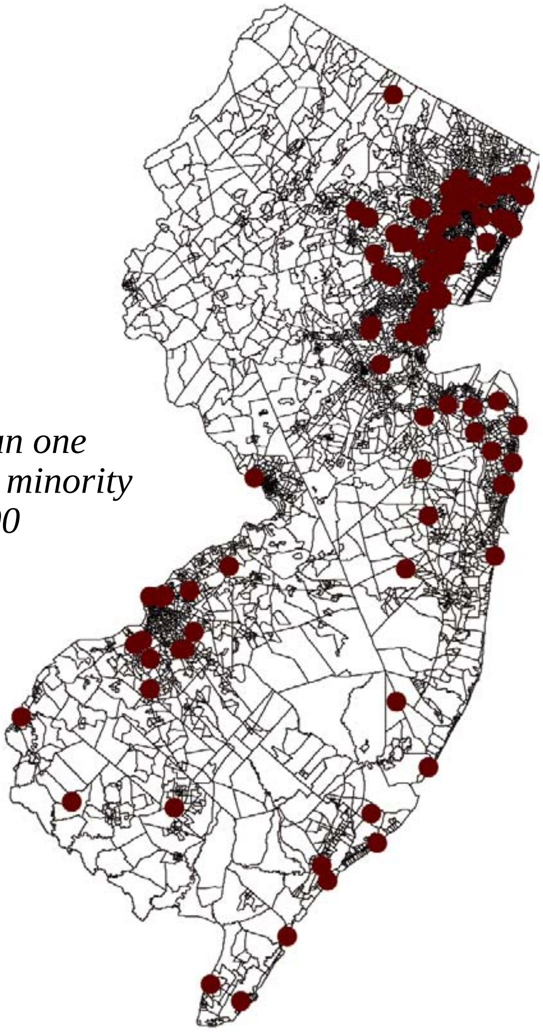
United States
Environmental Protection
Agency

Office of Research and
Development
Washington, D.C. 20460

EPA/600/R-01/078
October 2001

Guidance for Statistical Determination of Appropriate Percent Minority and Percent Poverty Distributional Cutoff Values Using Census Data for an EPA Region II Environmental Justice Project



Q: *In a random location, can one determine the level of % minority and % poverty within 100 contiguous census block groups?*

Q: *Does spatial distribution and nature of the census block groups dictate the clumping of the sample locations in a highly populated area?*

EPA/600/R-01/078
October 2001

Guidance for Statistical Determination of Appropriate Percent Minority and Percent Poverty Distributional Cutoff Values Using Census Data for an EPA Region II Environmental Justice Project

by

M.S. Nash, G.T. Flatman, D.W. Ebert, and C.L. Cross

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Las Vegas, Nevada

Notice

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD), funded and performed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

Preface

The purpose of this report is to assist Region II by providing a statistical analysis identifying the areas with minority and below poverty populations known as “Community of Concern” (COC). The aim was to find a cutoff value as a threshold to identify a COC using demographic data. Other consultants were also involved to provide similar information. Region II presented our method for the Senior Mangers on June 2000, as a comparison with another two methods: cluster-based cutoff and state averages. A decision was made to use the cluster-based cutoff and state average because they were easier to understand and to use at the community level. Although our method was not the preferred one, there was a significant amount of time and effort put forth by the authors to develop the methodology, and we feel the technique is a valid one with possible future uses.

Table of Contents

List of
Tables and Figures

Appendices

List of Abbreviations

TotPop90	Total Population in 1990; Universe persons
Pov_univ	All persons for whom poverty status is determine
Bel_Pov	Below poverty level; Universe: persons for whom poverty status is determine
P_belPov	Percent below poverty level; Universe: persons for whom poverty status is determine; Calculation: *Bel_Pov* / *Pov_univ* * 100
NhispWht	Non-hispanic White; Universe: persons
Nhispblk	Non-hispanic Black; Universe: persons
Nhispnat	Non-hispanic American Indian, Eskimo, or Aleut; Universe: persons
Nhispas	Non-hispanic Asian or Pacific Islander; Universe: persons
Nhispoth	Non-hispanic Other race; Universe: persons
Hisp_wht	Hispanic White; Universe: persons
His_blk	Hispanic Black; Universe: persons
Hisp_nat	Hispanic American Indian, Eskimo, or Aleut; Universe: persons
Hisp_as	Hispanic Asian or Pacific Islander; Universe: persons
Hisp_oth	Hispanic Other race; Universe: persons
Per_min	Percent minority; Universe: persons; Calculation: [(*Hisp_wht* + *His_blk* + *Hisp_nat* + *Hisp_as* + *Hisp_oth* + *Nhispblk* + *Nhispnat* + *Nhispas* + *Nhispoth) / TotPop90*] * *100*

Section 1

Introduction

The goal of this project is to identify a GIS and a statistical procedure which will objectively, reproducibly, and statistically identify a "Community of Concern" (COC) which is defined as a community with a "minority" or "below-poverty" population. We shall demonstrate the procedure using the census data for the state of New Jersey and New York located in EPA's Region II. This exercise in classification sounds straightforward and doable, but the choice of threshold values or cutoff values and changes of scale (e.g., census block groups to counties) changes the number and location of the COC, and may raise questions and criticism. An objective statistical algorithm is needed for identifying and locating the COC on the map of the Region. This is a non-trivial statistical problem. Because the data have time and space dimensions and skewed probability distributions, hypothesis testing, confidence intervals, and ratios and proportions are inappropriate and hence have the potential to mislead decision-makers.

Descriptive analyses of the probability distribution of the data when aggregated to the appropriate scale (census block or group, census tract, town, township, county, state, or region) is an appropriate approach for the data and will give the desired quality for identification of a COC. Decisions will be made from the probability of the cutoff, not from arbitrary cutoff. In this context, it is important to define units and scale. The basic (indivisible) sampling unit of data or information is the census "block group." The decision unit changes (e.g., census block group, census tract, township, county, or state) and is chosen by the specific question to be answered. To change scale to a different decision unit other than the census block group (sampling unit), all of the spatially included sampling units in the new decision unit must have the counts of their characteristics summed over the desired decision unit and the desired percentages recomputed. The counts or frequencies are additive but the percentages or relative frequencies (probabilities) are not.

The probability distribution is a useful statistical tool to measure the population of all decision units of a given scale (e.g., census tract, township, county, . . .). By choosing the cutoff probability at the 80th percentile for the characteristic of "minority" and the characteristic of "below poverty" in the population of all ***census block groups*** *decision unit*, the cutoff values associated with the cutoff probability are 48% and 68% for minority and 12% and 22% for below poverty, for New Jersey and New York, respectively (Table 1; Figures 1 & 2). It is not obvious that these cutoff values have anything in common, and they sound arbitrary, but in the probability of the population distribution they are determined (back transformed) by equal probability (80th percentile). It is important to note that the cutoff values associated with the equal probability decrease with a growth in area of the decision unit; this is to be expected from spatial statistics. It is also important to note that the cutoff values depend on locations of the area where the samples were taken. The cutoff values for the same probability (80th percentile) for the distribution of ***census tracts*** *decision units* are 56% and 77% for minority and 13% and 22% below poverty for New Jersey and New York, respectively (Table 1; Figures 3 & 4). The cutoff values for the same probability for the distribution of the county decision unit are 31% and 14% for minority and 10% and 13% for below poverty, for New Jersey and New York, respectively (Table 1; Figures 5 & 6). The commonality is their equal probability of the 80th percentile of their respective distributions. Thus the choice of COC will be based on a cutoff of “equal probability” instead of a cutoff of an arbitrary value (e.g., 50% minority or 50% below poverty). In summary, equal probability, as measured by the chosen highest percentile of the distribution of the data aggregated to the decision unit, will give the COC areas without using arbitrary cutoff values or percentages of “minority” or “below poverty.”

