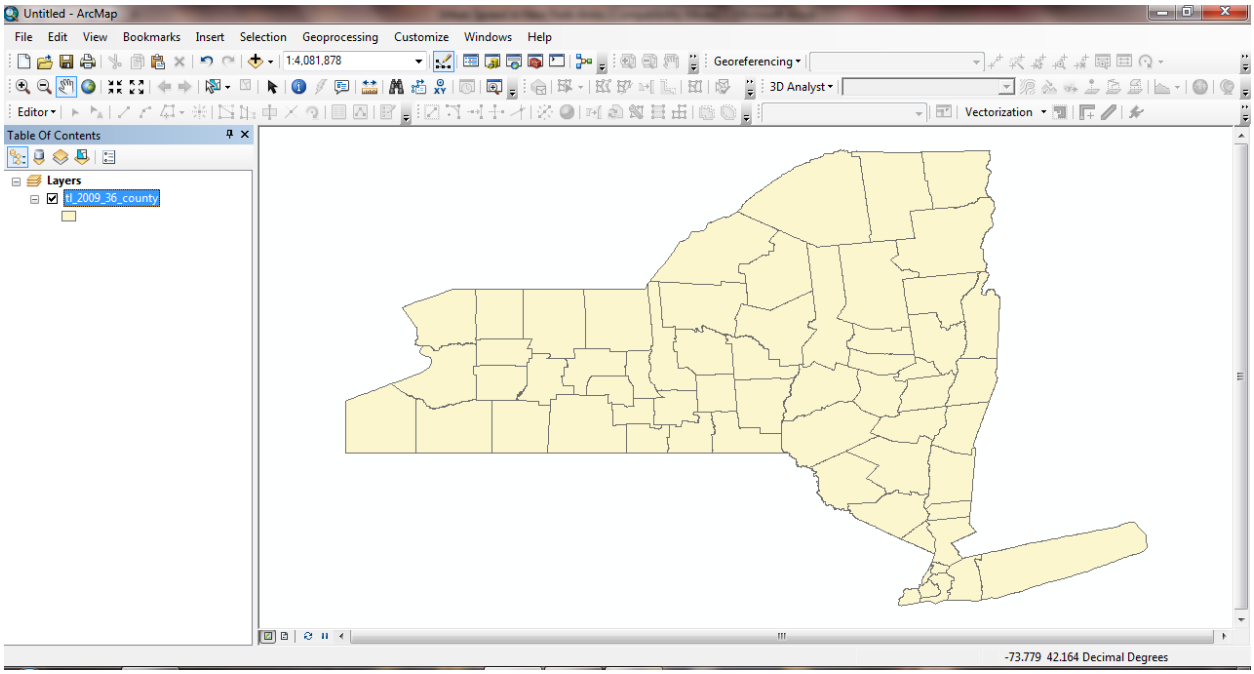


Fig.4.3: Blank map of New York

- Under properties counties are named using label option and selecting 'Name', from the attribute table

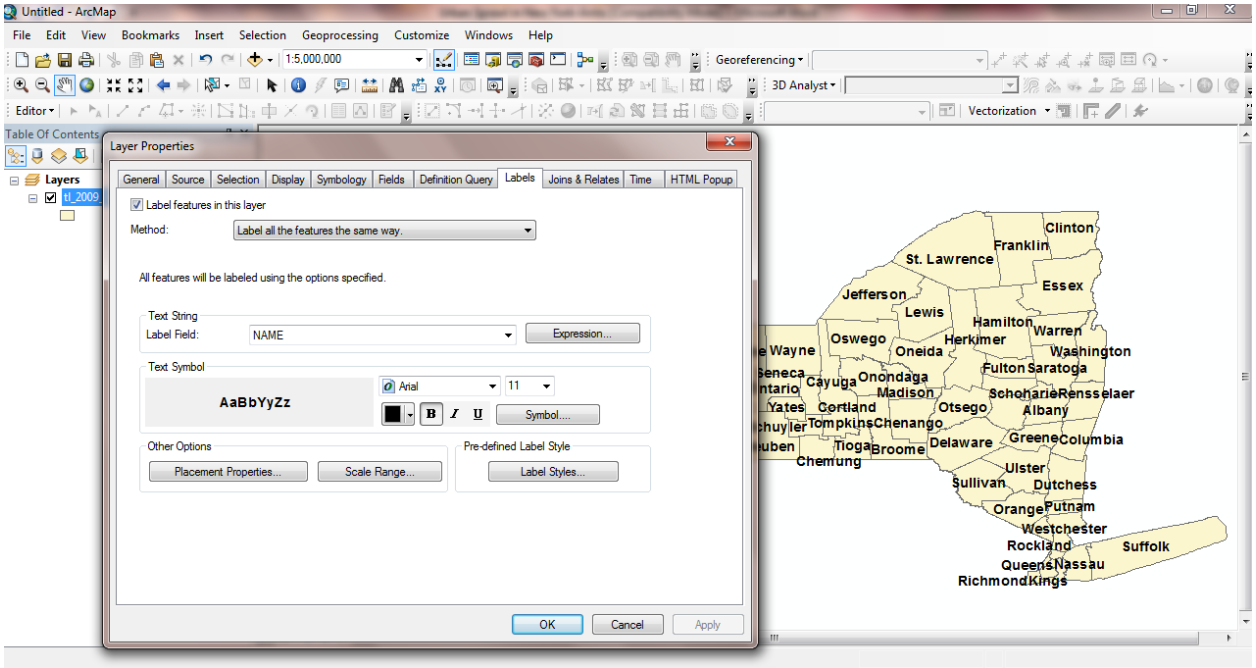


Fig.4.4: Naming the Counties

- Using Join option the targetted excel file is joined with the existing attribute table of the New York

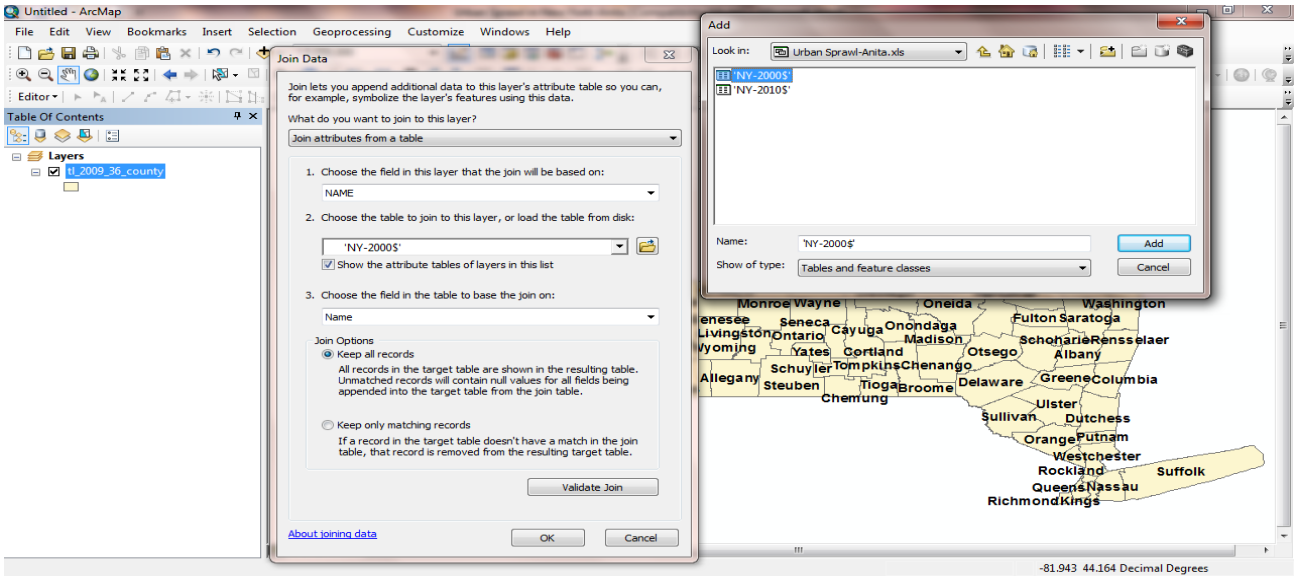


Fig.4.5: Joining data with shape file

- Two separate maps displaying Urban Sprawl in year 2000 and year 2010 are prepared

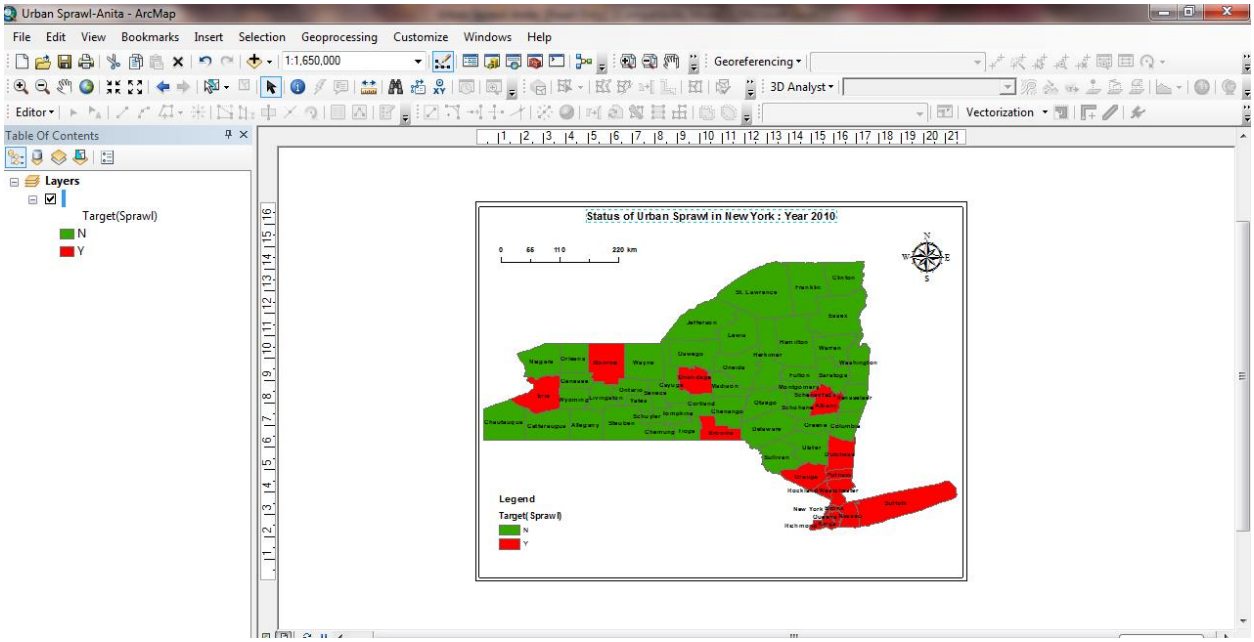


Fig.4.6: Urban sprawl for New York State for the year 2010

These are the maps prepared using ArcGIS. Red color indicates sprawl occurring counties and green color represents absence or less intensity sprawl. In the figure we can see that 5 more counties were affected with urban sprawl than from 2000. This is what exactly what we meant by suburban sprawl. Extending the sprawl to the nearby low density area. In the map for 2000 counties Putnam, Orange, Ditches are not in the list of sprawl affected county. But the nearby highly dense counties expansion affects these counties in 10 years (Fig.4.7)

