A Data Mining Approach Using Geographic Information Systems Data for Predicting Voting Behavior in the Presidential Election of 2004

Hamid Nemati, UNCG, USA; E-mail: nemati@Uncg.edu
Ian Andrew, UNCG, USA; E-mail: iandrew@watershedconcepts.com
Peter Amidon, UNCG, USA; E-mail: P_AMIDON@uncg.edu

ABSTRACT

Throughout the last few decades, the state of North Carolina has voted to put both Republicans and Democrats in elected positions. The election of 2004 was no exception. North Carolina voted for a Republican president and a Democratic governor. This research seeks to understand the reasons for this voting pattern by focusing on one the largest counties in North Carolina, Guilford County. Guilford County was used for this study since it was one of the few counties in North Carolina that votes democratic. This research is an attempt to discover pattern and insight on voting habits based on the demographics of homeowners, house values and the age of property. This research will be beneficial not only to political parties but also to the citizens of Guilford County who want to see the data and unique comparisons between voting records and housing information

INTRODUCTION

Throughout the last few decades, the state of North Carolina has voted to put both Republicans and Democrats in elected positions. The election of 2004 was no exception. North Carolina voted for a Republican president and a Democratic governor. Since 1964, North Carolina has only given its electoral votes to the Democratic Party twice – 1964 (Johnson) and 1976 (Carter). 2004 continued this same voting behavior from previous elections. George W. Bush (R) garnered 1.96 million votes while Mike Easley (D) won nearly as many votes as the president with 1.94 million. Guilford County is one of few counties in North Carolina that voted Democratic in both the presidential and gubernatorial 2004 elections. Overall exit polls for North Carolina indicated that income and demographics played a strong role.

A County-by-County election return data from *USA Today* (Vanderibei, 2004) together with County boundary data from the US Census' Tiger database. Blue for Democratic, red for Republican, and green is for all other. Each county's color is a mix of these three-color components in proportion to the results for that county (Vanderbei, 2004). Clearly this map indicates that there are pockets of voters who are more likely to vote one way of the other solely based on their locations. It is also interesting to see that the density of voter participation is also closely related to what is shown in Figure 1. In Figure 2, the voter density of most urban areas shows a distinct propensity to vote democratic. It is interesting to note that the concentration of minority voters in urban areas is much higher than that of rural area and as seen in both figures 1 and 2, these voters tend to vote democratic.

This research, we were interested in understanding the voting patterns for a specific southeastern county in the United States and to see whether the broader presented reached earlier holds for this county as well.

Specifically we were interested in:

- Map and validate the North Carolina exit poll data to Guilford County for the 2004 election

Figure 1. County-by-county election returns in Presidential Election of 2004

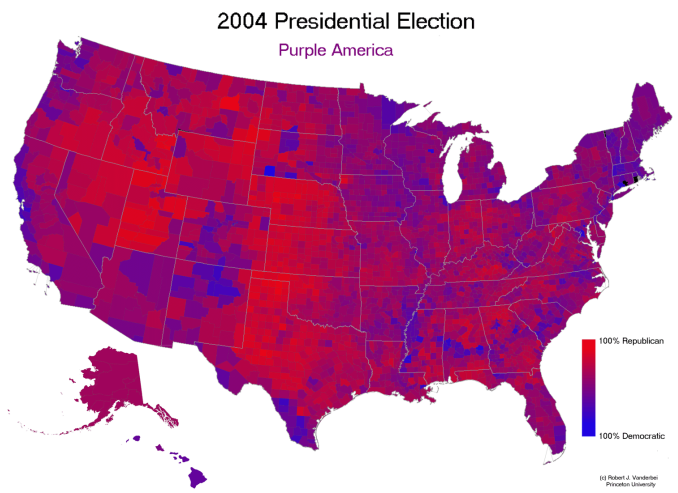
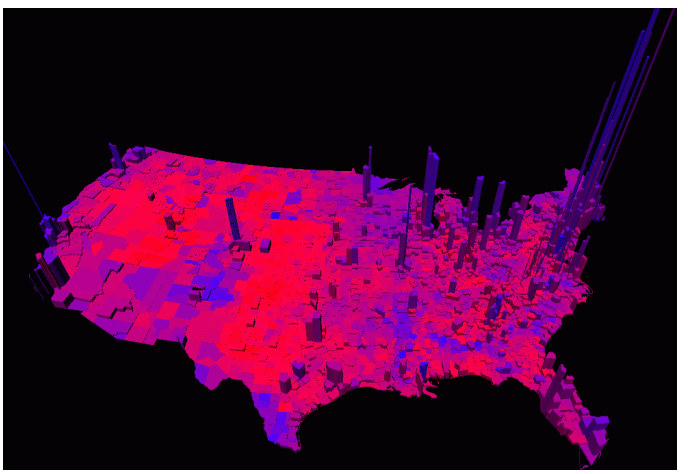