Table 1. Summary of cutoff values associated with the three decision units from the “minority” and “below poverty” statistical analysis of the US EPA Region II (New York and New Jersey) Environmental Justice Study. In all cases, the sampling unit is the census block group. A “*” indicates the state cutoff values are not based on the 80th percentile; these are the values for state as decision unit.

	Minority Cutoff (%)		**Below Poverty Cutoff (%)**	
Decision Unit	**New Jersey**	**New York**	**New Jersey**	**New York**
Census Block Group	48	68	12	22
Census Tract	56	77	13	22
County	31	14	10	13
State*	26	31	8	13

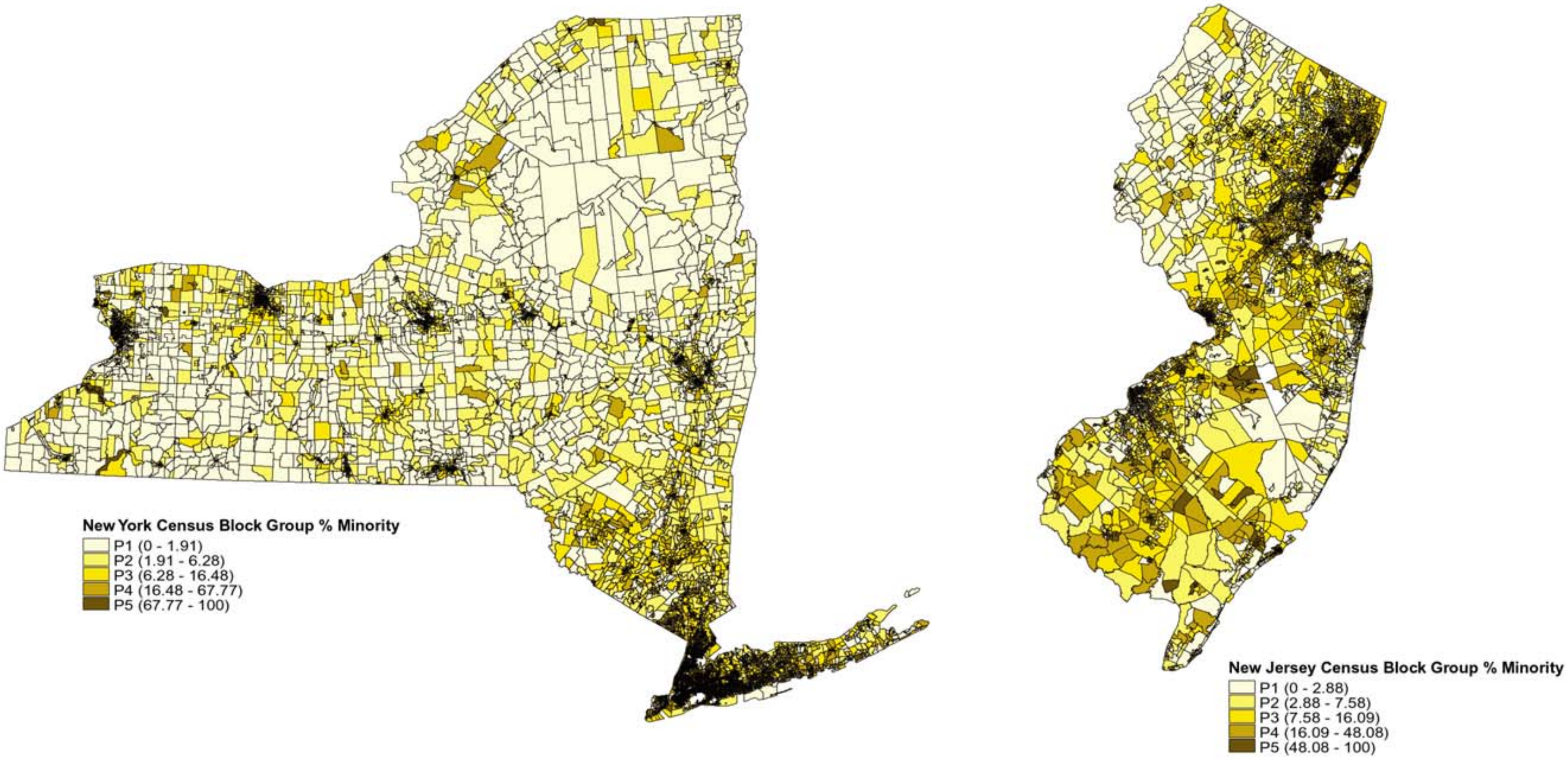


Figure 1. The Five Percentiles and Their Values for % Minority by Block Group.

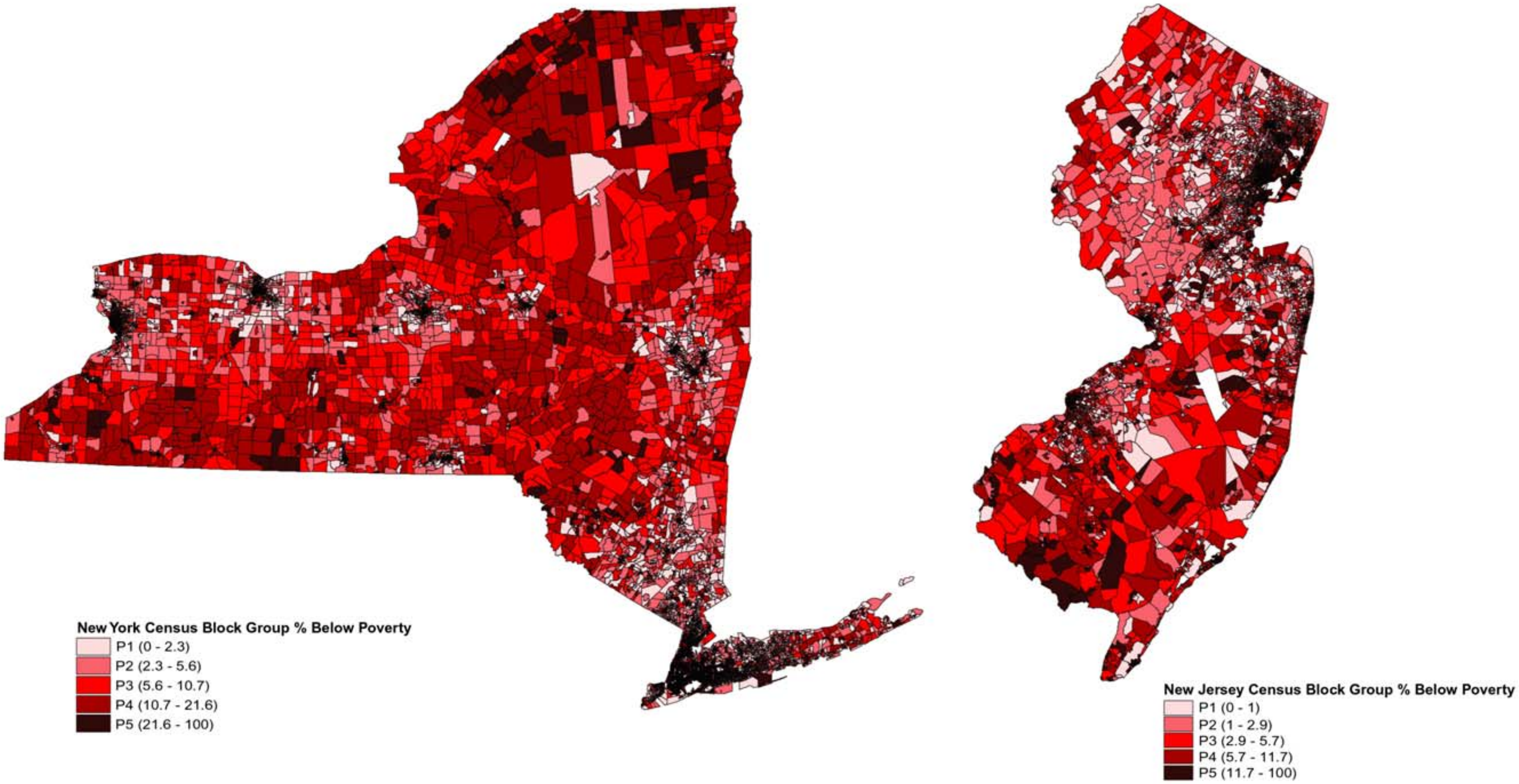


Figure 2. The Five Percentiles and Their Values for % Below Poverty by Block Group.