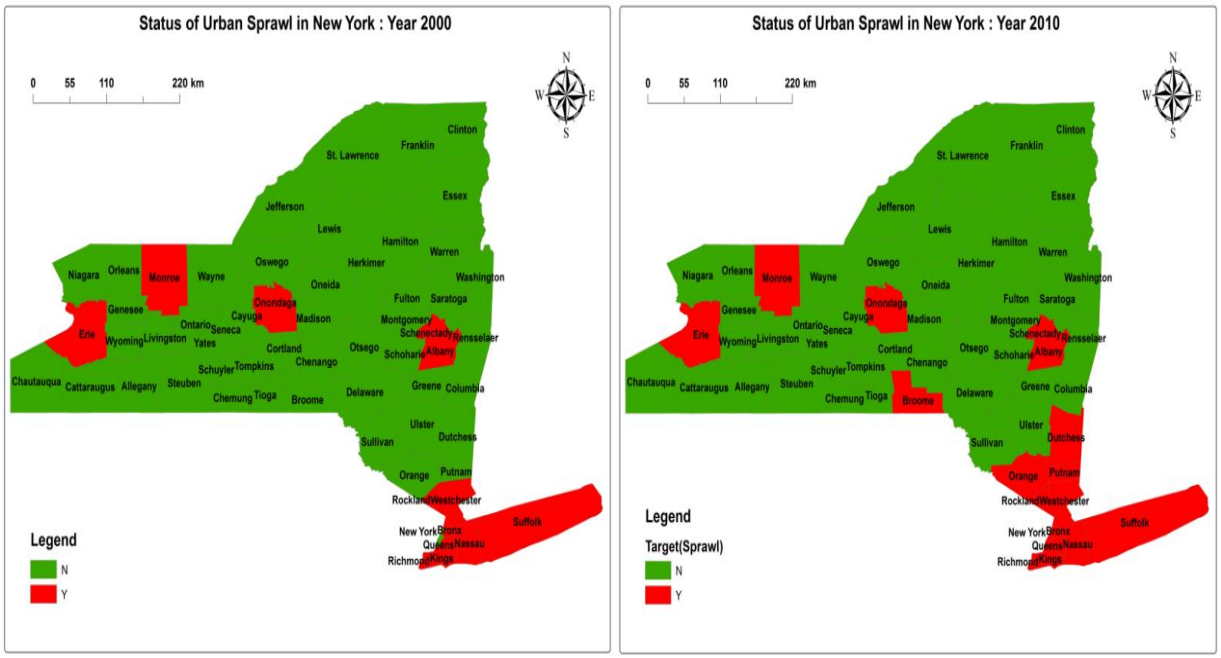


Fig.4.7: Maps representing the presence of sprawl for the years 2000 and 2010

5. Implementation and Evaluation

We discuss the implementation of our approach along with a summary of its experimental evaluation.

5.1 Implementation

5.1.1 Approach 1: Apriori for association rules using Java

5.1.1.1 Preprocessing of data

For running it in Apriori algorithm continuous data have to be converted to binary data. For that, first data was discretized, i.e.; converted the continuous data to nominal (categorical) data by arranging the continuous values into ranges for each variable. After that by using the data mining tool WEKA the ranged nominal data was converted to a binary file. The java code for Apriori was coded in such a way that the input file should be in .dat format file. So the binary output from WEKA was converted into .dat format files (Fig.5.1).

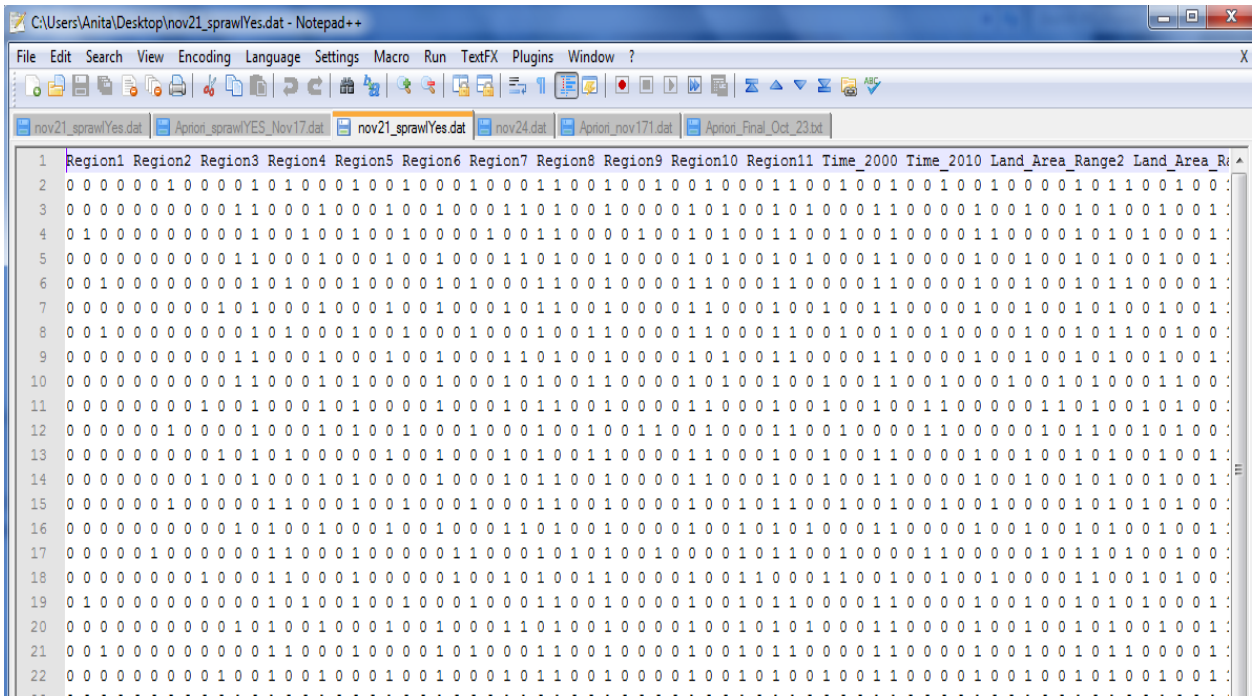


Fig.5.1: Binary Data

5.1.1.2 Implementation of Apriori Algorithm in Java

The overview of implementation of this approach is illustrated in (Fig. 5.2).
The algorithm involves two stages:

a. Identifying all item sets satisfying minimum support (frequent item set generation)
b. Identifying all rules meeting minimum confidence.

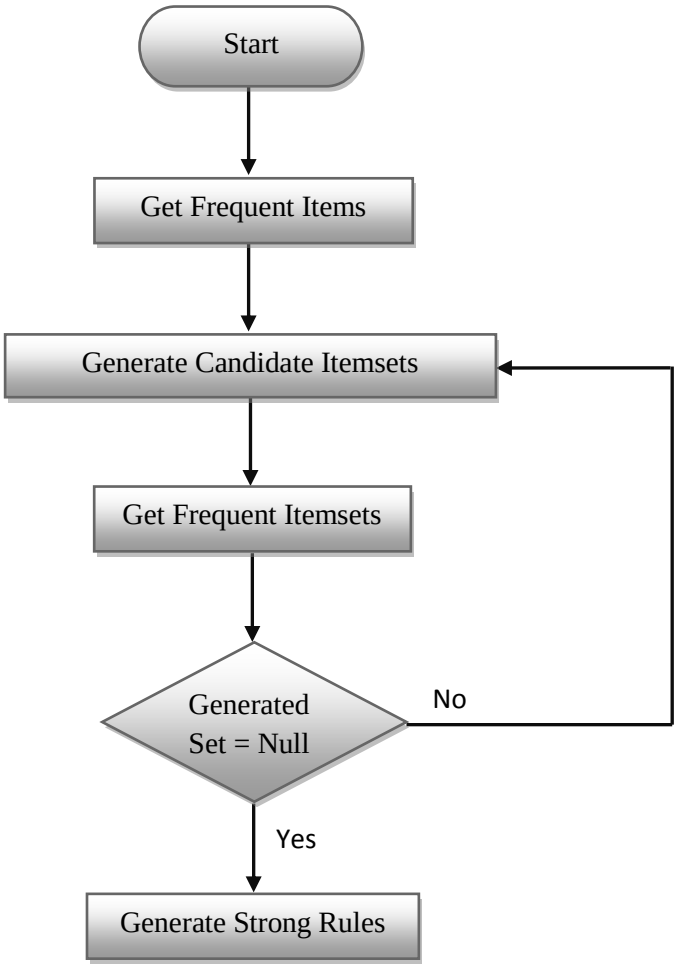


Fig.5.2: Flowchart for the Implementation of Apriori

Frequent item set generation

While identifying the item sets satisfying minimum support, the algorithm must ensure that the number of candidates generated and the number of comparisons involved are also minimal. Separating the algorithm into the following two stages ensures that both the stated requirements are satisfied simultaneously:

1. Candidate set generation

2. Frequent item set generation

Candidate item set generation mainly involves identification of probable candidates for the next iteration based on the frequent item set generated in the current iteration [26, 18]. The only exception in this case is the generation of the first frequent itemset where it relies only on the number of items involved in the transaction. This step reduces the number of candidates for which support has to be calculated [26].

Frequent itemset generation figures out the candidates meeting the minimum support. For this, a complete iteration of the transaction database is normally required for each candidate. The comparisons can be reduced by using advanced data structures like HashTree.

Essentials for implementing Apriori frequent item set generation in Java

1. Interface to connect to and read from the transaction database.
2. Interface to validate the transaction data.
3. Interface to transform the transaction data into optimized format.

Transaction data could reside on any data store. Hence, there should be an interface available that can connect to the data store irrespective of its type. This interface requires methods to connect to initialize the connections to the data store, read the next transaction data from the data store and close the open connections once the data is read. In addition it also should expose methods to identify the names of the items involved in the transaction. In the current implementation the transaction data is stored in a file. So a file reader implementation would be sufficient to meet this requirement. The input data which is stored in the form of an M x N matrix where M is the number of items and N is the number of transactions. Each row indicates a transaction. Each value in the row is either 0 or 1 indicating whether the item is involved in the transaction or not. This representation is a compressed form of representing the items involved in the transaction.

Transaction data which is read from the data store needs to be validated. The only validation that is carried out in the implementation is ensuring that the transaction data contains the required number of items.

The transaction data may not be stored in the appropriate format which can directly be used by the Apriori algorithm. So it needs to be transformed into the required format. In the implementation each row is converted to a String indicating the items involved in the transaction. Hence all the items with value 0 are discarded.

Apriori frequent item set generation using Java

For the implementation of Apriori in java, we follow an algorithm (Fig.5.3). Various support counting techniques are devised in data mining. However most of them require candidate items, number of items involved in transaction, method to increment the support count for a transaction and method to prune the candidates based on minimum support. Hence for the implementation of this step an abstract class is sufficient which would leave the concrete implementation of support counting and pruning to the implementing class. In the default implementation of support counting an advanced data structure called Hash Tree is used. Hash tree node is created by using the basic data structures like HashMap and List. In addition, it contains an attribute to indicate whether the node is an intermediate node or leaf node. If the node is an intermediate node then the HashMap would indicate an item and its children. If the node is a leaf node then the List would indicate the item set and its support. The support attribute is updated each time a transaction is considered for support counting. By using Hash tree we minimize the number of comparisons involved in support counting. Pruning of the candidate elements is carried out by traversing the hash tree to the leaf nodes and checking whether the item sets involved meet the minimum support. This is a pretty straight forward step. Traversing the hash tree during support counting and pruning is achieved using recursive algorithms.

Frequent itemset generation of the Apriori algorithm.

