Using Free Demographic Datasets to Support Teaching and Researching Projects

German Diaz
GIS/Data Center Support Specialist
Rice University
gdiaz@rice.edu

Twenty-Sixth Annual ESRI International User Conference. San Diego, CA.

Abstract

This paper provides an introduction to the use of Census 2000 resources freely available through the Internet for the generation of custom maps. It also illustrates some tips and tricks to address common problems and create effective displays. Finally, this paper illustrates some examples of how Rice University researchers are using census information to support their research projects.

Introduction

The study of the spatial distribution of demographic characteristics has drawn the attention of scholars from many disciplines including social and political science, urban planning, architecture, history, and anthropology.

Geographic information systems have become valuable instruments in the analysis of demographic and sociological characteristics. The rapid growth of the U.S. population and the high grade of complexity of census content have forced researchers to use robust computer systems. The effort of the U.S. Census Bureau to prepare and make accessible census data along with their respective geographic boundaries, and the facility of integrating this information in a GIS environment have made it easy for scholars to spatially analyze census characteristics and integrate the results in their research projects.

A variety of documentation related to the understanding of census data structure and its manipulation in a GIS environment is already available. Through the Internet people can access technical documentation, reports, tutorials, hands-on exercises, and digital books. Many books addressing the topic have been published. Alan Peters and Heather MacDonald in *Unlocking the Census with GIS*, [1] published by ESRI Press, present a comprehensive guide to the 2000 Census Population and Housing.

However, if the process of analyzing census information were as easy as following certain instructions, this paper would only refer to the existing literature. Changes and additions of new options in web delivery interfaces as well as modifications in the data structure have made most tutorials obsolete. Also, the high amount of data available, the complexity of the information, and several commonly faced problems make obtaining data, analyzing the information, and generating a map a very complicated process. Moreover, many decision makers demand high fidelity maps, and the process of generating an effective geographic display requires deep data examination.

This paper presents an overview of census data concepts making reference to the most relevant existing literature. Using an example, it explains how to obtain and integrate, in a GIS environment, census data that is freely available while highlights commonly faced problems. With some illustrations, it explains techniques for effective census data analysis and display. Finally, it presents some applications of the use of GIS in the integration of census data to support Rice University research projects.

2000 Census Population and Housing (Overview)

Besides providing basic counts of the U.S. population for apportionment of seats in the House of Representatives among the States, the U.S. Decennial Census provides data for a variety of demographic and sociological characteristics including races, education levels, and economic features. This information is valuable to redrawing State legislative and congressional districts, distributing funds for government programs, planning the right locations of public facilities, help real estate agents and potential residents learn about a neighborhood, and identifying trends over time that can help predict future needs [2]

In the Census 2000, people answered two kinds of questionnaires: A short form or 100-percent data containing basic questions of population and housing characteristics and considered filled out by all the population, and a long form or sample data containing more detailed information and filled out by 1 in 6 people. (1 in 2 in rural areas and 1 in 8 in more dense areas) [1]. Table 1 shows the population and housing characteristics contained in the short and long form questionnaires.

Short Form or 100 Percent Data	Long Form or Sample Data
Name	Population: Social Characteristics
Household relationship	Marital Status
Sex	Place of birth, citizenship, year of entry
Race	School enrollment
Hispanic or Latino origin	Educational attainment
Age	Language spoken at home
Tenure (whether home is owned or rented)	Ancestry
	Place of residence 5 years ago
	Veteran status
	Disability
	Grandparents as caregivers
	Population: Economic Characteristics
	Labor force status
	Place of work and journey to work
	Occupation, industry, class of worker
	Work status in 1999
	Income in 1999
	Housing: Physical Characteristics
	Units in structure
	Year structure built
	Number of rooms and bedrooms
	Year moved into residence
	Plumbing and kitchen facilities
	Telephone service
	Vehicles available
	Heating fuel
	Farm residence
	Housing: Financial Characteristics
	Value of home or monthly rent paid
	Utilities, mortgage, taxes, insurance, and fuel costs

Table 1. Population and housing characteristics contained in the Census 2000 short and long form questionnaires. Source: U.S. Census Bureau.

Along with census data, the U.S. Census Bureau has prepared geographic boundaries. Table 2 shows census definitions for the areas frequently used. A comprehensive definition of census geographies can be found at *Census 2000 Geographic Terms and Concepts* (http://www.census.gov/geo/www/tiger/glossry2.pdf) or by searching in the *America FactFinder Glossary* (http://factfinder.census.gov)

State and equivalent entity The primary legal subdivision of the United States. The District of Columbia, Puerto Rico and the Island Areas (the U.S. Virgin Islands, Guam, American Samoa, and the Northern Mariana Islands) are each treated as the statistical equivalent of a state for census purposes.
County and equivalent entity The primary legal subdivision of most states. In Louisiana, these subdivisions are known as parishes. In Alaska, which has no counties, the county equivalents are boroughs, a legal subdivision, and census areas, a statistical subdivision. In four states (Maryland, Missouri, Nevada and Virginia), there are one or more cities that are independent of any county and thus constitute primary subdivisions of their states. The District of Columbia has no primary divisions, and the entire area is considered equivalent to a county for statistical purposes. In Puerto Rico, municipios are treated as county equivalents.
County subdivision A legal or statistical division of a county recognized by the Census Bureau for data presentation. The two major types of county subdivisions are census county divisions and minor civil divisions. County subdivision not defined- The name assigned to an area of unpopulated coastal water within a county that belongs to no county subdivision.
Census tract A small, relatively permanent statistical subdivision of a county delineated by a local committee of census data users for the purpose of presenting data. Census tract boundaries normally follow visible features, but may follow governmental unit boundaries and other non-visible features in some instances; they always nest within counties. Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants. They may be split by any sub-county geographic entity.
Block group (BG) A subdivision of a census tract (or, prior to 2000, a block numbering area), a block group is the smallest geographic unit for which the Census Bureau tabulates sample data. A block group consists of all the blocks within a census tract with the same beginning number. Example: block group 3 consists of all blocks within a 2000 census tract numbering from 3000 to 3999. In 1990, block group 3 consisted of all blocks numbered from 301 to 399Z.
Block A subdivision of a census tract (or, prior to 2000, a block numbering area), a block is the smallest geographic unit for which the Census Bureau tabulates 100-percent data. Many blocks correspond to individual city blocks bounded by streets, but blocks – especially in rural areas – may include many square miles and may have some boundaries that are not streets. The Census Bureau established blocks covering the entire nation for the first time in 1990. Previous censuses back to 1940 had blocks established only for part of the nation. Over 8 million blocks are identified for Census 2000.
Congressional district (CD) An area established by law for the election of representatives to the United States Congress. Each CD is to be as equal in population to all other CDs in the state as practicable, based on the decennial census counts. The number of CDs in each state may change after each decennial census, and the boundaries may be changed more than once during a decade.
Voting District (VTD) Any of a variety of areas, such as election districts, precincts, legislative districts, or wards, established by states and local governments for voting purposes.
Urban Area Collective term referring to all areas that are urban. For Census 2000, there are two types of urban areas: urban clusters and urbanized areas.
Metropolitan area (MA) A collective term, established by the federal Office of Management and Budget, to refer to metropolitan statistical areas, consolidated metropolitan statistical areas, and primary metropolitan statistical areas.
Place A concentration of population either legally bounded as an incorporated place, or identified as a Census Designated Place (CDP) including comunidades and zonas urbanas in Puerto Rico. Incorporated places have legal descriptions of borough (except in Alaska and New York), city, town (except in New England, New York, and Wisconsin), or village.
ZIP Code Tabulation Area (ZCTA) A ZIP Code Tabulation Area is a geographic area that approximates the delivery area for a five-digit or a three-digit ZIP Code. ZCTAs do not precisely depict the area within which mail deliveries associated with that ZIP Code occur. A five-digit ZCTA ending in "HH" (e.g., "006HH") represents the water area within a three-digit ZCTA that is not associated with any mail delivery route. A five-digit ZCTA ending in "XX" (e.g., "006XX") represents the land area within a three-digit ZCTA that is not associated with any mail delivery route.

Table 2. Census definitions for the areas frequently used. Source U.S. Census Bureau

Although census geographic areas can be broken down into smaller subdivisions, not all the geographies lie entirely inside a larger geographic area. There is a hierarchy establishing how an area can be completely subdivided. In Figure 1 for example, just a portion of a Census Tract falls inside a Zip Code Tabulation Area. Figure 1 also shows the Census 2000 geographic hierarchy presenting the information as a series of nesting relationships.

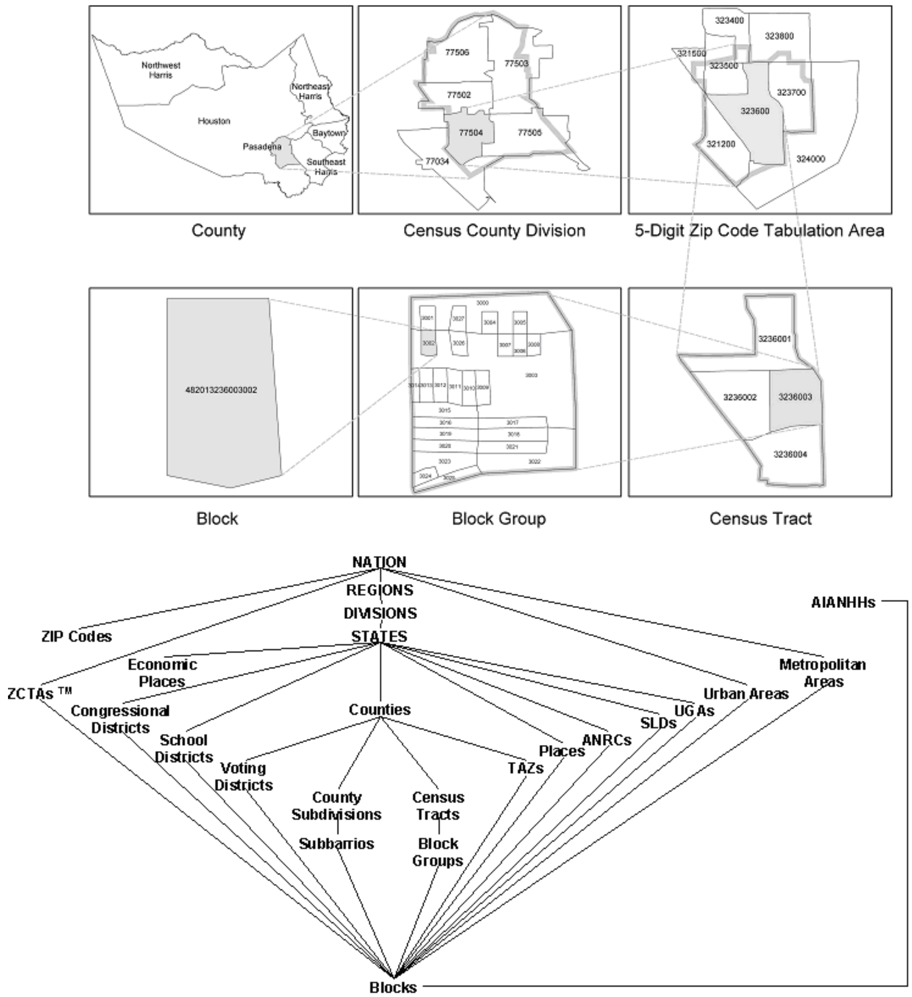


Figure 1 Census 2000 geographic hierarchy. Adapted from the U.S. Census Bureau.

Making Census 2000 information freely available, the U.S. Census Bureau has published the data in four releases: Summary File 1 (SF1) and Summary File 2 (SF2) presenting information from the 100-percent data or Short Form, and Summary File 3 (SF3) Summary File 4 (SF4) presenting information from the sample data or Long Form.

Summary File 2 and Summary File 4 contain similar information to Summary File 1 and Summary File 3 respectively but are cross-tabulated with 250 population groups: total population, 132 race groups, 39 Hispanic or Latino groups, and 78 American Indian and Alaska Native tribe categories. For data to be shown in Summary File 2 a population category must meet the criteria of a population size threshold of 100 or more people of that specific population category in a specific geographic area. For data to be shown in Summary File 4 a population category must meet a criteria of a population size threshold of 50 unweighted sample cases of a specific population category in a specific geographic area and the specific geographic area also has to be available in Summary File 2.