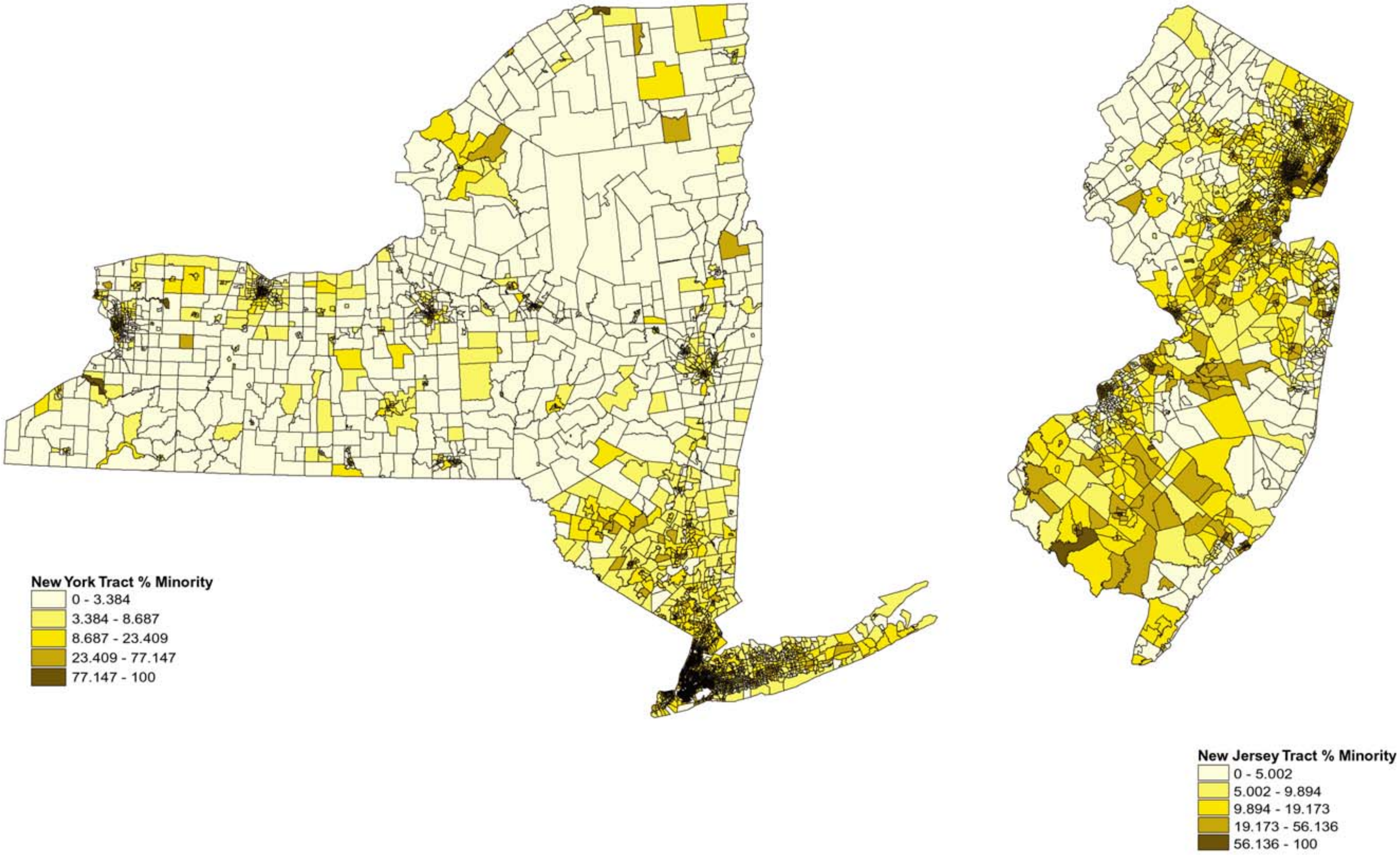


Figure 3. The Five Percentiles and Their Values for % Minority by Tract.

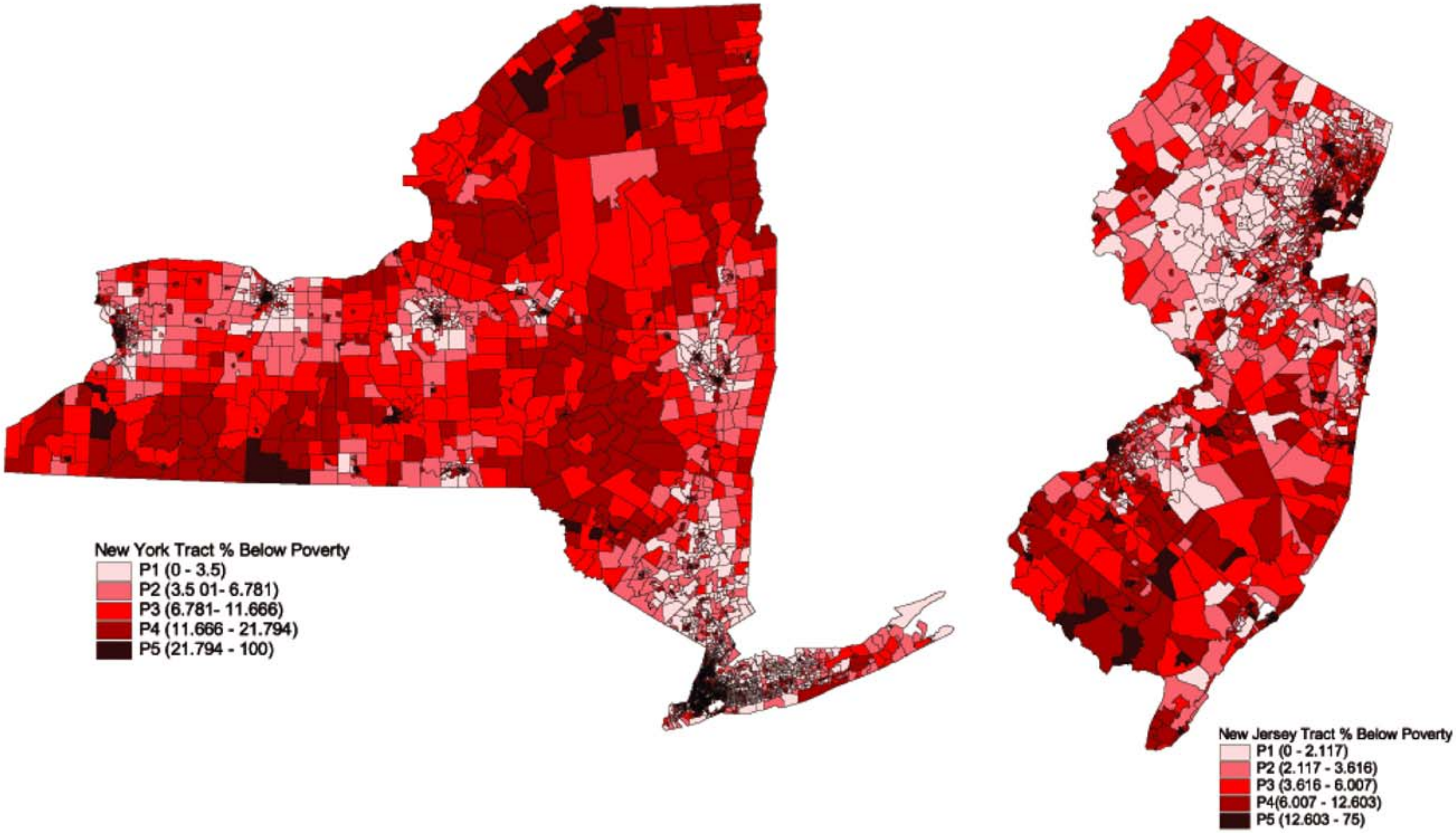


Figure 4. The Five Percentiles and Their Values for % Below Poverty by Tract.