1: $k = 1$.
2: $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times minsup \}$. {Find all frequent 1-itemsets}
3: **repeat**
4: $k = k + 1$.
5: C_k = apriori-gen(F_{k-1}). {Generate candidate itemsets}
6: **for** each transaction $t \in T$ **do**
7: C_t = subset(C_k, t). {Identify all candidates that belong to t}
8: **for** each candidate itemset $c \in C_t$ **do**
9: $\sigma(c) = \sigma(c) + 1$. {Increment support count}
10: **end for**
11: **end for**
12: $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times minsup \}$. {Extract the frequent k-itemsets}
13: **until** $F_k = \emptyset$
14: Result = F_k.

Fig.5.3: Algorithm for frequent itemset generation [26]

Rule Generation in Apriori algorithm

In the second phase of Apriori algorithm all the association rules of the form X => Y are supposed to be identified out from the frequent item sets. Each such association rule should meet the minimum confidence level. For this phase of implementation we followed two algorithms

(Fig. 5.4, 5.5). In order to find the confidence level the support count for each item set in the antecedent part (X) and the support for (X U Y) need to be calculated. These measures are already calculated during the frequent item set generation stage. So no more iteration over the transaction data base is required during this phase. The algorithm starts by identifying all the one item consequents for each frequent item set which meets the minimum confidence. This implementation is straight forward. This involves iteration over all the frequent item sets, generation of each probable one item consequents, and calculation of confidence for each such one item consequent and comparison of the calculated confidence measure against the minimum confidence measure. After identifying each one item consequents all the association rules from the corresponding frequent item set needs to be identified. The consequent part of the association rules generated from the frequent item set will grow after each iteration. It is also interesting to notice that the rule generation from frequent n-itemset depends on frequent n+1-itemset. Hence, the n+1-item set is generated based on Apriori candidate generation algorithm. The code for candidate generation is reused for this purpose.

```
Rule generation of the Apriori algorithm.
1: for each frequent k-itemset f_k , k ≥ 2 do
2: H_1 = { i | i ∈ f_k }        { 1-item consequents of the rule. }
3: call ap-genrules(f_k , H_1 .)
4: end
```

Fig.5.4: Algorithm for rule generation [26]

```
Procedure ap-genrules(f_k , H_m).
1: k = |f_k |        { size of frequent itemset. }
2: m = |H_m |        { size of rule consequent. }
3: if k > m + 1 then
4: H_{m+1} = apriori-gen(H_m ).
5: for each h_{m+1} ∈ H_{m+1} do
6: conf = σ(f_k )/σ(f_k − h_{m+1} ).
7: if conf ≥ minconf then
8: output the rule (f_k − h_{m+1} ) ⟶ h_{m+1} .
9: else
10: delete h_{m+1} from H_{m+1} .
11: end if
12: end for
13: call ap-genrules(f_k , H_{m+1} .)
14: end if
```

Fig.5.5: Algorithm for rule generation [26]

Inputs to the Java implementation of Apriori algorithm

1. Number of items per transaction
2. Minimum support

3. Location of input file path
4. Minimum confidence

These parameters are passed as program arguments.

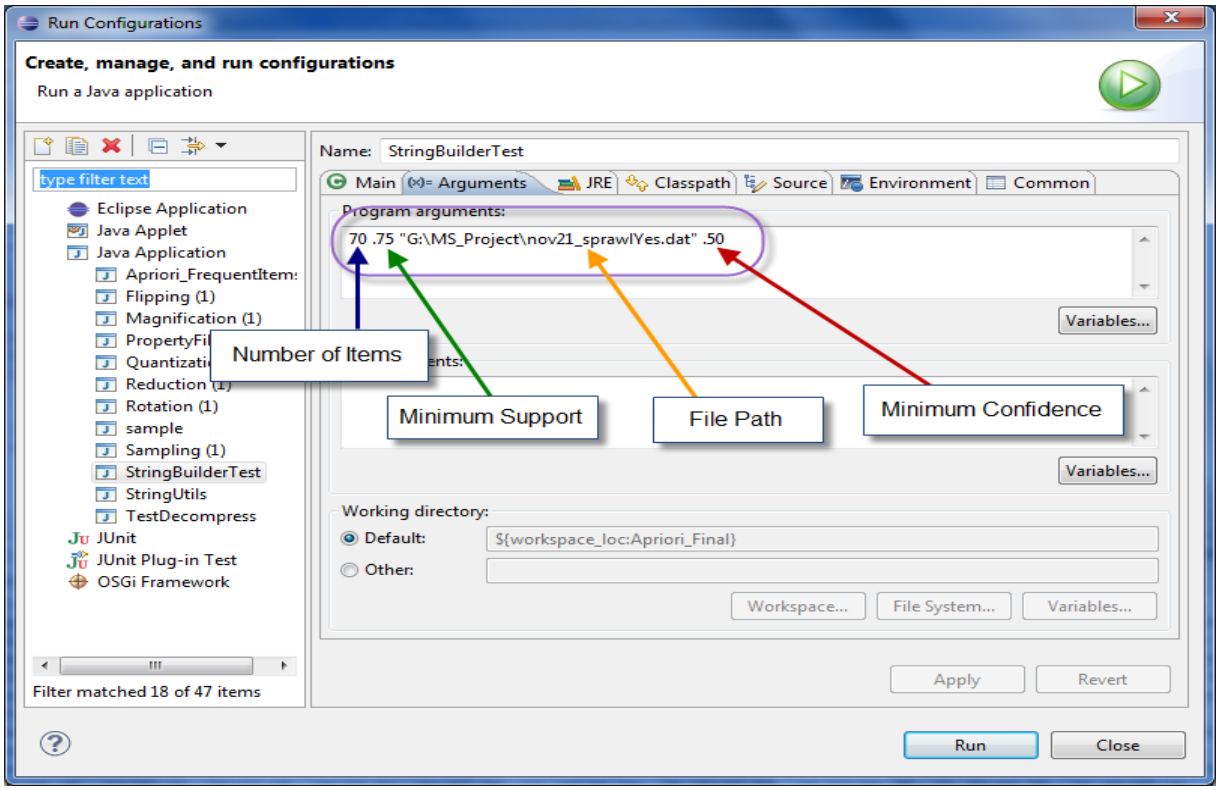


Fig.5.6: Run configuration

Output of the Apriori algorithm

The algorithm generates the frequent item sets satisfying the minimum support and association rules meeting the minimum confidence level (Fig.5.7). For example, when we give minimum support as 75 and minimum confidence as 50 along with our collected data for New York counties sprawl, we get a set of association rules (Fig.5.8).

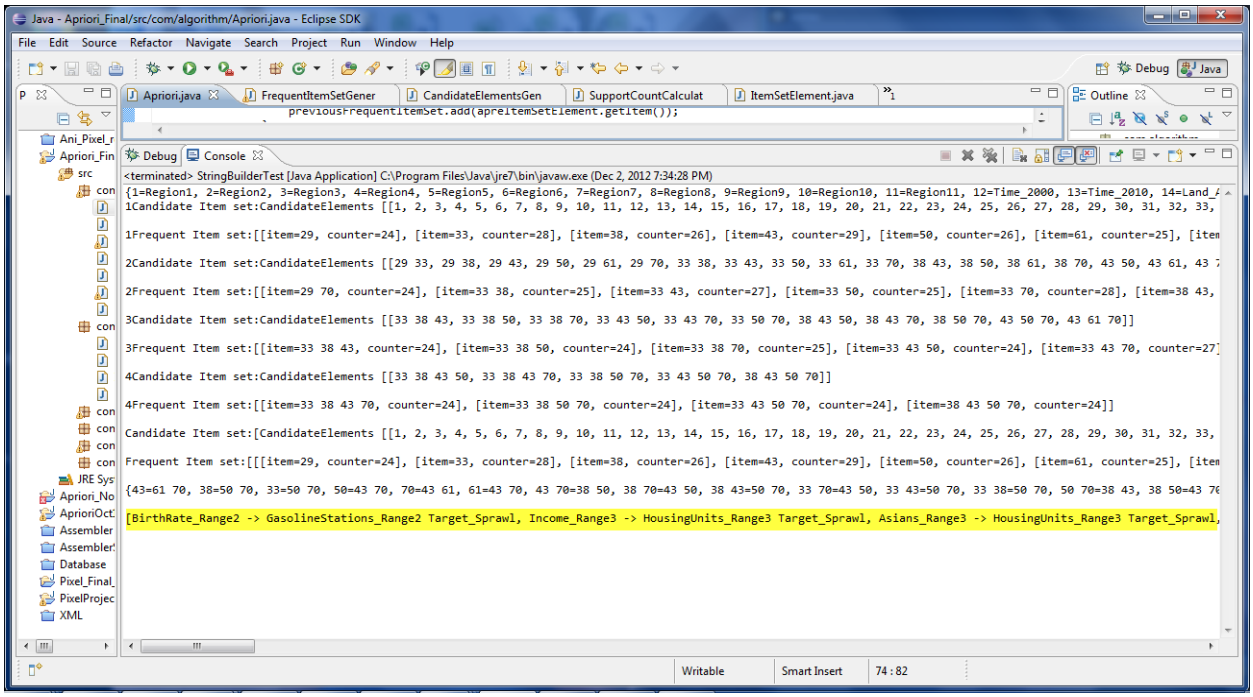


Fig.5.7: Output console while executing Apriori

[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]

Fig.5.8: Association rules from the Output of Apriori

5.1.2 Approach 2: J4.8 for decision tree classification using WEKA

5.1.2.1 Preprocessing of data

Data preprocessing is an important step in our experiments. The technique for extracting decisions trees is essentially based on WEKA [1] software. The method is general, but requires appropriate format for each file, which is readable by WEKA. WEKA can take .csv (comma separated value) and .arff (attribute-relation file format) files. So the collected data has to be preprocessed so as to make it compatible with WEKA formats. Whole variables contained in the dataset were continuous attributes except the target attribute which is binary and is given as input to WEKA. (Fig.5.9)

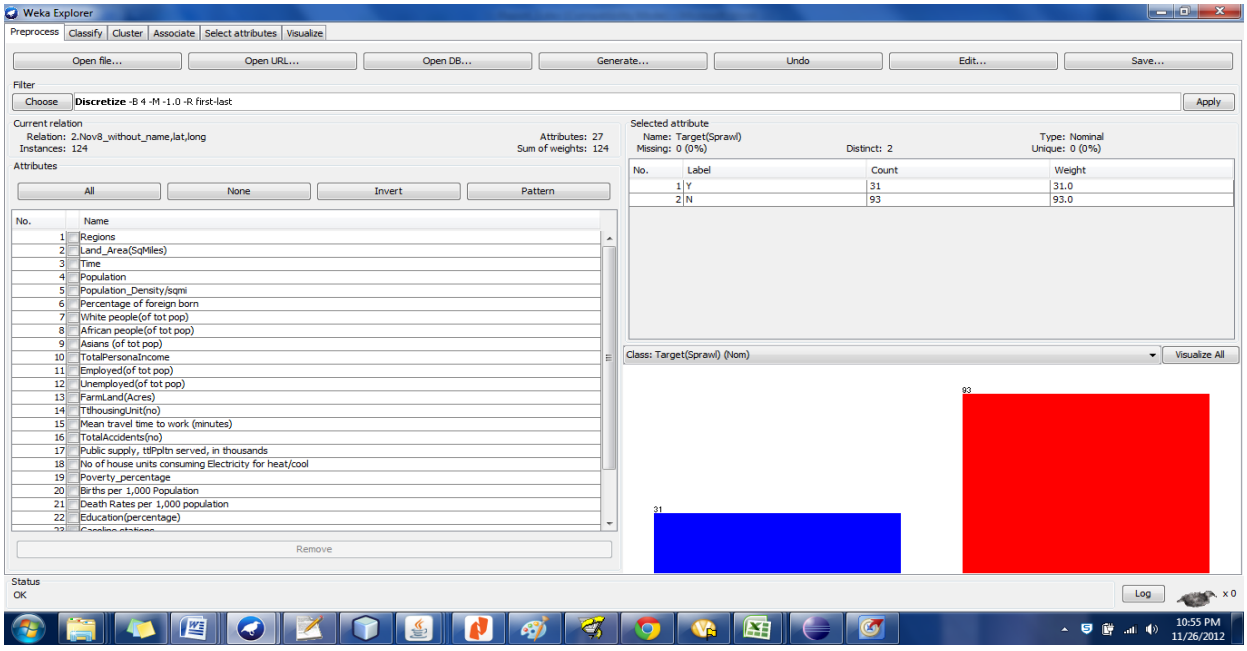


Fig.5.9: Data in WEKA

5.1.2.2 Experimentation

After preprocessing data we have started our experiments. The purpose is to gather decision patterns from the New York county dataset. We follow three different approaches to gather the patterns. J4.8 from classification tree algorithms is used to analyze the data for generating decision trees. And secondarily, bagging and boosting is also applied.