Most Summary File 1 and Summary File 2 data is available to the Census Block level and for Summary File 3 and Summary File 4 the smallest area is Census Block Group. Table 3 shows the subject content for each one of the census 2000 releases.

Summary File 1	Summary File 2
Subject content 171 Population tables (P) at the block level 56 Housing tables (H) at the block level 59 Population tables (PCT) at the census tract level **Geographic content Available in the American FactFinder** Nation Region Division State County County Subdivision Subbarrio Census Tract Block Group Block Place Consolidated City Alaska Native Regional Corporation Congressional District - 106th American Indian Area/Alaska Native Area/Hawaiian Home Land Reservation or Statistical Entity Only Off-Reservation Trust Land Only/Hawaiian Home Land Tribal Census Tract Tribal Block Group Tribal Subdivision/Remainder Metropolitan Statistical Area/Consolidated Metropolitan Statistical Area Primary Metropolitan Statistical Area New England County Metropolitan Area Urban Area	**Subject content** 36 Population tables (PCT) at the census tract level 11 Housing tables (HCT) at the census tract level Repeated or iterated for 250 Population Groups: Total population 32 Race groups 78 American Indian and Alaska Native tribe categories 39 Hispanic or Latino groups **Geographic content Available in the American FactFinder** Nation Region Division State County County Subdivision Subbarrio Census Tract Place Consolidated City Alaska Native Regional Corporation American Indian Area/Alaska Native Area/Hawaiian Home Land Reservation or Statistical Entity Only Off-Reservation Trust Land Only/Hawaiian Home Land Tribal Census Tract Tribal Subdivision/Remainder Metropolitan Statistical Area/Consolidated Metropolitan Statistical Area Primary Metropolitan Statistical Area New England County Metropolitan Area Urban Area
Summary File 3	**Summary File 4**
Subject content 288 Population (P) tables summarized to Block Group Level 196 Population tables (PCT) summarized to Census Tract Level 121 Housing tables (H) summarized to Block Group Level 208 Housing tables (HCT) summarized to Census Tract Level **Geographic content Available in the American FactFinder** Nation Region Division State County County Subdivision Subbarrio Census Tract Block Group Place Consolidated City Alaska Native Regional Corporation Congressional District - 106th American Indian Area/Alaska Native Area/Hawaiian Home Land Reservation or Statistical Entity Only Off-Reservation Trust Land Only/Hawaiian Home Land Tribal Census Tract Tribal Block Group Tribal Subdivision/Remainder Metropolitan Statistical Area/Consolidated Metropolitan Statistical Area Primary Metropolitan Statistical Area New England County Metropolitan Area Urban Area 3-Digit ZIP Code Tabulation Area 5-Digit ZIP Code Tabulation Area	**Subject content** 213 Population tables (PCT) at the census tract level 110 Housing tables (HCT) at the census tract level Repeated or iterated for 336 Population Groups: Total population 132 Race groups 78 American Indian and Alaska Native tribe categories 39 Hispanic or Latino groups 86 Ancestry groups **Geographic content Available in the American FactFinder** Nation Region Division State County County Subdivision Subbarrio Census Tract Place Consolidated City Alaska Native Regional Corporation American Indian Area/Alaska Native Area/Hawaiian Home Land Reservation or Statistical Entity Only Off-Reservation Trust Land Only/Hawaiian Home Land Tribal Census Tract Tribal Subdivision/Remainder Metropolitan Statistical Area/Consolidated Metropolitan Statistical Area Primary Metropolitan Statistical Area New England County Metropolitan Area Urban Area

Table 3. Subject and geographic content for each one of the Census 2000 releases

Detailed technical documentation about the Census cartographic boundaries and census releases is available at:
Census Cartographic Boundaries: http://www.census.gov/geo/www/garm.html
Census 2000 Summary File 1: http://www.census.gov/prod/cen2000/doc/sf1.pdf
Census 2000 Summary File 2: http://www.census.gov/prod/cen2000/doc/sf2.pdf
Census 2000 Summary File 3: http://www.census.gov/prod/cen2000/doc/sf3.pdf
Census 2000 Summary File 4: http://www.census.gov/prod/cen2000/doc/sf4.pdf

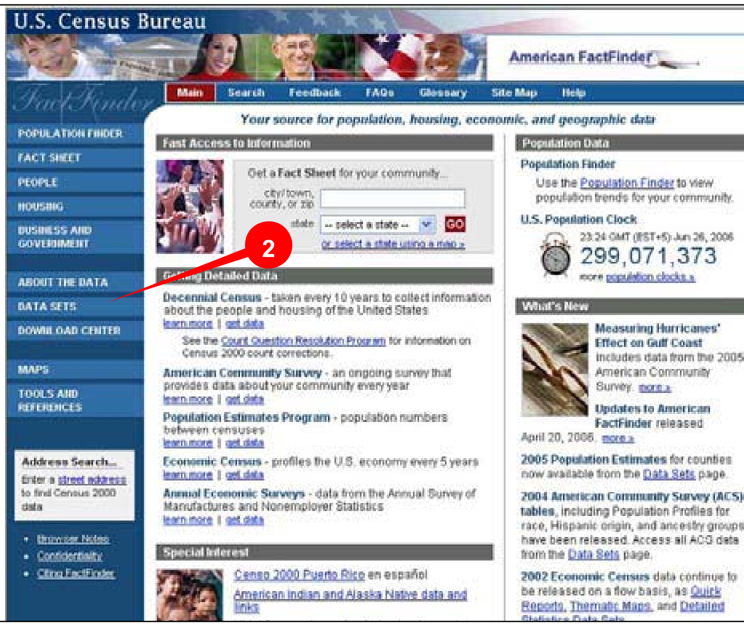
Acquiring Census 2000 Data From the Census Bureau-American FactFinder Portal And Geographic Boundaries from the ESRI Data Portal

Researchers frequently acquire Census 2000 data by obtaining, upon request, Census CD's directly from the U.S. Census Bureau (http://www.census.gov/mp/www/), downloading raw files from the U.S. Census Bureau FTP site (ftp://www2.census.gov/census_2000/datasets/), visiting depository libraries, or purchasing datasets fixed by particular vendors. As the preparation of the data requires a long processing time or a large investment of capital, this option must be considered only in the circumstance of large-scale census-related projects or in the case the researcher plans to use the datasets in further investigations.

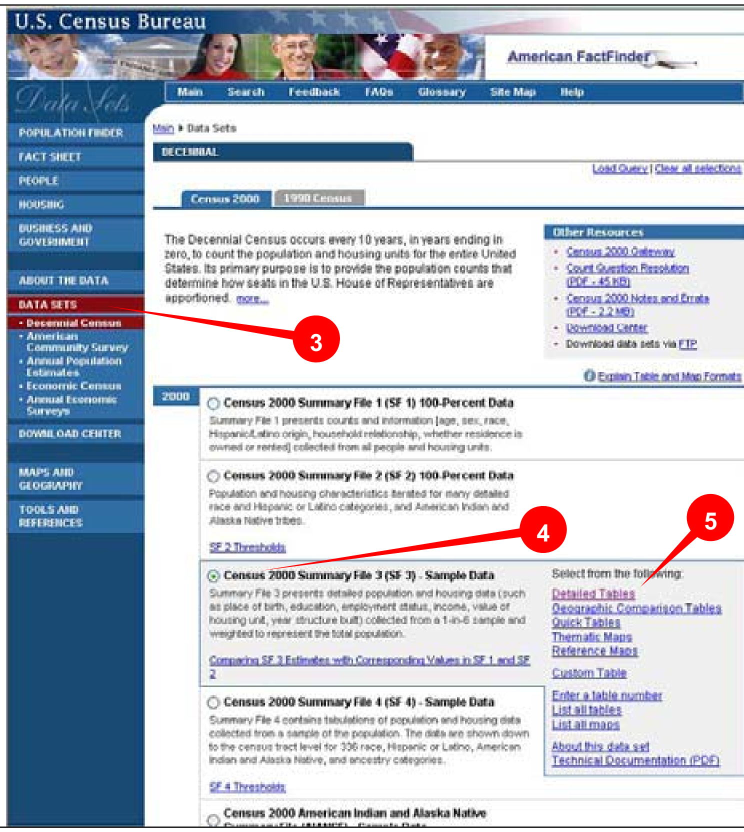
For sporadic census analysis projects or for projects that only use census information to support a researching idea, the U. S. Census Bureau American FactFinder website and the ESRI Data portal provide a solution to obtain census data at no cost and using very intuitive interfaces. After preparation, those datasets can be integrated in a GIS environment for the generation of custom maps. In the following procedures, this paper explains how to use those web portals to obtain census geographic data and cartographic boundaries.

Obtaining Census 2000 Tabular Data From the Census Bureau American FactFinder

	1. Navigate to the U.S. Census Bureau website (http://www.census.gov/) Locate and click on **American FactFinder**



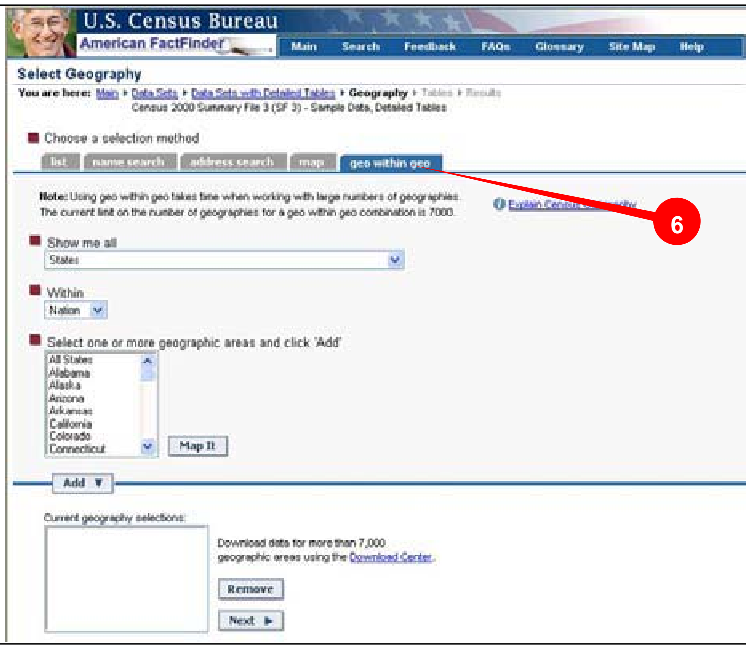
2. Click on **Data Sets**



3. Click on **Decennial Census**

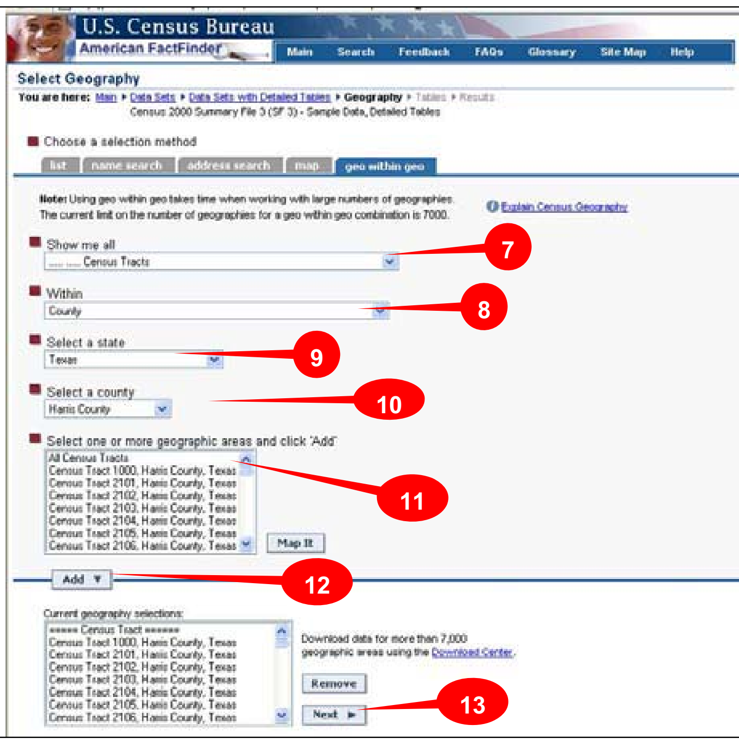
4. Select the appropriate census data release option (for this exercise, click on **Census 2000 Summary File 3 (SF3)- Sample Data**)