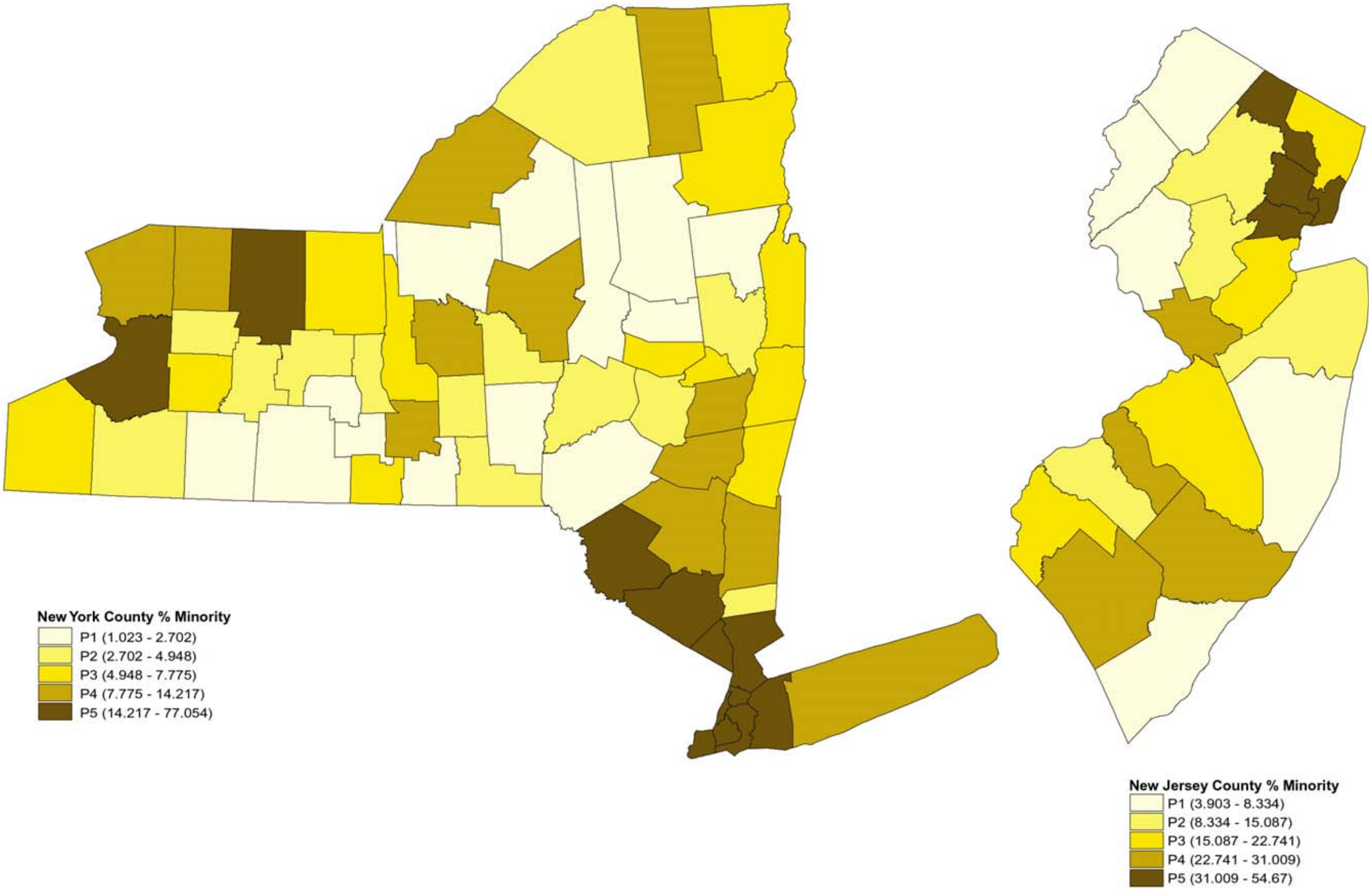


Figure 5. The Five Percentiles and Their Values for % Minority by County.

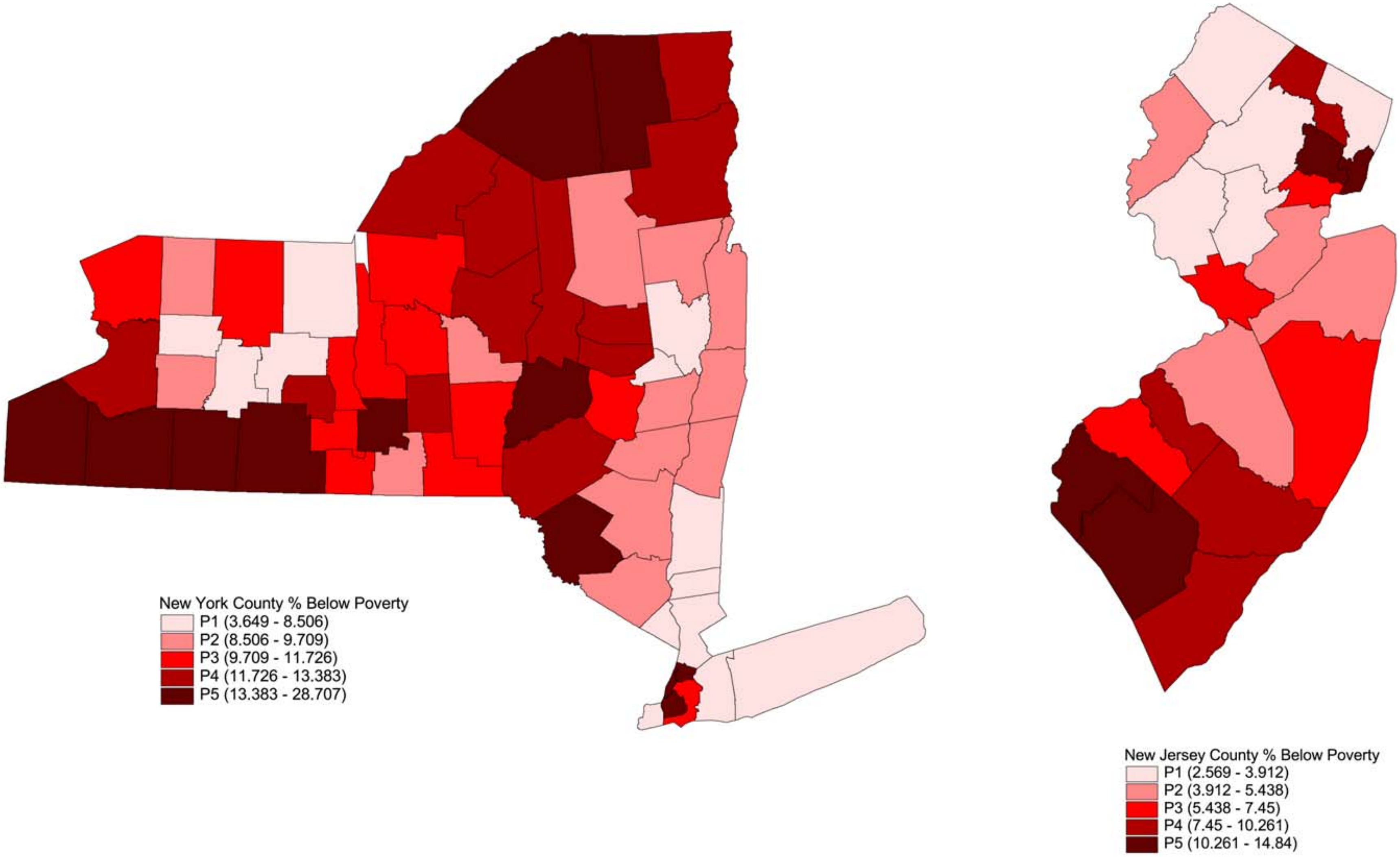


Figure 6. The Five Percentiles and Their Values for % Below Poverty by County.