Among the 27 attributes, since the population density and percentage of Asians have most impact on sprawl and had comparatively very high impact when compared with other variables, only those patterns were displayed in the result (Fig5.10). Bagging and boosting [26] are applied on the data in order to get other hidden relations (Fig.5.13, 5.14), even though it is not that

significant as population density or percentage of Asians. Attribute selection is also done since we realize some of the variables doesn't have much impact on the target attribute, this gave us more results with more patterns (Fig.5.11). Data is input to this algorithm in both ways: as continuous attributes and nominal attributes (ranged data). After converting continuous data to nominal data not only sprawl but we tried some other variables also as target variable since they are also discrete attributes. For example, considering mean time to travel as the target attribute (Fig.5.12).

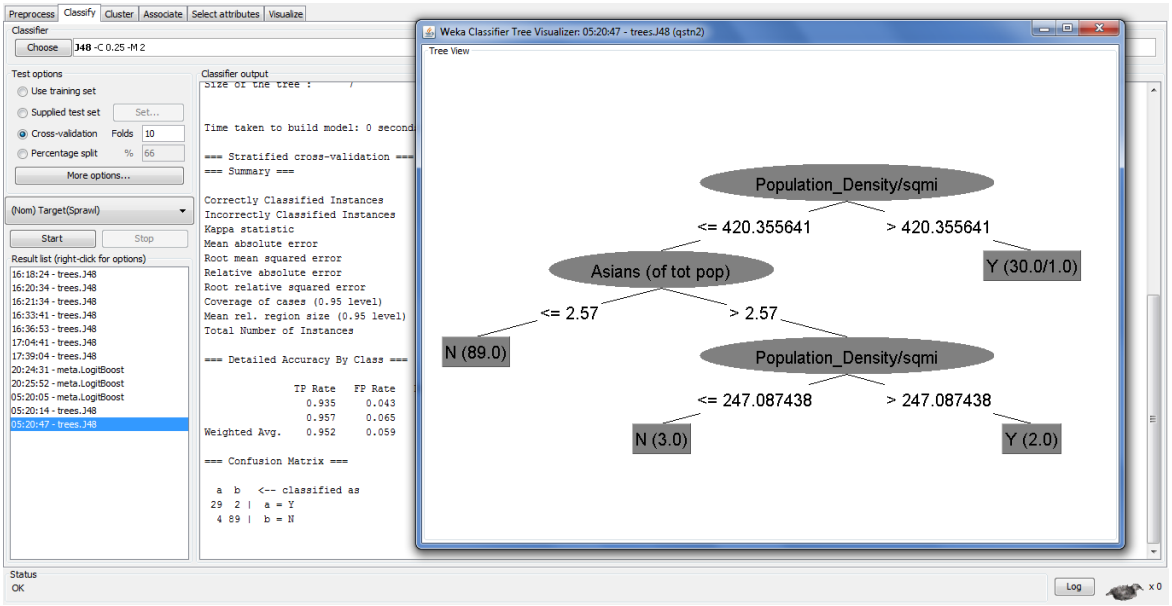


Fig.5.10: Initial output from J4.8

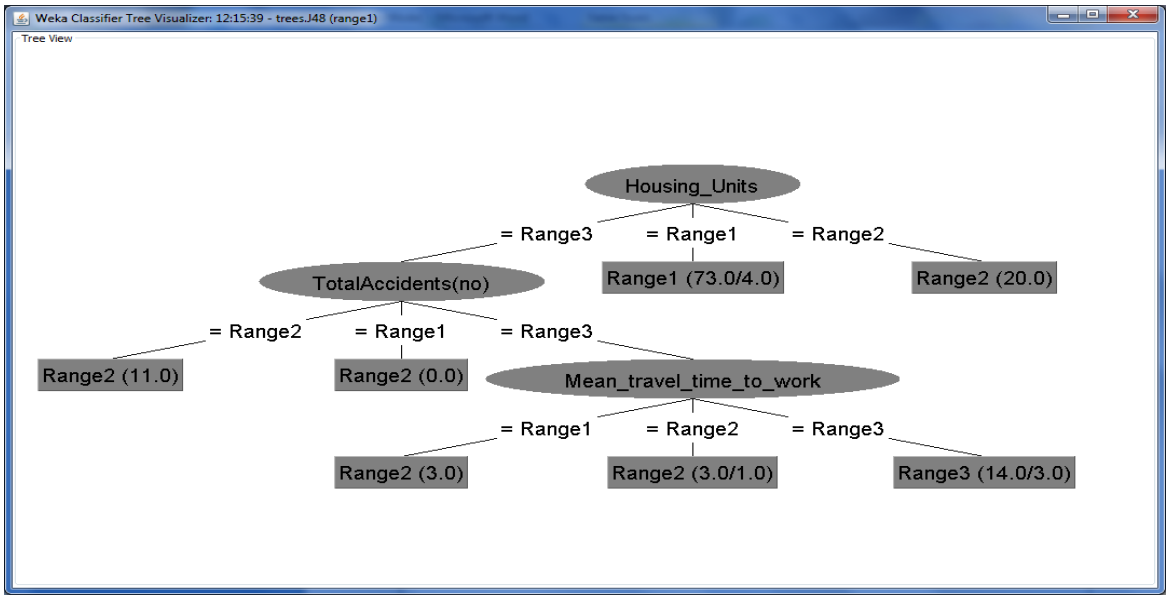


Fig.5.11: Output after attribute selection

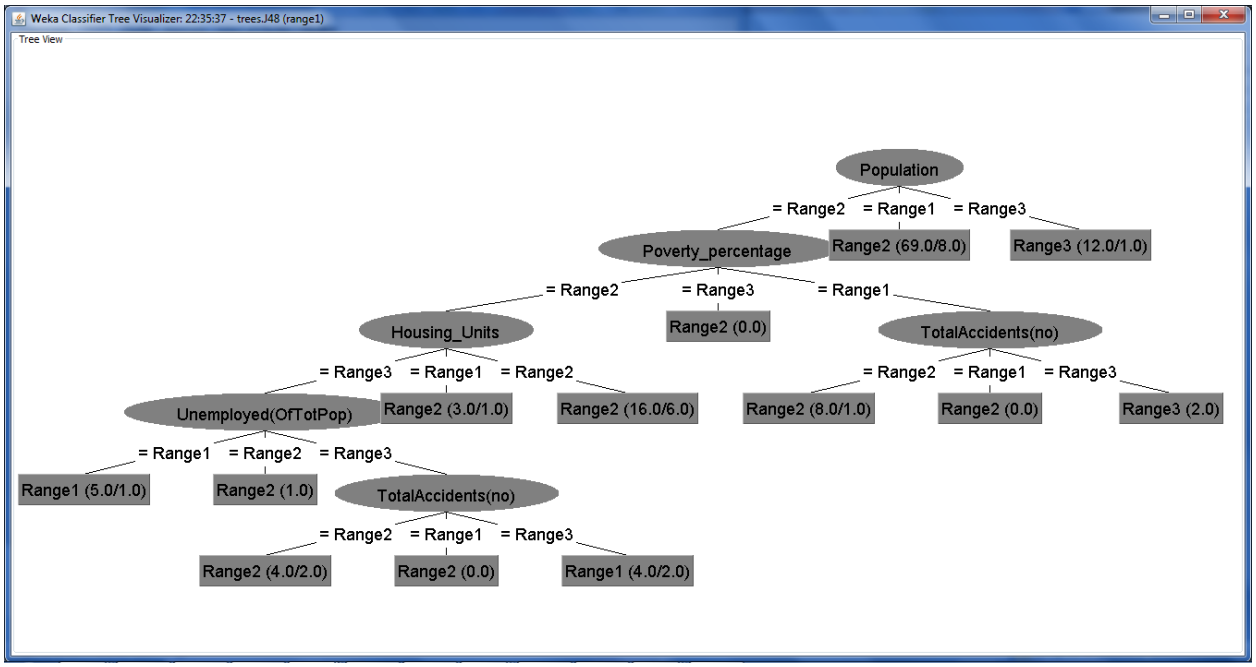


Fig.5.12: Mean Travel time to work

Iteration 7
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
Iteration 8
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
Iteration 9
        Class 1 (Target(Sprawl)=Y)
Decision Stump
Classifications
White people(of tot pop) <= 87.79499999999999 : 0.32519636604214447
White people(of tot pop) > 87.79499999999999 : -1.1231869975522928
White people(of tot pop) is missing : 0.011675003656158107
        Class 2 (Target(Sprawl)=N)
Decision Stump
Classifications
White people(of tot pop) <= 87.79499999999999 : -0.32519636604214475
White people(of tot pop) > 87.79499999999999 : 1.123186997552293
White people(of tot pop) is missing : -0.01167500365615946
```

Fig.5.13: Output after boosting

```
=== Classifier model (full training set) ===
All the base classifiers:
REPTree
============
TotalPersonaIncome < 11713160
|   Employed(of tot pop) < 18.88 : Y (2/0) [1/0]
|   Employed(of tot pop) >= 18.88 : N (61/1) [30/0]
TotalPersonaIncome >= 11713160 : Y (19/0) [11/0]
Size of the tree : 5
REPTree
============
African people(of tot pop) < 6.47 : N (58/0) [29/0]
African people(of tot pop) >= 6.47
|   FarmLand(Acres) < 74.6 : Y (21/1) [12/1]
|   FarmLand(Acres) >= 74.6 : N (3/0) [1/0]
Size of the tree : 5
REPTree
============
```