5. Click on **Detailed Tables**



6. Chose the selection method

The suggested method is the ***geo within geo*** *option. Note that the maximum of combinations for this method is 7000.*



7. Select the desired subdivision (**Census Tract**)

8. Select the desired area (**County**)

9. Select the **State**

10. Select the **County**

11. Select the **Census Tracts**

There is the option to select all the subdivisions or to select more than one by holding the shift key

12. Use the **Add** button to list the desired subdivisions

13. Once you finish listing the subdivisions, click on **Next**

14. Chose the selection method. The **by subjects** option provides an easy way to find the desired information.

15. Select the subject

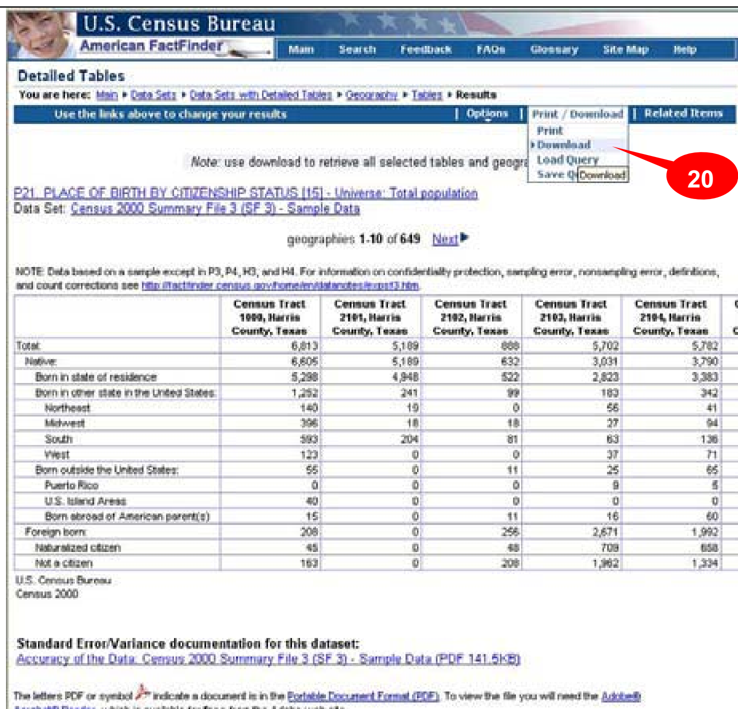
16. Click on **search**. A list of tables for the specified subject will be displayed

17. Select the tables

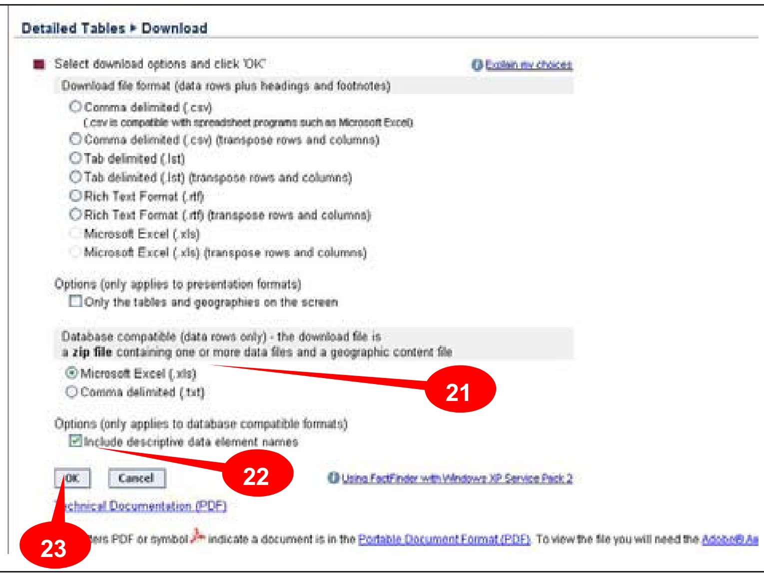
18. Click on the **Add button**

Commonly, unfamiliarity with the census content limits the search to basic counts. By taking a deep look at the list of variables, scholars can find more variables that can be used for a better support of their research projects or for further projects.

19. Once you finish selecting the tables click on the **show results** button. A report will be generated



20. Select the **download** option



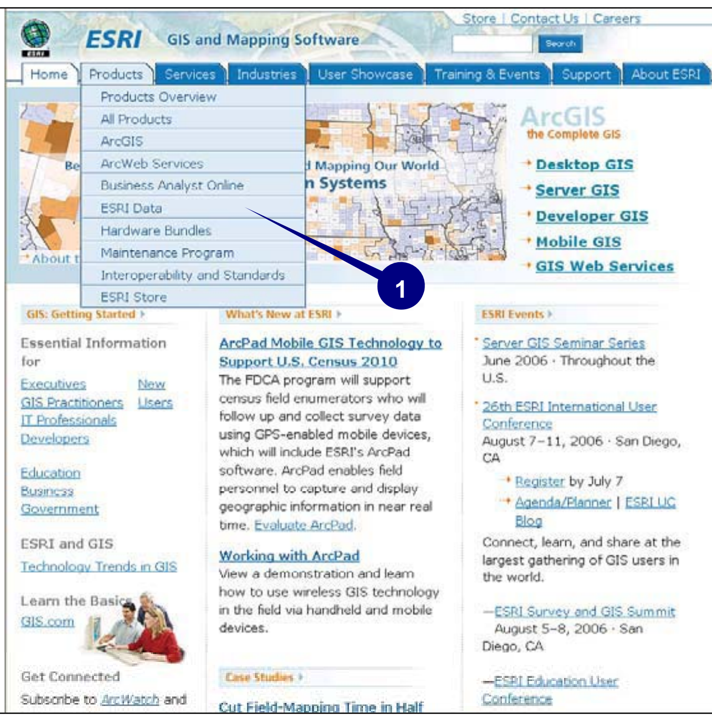
21. Select the **Microsoft Excel** option

22. Check the **Including descriptive data elements names** option

23. Click on the **OK** button

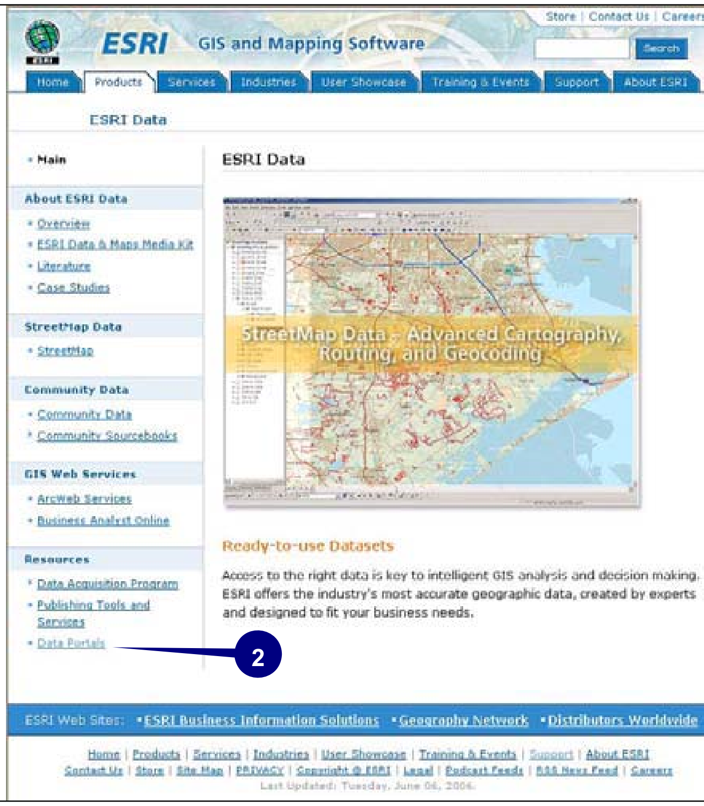
Save the files in your computer and uncompress the files

Obtaining Census 2000 Cartographic Boundaries From the ESRI Data Portal



Navigate to the ESRI website (http://www.esri.com)

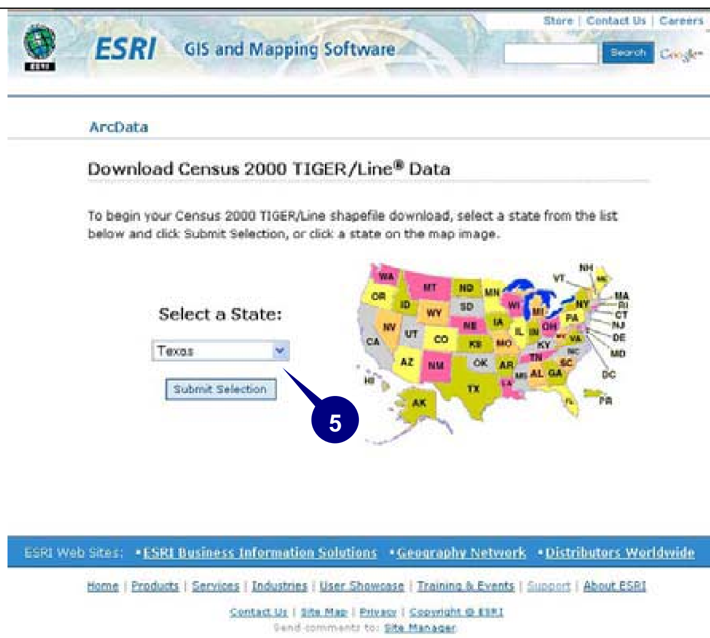
Locate **Products - ESRI Data**



2. Click on **Data portals**

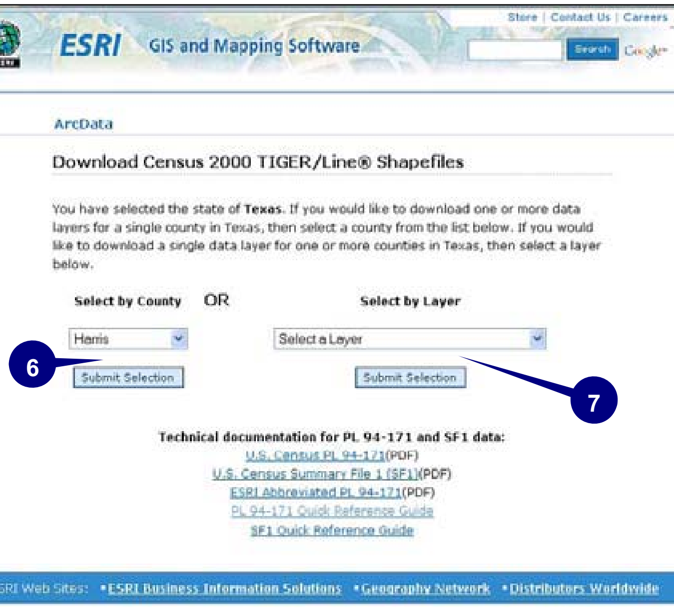
3. Click on **Census 2000 TIGER/Line Data**

4. Click on **Preview and download**



5. Select The **State**

Besides of census 2000 cartographic boundaries, ESRI Data - TIGER/Line Files provides a comprehensive resource and geographic features such as roads, railroads and drainage systems as well as Census 1990 cartographic boundaries.