Figure 2. Voter density in 2004 Presidential Election



- Determine presidential and gubernatorial voting behaviors for precincts in Guilford County
- Map home values in precincts to voting behaviors
- Map age of the home to voting behaviors

In 2004, Guilford County broke from the "North Carolina pattern" in a slim fashion to vote for John Kerry, the Democratic challenger. Guilford County was one of a minority of counties in the state to vote for Kerry over George W. Bush. In addition, Guilford County was one of the 14 counties that made up 50% of the presidential vote in 2004. In addition to Guilford County, Orange County and Durham County were three of these fourteen counties that voted for Kerry (Guillory, April 2005). It has been widely held that Guilford County tends to lean democratic due to the combination of several factors Paul Shumaker, a political consultant for the Republican Party, states "Large urban counties and counties with a large public university presence provided the best results for Democrats." (Shumaker, April 2005). It can certainly be argued that Guilford Country has a strong educational infrastructure with schools such as UNC-G, NC A&T, Guilford College and Greensboro College. In addition, an argument can be made that this county has an active Democratic contingent and did an excellent job of promoting early voting. Most importantly, Guilford County has a sizable minority population. But what are other factors that impact this pattern of voting. According to 2004 exit polls conducted by Edison/Mitofsky, it appears that in general, suburbs in North Carolina vote strongly Republican while urban areas tend to vote evenly for both sides. In addition, voting by income reveals that incomes under 30K vote democratic versus incomes over 30K that vote republican (*Southnow*, April 2005). Is there a relationship between the value of a house and the voting behavior of the homeowner? Are there correlations between the voting behaviors of precincts and the value of the homes in those precincts?

Focusing on present housing statistics, voter demographics and votes cast in Guilford County in the 2004 election, this research project will seek to validate the exit poll results from Edison/Mitofsky as they pertain to Guilford County? Our research group intends to validate these findings in Guilford County by matching home value to income through a 40% rule of mortgage to income as well as define precincts as being in the city of Greensboro – urban – or outside of Greensboro but within Guilford County - rural. We anticipate that while most of the results match this exit polling, our research will also support the data that "In terms of electoral politics, the attitudes of both groups of voters (Republicans and Democrats) reflect the divided nature of an electorate in which neither major party commands a majority of popular support" (*Southnow*, April 2004).

This research is an attempt to discover pattern and insight on voting habits based on the demographics of homeowners, house values and the age of property. This research will be beneficial not only to political parties but also to the citizens of Guilford County who want to see the data and unique comparisons between voting records and housing information

DETAILED DATA MINING USING GIS DATA

In order to validate that North Carolina exit poll data (Edison/Mitofsky 2004) applies to Guilford County, this research project will focus on the following data points:

- present housing statistics
- voter demographics
- votes cast in Guilford County in the 2004 election

In addition the data mining of the GIS data would allow us to have a better understanding to the following questions:

- Do higher home values translate to Republican votes?
- Do lower home values translate to Democratic votes?
- Do precincts that are considered urban or metro vote democratic? Are there strong levels of home ownership in these districts and do they vote democratic as well?
- Do the "rural" precincts in Guilford County – outside of Greensboro city limits – vote republican?
- Do residents of identically valued houses in different neighborhoods vote the same or different?
- Are there particular precincts that split their votes or vote straight ticket?
- Are there particular home values that split their votes or vote straight ticket?
- Is there a correlation between home values and presidential and gubernatorial election results?
- How should precincts with lower numbers of homes for sale be scrutinized for other forms of housing such as rental apartments?