Section 2
Sampling and Decision Units

Two statistical units were identified: (1) decision units and (2) sampling units. These units were used to determine whether a community was/was not a minority and/or below poverty. The sampling unit is the census block group and the decision unit can be any unit that is equal to or larger than the census blocking group. For a preliminary attempt, we used census block group, tract, and county units as decision units. We used three combinations of sampling and decision units to examine the relative frequency of minority and below poverty. The three combinations were:

1. Decision unit is census block group and sampling unit is census block group,
2. Decision unit is census tract and sampling unit is census block group, and
3. Decision unit is county and sampling unit is census block group.

Section 3

Distribution and Cutoff Value

Initially, a histogram was developed using blocking group percent minority *(Per_Min*) and percent below poverty (*P_belPov*) for each county and state (Appendices 1a - 1d). We visually examined the distribution of each histogram, and along with the five equal probability percentiles of the ARCview maps, a decision cutoff value was defined. A different cutoff value for each of these two variables was made. Mathematical derivation of the percent minority and percent below poverty for decision units is explained below:

3.1 Decision Unit Is Census Block Group and Sampling Unit Is Census Block Group

For this we used the *Per_min* and *P_belPov* variables that were provided to us by Region II and subsequently verified and recalculated by scientists in Las Vegas prior analysis (See “GIS Remediation” and Appendices 2a - 2c).

3.2 Decision Unit Is Census Tract and Sampling Unit Is Census Block Group

To calculate % minority and % below poverty at the tract level, counts must be used rather than census block group percentages. Counts of minority (summation of *Hisp_wht, Hisp_blk, Hisp_nat, Hisp_as, Hisp_oth, Nhispblk, Nhispnat, Nhispas, and Nhispoth*), *TotPop90, Bel_Pov, and Pov_Univ* from each census block group were used. Relative frequencies for minority and below poverty at the level of the census tract were calculated. Tract percent minority and percent below poverty are the relative frequencies times 100. Calculations were done as follows:

a) Tract % minority

$$\text{Tract \% Minority} = \frac{\sum_{i=1}^{t} m_i}{\sum_{i=1}^{t} T_i} \times 100$$

Where,

∑ = summation,
t = total number of block groups in a given census tract,
i = census block group (i = 1, 2, ..., t),
m = counts of minority in each census block group, and
T = *TotPop90* = count of total population in a census block group.

b) Tract % below poverty:

$$\text{Tract \% Below Poverty} = \frac{\sum_{i=1}^{t} (BP)_i}{\sum_{i=1}^{t} P_i} \times 100$$

Where,

∑ = summation,
t = total number of block groups in a given census tract,
i = census block group (i = 1, 2, ..., t),
BP = *Bel_Pov* = count of Below Poverty in each census block group, and
P = *Pov_Univ* = count of all people who reported their income in each census block group.

3.3 Decision Unit Is County and Sampling Unit Is Blocking Group

To calculate % minority and % below poverty at the county level, counts must be used rather than percentages. Counts of minority (summation of *Hisp_wht*, *Hisp_blk*, *Hisp_nat*, *Hisp_as*, *Hisp_oth*, *Nhispblk*, *Nhispnat*, *Nhispas*, and *Nhispoth*), *TotPop90*, *Bel_Pov*, and *Pov_Univ* from each census block group were used. Relative frequencies for the minority and below poverty at the level of the county were calculated. County percent minority and percent below poverty are the relative frequencies times 100. Calculations were done as follows:

a) County % minority:

$$\text{County \% Minority} = \frac{\sum_{i=1}^{c} m_i}{\sum_{i=1}^{c} T_i} \times 100$$

Where,

∑ = summation,
c = total number of census block groups in a given county,
i = census block group (i = 1, 2, ..., c),
m = counts of minority in each census block group, and
T = *TotPop90* = count of total population in a census block group.

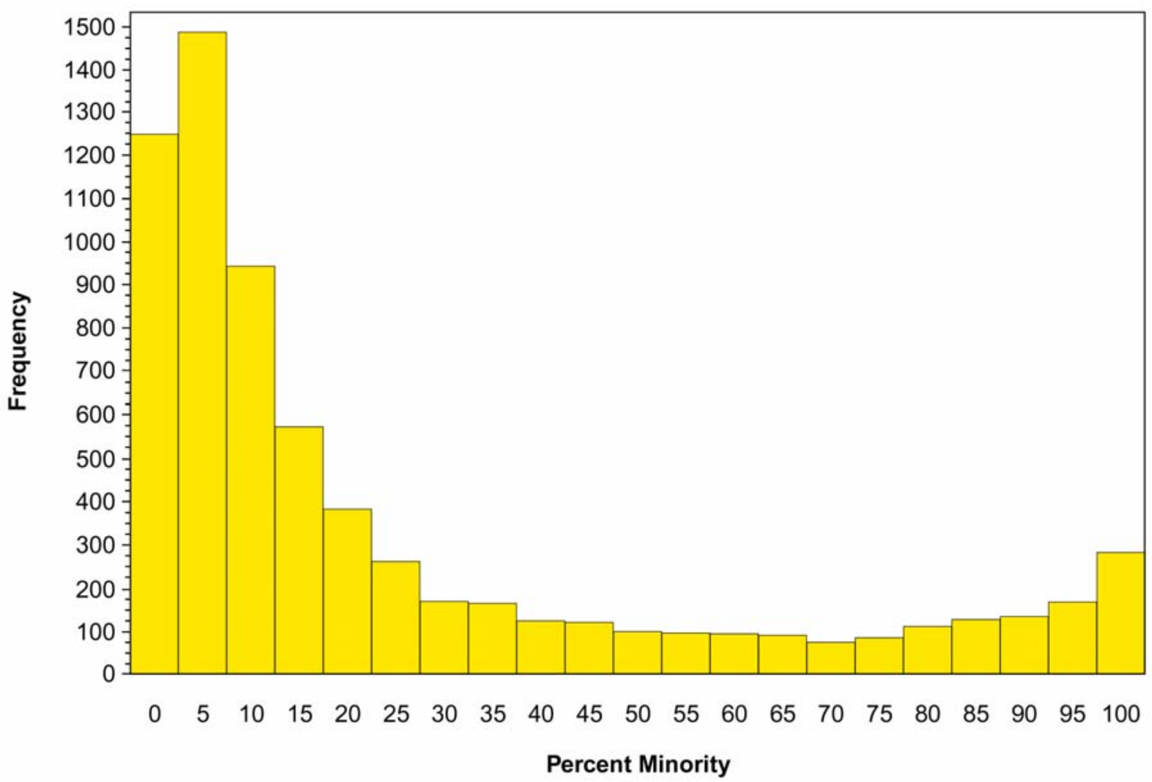
b) County % Below Poverty:

$$\text{County \% Below Poverty} = \frac{\sum_{i=1}^{c} (BP)_i}{\sum_{i=1}^{c} P_i} \times 100$$