6. Select the county to obtain multiple layers of the same county and click on **Submit Selection***

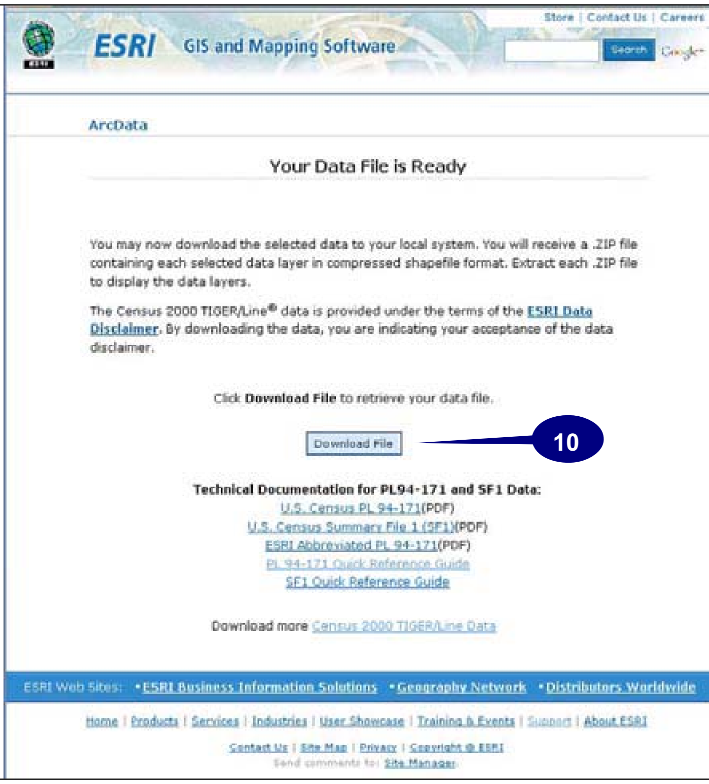
Or
7. Select a layer to obtain a feature dataset for multiple counties and click on **Submit Selection**

*Two options are available in this window: The **Select by County** option allowing obtaining multiple features for a specific county and the **Select by layers** option allowing obtaining a geographic feature for the multiple counties in the specified State.*

** Skip step 7 in the exercise, use step 6 instead*

8. Select the features for the specific county

9. Proceed to download the file



10. Download the File

The downloaded product is a single compressed file containing all the selected features. Each feature is also a compressed file grouping the components of a shapefile (dbf, shp, shx) and also needs to be uncompressed. The file name structure consists on the State FIPS, the county FIPS and the abbreviation of the geographic feature. Table 4 shows the abbreviations of the geographic features available in the TIGER/Line files.

Abb	Geographic Area
lkA	Line Features - Roads
lkB	Line Features - Rails
lkC	Line Features - Misc. Transport
lkD	Line Features - Landmarks
lkE	Line Features - Physical
lkF	Line Features - Non-visible
lkH	Line Features - Hydrography
lkX	Line Features - Unknown
cty	County 1990
ctycu	County - Current
cty00	County 2000
trt	Census Tracts 1990
trt00	Census Tracts 2000
grp	Block Groups 1990
grp00	Block Groups 2000
blk	Census Blocks 1990
blk00	Census Blocks 2000
plc	Designated Places 1990
plc00	Designated Places 2000
plccu	Designated Places - Current
ccdcu	County Census Divisions - Current
ccd00	County Census Divisions 2000
vot	Voting Districts
vot00	Voting Districts 2000
air00	Indian/Alaska Native Areas
aits	American Indian Tribal Subdivisions
air	American Indian/Alaska Native Areas
aircu	American Indian/Alaska Native Areas - Current
arc	Alaskan Native Regional Corporations

Abb	Geographic Area
kgl	Key Geographic Locations
lpy	Landmark Polygons
lpt	Landmark Points
taz	Traffic Analysis Zones
urb	Urban Areas
city	Consolidated Cities
elm	School Districts - Elementary
mid	School Districts - Middle
sec	School Districts - Secondary
uni	School Districts - Unified
wat	Water Polygons
msa00	CMSA/MSA Polygons 2000
pms00	PMSA Polygons 2000
cd106	Congressional Districts - 106th
cdc	Congressional Districts - Current
hse	State House Districts
sen	State Senate Districts
uga	Oregon Urban Growth Area
colblk	Census 2000 Collection Blocks
zcta	ZIP Code Tabulation Areas
sldl	State Legislative District Lower Chamber
sldu	State Legislative District Upper Chamber
alt	Alternate Feature Names
add2	Address Matching Info
zip	ZIP+4 Left and Right Info
add	Key Geographic Location Addresses
lpy2	Landmark Polygon Names
lpy3	Landmark Polygons - Multi-landmark
wat2	Water Polygons - Multi-names

Table 4. TIGER/Line File Abbreviations

Preparing and Exporting the Microsoft Excel Table to a DBF Format

The use of a table in Data Base Format (DBF) makes census information accessible to people without the latest versions of the ESRI GIS software. The following are some tricks for exporting a Microsoft Excel file downloaded from American FactFinder to a DBF format.

- When downloading the file from American FactFinder use the option “Include descriptive data element names”. After uncompressing the files make a copy of the files containing the data. In this way you will have a table with the data and another table with the field descriptions

- Open the table containing the data (…data1.xls, …data2.xls,) and delete the row with the descriptive data element names

- Rename the fields. Do not use more than 10 characters or special characters in the field names

- Format the fields (numeric, text, date). If the field is numeric with decimals make sure the field format shows the number of decimal places

- Extend the field widths to make sure the character strings are completely displayed. Otherwise Microsoft Excel will truncate the string to the column width

- Before exporting to DBF format, save the files as Microsoft Excel. In case of problems, you have a backup

- Copy the information to a new Microsoft Excel file using the Paste Special – Values option

- Select the information to export by dragging in the cells individually. Do not select the data by dragging in the field headers

- Save the file as DBF IV using no more than 10 characters or special characters. Avoid using spaces in the destination folder path

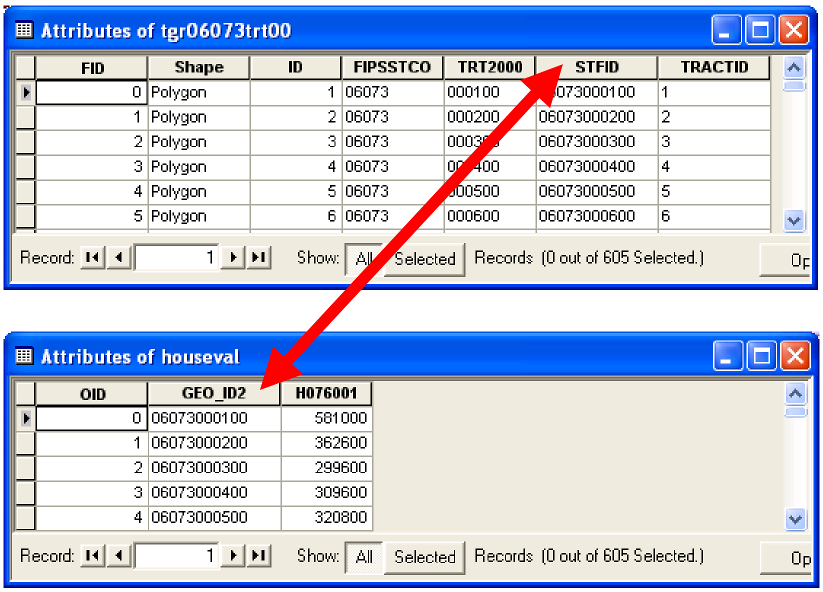
- After exporting the file close it in Microsoft Excel. The file cannot be manipulated in two programs simultaneously

- Review the content using ArcGIS in case of problems (commonly the lost of decimal places) remember, you have a backup in Microsoft Excel format. Open it, review what is wrong, and export the file again.

- More information about DBF files is available at the Microsoft Excel Help.

Dealing With the Census Geographic Identifier

Joining relational tables requires that both tables share a common field. In the case of census datasets the geographic identifier, usually named GEO_ID, STFID, FID, FIPS, or AREAKEY, is the key to joining census data with geographic features. To identify the common field open both tables (figure 2) and make an inspection looking for fields with similar information that is also consistent with the dataset geographic hierarchy. Figure 3 for example, shows the nomenclature for a specific census block in San Diego County, CA. In the new American FactFinder files the GIS-compatible geographic identifier is GEO_ID2.



Attributes of tgr06073trt00

FID	Shape	ID	FIPSSTCO	TRT2000	STFID	TRACTID
0	Polygon	1	06073	000100	06073000100	1
1	Polygon	2	06073	000200	06073000200	2
2	Polygon	3	06073	000300	06073000300	3
3	Polygon	4	06073	000400	06073000400	4
4	Polygon	5	06073	000500	06073000500	5
5	Polygon	6	06073	000600	06073000600	6

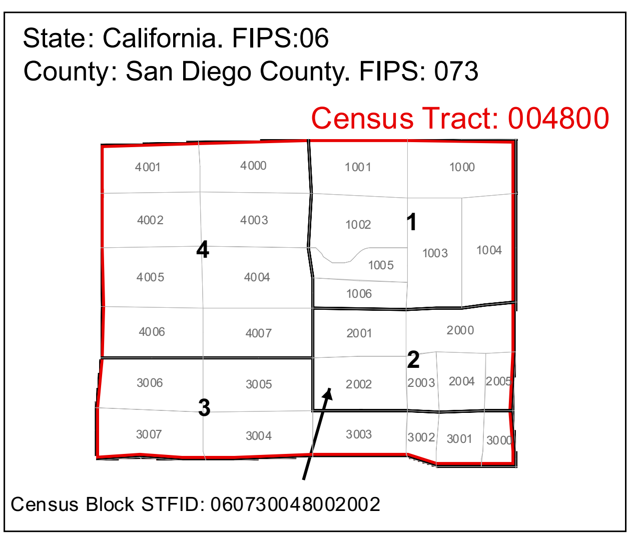
Record: 1 Show: All Selected Records (0 out of 605 Selected.)

Attributes of houseval

OID	GEO_ID2	H076001
0	06073000100	581000
1	06073000200	362600
2	06073000300	299600
3	06073000400	309600
4	06073000500	320800

Record: 1 Show: All Selected Records (0 out of 605 Selected.)

Figure 2. Detecting Common geographic Identifiers.



STFID= SSCCCTTTTTTGBBB
06 073 004800 2 002

Where:
SS =State (06, California)
CCC= County (073, San Diego County)
TTTTTT= Census tracts (004800)
G=Block group (2)
BBB= Census block (002)

Figure 3. STFID nomenclature for a specific census block in San Diego County, CA.

Two common problems in joining census datasets are the difference in the field type and variations in the geographic identifier nomenclature. Sometimes the nomenclature apparently looks the same but in one table the field type is numeric and in the other table the filed type is text. To correct this problem create a new text type field in the table containing the numeric type field (the one showing the characters at the right) and calculate the values using the expression:

[Text_Field]=[Numeric_Field]

In some cases, the strings look very similar but are not exactly the same, usually missing zeros on the left (6 instead 06), containing some extra characters at the beginning or end of the string (SS06 instead 06) or having additional zeros in some concatenations of geographic identifier (060730048002 instead 06073 48002). This problem can be fixed using text format in the fields and string functions. Table1 lists some useful string functions to solve problems with strings in census datasets.

OPERATION	CONCATENATING SYNTAX	EXAMPLE
Removing extra characters (EC) in the left	[Text_Field] = Right([Text_Field] , len([Text_Field])-number of EC)	Right("SS06076004800" , len("SS06076004800")-2) = "06076004800"
Removing extra characters (EC) in the right	[Text_Field] = Left([Text_Field] , len([Text_Field])-number of EC)	Left("0607600480000" , len("0607600480000")-2) = "06076004800"
Removing extra characters (EC) inside the string	[Text_Field] = Left([Text_Field], characters before the EC) & Right([Text_Field], characters after the EC)	Left("06076004800",5) & Right("06076004800",4) = "060764800"
Adding extra characters (EC) to the left	[Text_Field]= "EC" & [Text_Field]	"06"& "76004800" = "0676004800"
Adding extra characters (EC) to the Right	[Text_Field]= [Text_Field] & "EC"	"06760048" &"00" ="0676004800"
Adding extra Characters (EC) inside the string	[Text_Field] = Left([Text_Field], [characters before the EC]) & "EC" & Right([Text_Field], [characters after the EC])	Left("06076004800",5) &"00" &Right("06076004800",4) = "06076004800"
Recovering missing Zeros in the first 9 digits	Use the Select by attribute tool applying the following SQL sentence: Select * from [Table] where len([Text_field])= [length of the string with missing zeros] Recalculate the value using the expression: [Text_Field]= "0" & [Text_Field]	"0" & "6076004800" = "06076004800"

Table 5. Useful string functions to solve problems with strings in census datasets

Once the common fields share the same format and the same string structure the feature dataset is ready to join with the census data table based in the common ID (figure 4). A good practice is to export the joined map to keep the join permanently.

