Paper 124-2010

Using SAS® Enterprise Miner 5.3® to Examine the Factor Such as Poverty Level in Different School Districts.

Khurram J Tanwir, Patricia Cerrito, University of Louisville, Louisville, KY

ABSTRACT

The U.S. Congress appropriates revenues for elementary and secondary education in the federal government's annual budget in the form of an appropriation. Title I is part of the federal Elementary and Secondary Education Act and provides money meant to improve the academic performance of students whose families are living below the poverty level, students who are learning English as a second language, students who are not meeting state education standards sanctioned by the federal government, and whether a program school district has two percent or fewer poor children, etc.

The data set is collected from the Census Bureau, which includes data from 13,753 school districts in 50 states showing the number of children living in poverty in each school district. Text Miner software in SAS 9.1® is used to investigate the information related to the school districts. The association of different states with regards to the poverty level to each state's school districts was found. This paper shows that by using the Regression node within SAS, correlations can be made between different variables that can predict the most important variables.

INTRODUCTION

Should all schools be funded equally by the government? In today's society, we often wonder how the money we pay in taxes is being distributed amongst educational facilities. If we pay equal taxes within our school district, should the schools we send our children to be funded equally too? According to an American Youth Policy Forum that was held on June 8, 2001 on Capitol Hill, students from poor families, who live in an under privileged school district, should have the right to gain access to equal education opportunity. If a student lives in a bad school district and is sent to a private school, should they still suffer? A poor school district may not have resources available to students while a private school may. If schools are equally funded, all students will have an equal opportunity to learn.

Public schools are funded through federal, state and local taxes and do not charge tuition. The U.S. Congress appropriates revenues for elementary and secondary education in the federal government's annual budget. Most federal money for elementary and secondary schools is made available in the form of an appropriation to the U.S. Department of Education. The department then sends the budgeted funds to the state departments of education, which distribute the funds to their states' schools. Title I is part of the federal Elementary and Secondary Education Act, originally passed in 1965 (the act was reauthorized in 2002 as the No Child Left Behind Act of 2001). Title I provides money meant to improve the academic performance of students who face a variety of specific potential educational challenges. The law includes money for schools with children whose families are living below the poverty level, children who are learning English as a second language, children who face cultural and linguistic barriers, children who are not meeting state education standards sanctioned by the federal government, or children who are neglected or delinquent.

The resources available between different states' education departments vary. The poverty level in an area and the number of people living under the poverty level is very important in channeling these resources to the right people. To allocated proper resources, a correlation and association can be found among different states based on the number of school districts in each state and the number of children living in poverty. To determine correlations and association for this large dataset, the Association Node is very helpful. These data could prove very useful in analyzing the distribution of federal, state and local money, and can give us a better understanding of how the funds are allocated to each state. For example, in Kentucky, there are 176 school districts with 149,095 children living in poverty. Alabama has 133 school districts with 174,665 children living in poverty, and New York has 685 school districts with a total of 575,576 children living in poverty. It appears obvious that more funding is required for New York as compared to Kentucky and Alabama.

METHOD

It is the purpose of this study to use SAS to find associations within the data and show that by analyzing the data the

patterns can be found that could lead to better understanding of how funding vary between states. The results show that by finding associations within these data, patterns are found throughout the school district data for the number of students living in poverty.

The data analyzed in this paper is a table of School district estimates of data extracted from the Census Bureau web site in an Excel sheet, transformed into a SAS table. [1] It contains data collected from different school districts from all 52 states. The data contains information such as State Postal Code, the School District ID, number of children living in poverty and total population of the school. The data are comprised of the number of children living in poverty from over 13753 school districts in 52 states.

The first step to analyze the data was to change how SAS interpreted the variables within the dataset so the Association node could process it. The Association node requires the data set that it is processing to have one ID variable and one target variable. In the data set, the roles of the variables were changed to match what was needed, changing the "**State Postal Code**" variable to the target, and the "**Number of students under 18 living in poverty**" variable to the ID. Since this is a large dataset, we connected the dataset to Sample Node. The Association node was then run to find correlations and patterns within the states' data.