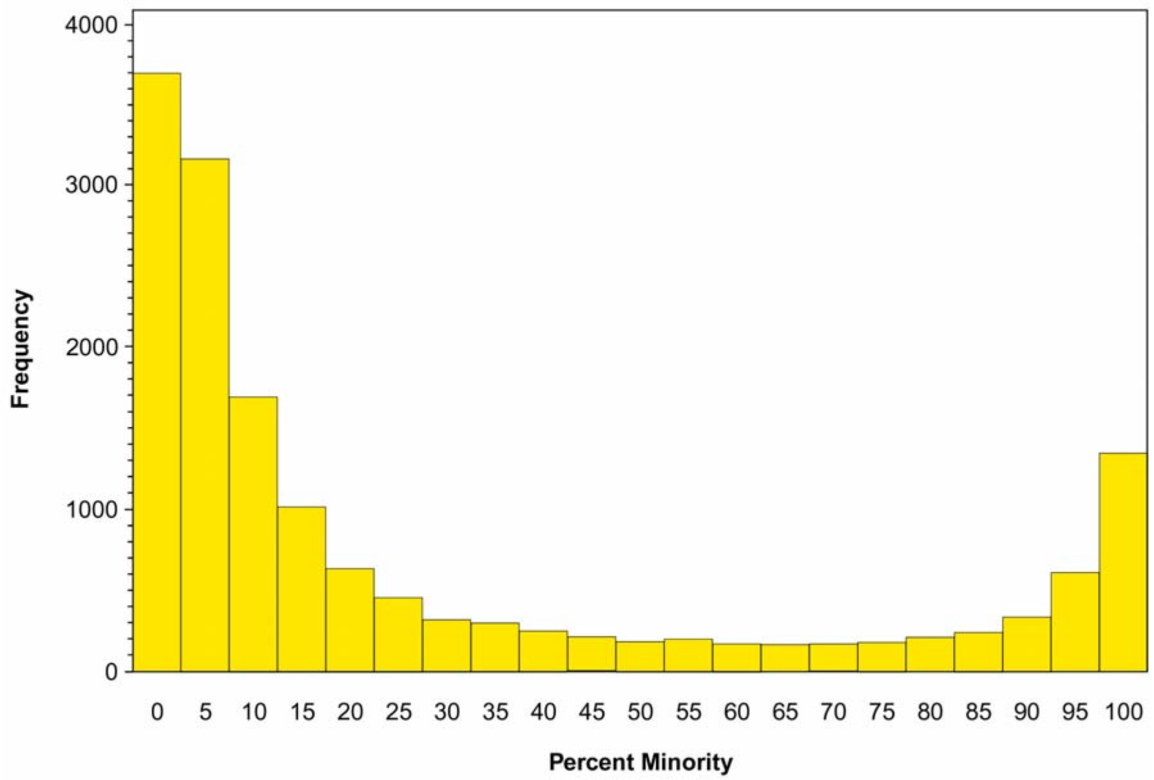
Where,

$\sum$ = summation,
c = total number of census block group in a given county,
i = census block group (i = 1, 2, ..., c),
BP = *Bel_Pov* = count of Below Poverty in each census block group, and
P = *Pov_Univ* = count of all people who reported their income in each census block group.

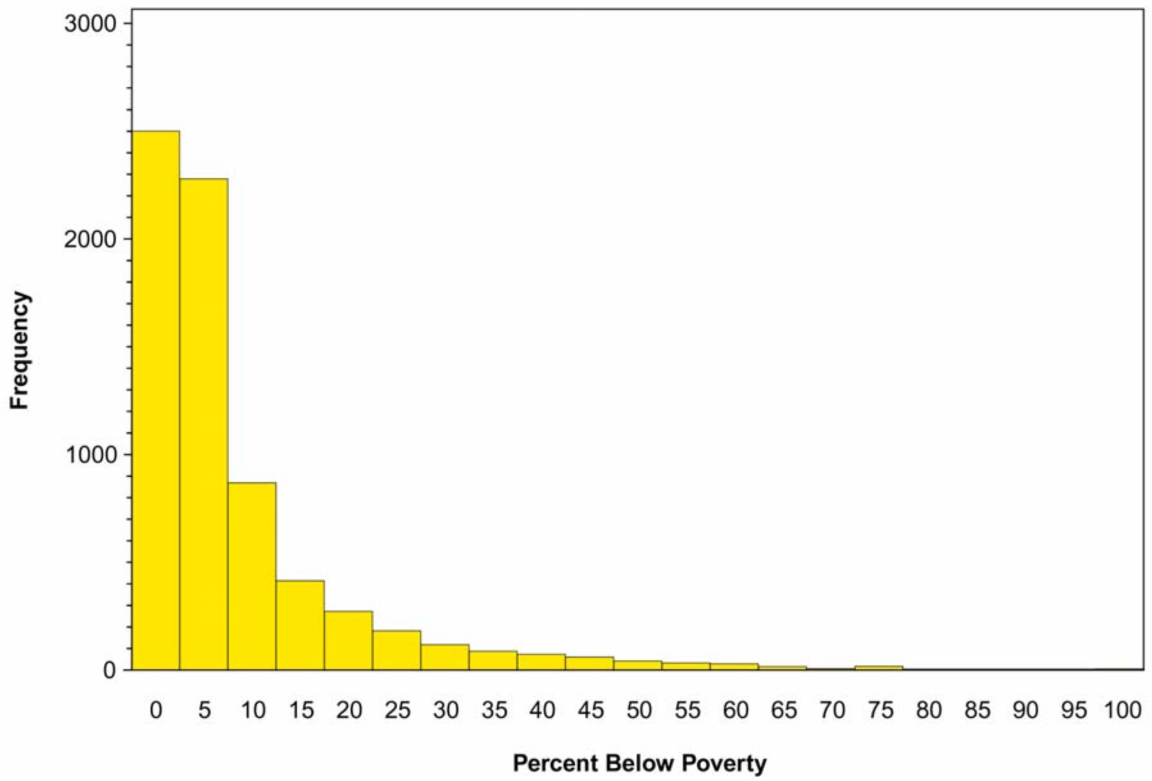
It is important to note that we excluded block groups with *TotPop90* and *Pov_univ* of zero value prior posting their five percentile values on maps. This also has to be considered in any other analyses such as clusters and averages; otherwise, different analyses will result in non comparable results.



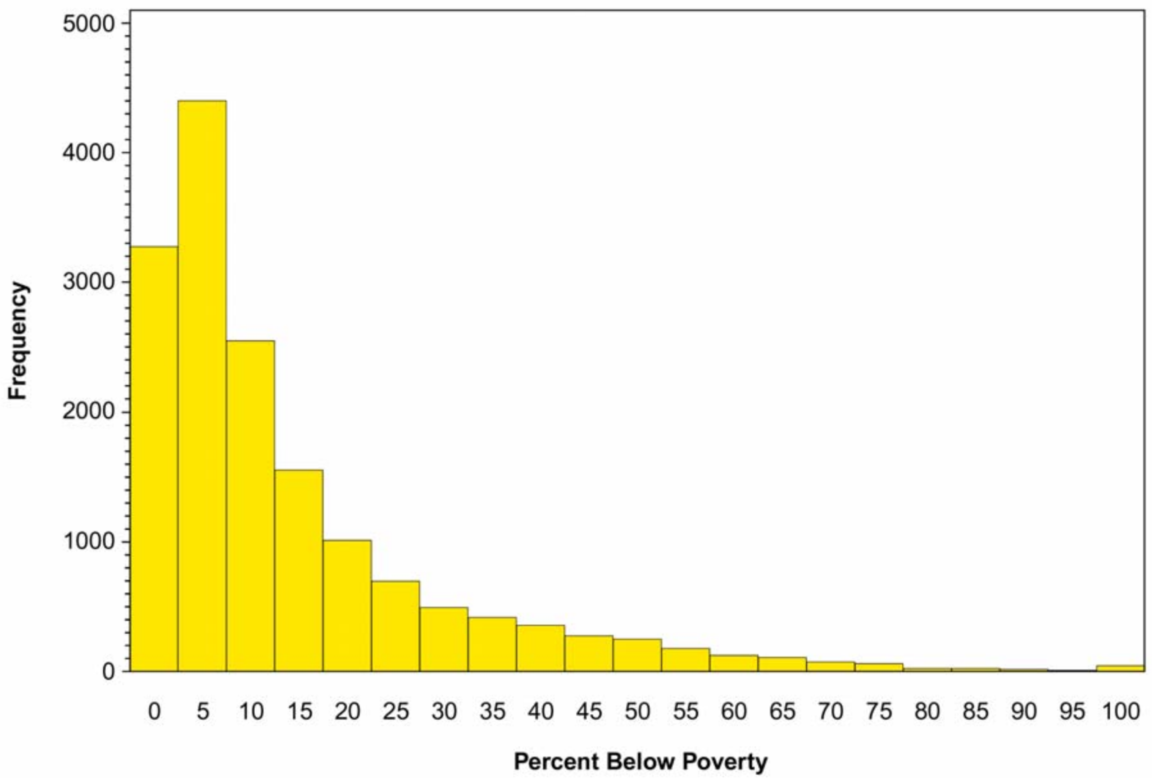
Appendix 1a. Percent minority in New Jersey. Values are from the sampling unit (a census block group).