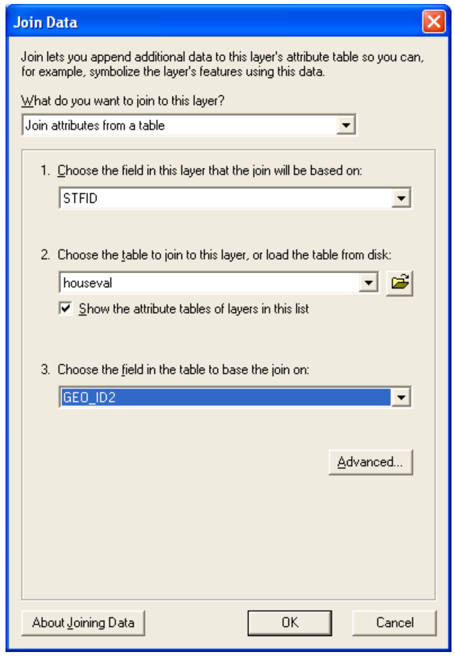


Figure 4. Joining by attributes

Making Effective Map Displays

Researchers have to understand and carefully examine map display alternatives to provide the high-grade of fidelity of geographic data analyses demanded by decision-makers. Multiple options in the grade of detail, relations with other features, and classification methods allow analyzing and interpreting the same characteristic in different ways. Also, similarity in terms and definitions confuses the effect of the chosen option in the map display.

- **Determining the Grade of Detail**

It is important to examine the grade of detail of the dataset subdivisions. A generalized subdivision is very useful to determine global trends or to focus the results in the studied areas. A more detailed subdivision unmasks regional patterns, and a very highly detailed subdivision determines some local patterns, or clusters, in the dataset. Any grade of detail also produces a masking effect in the results; a generalized subdivision masks local patterns, the detailed subdivision masks the limits of the studied areas and the very high subdivision masks even more the limits of the studied areas.

The maps in Figure 5 display the same demographic characteristic (Hispanic population) but in different grades of detail (states, counties, a and census tracts) to analyze patterns at the state level (the studied areas). As the value intervals change according to the grade of detail, the classification was fitted for each level using 6 quantiles. The map subdivided by states shows a regional distribution of the Hispanic population (high in the Southeast and South of the country and low in the North) and shows the Hispanic population distribution by state (High for the states of California, Arizona, Texas, Florida, Illinois, and New Jersey and low for the states of Montana, North Dakota, South Dakota, Wyoming, Kentucky, West Virginia, Mississippi, and Maine). On the other hand, the map subdivided by counties shows how the high concentration of Hispanic population in some counties masks the distribution by state (i.e. Cook County, IL and surrounding counties in the State of Illinois, and New York County in the New York State). Finally, the map at the census tract level shows some cluster in the concentration of Hispanic population affecting its global distribution. (i.e. South of Texas, South of Florida, Chicago, and New York).

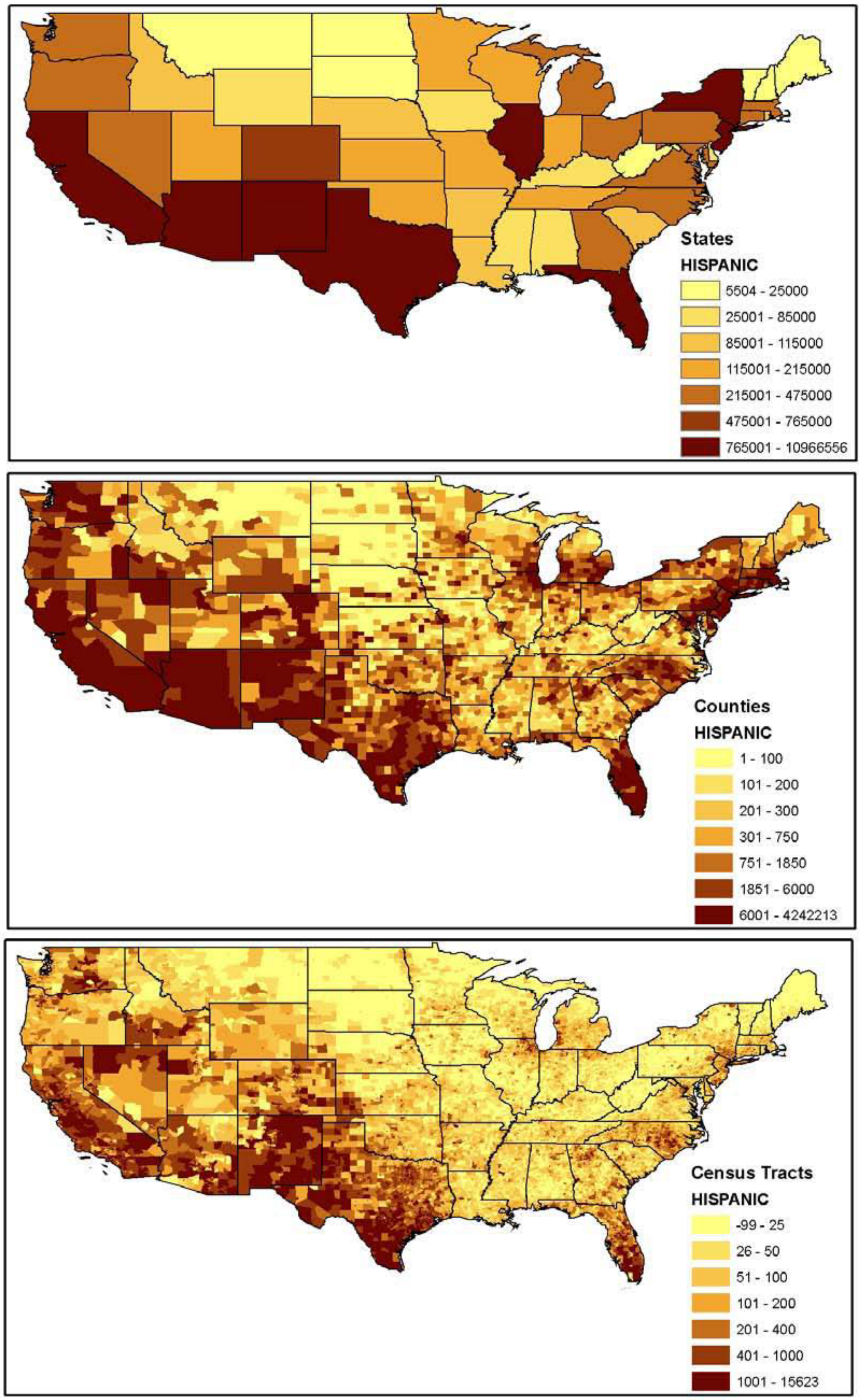


Figure 5. National wide distribution of the Hispanic/Latino population at state, county, and census tract levels.

- **Percentage, Percentage of the Total Population/Housing, Value as a Percentage of the Total, Value as a Percentage of the Total Universe, and Density.**

Scholars find in proportions and densities a better explanation of phenomena. Subdivisions with large size, high concentration of the universe or highly concentration of population/housing have the tendency to contain more people with the analyzed demographic or social condition. However, similarity in terms and definitions of proportions results in misuse of concepts.

The term **percentage** is defined as the counts in a subdivision divided by the universe in the same subdivision and then multiplied by 100. However, not always the universe is the total population or the total house units, for example analyzing citizenship (e.g. Naturalized) for the foreign-born population, (figure 6A), the universe is the foreign-born population and not the total population. The resulting value of the **percentage** is the percent of foreign-born naturalized population in relation to the foreign-born population (figure 6B), while the resulting value using the **percentage of the total population** is the percent of foreign-born naturalized population in relation to the total population (figure 6C).

The similarity of the term **value as a percentage of the total** to the term **percentage** often produces a misinterpretation. To calculate the **value as a percentage of the total** ArcGIS divides the counts in a subdivision by the sum of the counts in all the subdivisions and then multiplies by 100. In the example, the resulting **value as a percentage of the total** is the foreign-born naturalized population as a percentage of the total foreign-born naturalized population (figure 6D).

The **value as a percentage of the total universe** is the result of dividing the counts in a subdivision by the sum of universes in all the subdivisions and multiplying by 100. In the example, **the value as a percentage of the total universe** is the foreign-born naturalized population as a percentage of the total foreign-born population (figure 6E).

Density maps shows where values are concentrated. The density is the result of dividing the counts in a subdivision by the area (generally expressed in square miles). In the example the **density** is equal to the foreign-born naturalized population per square mile (figure 6F).

Maps displaying **values as a percentage of the total** and **values as a percentage of the total universe** look similar to maps showing the single value because the value is divided by a constant. Only the ranges of values change reflecting the proportion.

Proportions and densities calculation requires a division operation. When the software finds a divisor equals or rounded to 0 it stops the field calculation and sends an error message. To avoid this issue, before making the calculation select by attributes the values different to 0 and the calculation will be applied just to the selected features.

```
Select * from [layer] where [divisor] < > 0
```

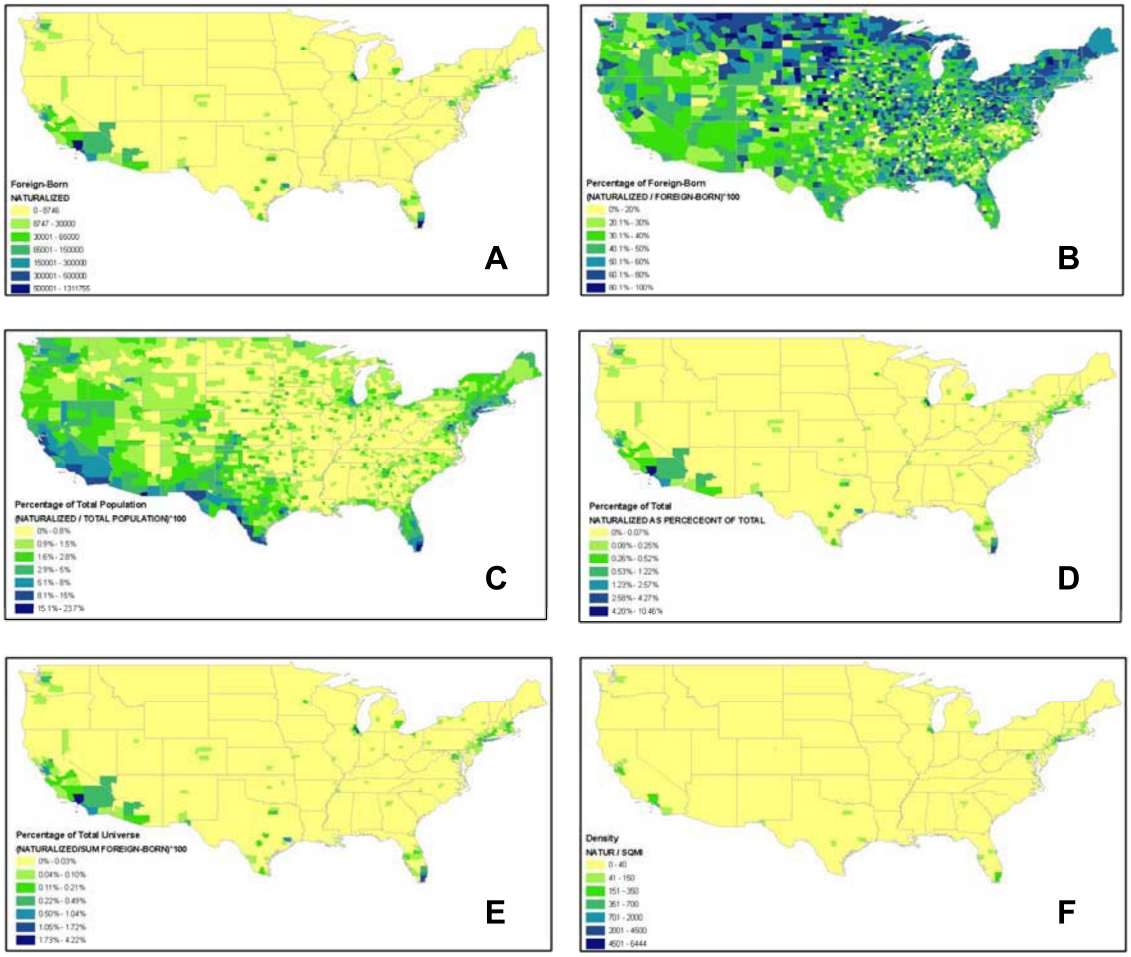


Figure 6. Comparison of relation terms applied to plot nationalized foreign-born population.