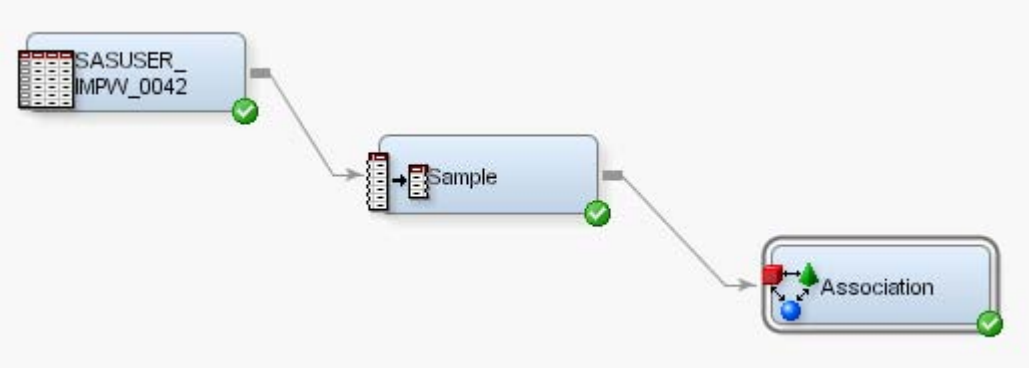


Figure 1.1

The default setting on the Association node was changed to find only the association rules between two states in order to overcome the out-of-memory issues. This was done by changing the Maximum item value from the default 3, to 2. After analyzing the data found by using the default values, the minimum confidence level, and the sorting was adjusted to give a stronger association of the items.

RESULTS

Running the Association node on the data set with the default setting was not possible due to an out-of-memory issue. To ensure that all associations are examined, the data were sorted by state postal code. The Maximum item value was changed from the default 3 to 2. The initial results appear in figure 1.2.