Appendix 1b. Percent minority in New York. Values are from the sampling unit (a census block group).



Appendix 1c. Percent minority in New Jersey. Values are from the sampling unit (a census block group).



Appendix 1d. Percent minority in New York. Values are from the sampling unit (a census block group).

Section 4
Distribution and Cutoff for Re-Sampling

There was a need to demonstrate the application of the above analysis on randomly aggregated numbers of contiguous census blocks in each state. This was done to simulate the results for larger decision units than that of the census block group, decision units such as townships, tract, and/or county. We generated 100 samples of contiguous census block groups at the following size groupings: 50, 100, 150, 200, and 250 contiguous census block group. The %minority and %below poverty were calculated for these simulated groups and their corresponding 80th percentiles were determined (Table 2). The overall trend was for cutoff values to decrease as the number of neighbors increased.

Table 2. Number of neighboring census block groups (No.) from random selection, and five percentiles for percent minority for New Jersey and New York. 100th percentile is the maximum value.

State	No.	20th	40th	60th	80th	100th
New Jersey	50	9.35	15.24	30.71	57.03	93.92
	100	10.96	17.27	26.72	55.30	96.71
	150	10.96	17.16	29.26	55.62	93.40
	200	13.67	20.34	26.02	47.13	87.38
	250	13.52	22.05	26.70	41.40	82.74
New York	50	5.93	13.66	25.54	56.94	99.42
	100	8.18	13.14	27.53	61.41	98.80
	150	6.48	11.73	22.12	48.92	98.37
	200	6.57	14.43	24.63	60.12	98.55
	250	9.53	16.25	26.36	54.84	96.83

The locations of the central block group for the 100 samples of the 100 contiguous block group simulation for New Jersey and New York are shown in Figure 7. The apparent clumping of the sample locations in highly populated areas is due to the spatial distribution and nature of the block groups. In New York, sample locations were mostly in New York City and Buffalo, and in New Jersey, they were mostly in Jersey City, Newark, Staten Island, Hackensack and Camden (Figure 7). Block groups are drawn to include approximately an equal number of people. Therefore, block groups in densely populated areas are smaller in size and occur in greater numbers than in rural areas. It follows then that if 90% of the block groups occur in urban areas, then 90% of randomly selected groups will fall within these same areas.

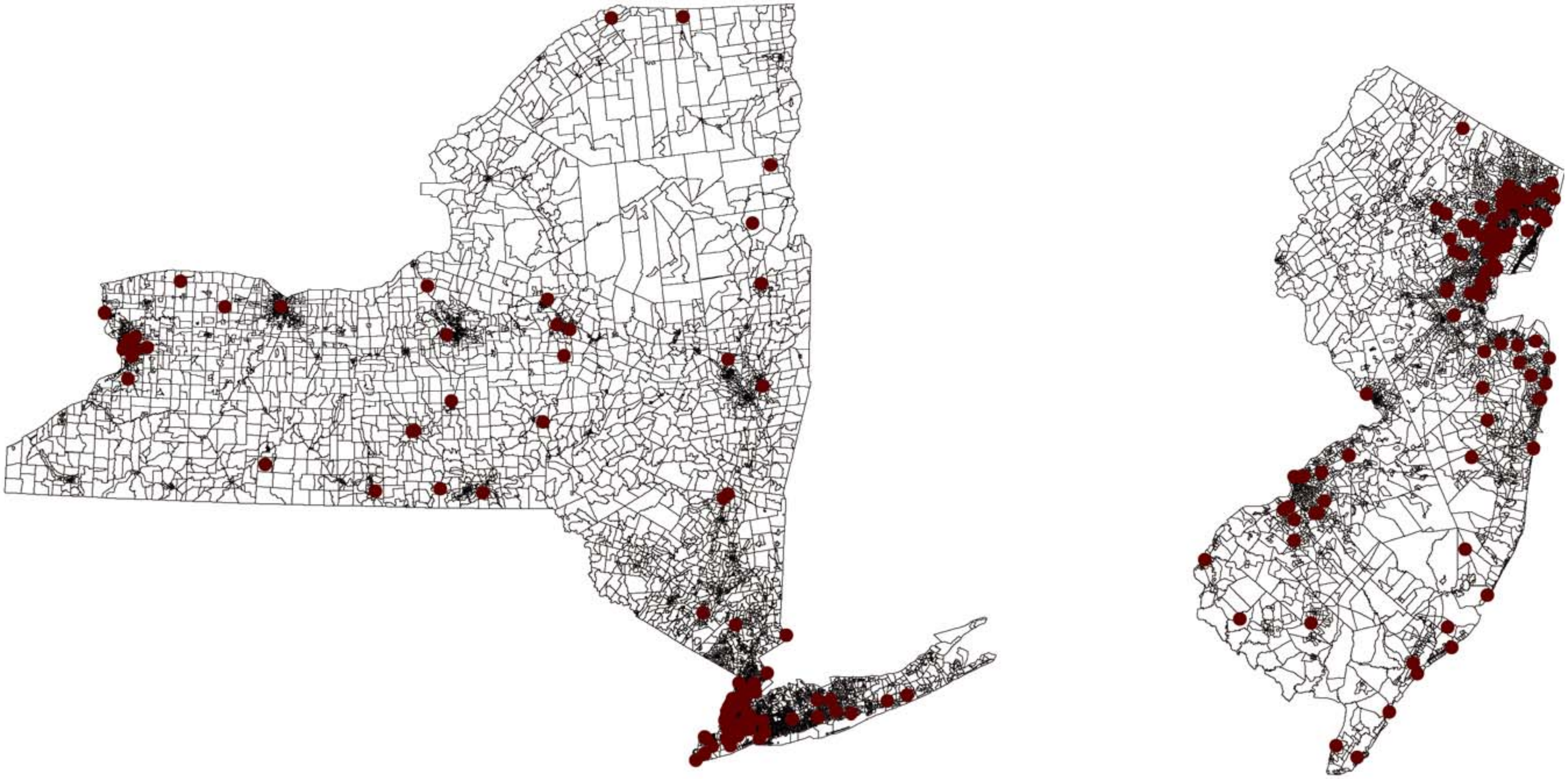


Figure 7. Sample Locations (red circles) of the 100 Contiguous Census Block Groups.