- **Classification Methods**

ArcMap provides 4 automatic methods of classification: natural breaks, equal intervals, quantiles, and standard deviation. Depending on the data values distribution, all the methods reflect different patterns affecting the map display (Figure 7). Andy Mitchell, in *The ESRI guide to GIS Analysis* and published by ESRI Press [3], presents a detailed description of the ArcMap automatic classification methods explaining how they work, and describing their advantages and disadvantages (table 5).

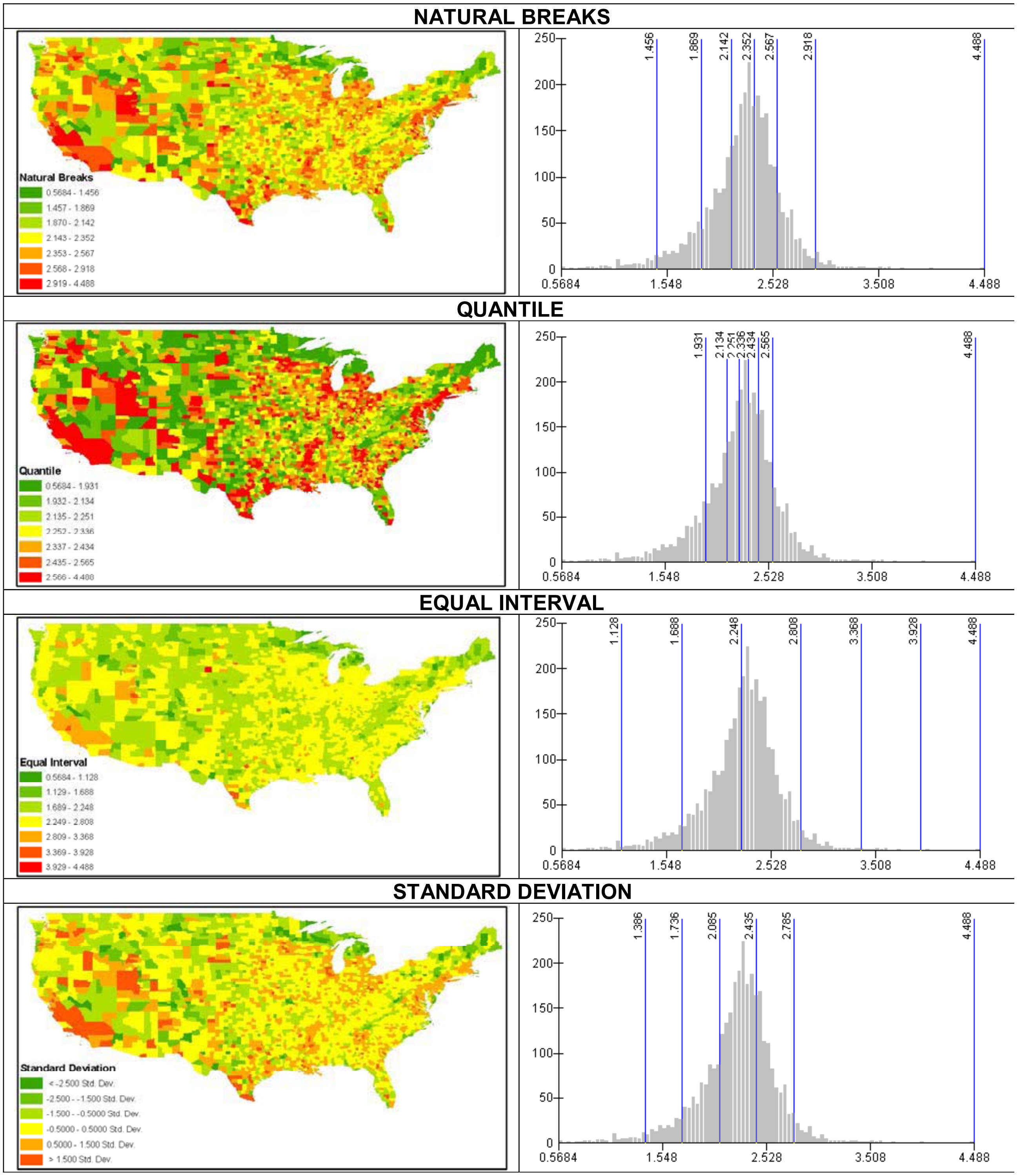


Figure 7. ArcGIS Automatic methods of classification patterns affecting the map display

Natural breaks

How it works
The GIS automatically determines the high and low value for each class, using a mathematical procedure to test different class breaks. It picks the class breaks that best group similar values and maximizes the differences between classes.

Advantages
Mapping data values that are not evenly distributed, since it places clustered values in the same class.

Disadvantages
- Since the class ranges are specific to the individual dataset, it's difficult to compare the map to other maps.
- Choosing the optimum number of classes is difficult, especially if the data is evenly distributed.

Quantile

How it works
The GIS orders the features, based on the attribute value-from low to high-and sums the number of features as it goes. It divides the total by the number of classes you've specified to get the number of features in each class. It then assigns the first features in the order to the lowest class until that class is filled, then moves on to the next class, fills it up, and so on.

Advantages
- Comparing areas that are roughly the same size.
- Mapping data in which the values are evenly distributed.
- Emphasizing the relative position of a feature among other features. For example, you can show which counties in a state are in the top
20 percent, in terms of median income (those in the highest of five categories).

Disadvantages
Features with close values may end up in different classes, especially if values cluster. This may exaggerate the differences between features. Conversely, a few widely ranging adjacent values may end up in the same class, minimizing the differences between these features.

Equal interval

How it works
The GIS subtract the lowest value in the dataset from the highest. It then divides that number by the number of classes specified. It adds that number to the lowest data value to get the maximum value for the first class. It then adds to each maximum value to set the brakes for the rest of the classes.

Advantages
- Presenting the information to a no technical audience.
- Easier to interpret.
- Mapping continuous data.

Disadvantages
If the data values are clustered rather than evenly distributed, there may be many features in one or two classes and some classes with no features.

Standard deviation

How it works
The GIS first finds the mean value by adding all the data values and dividing by the number of features. It then calculates the standard deviation by subtracting the mean from each value and squaring it (to make sure it's positive), summing these numbers, and then dividing by the number of features. It then takes the square root to get the final number. The formula looks like this:

$$S= \sqrt{\frac{\Sigma(x - X)^2}{n}}$$

Where s is the standard deviation, x is the value of a feature, X is the mean, and n is the number of features.You can think of this as the average amount the data values vary from the mean. The GIS creates class breaks above and below the mean based on the number of standard deviations you specify, such as 1/2 or 1 standard deviation.

Advantages
- Seeing which features are above or below an average value.
- Displaying data that has many values around the mean, and few further from the mean (a bell curve, or normal, distribution).

Disadvantages
- The map doesn't show the actual values of the features, only how far their value is from the mean
- Very high or low values (outliers) can skew the mean so that most features will fall in the same class.

Table 5. ArcGIS automatic classification methods description. [3]

In most of the cases, census data does not fit in any standard classification model and often presents values that are much larger or much smaller than most of the other observations (outliers). The manual classification allows using the data values distribution graph to add or modify breaks in the classification.

The detection of outliers is very helpful to highlight a phenomenon or identify errors in the datasets. Hogg and Tanis in *Probability and Statistical inference*, refers to a method suggested by Tukey (1977) for defining outliers that is resistant to the effect of one or two extreme values and makes use of the Interquartile Range (IQR)[4]. This method can also be used to determine breaks classification: Classify the data using 4 quantiles (quartiles) and measure the distance between the first and third quartile (IQR) (Figure 8). Insert breaks at one distance of 1.5 times the IQR left to the first quartile and right of the third quartile (Inner fences). Insert breaks at one distance of 3 times the IQR left to the first quartile and right of the third quartile (Outer fences). The values that lie between the inner and the outer fences are called suspected outliers. Values beyond the outer fences are called outliers. Suspected values require a closer look and the outliers should be looked at very carefully. This method can also be used to assign the distribution of the breaks in the classification setting more breaks in the Interquartile Range and not inserting brakes beyond the outer fences.

There is not a standard definition of the number and distribution of the breaks. Andy Mitchell [3] suggest no less than three or four: “Using fewer than three or four classes doesn’t show much variation and hence no clear patterns”, and not more than seven: “Most map-readers can distinguish up to seven colors on a map, more than that it make hard to find features with similar values”.

After selecting the classes, round or approximate the break values to an easily understood number. This makes the map interpretation easier. Also, if the map shows a classification based on the equal interval, a classification displaying only the minimum and maximum is more effective. Finally always use a color-graded classification scheme; people usually associate darker colors with high concentration and light colors with low.

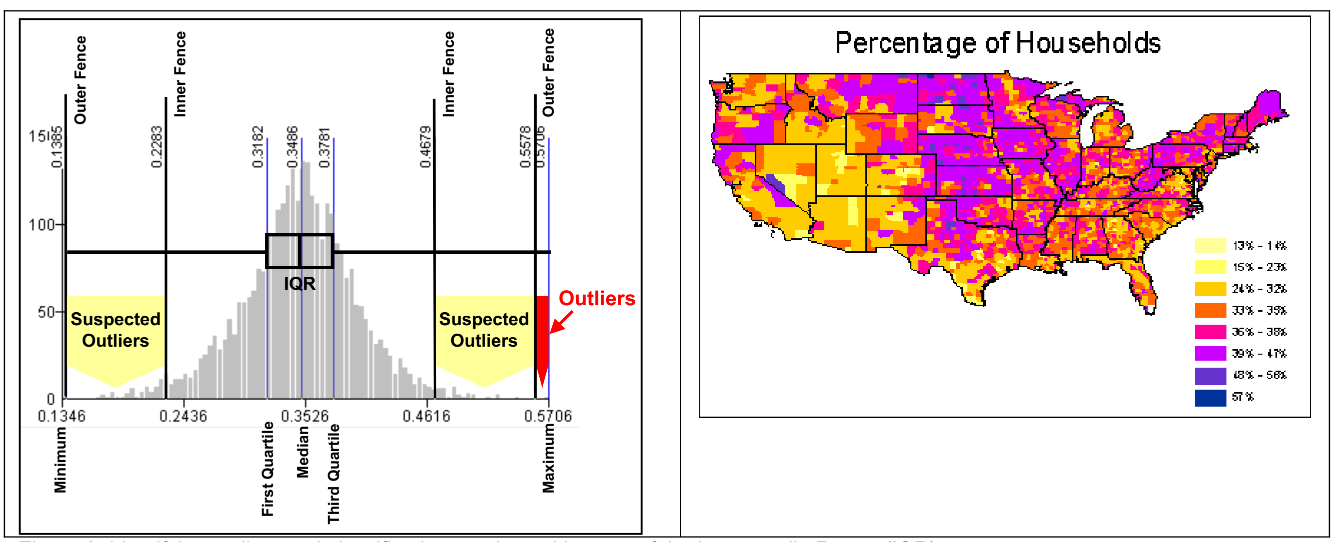


Figure 8. Identifying outliers and classification creaks making use of the Interquartile Range (IQR).