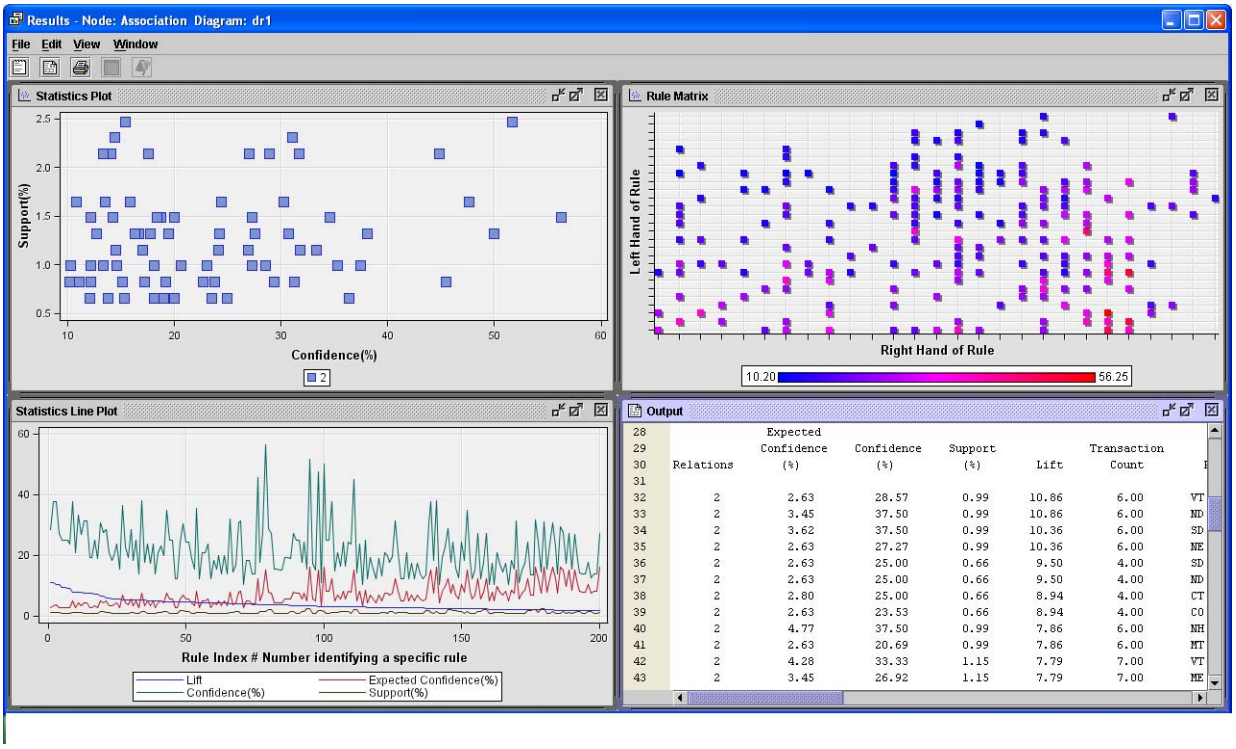


Figure 1.2

Figure 1.2 shows the Association rules and confidence for the data. Association rules examine the strength of the item combinations. The confidence is one of the numbers along with support that expresses the degree of uncertainty about the association rule. The confidence is the ratio of the number of a particular transaction to the number of total transactions. The default lists the first 200 rules.

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Index # Number identifying a specific rule
2	2.63	28.57	0.99	10.86	6.00	VT ==> ND	VT	ND	VT	==========>	ND	1
2	3.45	37.50	0.99	10.86	6.00	ND ==> VT	ND	VT	ND	==========>	VT	2
2	3.62	37.50	0.99	10.36	6.00	SD ==> NE	SD	NE	SD	==========>	NE	3
2	2.63	27.27	0.99	10.36	6.00	NE ==> SD	NE	SD	NE	==========>	SD	4
2	2.63	25.00	0.66	9.50	4.00	SD ==> ND	SD	ND	SD	==========>	ND	5
2	2.63	25.00	0.66	9.50	4.00	ND ==> SD	ND	SD	ND	==========>	SD	6
2	2.80	25.00	0.66	8.94	4.00	CT ==> CO	CT	CO	CT	==========>	CO	7
2	2.63	23.53	0.66	8.94	4.00	CO ==> CT	CO	CT	CO	==========>	CT	8
2	4.77	37.50	0.99	7.86	6.00	NH ==> MT	NH	MT	NH	==========>	MT	9
2	2.63	20.69	0.99	7.86	6.00	MT ==> NH	MT	NH	MT	==========>	NH	10
2	4.28	33.33	1.15	7.79	7.00	VT ==> ME	VT	ME	VT	==========>	ME	11
2	3.45	26.92	1.15	7.79	7.00	ME ==> VT	ME	VT	ME	==========>	VT	12
2	4.28	31.25	0.82	7.31	5.00	SD ==> ME	SD	ME	SD	==========>	ME	13
2	4.28	31.25	0.82	7.31	5.00	NH ==> ME	NH	ME	NH	==========>	ME	14
2	2.63	19.23	0.82	7.31	5.00	ME ==> NH	ME	NH	ME	==========>	NH	16
2	2.63	19.23	0.82	7.31	5.00	ME ==> SD	ME	SD	ME	==========>	SD	15
2	3.45	25.00	0.66	7.24	4.00	CT ==> VT	CT	VT	CT	==========>	VT	17
2	2.63	19.05	0.66	7.24	4.00	VT ==> CT	VT	CT	VT	==========>	CT	18
2	4.77	30.77	1.32	6.45	8.00	ME ==> MT	ME	MT	ME	==========>	MT	19
2	4.28	27.59	1.32	6.45	8.00	MT ==> ME	MT	ME	MT	==========>	ME	20
2	4.77	28.57	0.99	5.99	6.00	VT ==> MT	VT	MT	VT	==========>	MT	21
2	3.45	20.69	0.99	5.99	6.00	MT ==> VT	MT	VT	MT	==========>	VT	22
2	3.29	18.18	0.66	5.53	4.00	AZ ==> OR	AZ	OR	AZ	==========>	OR	23
2	3.62	20.00	0.66	5.53	4.00	OR ==> AZ	OR	AZ	OR	==========>	AZ	24
2	4.28	23.53	0.66	5.50	4.00	CO ==> KS	CO	KS	CO	==========>	KS	25