Our research intends to validate these findings in Guilford County by matching home value to income through a 40% rule of mortgage to income as well as define precincts as being in the city of Greensboro – urban – or outside of Greensboro but within Guilford County - rural. We anticipate that while most of the results match this exit polling, our research will also support the data that "In terms of electoral politics, the attitudes of both groups of voters (Republicans and Democrats) reflect the divided nature of an electorate in which neither major party commands a majority of popular support" (*Southnow*, April 2004).

Through this analysis, we intend to prove that in addition to the more widely understood, race and as predictor of voting preference, home value also plays somewhat of an indicating role in how precincts vote. Certain home value ranges will prove to be a wash. Precincts with lower home values will prove to vote democratic. Precincts with higher home values will prove to vote republican.

There are several assumptions that need to be considered in this project. We assume that:

- 40% of the combined income of individual or family translates to the assumable mortgage of a home
- The value of a house translates roughly to a home mortgage financed at 90%
- The value of a home is a close indication of the income of the owner
- Cities with universities tend to vote Democratic
- Precincts within the city limits of Greensboro will be considered urban
- Precincts within Guilford County but outside of the city limits of Greensboro will be considered rural

We were interested in studying whether using the age of the house, the tax value of the house and demographics of the homeowner can be used to predict the voting behavior of the homeowners.

We used the following commonly understood relationship between the value of the house and the income of the occupants. Based on the formula:

Value of house = 40% of (combined) income

Using this formula, we construct the following propositions:

- **< $30,000 income translates to $75,000 home and these voters tend to vote Democratic**
- **$30,000 – $50,000 = $75,000 - $125,000 home with no differentiation between parties**
- **$50,000 –$ 75,000 = $125,000 - $187,500 home with no differentiation between parties**
- **$75,000 – $100,000 = $187,500 - $250,000 home with no differentiation between parties**
- **> $100,000 translates to $250,000+ house and these voters tend to vote Republican**

IMPLEMENTATION

The data was analyzed and mined to determine the validity of the conventional wisdom. In addition to using Microsoft Excel, the project team also used mining tools provided by Microsoft Analysis Services in the SQL 2005 suite. In combination with these analysis tools, the project team also relied on data visualization tools through the use of GIS – geographic information system.

A data warehouse was developed using dimensional modeling approach to be the feeder for our data mining tools. This data warehouse is used to store:

- Precinct data including precinct ID, city, county, and precinct name,
- Voter registration statistics including voter registration, sex and ethnic background
- Presidential and gubernatorial voting data from 2000 and 2004
- Housing data including the average home value in the precinct, average value per square foot, age of home, and average year built.