```
African people(of tot pop) < 6.31 : N (60/1) [32/2]
African people(of tot pop) >= 6.31
|   TotalPersonaIncome < 3520698.5 : N (2/0) [2/0]
|   TotalPersonaIncome >= 3520698.5 : Y (20/1) [8/0]
Size of the tree : 5
REPTree
============
TotalPersonaIncome < 11286795 : N (65/3) [33/1]
TotalPersonaIncome >= 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
|   |   TotalPersonaIncome < 373943200 : Y (15/0) [5/0]
|   |   TotalPersonaIncome >= 373943200 : N (2/0) [1/0]
|   FarmLand(Acres) >= 44.7 : N (9/2) [9/3]
Size of the tree : 7
REPTree
============
TotalPersonaIncome < 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]
TotalPersonaIncome >= 10641129 : Y (19/0) [10/1]
Size of the tree : 7
REPTree
============
Public supply, ttlPpltn served, in thousands < 128.98 : N (58/0) [31/1]
Public supply, ttlPpltn served, in thousands >= 128.98
|   White people(of tot pop) < 88.99 : Y (21/3) [9/1]
|   White people(of tot pop) >= 88.99 : N (3/0) [2/0]
Size of the tree : 5
```

Fig.5.14: Output after Bagging

5.2 Experimental Evaluation and Results

After running data using the two approaches Apriori and J4.8, the models which are received as outputs are saved. From the output models we realize that, population density is the major cause or has most impact on urban sprawl and because of the increase in population, many other factors increased. Other than the common known facts like population density and farm lands we are able to find some interesting results like how the number of housing units, number of truck transportations and number of gasoline stations are related in a place which has sprawl.

Based on these saved models a mini user interactive prototype application is developed which consists of ten pieces of information. This application can be used by decision makers to know about the chances of sprawl occurrence or the significance of a variable or variables based on another variable or variables. For example, a city planner wants to construct a new flat and the population he expects to accommodate in that flat can be added to the present population of that particular county and can see how it would affect other variables or thereby sprawl.

This application is developed for the non-expert decision makers so that they can utilize the application without nerve-racking about how the patterns are made. From the results we receive, we select the ten best patterns and constructed the prototype based on that.

5.2.1 SDSS prototype

Prototype consists of ten user interactive questions. It has a description about itself, for the first time users, which is given under 'for new user information' button (Fig.5.15).

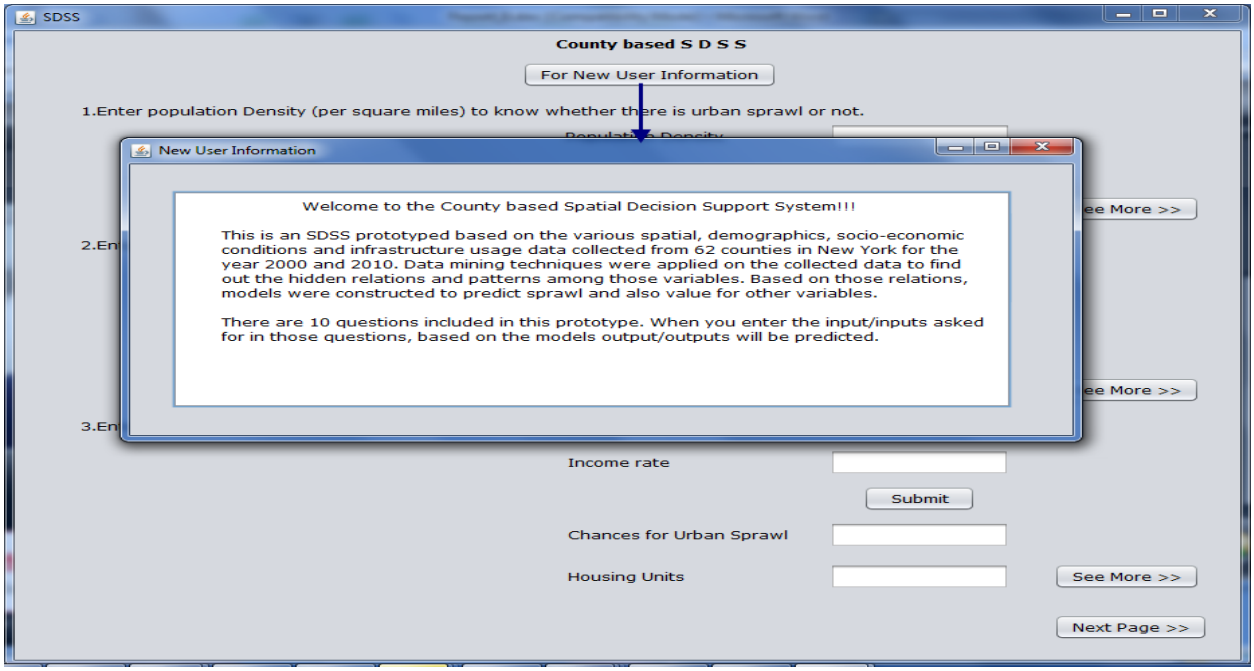


Fig.5.15: SDSS prototype: New User Information

1. Relation between population density (per square miles) and sprawl occurrence.

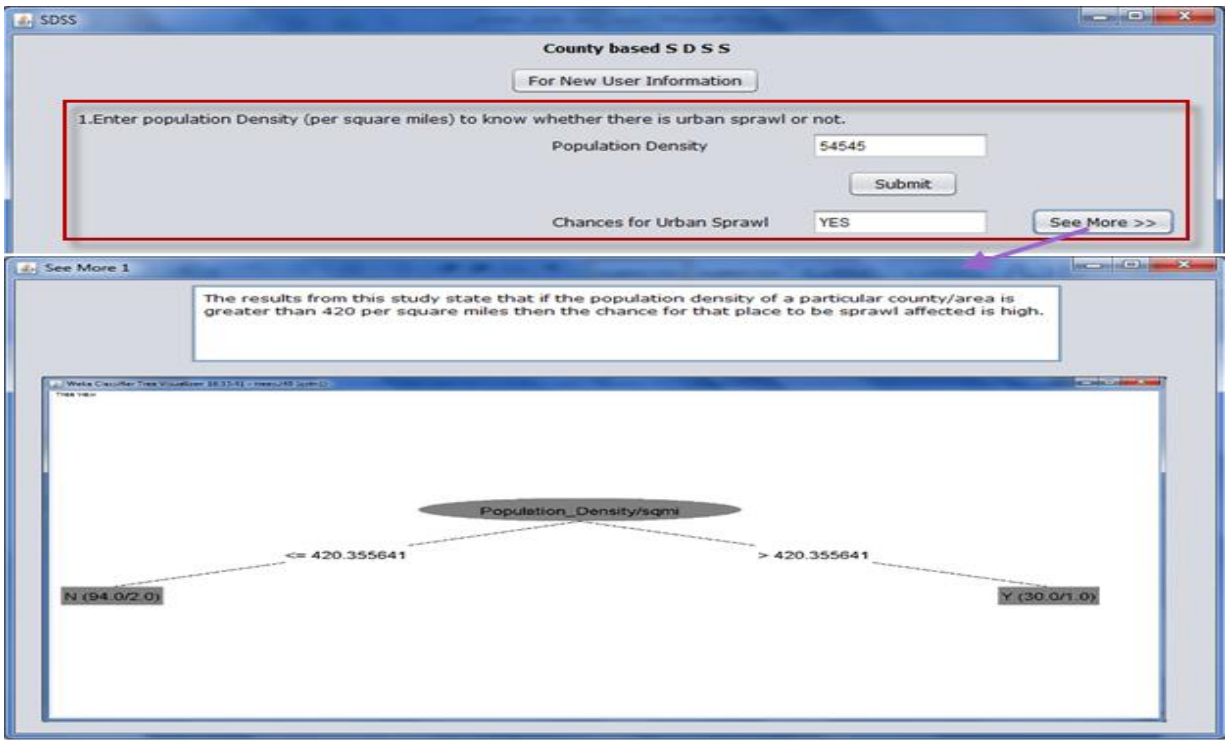


Fig.5.16: Relation 1

2. Relation between population density, percentage of Asians and sprawl occurrence.

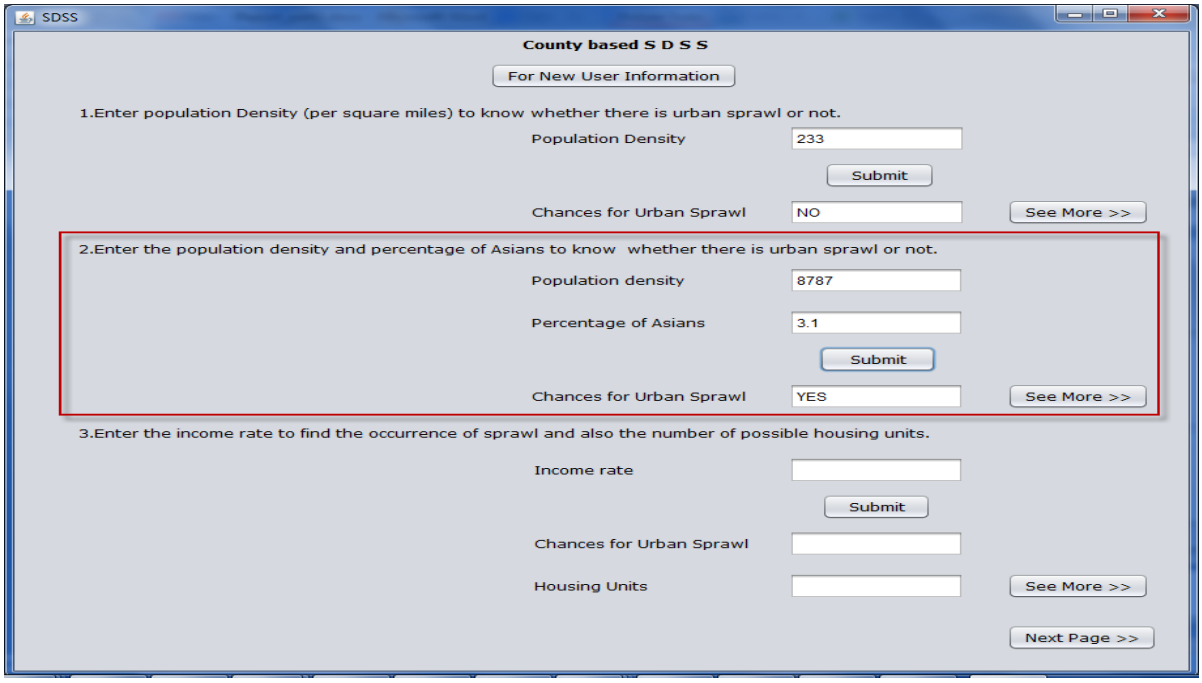


Fig.5.17: Relation 2

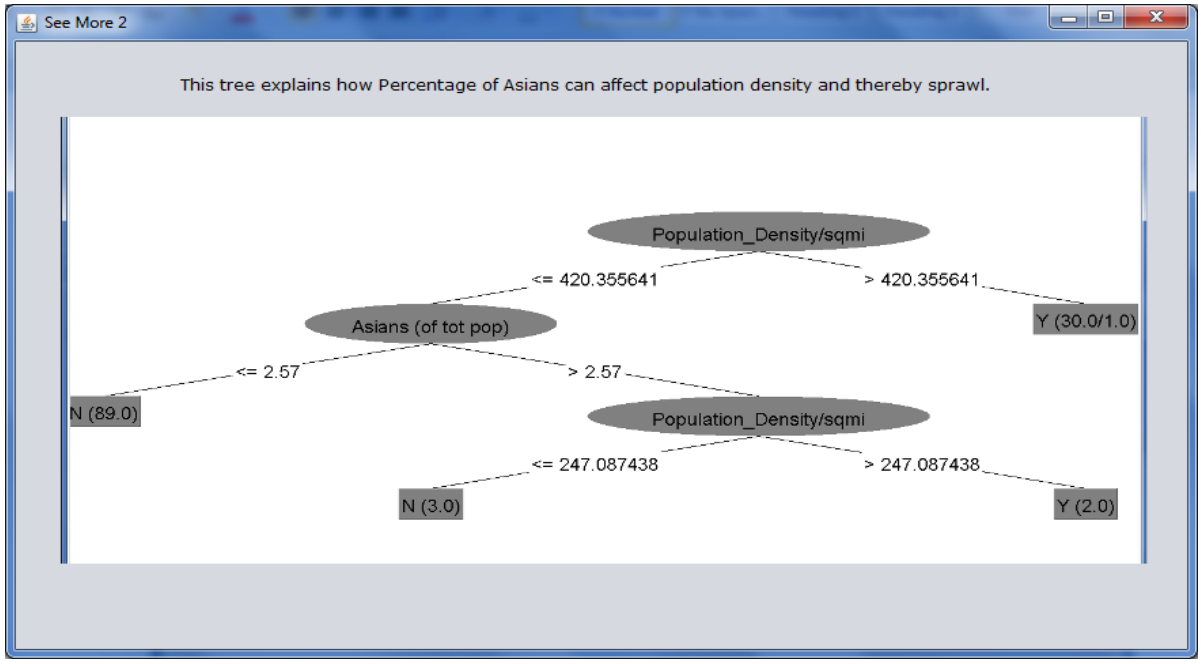


Fig.5.18: See more for Relation 2

3. Relation between Income rate and sprawl.

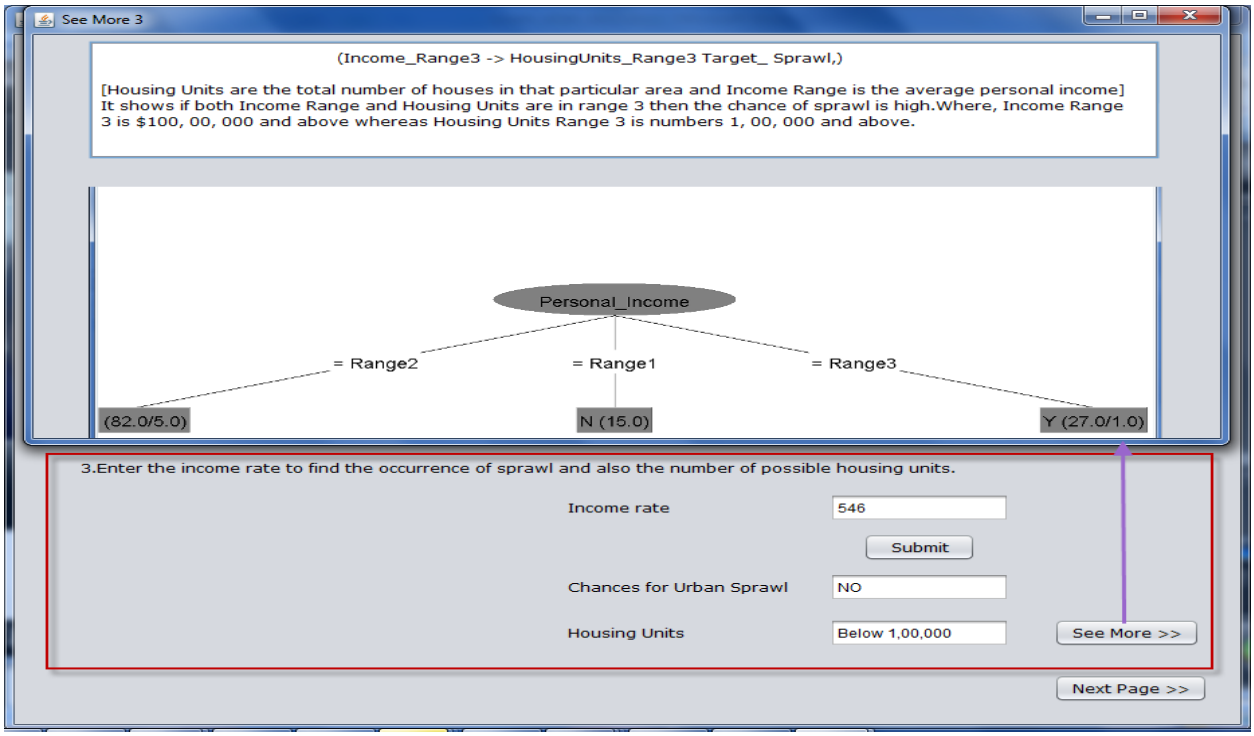


Fig.5.19: Relation 3

4. Relations among percentage of Asians, birth rate, number of housing units, and sprawl.

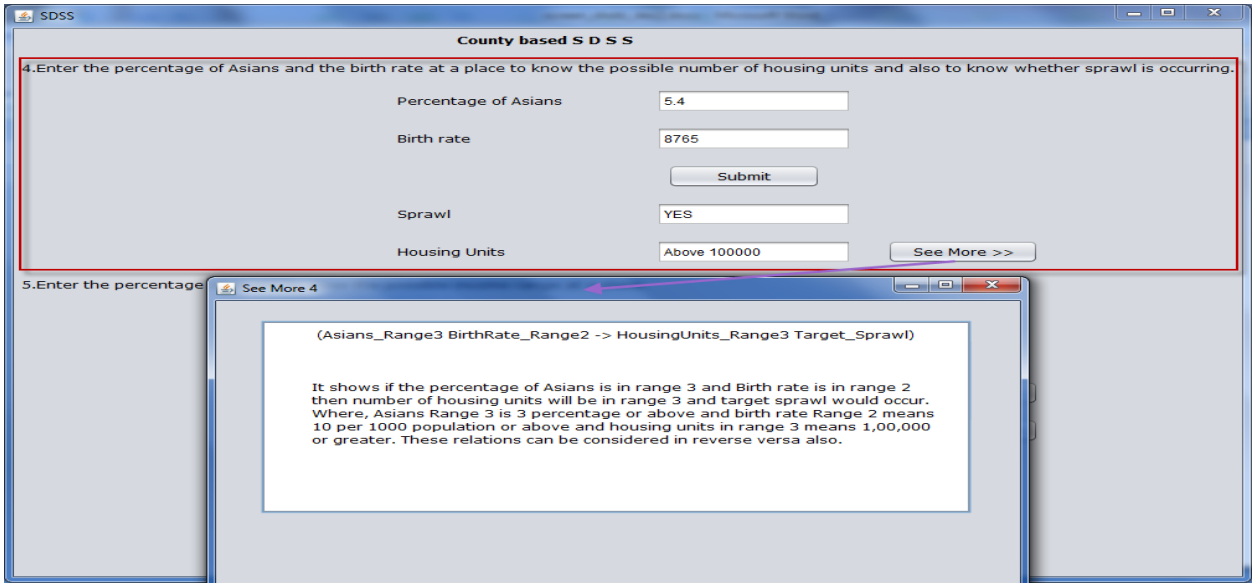


Fig.5.20: Relation 4

5. Relation between percentage of Asians and income range.

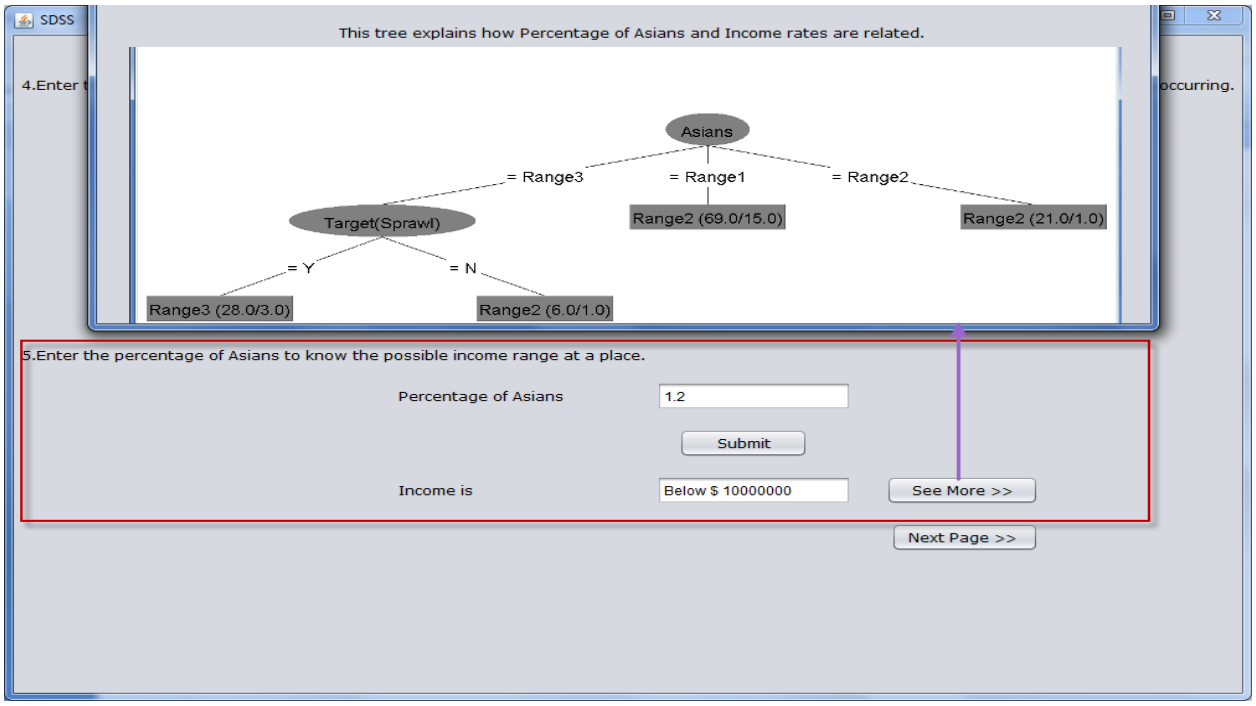


Fig.5.21: Relation 5