Figure 1.3

Figure 1.3 shows the first rules in the table for the two parameters. The aim of the analysis is to determine the strength of all the association rules among a set of items. Confidence in Figure 1.3 measures the percentage of an event or item occurs given the condition that another event or item occurs. The confidence of an association rule $A \Rightarrow B$ is the conditional probability of a transaction containing item set B given that it contains item set A so the confidence level in rule ND ==> VT is 37.5%. The rules are summarized in Table 1.1.

Table 1.1

RULE	CONFIDENCE (%)	Transaction count
ND ==> VT	37.5	6
SD ==> NE	37.5	6
NH ==> MT	37.5	6
VT ==> ME	33.33	7
SD ==> ME	31.25	5
NH ==> ME	31.25	5
ME ==> MT	30.77	8
VT ==> ND	28.57	6
VT ==> MT	28.57	6
MT ==> ME	27.59	8
NE ==> SD	27.27	6
ME ==> VT	26.92	7
SD ==> ND	25	4
ND ==> SD	25	4
CT ==> CO	25	4
CT ==> VT	25	4
CO ==> CT	23.53	4
CO ==> KS	23.53	4
MT ==> NH	20.69	6
MT ==> VT	20.69	6
OR ==> AZ	20	4
ME ==> NH	19.23	5
ME ==> SD	19.23	5
VT ==> CT	19.05	4
AZ ==> OR	18.18	4

Table 1.2 shows the support. Support is the number of times that the combination of items appears. It is derived by taking the number of transactions that two products have in common and then dividing that number by the total number of transactions. So, the number of transactions (school districts with an equal number of children living in poverty) in one state divided by all transactions will give us the support level for the rule A $\Rightarrow$ B. In our dataset, ND $\Rightarrow$ VT support level is 0.99%.

Table 1.2

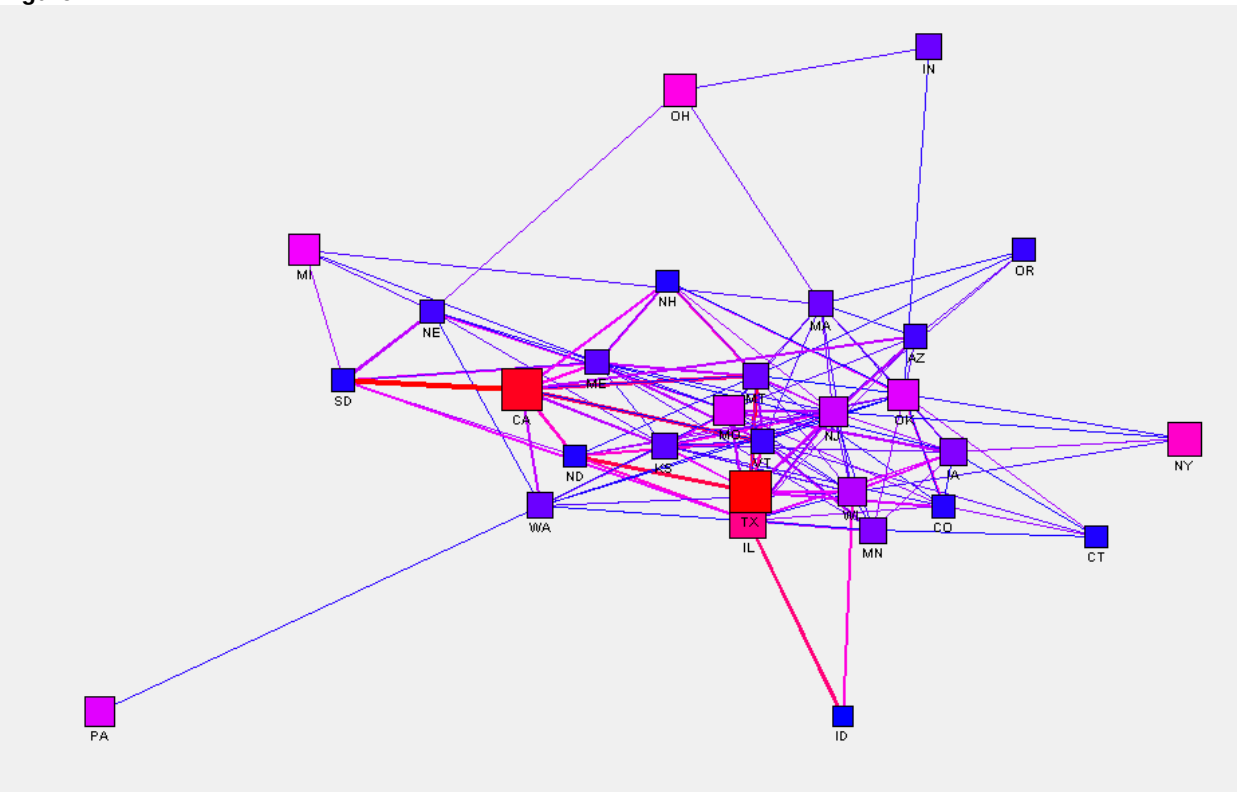
RULE	SUPPORT (%)	Transaction count
ME ==> MT	1.32	6
MT ==> ME	1.32	6
VT ==> ME	1.15	6
ME ==> VT	1.15	7
VT ==> ND	0.99	6
ND ==> VT	0.99	6
SD ==> NE	0.99	6
NE ==> SD	0.99	7