Figure 3. Dimensional model used for data warehouse

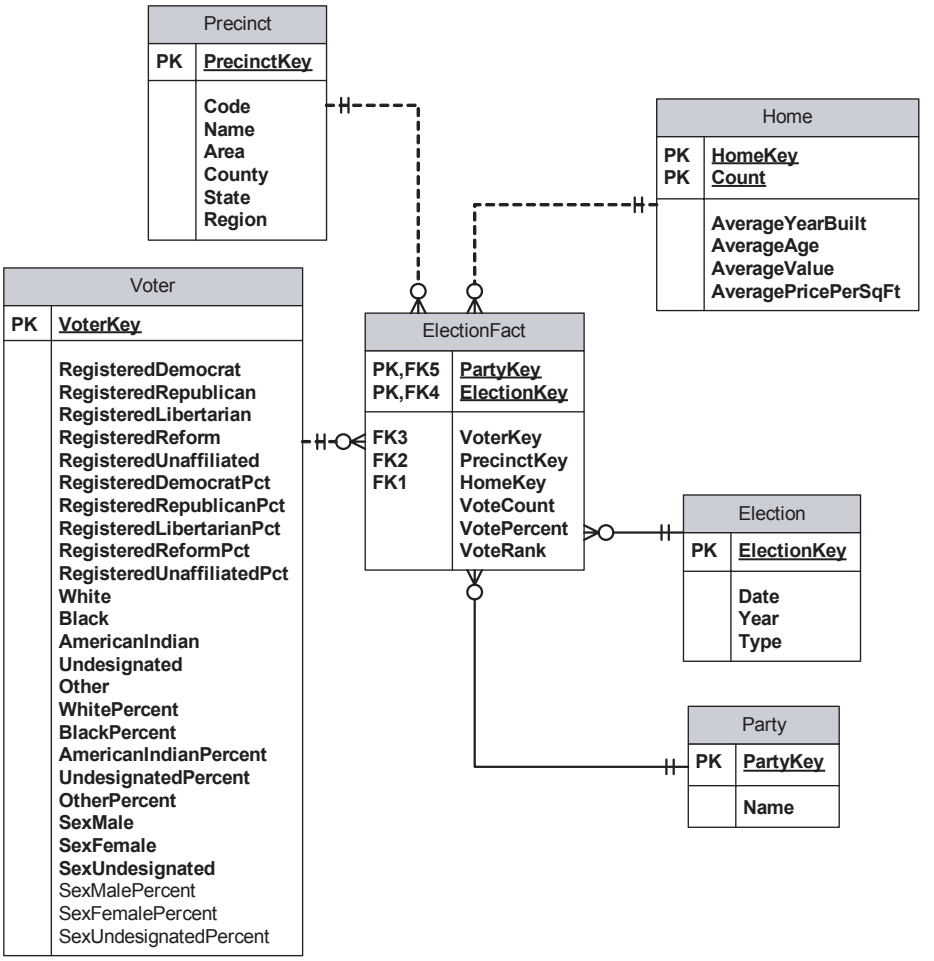


Figure 3 depicts the dimensional model used for this data warehouse.

To populate this data warehouse, the data was collected from various sources including:

- Triad Multiple Listing Service (MLS) – This is a powerful real estate listing site that provides detailed data on the local housing market
- Guilford County Taxation Department
- U.S Census Bureau
- Guilford County Board of Elections
- North Carolina State Board of Elections
- Triad MLS – A detailed report containing the fields in the previous section was created. A search was performed on active and pending houses in Guilford County for all areas of the county. Approximately 3692 records were created and displayed in the report. The report was exported from the site as a tab-delimited text file.

Data on the 2004 General Election Results were obtained from the Guilford County Board of elections. This data was in the form of a spreadsheet. Unfortunately, the 6 precincts were split into two precincts after the Precinct TIGER shapefile was created. Data for these two precincts was manually combined prior to further processing. A VBA routine was created to copy the appropriate data from the results into the database. This routine also generated and populated calculated data such as percent of vote. Data on Precinct Voter Registration was obtained from the Guilford County Board of elections. Finally a routine was developed to move the data to the data warehouse. This routine generated and populated calculated results such as translating counts into percentages. Data on Home Values was imported into the data warehouse. This method generated and populated the average home price, the average price per square foot, number of homes in the sample for the precinct, average age, and average year built.

GEOGRAPHICAL INFORMATION SYSTEMS

Data visualization provides a graphical interpretation of a data so that it can be analyzed from different perspectives. This study will combine data warehousing analysis techniques with spatial analysis to generate visual ads to assist in data visualization. Spatial data is data describes a location (point), line, or a shape (polygon) such as. A point object could represent a poll location, a line object could represent a street segment and a polygon could represent an area such as an election precinct. Spatial Analysis allows for the analysis based on the spatial feature as well as the relationship between two or more objects. For example what is the greatest distance from a polling location to the edge of the precinct.

The primary software package used for this phase was ESRI's ArcGIS ArcMap function. This application enabled the project team to:

- Load and view spatial data stored in a variety of formats.
- Manipulate and modify spatial objects and their attributes.
- Manipulate display graphics by symbolizing, classifying, and labeling spatial objects.
- Identifying, selecting, and finding features by attributes or location.
- Preparing data for analysis by removing unwanted features and combining others.
- Analyze spatial data by buffering and overlaying features.
- Generating the final output Maps