Using Census Data to Support Rice University Researching Projects (Examples)

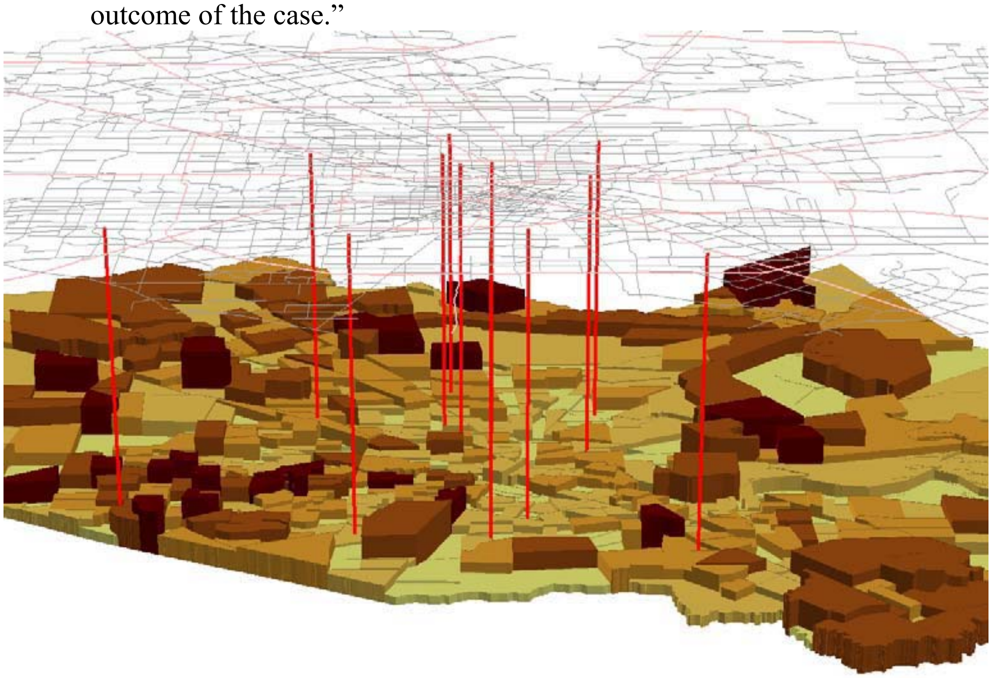
At Rice University, scholars use Census data for research or to support research projects; architecture students to examine demographic patterns in their area of studies and have a better understanding about the social characteristics of their sites; political and social science researchers to analyze demographic and social patterns influencing electoral and social behaviors; history students to determine migration patterns, and civil engineering students to evaluate the social impact of development areas. All of them are using the same dataset with a different focus.

At Rice University, the GIS/Data Center has demonstrated the effectiveness of a centralized center providing access to GIS, making demographic datasets accessible, and offering technical support. Through the following examples some scholars share their experiences in the use of GIS/Data center services for the integration of census information into their researching projects.

LINKING CAPITAL MURDER DEFENDANTS AND VICTIMS TO CENSUS

Dr. Scott Phillips
Assistant Professor, Department of Sociology, Rice University, 2003-2005, now at the Department of Sociology and Criminology, University of Denver.

> "I have linked capital murder defendants and victims to census block groups to see if the social characteristics (e.g. median household income) of the parties' neighborhoods influence the outcome of the case."



MAPPING JUSTICE DEPARTMENT OBJECTIONS TO RACIALLY DISCRIMINATORY ELECTION PROCEDURES IN STATES COVERED BY SECTION 5 OF THE VOTING RIGHTS ACT

By Chandler Davidson

Radoslav Tsanoff Professor of Public Affairs Emeritus

As a member of the National Commission on the Voting Rights Act in 2005-6, I had the task of examining the extent of racial discrimination in the electoral process during the almost 25 years since the temporary features of Act were last renewed by Congress in 1982. This was one small but important part of the Commission's published report, of which I was the primary author.

Through a Freedom of Information request to the U.S. Department of Justice (DOJ), I obtained several data sets indicating vote discrimination against racial minorities.

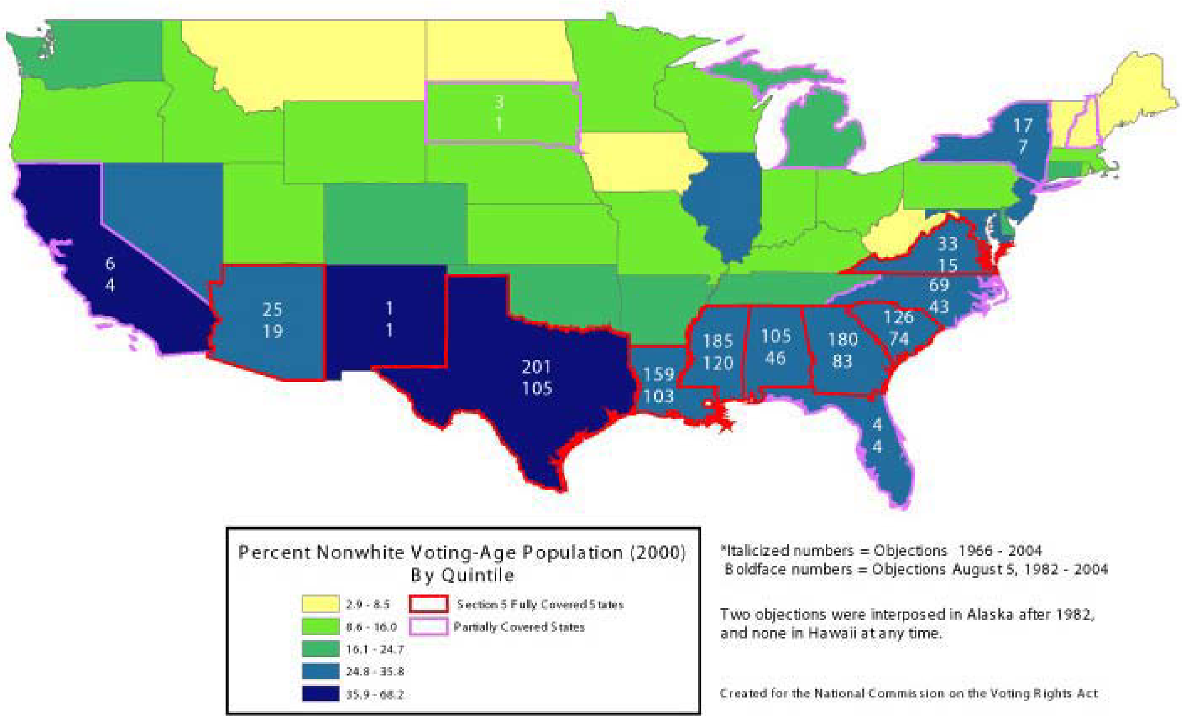
One of the most useful lists concerned DOJ communication with various political jurisdictions regarding proposed changes in electoral procedures. Under Section 5 of the Act, any state government or political subdivision within the state such as a city or county in 9 entire states, mostly in the Deep South, and various counties or townships in several other states, are required to submit either to the DOJ or the U.S. District Court for the District of Columbia any proposed changes in electoral procedures for approval or "preclearance." Only if approval is received can the changes be implemented. If the Justice Department believes the changes have a racially discriminatory intent or would have a discriminatory effect it sends the submitting entity an "objection letter" prohibiting the change (or changes) from being implemented. An objection letter, in other words, is an indication that one or more discriminatory changes a covered jurisdiction had hoped to implement were prevented by the DOJ from occurring.

I wanted to discover the geographical spread of the 626 objection letters sent to the states covered by Section 5, and to display this information in an interesting and easily understandable way. Moreover, I wanted to show how the location of the political subdivisions within the various states correlated with the percent nonwhite population of voting age in 2000. I decided to create GIS maps of the DOJ objection letters sent between 1982 and 2004, showing first the number of such letters statewide and then the number of letters by county within each state. I was very lucky to have a number of Rice undergrad and grad research assistants who were taught to use GIS for making these maps. German Diaz in the Rice GIS Lab was a great help in teaching them and helping them with problems that continually arose over the course of the project. Eva Garza, the GIS Lab director, was also a great help.

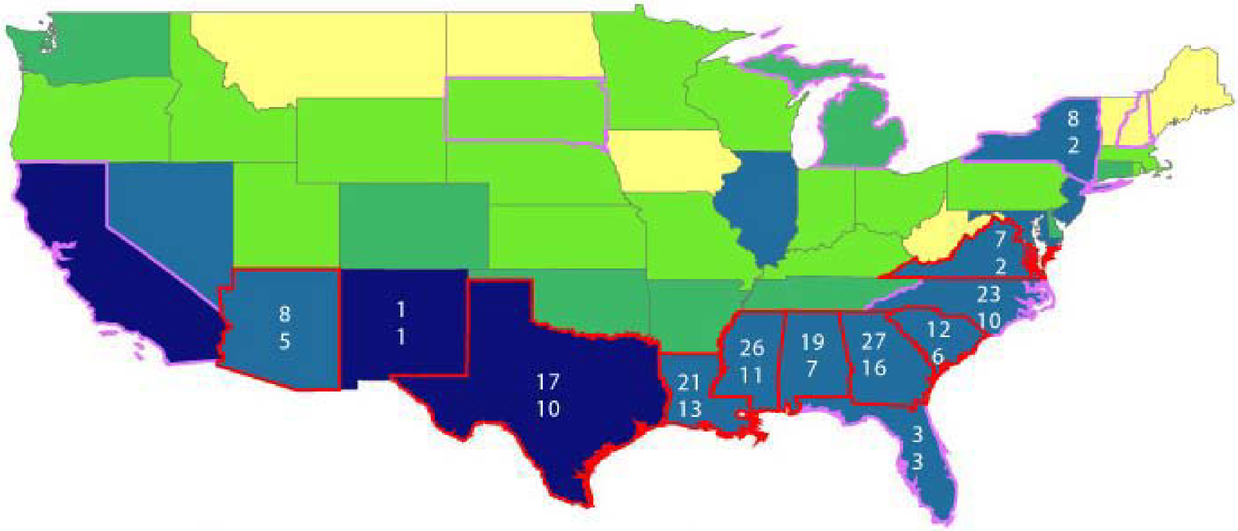
Data on the percent non-white voting-age population, both at the state and county level, was calculated from the 2000 census. "Nonwhite" was defined for our purposes as all people of voting age (not merely citizens) who were of any race except non-Hispanic whites.

The maps I have described, as well as a great many others which show the geographical distribution of other indicators of racial discriminatory voting behavior, were published in the report of the National Commission on the Voting Rights Act in February 2006. This report, entitled Protecting Minority Voters: The Voting Rights Act at Work 1982-2005, can be accessed at http://www.votingrightsact.org. The GIS maps are found in the map section at the back of the report, and the maps relating to DOJ objection letters, described above, are Maps 5A-5K. These same maps are contained in pdf files I have sent along with this description.