RULE	SUPPORT (%)	Transaction count
NH ==> MT	0.99	6
MT ==> NH	0.99	8
VT ==> MT	0.99	4
MT ==> VT	0.99	5
SD ==> ME	0.82	4
NH ==> ME	0.82	4
ME ==> NH	0.82	4
ME ==> SD	0.82	4
SD ==> ND	0.66	5
ND ==> SD	0.66	5
CT ==> CO	0.66	8
CO ==> CT	0.66	6
CT ==> VT	0.66	4
VT ==> CT	0.66	4
AZ ==> OR	0.66	5
OR ==> AZ	0.66	4
CO ==> KS	0.66	4

The *lift* of the rule $A \Rightarrow B$ is the confidence of the rule divided by the expected confidence, assuming the item sets are independent. The lift can be interpreted as a general measure of association between the two item sets. Values greater than 1 indicate positive correlation, values equal to 1 indicate zero correlation, and values less than 1 indicate negative correlation. The Lift in the association rule ND $\Rightarrow$ VT is 10.86. The lift measures the strength of association.

SAS's link graph tool gives a graphical representation of the association between the items. Each line on the graph represents a rule connecting one node to another. Figure 1.4 shows the default link graph for the data.

Figure 1.4



As seen in figure 1.4, link graphs can often be hard to read, especially when there are many nodes and connections. SAS uses the size of the nodes and the color to show the strength of association, as well as lines to show the actual relations between items. Here, we see that CA has the strongest confidence with SD. The relationship between TX and ND is shown with a large red line, showing that the confidence between these two is also very high.