6. Relations among the number of housing units, number of trucks used for transportation, number of gasoline stations and sprawl.

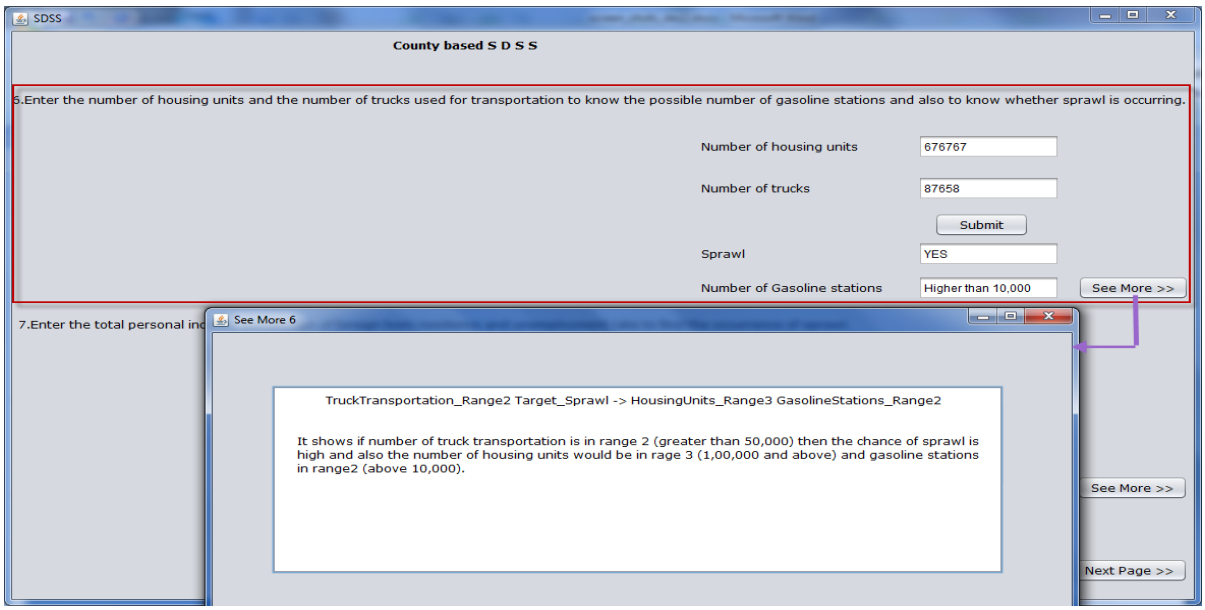


Fig.5.22: Relation 6

7. Relations among the personal income, percentage of foreign born residence, unemployment rate and sprawl.

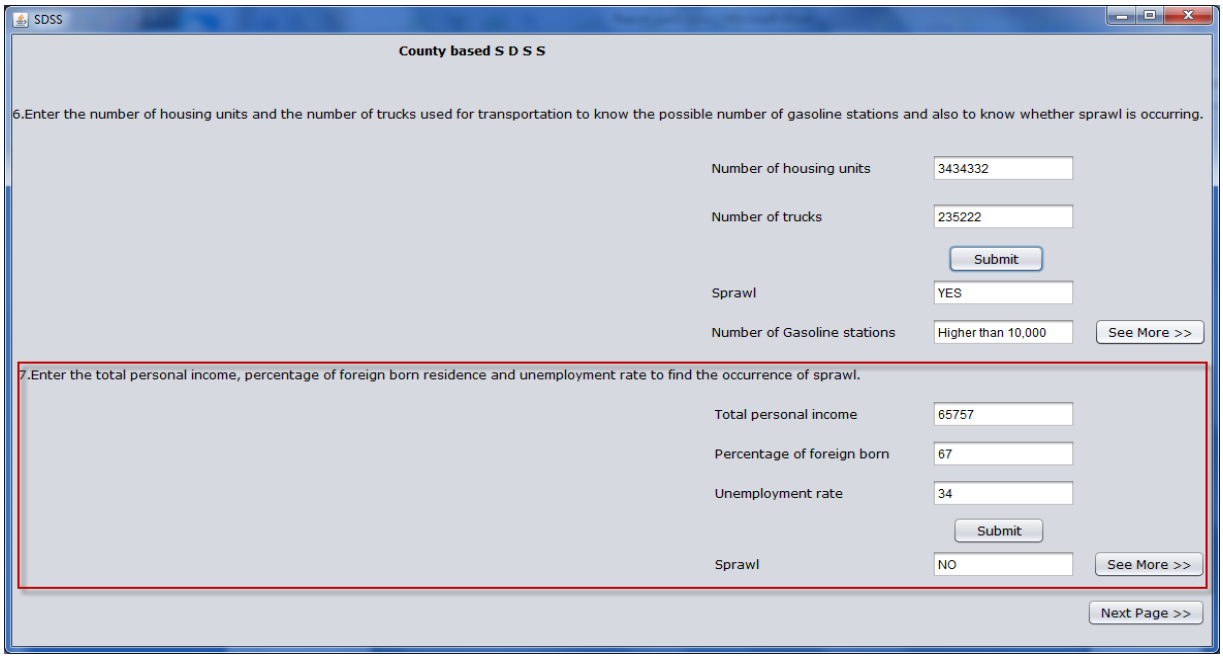


Fig.5.23: Relation 7

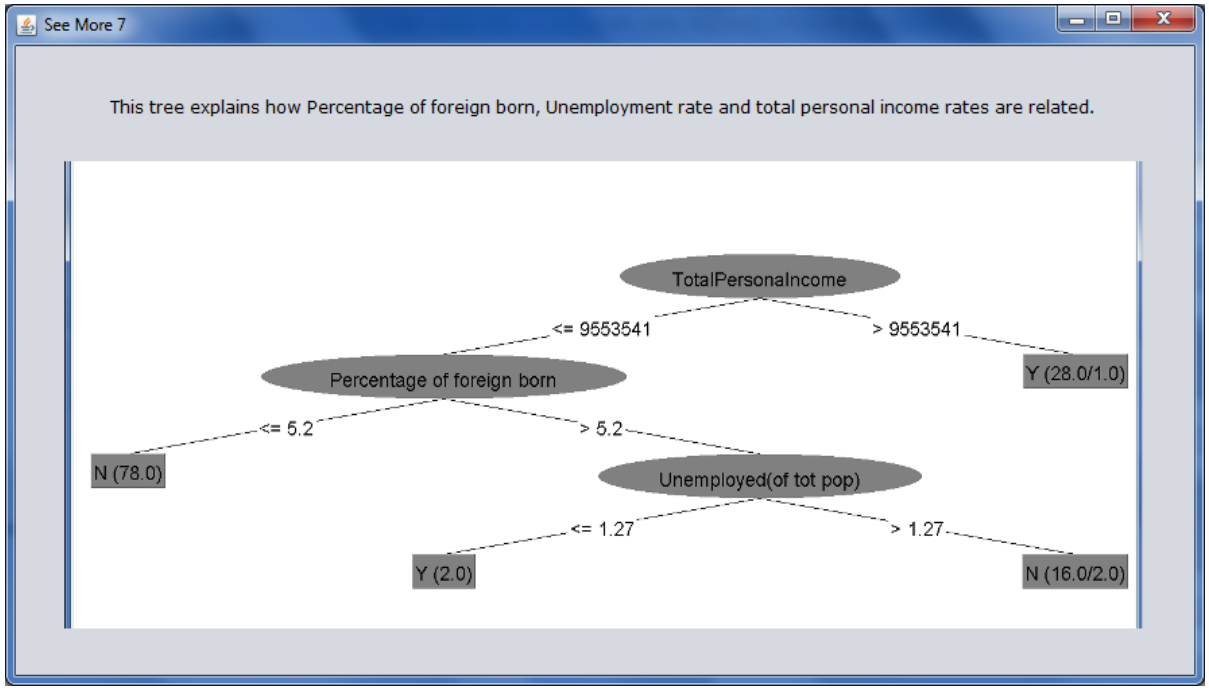


Fig.5.24: See more for Relation 7

8. Relation between employment rate and sprawl.

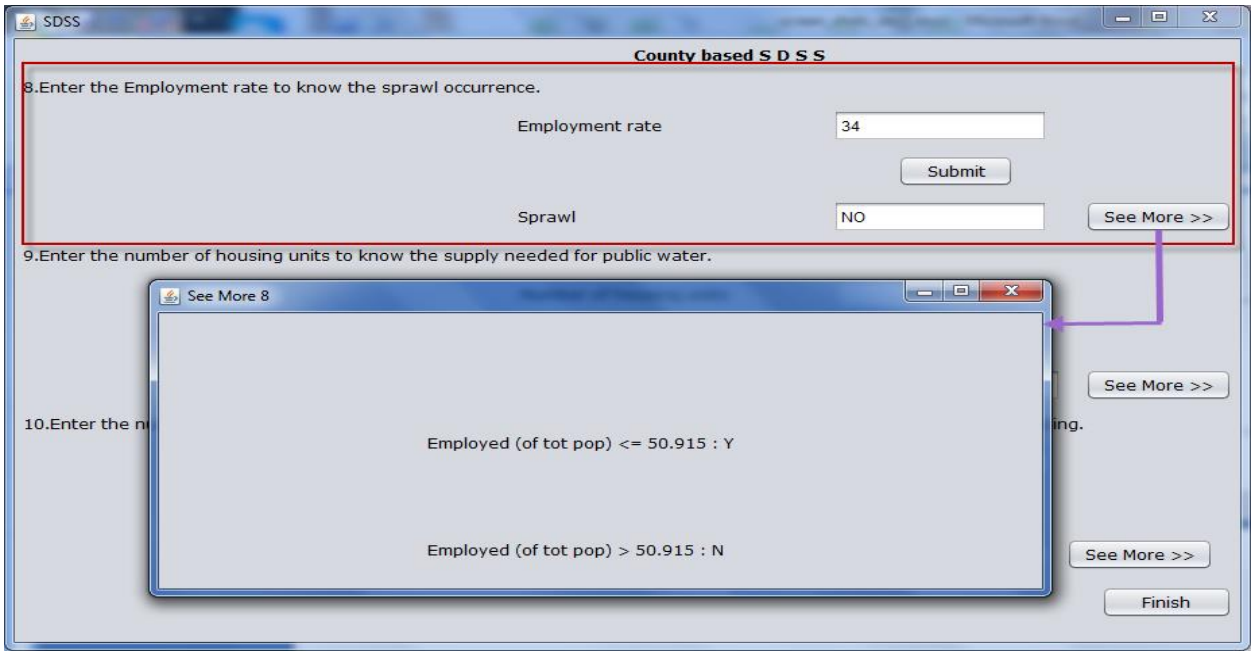


Fig.5.25: Relation 8

9. Relation between the number of housing units and the supply needed for public water.

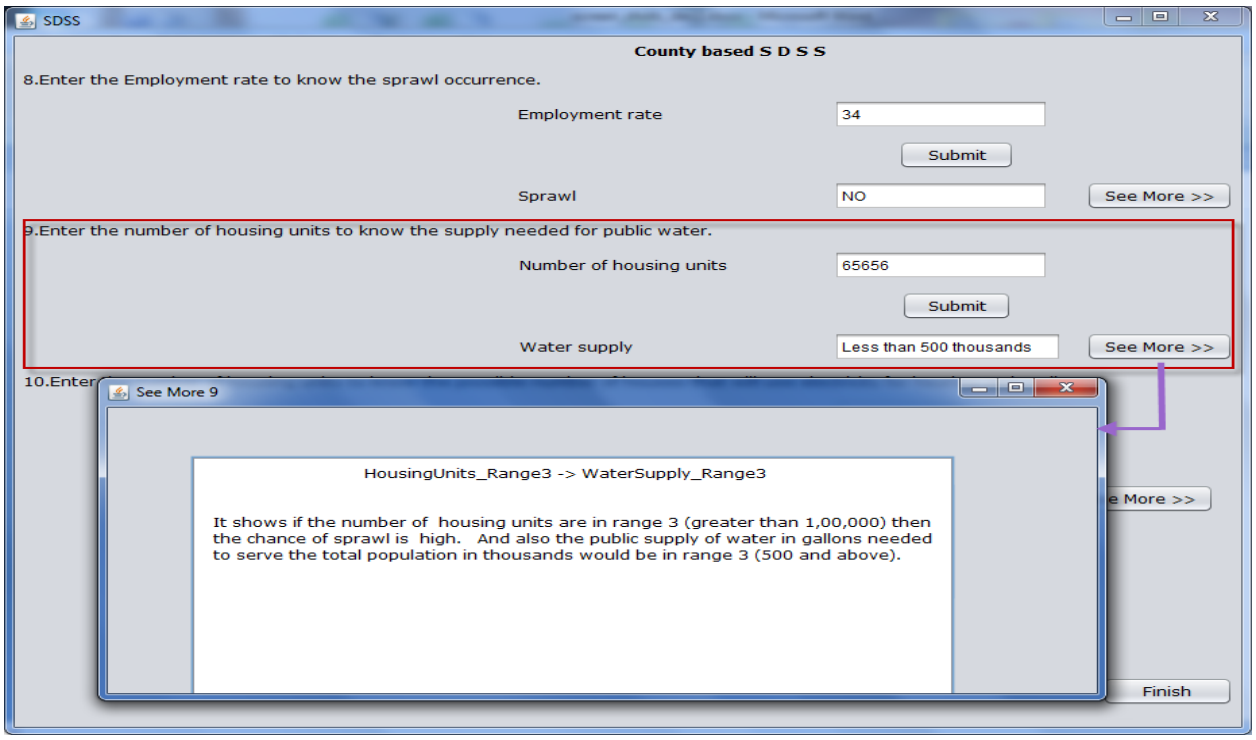


Fig.5.26: Relation 9

10. Relation between the number of housing units and the possible number of houses that will use electricity for heating and cooling among them.

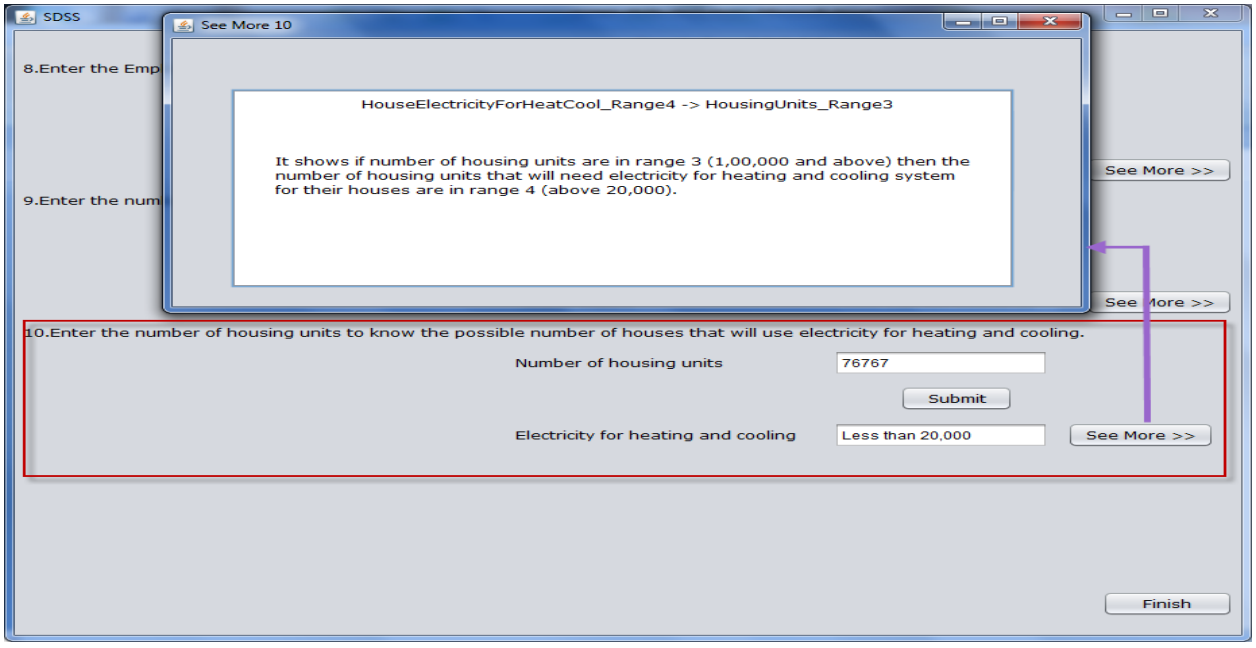


Fig.5.25: Relation 10

6. Related Work

In recent years technology is developing in an exponential manner. 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 their effort to topics like urban sprawl and sustainability, which is now a big concern among geologists, environmentalists and city planners. We would like to cite some of the related works here.

Measuring urban sprawl in Beijing with geo-spatial indices

In Jiang el at., 2007 they took Beijing as a case 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 [14]. So basically the authors were trying to measure sprawl or rate the sprawl based on the data they collected but in our case we were trying to find the patterns which caused sprawl.

Urban sprawl: Metrics dynamics and modeling using GIS

In the study done by Sudhira el at., 2003, the authors tried to find sprawl patterns using remote sensing spatial data along with other attributes. The study area was India, and their study attempted to identify sprawls, quantify by defining new metrics, understand the dynamic process and subsequently model the same to predict for the future. The authors in this study have used statistical analysis and methods to find the patterns, whereas we are using data mining algorithms to find the sprawl causing patterns.

Modeling Urban Land Use Change and Urban Sprawl: Calgary, Alberta, Canada

The authors of Sun el at., 2007 have implemented land use classification for the City of Calgary, Alberta, Canada, using an object-oriented approach and simulates the land use pattern in the future using Markov Chain analysis and Cellular Automata analysis based on the interactions between these land uses and the transportation network. This research proves that an object-oriented approach can produce satisfactory classification results whereas in our project we analyze data using strong data mining techniques. Also, this project reveals the manner in which land use is likely to develop in the future, and demonstrates that urban sprawl continued to grow in Calgary during the years between 1985 and 2001; similarly we explained urban sprawl using maps for 2000 and 2010 years.

7. Future Studies

- The increasing interest and popularity of data mining techniques in GIS data and its uses has led some geologists and data mining professionals to introduce Apriori in map reduce [18] frame work and to explore the huge datasets which can be of huge sizes as terabytes and petabytes.
- Developing a full-fledged SDSS considering various aspects of causes of sprawl with several decision-making scenarios that would head towards performing the forecasting of urban land use dynamics.

8. Conclusion