Figure 4. Guilford County 2004 Presidential Election results

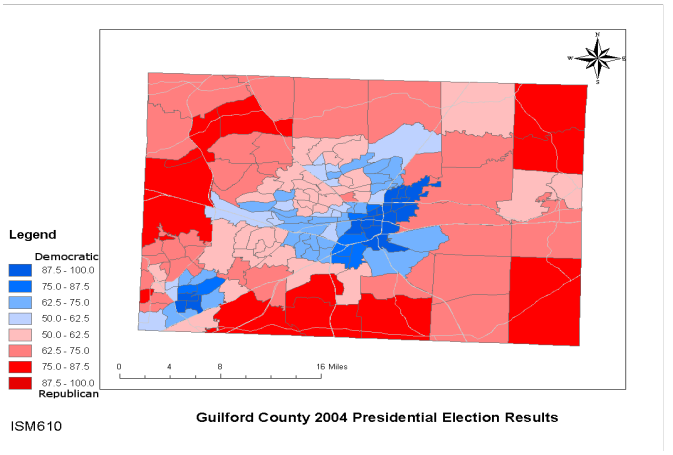


Figure 5. Guilford County registered minority (2004)

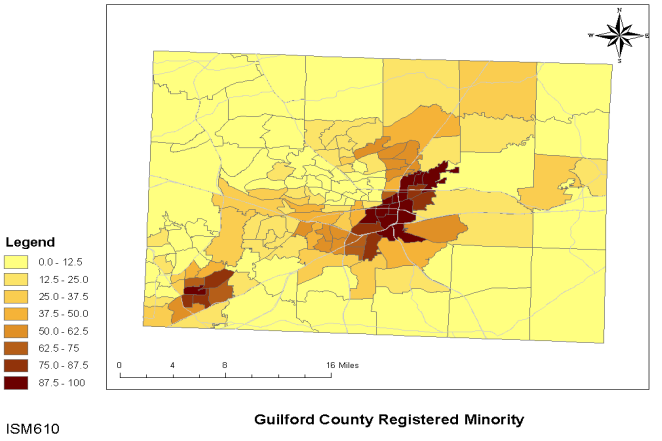


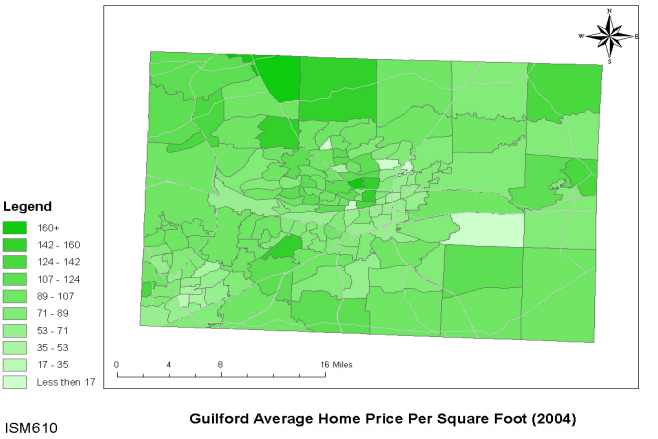
Figure 4 shows Guilford County 2004 presidential election results and Figure 5 shows Guilford County Registered Minority in year 2004. This Figure depicts the percentage of minority voters in each precinct based on voter registration information obtained from the Guilford County Board of Elections for 2004. Precincts rendered with darker values have a higher percentage of voters identifying themselves as being of minority background. As one can clearly see, the two figures are almost identical, indicating a high correlation between the two.

On the other hand, Figure 6 features the average value per square foot of the homes sold in the precinct during the study period. This information was imported into the data warehouse and rendered in ArcGIS application and the resulting map generated. Precincts with lighter shades of green experienced home sales of lower average value per square feet compared to those areas with darker shades of green. In this case, the relationship between the voting pattern and home prices per square foot is not clear.

Home Values

The group also analyzed housing data from MLS by grouping listings and data into the corresponding precincts. The group could not collect housing data for 4 of the 151 precincts. These precincts are Greensboro 28, Greensboro 45, Greensboro 8 and Jefferson 4.

Figure 6. Guilford average home prices per square foot (2004)



Average home value for the county from the sample was $140,583.89. We gathered housing information on 1,353 homes in Guilford County. Removing the four precincts where we did not have information, the average number of homes to precincts was 9.2.

Average age – 35 years.

Average $ / sqft - $91

No precinct that voted democratic had an average home value more than 211,200. The following graph represents how precincts with various average home values voted in the 2004 presidential campaign. The y-axis represents average home value per precinct. Each blue dot is a precinct. 0 represents a precinct that voted republican and 1 represents democratic ones.