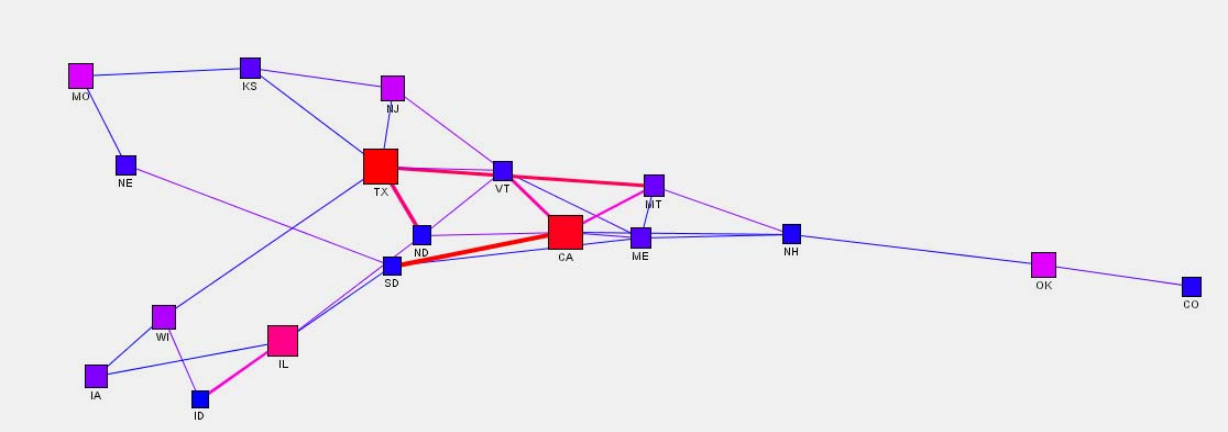
By changing the default that the Association node uses, the data can be much better represented in not only the link graph, but also in how the rules are formed. For example, if someone wanted to see a stronger relationship between items, they could do this by changing the minimum confidence level needed to “30” instead of the default “10”, and selecting to sort by “Default”. Tables 1.3 and Figure 1.5 show the results. The confidence values now range from 30.3% to 56.25%.

Table 1.3

RULE	CONFIDENCE (%)	SUPPORT (%)
SD ==> CA	56.25	1.48
MT ==> TX	51.72	2.47
ND ==> TX	50	1.32
VT ==> CA	47.62	1.64
ID ==> IL	45.45	0.82
MT ==> CA	44.83	2.14
VT ==> NJ	38.1	1.32
ND ==> VT	37.5	0.99
SD ==> NE	37.5	0.99
NH ==> MT	37.5	0.99
ND ==> IL	37.5	0.99
ND ==> CA	37.5	0.99
ID ==> WI	36.36	0.66

RULE	CONFIDENCE (%)	SUPPORT (%)
CO ==> OK	35.29	0.99
KS ==> NJ	34.62	1.48
VT ==> ME	33.33	1.15
NE ==> MO	31.82	1.15
SD ==> ME	31.25	0.82
NH ==> ME	31.25	0.82
NH ==> OK	31.25	0.82
SD ==> IL	31.25	0.82
ME ==> MT	30.77	1.32
KS ==> MO	30.77	1.32
IA ==> WI	30.3	1.64
IA ==> IL	30.3	1.64

Figure 1.5



By increasing the minimum level of Confidence, it makes SAS only show relationships that have a fairly strong connection. Recall that confidence represents the percentage of events/items that have the right hand side item among those who have the left side item. The new graph shows us the relationship between TX & MT and CA & SD. The link graph looks much cleaner and is easier to understand after narrowing down the results to only connections with higher relationships.

CONCLUSION

The Association works great for finding patterns in related data sets. Sometimes to get more precise data from a large data set, the data have to be filtered as it was here. By raising the minimum confidence, the data came back with fewer connections, and much stronger relationships.

Clearly from the research, it is easy to find the patterns of the levels of poverty in different school districts by analyzing past data. By analyzing these data, legislators can easily analyze which states have higher levels of child poverty in different school districts and require more funds.

REFERENCES

1. http://www.census.gov
2. http://www.ed.gov/pubs/NatAssess/sec6.html
3. Statistical Issues in Allocating Funds by Formula by Thomas A. Louis, Thomas B. Jabine, and Marisa A. Gerstein, Editors, National Research Council
4. Patricia Cerrito: Introduction to Data Mining, Using SAS Enterprise Miner

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name:	Patricia Cerrito
Enterprise:	University of Louisville
Address:	Department of Mathematics
City, State ZIP:	Louisville, KY 40292
Work Phone:	502-742-0889
Fax:	502-852-7132
E-mail:	pcerrito@gmail.com
Web:	lulu.com/CECS694 http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Dstripbooks&field-keywords=patricia+cerrito&x=0&y=0