We have addressed the issue of discovering the inherent relations and patterns among some of the causes of urban sprawl, based on New York counties' data. i.e., by gathering data for some variables which are directly or indirectly related to urban sprawl and by implementing data mining algorithm Apriori for association rule mining using Java and by experimenting with J4.8 for decision tree classification using WEKA (a data mining tool). Finally, based on those results we implemented a mini user interactive prototype using Java swings and converted the same to an executable .exe file which can be used by urban planners, city dwellers and also any users who would like to find out the chance of sprawl occurrence by entering some of the variables, and also to see how some of the variables can affect each other.

In conclusion, we can say this project on the whole contribute to both the computing community of spatial data mining, and the geosciences community of urban sustainable development. Since this field of data mining in GIS and Spatial Decision Support Systems using GIS data are all new, this project involved a great deal of exploratory work. With the help of this project, we were able to familiarize with many concepts pertaining to various data mining techniques and GIS software.

9. Acknowledgements

I would like to extend my heartiest gratitude to all of those persons who were very supportive and helpful to improve the quality of my work. I would like to thank Dr. Aparna Varde, my Project Advisor for her constant support and help. I received proper guidance at every step of this project. All the materials, books and web sites that she provided were very appropriate for my project.

I would like to thank Dr. Danlin Yu, Associate Professor in the Department of Earth and Environmental Studies, Amy Ferdinand, Director of Department of Environmental Health and Safety and Dr. Sushant K. Singh, Graduate Research Assistant in Department of Earth and Environmental Studies at Montclair State University. Their invaluable guidance and support throughout this project has been an immense help to us. In general, I would like to thank the faculty, staff and students in Department of Computer Science and Department of Earth and Environmental Studies at Montclair State University for their cooperation.

10. References

(All websites accessed till December 3rd)

[1] Aksenova Svetlana S., *Aksenova Machine Learning with WEKA*. WEKA Explorer Tutorial, 2004.

[2] Brueckner Jan K..*Urban Sprawl: Diagnosis and Remedies.* International Regional Science Review 23, 2: 160–171 (April 2000)

[3] Choi Yoon-Seok, Moon Byung-Ro and Seo Sang Yong. *Genetic Fuzzy Discretization with Adaptive Intervals for Classification Problem.* GECCO'05 Washington, DC, USA 2005.

[4] D. O'Sullivan and D.J. Unwin. *Geographic Information Analysis*. Wiley, NJ, 2002.

[5] Han Jiawei, Kamber Micheline and Pei Jian. *Data Mining: Concepts and Techniques*.2011.

[6] http://egsc.usgs.gov/isb/pubs/gis_poster/

[7] http://www.makingthemodernworld.org.uk/learning_modules/geography/04.TU.01/

[8] http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/What_is_ArcGIS_Desktop/

[9] http://en.wikipedia.org/wiki/Spatial_decision_support_system

[10] http://www.cs.waikato.ac.nz/ml/weka/

[11] http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

[12] http://maya.cs.depaul.edu/~Classes/Ect584/Weka/preprocess.html

[13] http://en.wikipedia.org/wiki/Urban_sprawl

[14] Jiang Fang, Liu Shenghe, Yuan Hong and Zhang Qing. *Measuring urban sprawl in Beijing with geo-spatial indices*. Journal of Geographical Sciences. pp 469-478, 2007.