NAÏVE BAYES DATA MINING BASED ANALYSIS

In analyzing the GIS data, we used the Naïve Bayes approach for data mining. Naïve Bayes is a classification and prediction algorithm that calculates probabilities for each possible state of the input attribute, given each state of the predictable attribute. This can then be used to predict an outcome of the predicted attribute based on the known input attributes. In order to better understand how precinct housing price data plays an indicating role in how precincts vote, our group needed to analysis the available data from several vantage points.

In combination with the voting data and voter demographic data, we looked to combine housing data from MLS. Through this analysis, we intended to prove that home value plays somewhat of an indicating role in how precincts vote. Although certain home value ranges would prove to be a wash, the group sought to prove that precincts with lower home values would prove to vote democratic while precincts with higher home values would prove to vote republican.

Based on the formula: *Value of house = 40% of (combined) income*

We analyzed the data within the following parameters. The analysis is included with each part.

< $30,000 income translates to $75,000 home and these voters tend to vote Democratic

40 precincts fell in this category. 34 of the 40 precincts voted democratic (85%) in the 2004 presidential campaign. The data strongly supported this position.

$30,000 – $50,000 = $75,000 - $125,000 home with no differentiation between parties

39 precincts qualified in this category. 23 of these precincts (59%) voted democratic. There did not appear to be a significant differentiation between this income range and voter outcome.

$50,000 –$ 75,000 = $125,000 - $187,500 home with no differentiation between parties

32 precincts composed this category. 6 of these precincts (19%) voted democratic. It appears that our data disproved this part. 4 out of 5 precincts with incomes in this range voted republican.

$75,000 – $100,000 = $187,500 - $250,000 home with no differentiation between parties

19 precincts fall in this category. 3 of these precincts (16%) voted democratic. Similar to the previous income range, the data that we collected disproved our initial thought that this range would be a wash. In fact, this income range strongly supported the republican candidates.

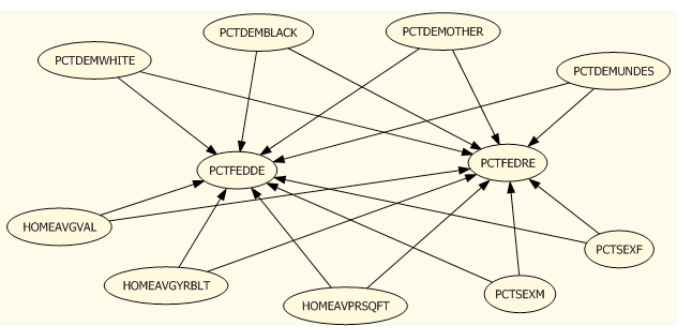
> $100,000 translates to $250,000+ house and these voters tend to vote Republican

17 precincts fall this category. Every one of these precincts voted republican. In this range, the data clearly supported the initial premise.

The Naïve Bayes algorithm was used to analyze the relationships between various voter attributes and the percentage of votes for each political party. The collected input variables included:

- Ethnic Background
  - o PCTDEMWHITE - percentage of registered voters identified as White.
  - o PCTDEMBLACK - percentage of registered voters identified as African-American.
  - o PCTDEMUNDES - percentage of registered voters undesignated.
  - o PCTDEMOTHER - percentage of registered voters identified as other.
- Male/Female
  - o PCTSEXM - percentage of registered voters designated as Male.
  - o PCTSEXF - percentage of registered voters designated as Female.
- Housing Information
  - o HOMVAVPRSQFT – Average home value per square foot.
  - o HOMEAVGVAL – Average home value.
  - o HOMEAVGYBUILT – Average year that home was constructed

The following diagram series of figures represents the dependencies between input values and output values that were used in the algorithm.



IMPORTANCE OF EACH DATA ELEMENT:

Each of the following graphics depicts the critical relationship between each of the attributes and the precinct percentage of votes for the democratic presidential candidate. These graphics are presented in order of significance from most to least. Based on analysis the importance of each input data type ranks as follows:

1. Percentage of registered voters identified as African-American.
2. Percentage of registered voters identified as White.
3. Percentage of registered voters identified as Female.
4. Percentage of registered voters identified as Male.
5. Average Home Value
6. Percentage of registered voters identified as other.
7. Average Home Value per Square Foot
8. Percentage of registered voters of unidentified Ethnic Groups
9. Average Age of Home.