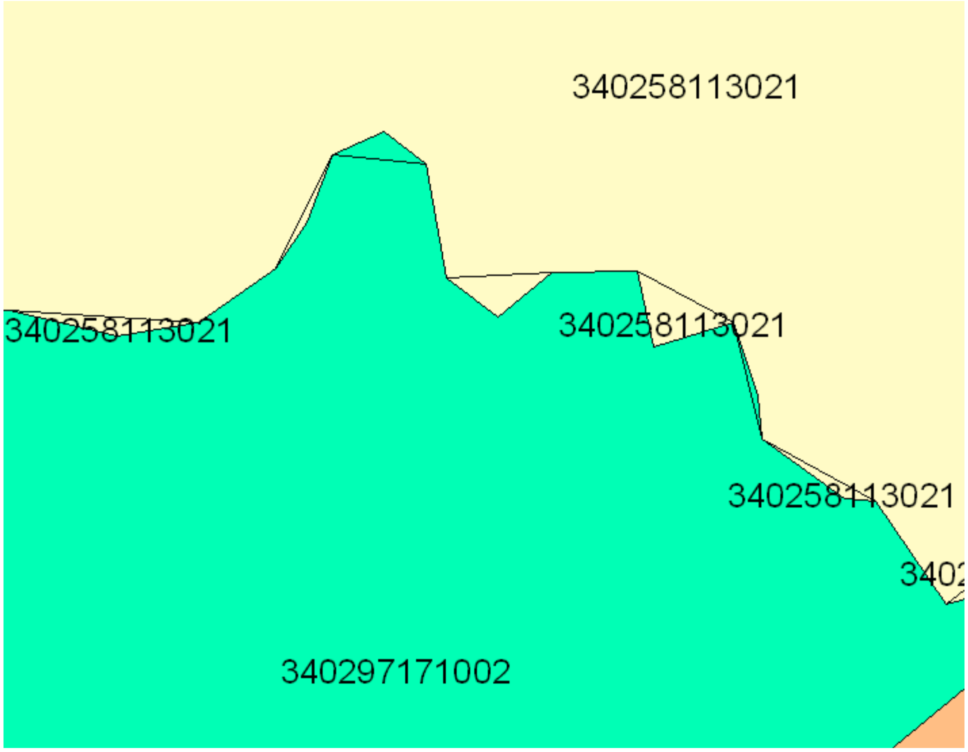
Section 5

GIS Remediation

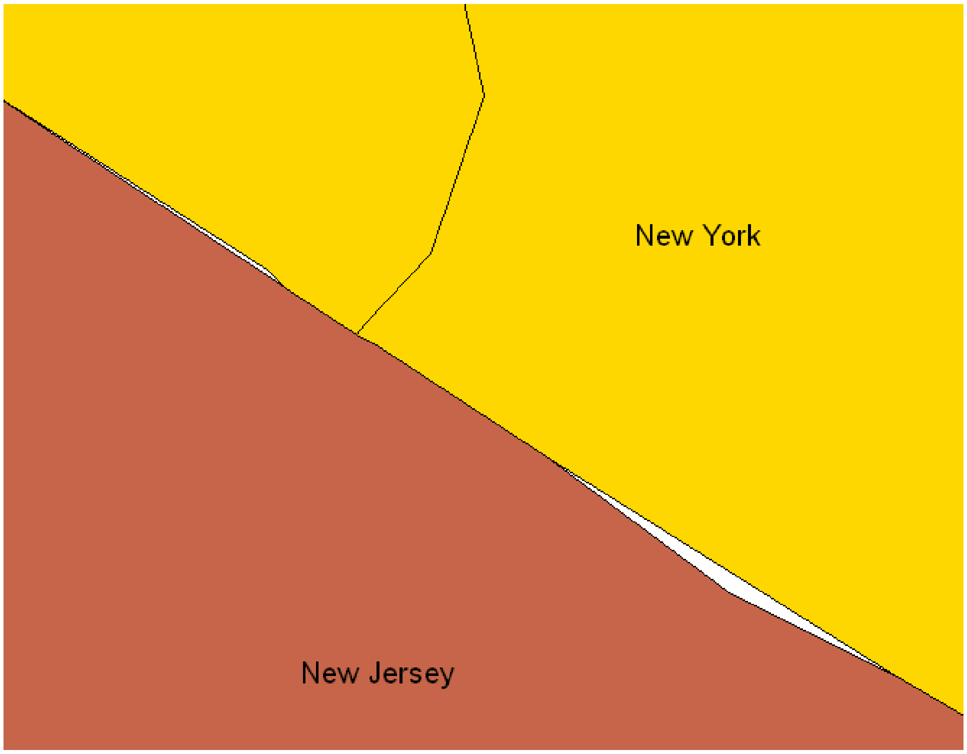
When we began the statistical analyses, we found errors in the data. These errors were:

1) Numerous block groups are comprised of several polygons where only one was necessary (Appendix 2a),
2) Several polygons are missing from the block group coverage obtained from Region II (Appendix 2b), and
3) Several polygons have erroneous id codes (see Appendix 2c).

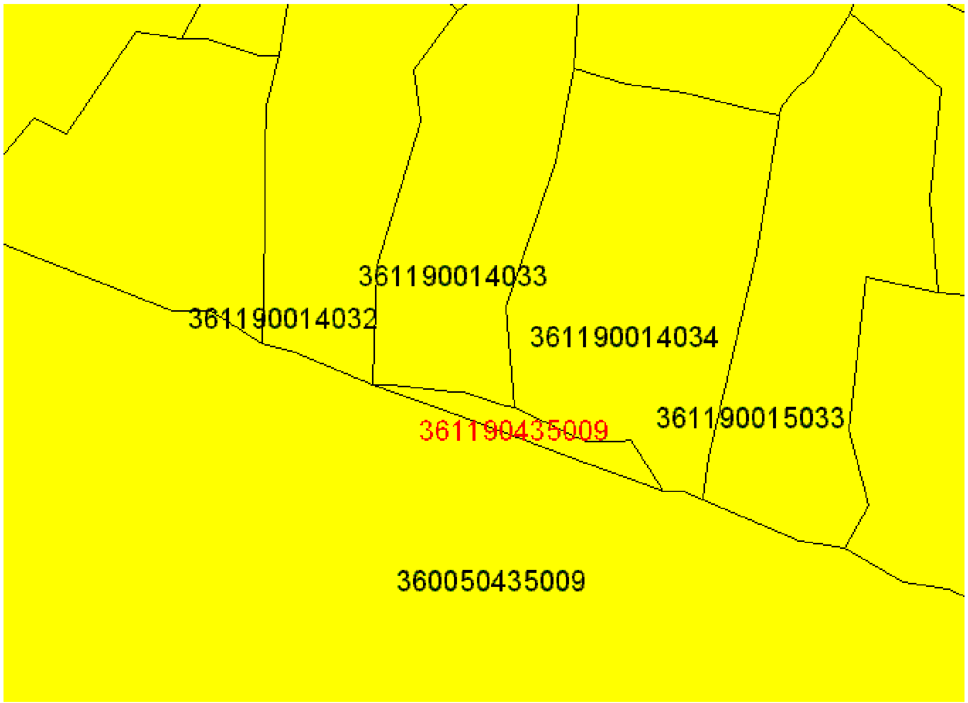
To remediate the errors so that both Region II and Las Vegas scientists could work on the same data set, the polygon data was downloaded from ESRI's ArcData Online site, internal boundaries between like block groups were dissolved, and the tabular demographic data supplied by Region II was joined to the polygons. Results were visually inspected for correctness.



Appendix 2a. Example of error 1.



Appendix 2b. Example of error 2.



Appendix 2c. Example of error 3.

Section 6
Summary and Conclusion

We demonstrated a simple descriptive method using the probability distribution of census and random sampling data sets that used to identify a COC based on a cutoff value. The cutoff value associated with cutoff probability at the 80th percentile in the population in the decision unit for the characteristic of "minority" and the characteristic of "below poverty" was used. For this analysis, it is important to define the sampling and decision units. The basic sampling unit was the census "block group." The decision unit may be equal to or larger than that of the sampling unit (e.g. county). If the decision unit is larger than that of the sampling unit, then all of the characteristics of the spatially included sampling units in the new decision unit must be recomputed. The above analysis, therefore, offers an easy method to evaluate a cutoff value based on the spatial proximity (scale) of the decision unit in order to determine if that is a COC. The choice of the scale is dependent on the degree of details that is required in answering a question and/or to make a managerial decision. In summary, this is one method that could be used to estimate distribution across the regional scale using census data.

References

SAS/STAT User's Guide (Version 6, 4th Ed.), Vol. 2. 1990. SAS Institute Inc., Cary, North Carolina, USA.