[15] J. Luo, D.L. Yu, and X. Miao. Modeling Urban Growth Using GIS and Remote Sensing, *GIScience & Remote Sensing*. 45(4): 426-442, 2008.

[16] J. R. Quinlan, 1986. Induction of Decision Trees. *Machine Learning* 1(1):81-106, 1986.

[17] Lv Zhi-qiang, Dai Fu-qiang and Sun Cheng. *Evaluation of urban sprawl and urban landscape pattern in a rapidly developing region.* Environmental Monitoring and Assessment. pp 6437-6448, 2012.

[18] Lin Ming-Yen, Lee Pei-Yu and Hsueh* Sue-Chen. *Apriori-based Frequent Itemset Mining Algorithms on MapReduce.* ICUIMC'12, Kuala Lumpur, Malaysia. 2012.

[19] McCann Barbara A., Ewing Reid. *Measuring the Health Effects of SPRAWL.* Smart Growth America Surface Transportation Policy Project 2003.

[20] R. Agrawal, T. Imielinski and A. Swami. Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD*, pp. 207-216, 1993.

[21] R. Agrawal, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499, 1994

[22] Sprague, R. H., and E. D. Carlson. *Building effective Decision Support Systems*. Englewood Cliffs, N.J.:Prentice-Hall, 1982.

[23] Squires Gregory D. *Urban Sprawl: Causes, Consequences, & Policy Responses.* The Urban Institute, 2002.

[24] Sun Heng, Forsythe Wayne and Waters Nigel. Modeling Urban Land Use Change and Urban Sprawl: Calgary, Alberta, Canada. Networks and Spatial Economics. pp 353-376, 2007.
[25] Tang Hong, McDonald Simon. *Integrating GIS and Spatial Data mining Technique for Target marketing of University Courses.* Symposium on Geospatial Theory, Processing and Applications. Ottawa
[26] Tan Pang-Ning, Steinbach Michael, Kumar Vipin. *Introduction to Data Mining*. Association Analysis. pp 327-414, 2006.
[27] Warner Mildred, Hinrichs Clare, Schneyer Judy, and Joyce Lucy. *Sustaining the Rural Landscape by Building Community Social Capital.*1997.
[28] Witten Ian H., Frank Eibe. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2005.
[29] Wu Xindong, Kumar Vipin. *The Top Ten Algorithms in Data Mining*. 2009.
[30] Wu Xindong, Kumar Vipin, Quinlan J. Ross, Ghosh Joydeep, Yang Qiang, Motoda Hiroshi, McLachlan Geoffrey J., NG Angus, Liu Bing, Yu Philip S., Zhou Zhi-Hua, Steinbach Michael, Hand David J. and Steinberg Dan. *Top 10 algorithms in data mining*. ICDM '06, Hong Kong, 2006

Images

[31] http://allegany.umd.edu/Allegany%20County%20AGNR.cfm
[32] http://culturehall.com/artwork.html?page=9823
[33] http://travel.yahoo.com/p-travelguide-191502048-melbourne_vacations-i
[34] http://www.airlinereporter.com/2011/12/destination-travelling-to-tokyo-japan/
[35] http://www.bobzworldcity.com/most-populous-cities-in-the-world-2011/
[36] http://www.embarq.org/en/problem/urban-sprawl
[37] http://www.oregonswashingtoncounty.com/About-the-Area/Visit
[38] http://www.tripadvisor.com/LocationPhotos-g60763-w2-New_York_City_New_York.html

Dataset

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/
b) http://www.health.ny.gov/statistics/vital_statistics/2010/
c) 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
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
b) http://www.health.ny.gov/statistics/vital_statistics/2010/

Land Area Details

a) http://www.health.ny.gov/statistics/vital_statistics/2000/table02.htm
b) 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

Total Housing Unit: http://esd.ny.gov/NYSDataCenter/Population_HousingData.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

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

Accident Data:

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