The Naïve Bayes analysis closely follows the general analysis above in terms of the strong influence the race and gender have on precinct voting. Housing data does not play as strong a role with average home value being a stronger attribute than home value per square foot or average age of home.

ATTRIBUTE DISCRIMINATION VALUES

Attribute Discrimination indicates the importance of a particular attribute category in determining the output values. The following charts show how the most important data attributes play a part in the final output values – the percentage votes for the republican and democratic parties.

The attributes favoring *high* democratic vote results in a precinct include:

- Home average value less then 95,000
- Percent of white voters in precinct of less then 17%
- Percent black voters greater then 85%
- Average home value per square foot of less then $67

The attributes favoring *low* democratic vote results in a precinct include:

- Percent African-American voters of less then 13%
- Percent White voters greater then 85%
- Registered male voters greater then 47%

Because there are only two major parties, the factors that favor one party tend to be ones that work against the other as indicated in the Attribute Discrimination chart for the Republican Party. The following graphics depict the information in the preceding bulleted lists:

CONCLUSION

As stated in the introduction, Guilford County was one of a minority of counties that voted not only for the Democratic candidate for governor but also the democratic candidate for president. The intent of this paper was to answer the question – why? In order to tackle this question, we needed to gather not only the basic information such as the voting data from 2004, but we also wanted to gather housing data to determine if there was a significant correlation between housing data and voting data.

The data supports one of the group's premises that higher home values translate to republican votes. Conversely, lower home values translate to democratic votes. In addition, there appears to be a strong correlation between home values and the presidential election results. Gubernatorial results are not as clear as the presidential results.

In terms of gender and racial demographics, our analysis supported the traditional thoughts around Guilford County. Race and gender play a critical role in determining voting trends in precincts. Housing played a complementary role in determine democratic and republican voting behaviors. Voters from precincts with an average home value of less than $70,000 strongly supported democratic candidates. Voters from precincts with an average home value of more than $125,000 voted more republican. All precincts with an average home value of $250,000 voted republican. In fact, no precinct that voted democratic had an average home value of more than $211,200.

In conclusion, our project team successful merged two sets of seemingly disparate data to reveal intriguing connections between voter demographics, voting results and housing statistics. While race and gender were the strongest attributes in the overall study, Naïve Bayes analysis revealed some strong connections between housing and voting results.

REFERENCES

Census TIGER/Line Data. 2000. http://www.esri.com/data/download/census2000_tigerline/

Edsall, Thomas B. "Voter Values Determine Political Affiliation". Washington Post. March, 2001. http://www.washingtonpost.com/ac2/wp-dyn/A56905-2001Mar25?language=printer

Gautschi, Eric. "Suburban Thirty-Somethings Make the Difference". Southnow, April 2005. http://southnow.org/pubs/ncdn/ncdn38.pdf

Guilford County General Election Results. 2000. http://www.co.guilford.nc.us/government/elections/gen11-07-00.xls

Guilford County General Election Results. 2004. http://www.co.guilford.nc.us/government/elections/vt110204.xls

Guillory, Ferrel ."Straight-Party and Split-Ticket Voting" Southnow, April 2005. http://southnow.org/pubs/ncdn/ncdn38.pdf

North Carolina Voter Registration Statistics. April 1998. http://www.app.sboe.state.nc.us/voterreg/vr0498p.pdf

Quinterno, John. "New Voters Altering Political Landscape" Southnow. April 2004. http://southnow.org/pubs/ncdn/ncdn36.pdf

Shumaker, Paul. "The Republican Base and Cross-Over Appeal". Southnow, April 2005. http://southnow.org/pubs/ncdn/ncdn38.pdf

U.S. PRESIDENT / NORTH CAROLINA / EXIT POLL. CNN. 2004. http://www.cnn.com/ELECTION/2004/pages/results/states/NC/P/00/epolls.0.html

Vanderbei, Robert J. 2004. http://www.princeton.edu/~rvdb/JAVA/election2004/

VR Statistics By Precinct. 2005. http://www.co.guilford.nc.us/government/elections/st053105.pdf