MAP 5A
Objections
(1966 - 2004 and August 5, 1982 - 2004)*



MAP 5B
Statewide Objections Only
(1966 - 2004 and August 5, 1982 - 2004)*



Percent Nonwhite Voting-Age Population (2000) By Quintile

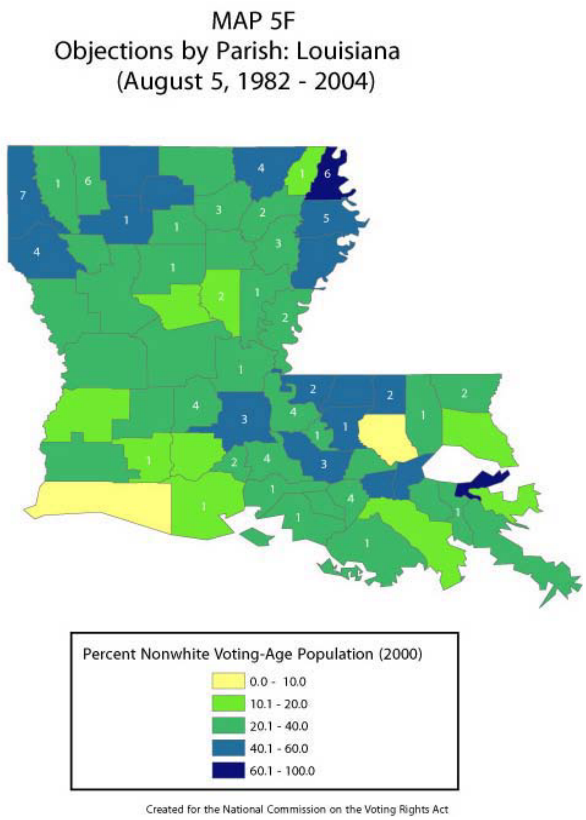
- 0.0 - 8.5
- 8.6 - 16.0
- 16.1 - 24.7
- 24.8 - 35.8
- 35.9 - 68.2

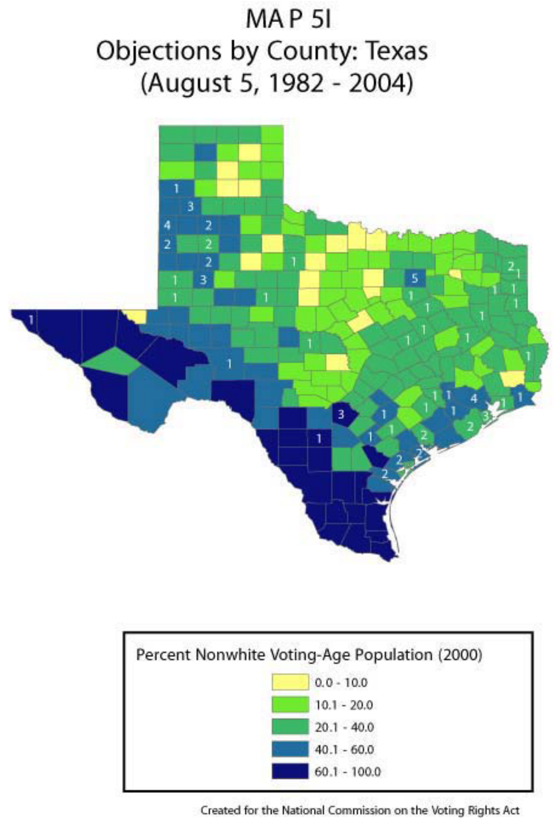
- Section 5 Fully Covered States
- Partially Covered States

*Italicized numbers = Objections 1966 - 2004
Boldface numbers = Objections August 5, 1982 - 2004

Two objections were interposed in Alaska after 1982, and none in Hawaii at any time.

Created for the National Commission on the Voting Rights Act

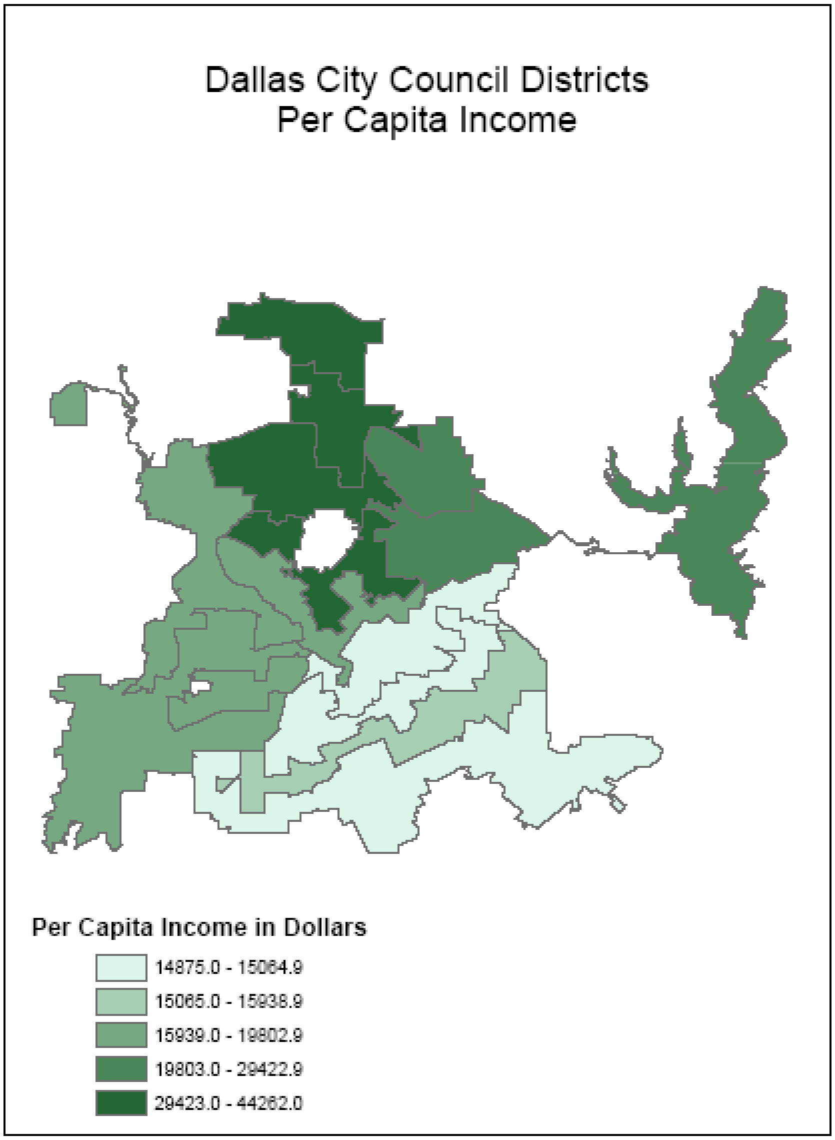




STUDY OF MINORITY INCORPORATION IN LOCAL POLITICS

Dr Melissa J. Marschall
Associate Professor, Department of political science, Rice University

"This map is part of a large-scale study of minority incorporation in local politics. The project considers multiple demographic and economic characteristics within local voting districts in order to find certain characteristics that may be related to the election of minority officials."



ANGLO VOTING ON NATIVIST BALLOT INITIATIVES: THE PARTISAN NATURE OF THE IMPACT OF SPATIAL PROXIMITY TO THE BORDER OF MEXICO

Regina Branton, Gavin Dillingham, Johanna Dunaway, Beth Miller
Department of Political Science, Rice University

"In this study, we examine how contextual factors influence voting behavior on nativist-oriented ballot initiatives using the case of Proposition 227. We argue that spatial proximity to the border, in addition to the ethnic environment in which a person resides, is associated with voting behavior on nativist

initiatives. The results indicate that spatial proximity to the border is significantly related to individual-level voting on nativist initiative. Further, we find that the impact of proximity to the border varies as a function of individual-level partisan affiliation."

Figure 2: Probability of Support for Proposition 227

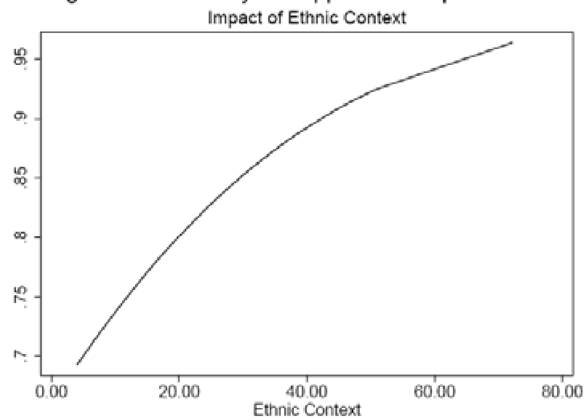
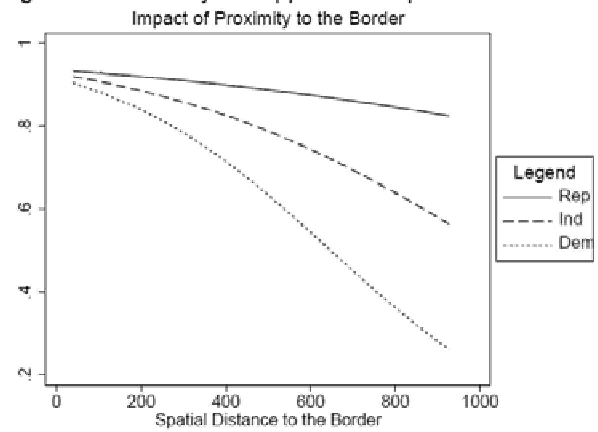


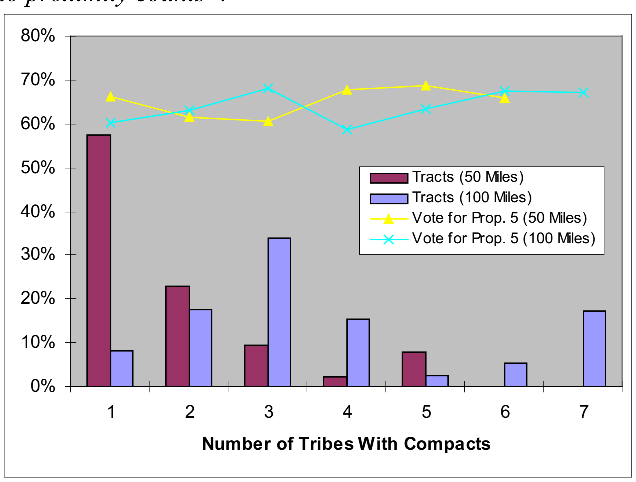
Figure 1: Probability of Support for Proposition 227



CLOSE ENOUGH FOR COMFORT: A SPATIAL ANALYSIS OF GAMING INITIATIVES IN CALIFORNIA

Frederick J. Boehmke, University of Michigan and University of Iowa
Gina Branton, Rice University
Gavin Dillingham , Rice University
Richard Witmer, Creighton University

"We suggest that Geographic Information Systems provides a unique opportunity to test information and economic explanations of voter choice on ballot measures. By examining voter preferences at the census block level during initiative elections in California in 1998 and 2000, we find that proximity is a key explanatory variable, suggesting that to voters, closeness, and the information that is available due to proximity counts".



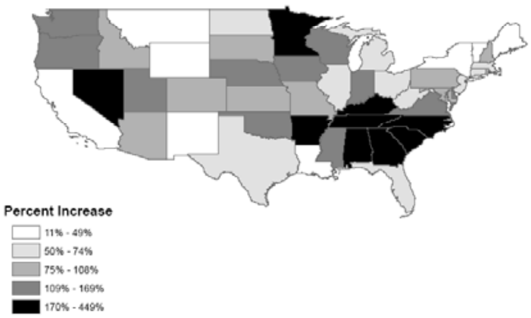
Distribution of Tracts and Mean Vote Share by Number of Tribes with Gaming and Distance

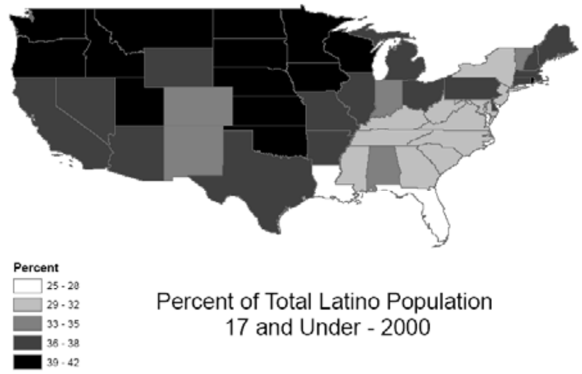
1990-2000 LATINO POPULATION CHANGES AFFECTING THE LATINO REPRESENTATION IN SCHOOL BOARDS

Paru Sahah
Doctoral Student, Department of political science, Rice University

"These maps were generated as part of project that examined how changes in the Latino population between 1990 and 2000 can partly explain different likelihoods of Latino representation in school boards. As argued in the project, while the Latino population overall matters for representation, it is in the end the eligible population that determines how likely it is that a Latino will be selected for the school board. Thus, we looked at the proportion of the population eligible to vote in school board elections (18+ years). In addition, we looked at how rapid changes in the Latino population posed greater challenges for some school districts over others. Our findings suggest that both these factors are important variables to consider when examining Latino representation."

Figure 2.1
Change in Latino Population 1990 - 2000





References

[1] Peters, A. and MacDonald, H., 2004, Unlocking the Census with GIS. Redlands, CA: ESRI Press.

[2] U.S. Census Bureau. 2006. Decennial Census – American FactFinder – Historic Snapshot of the Nation. http://factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_lang=en&_sse=on&_content=sp4_decennial.html&_title=Decennial+Census. (Accessed: July, 2006)

[3] Mitchell, A., 1999, The ESRI Guide to GIS Analysis, Vol. 1: Geographic Patterns & Relationships. Redlands, CA: ESRI Press.

[4] Hogg, R. and E. Tanis, 2001, Probability and Statistical Inference, 6th Edition, Prentice-Hall Inc.