Unlocking The Census Storehouse For Beginning Undergraduates

by William Bosworth [1]
Political Science Department,
Lehman College - City University of New York

As the industrial revolution gained momentum, the individual artisan gave way to organized, hierarchized factories. Now, academic computer users seem to be evolving backwards: data processing is swiftly moving from an organized mainframe environment to increasingly powerful PC's in the hands of independent computer users, just like individual artisans. But as we become able to store, manipulate, and display even the most complex forms of data, we can easily become isolated from fellow researchers. Manipulating data on one's own machine may have its advantages: for example, we can tailor data sets and programs to the specific needs of our students. However, when we deal with data sets as universally necessary as the US Census, there is a danger of "reinventing the wheel."

This paper discusses specific projects developed for beginning students in an urban setting, in the hope that others can find the work useful for their own academic projects. It is always desirable for the researchers in this field to suggest changes and improvements in their various projects. It really makes no sense for each of us to reinvent the same wheel; we may be independent artisans in our census-oriented research, but with a little cooperation we can perfect different approaches and, by sharing them, change that tiresome wheel into a complete, road-worthy vehicle.

This paper concentrates on data derived from the US decennial censuses, since such data has a number of special advantages for all researchers dealing with social questions in the US. Census items are uniformly labelled, so a program to access and transform them will work everywhere in the US. No data is more reliable. Census data reaches down to the level of city blocks, and, transformed into percentages, can enable us to compare almost any government units with any others, large or small. The Census Bureau itself provides us with hundreds of cross-tabulations so we can characterize in extraordinary detail the qualities of various age, racial, and economic categories of the population. And though most census data is based on geographic units, the PUMS file is based on a representative sample of people for each major county. Finally, census material is available on tape at least from 1960 onwards, so comparisons through time are facilitated. Where tracts and blocks have remained basically the same, such comparisons can be done for very small geographic units. Thus, those of us interested in getting undergraduates started in data analysis have in the census data our richest storehouse.

But a storehouse with a locked door is of no use. Confronted by mountains of census information coming raw from the government, each researcher is tempted to develop his own program to convert the data into usable form. Here we see the sinister danger of simultaneously re-inventing the wheel. At this point, we should inform one another what works in our experience so that others can profit from it.

First, a note about the specific needs and resources that influence our activities here. Lehman College is a public commuter college, 80% of whose students come from one county (The Bronx, a borough of New York City). Thus there is a built-in student interest in studying this area. The Bronx is separated by water and a greenbelt from surrounding counties, so it is easy to identify and analyze through time. It is one of those Northeast urban areas that have changed dramatically over the past thirty years, so demographic analysis through time is particularly rewarding. And when the Bronx is seen in detail (particularly when we examine each of its 4,132 blocks) we see economic and ethnic differences that allow for many other dimensions of analysis.

Resources for data analysis are available from our college and from the City University of New York as a whole (the artisan has his own workshop, but he can also whittle away in the modern factory). We have a powerful university mainframe computer, and individual faculty members can have terminals in their offices. The mainframe provides powerful statistical languages such as SPSS-X and SAS, as well as tape storage and disk space for real-time work. University membership in the Inter-university Consortium for Political and Social Research (ICPSR) enables us to get most census tapes without charge. As a US government depository, our library has most of the technical documentation needed to identify census items on the tapes. New York City's City Planning Commission has developed a mapping system for computer representation of City features down to the individual block. At the college we have a number of classrooms with networked PC's and a common elementary statistical language (ABC). There is also a

classroom with Unix-based PC's connected to an RT file server. Here students can display and manipulate census data directly on maps of the Bronx, down to the level of city blocks. For this work we use the History Machine program developed by Prof. David Miller at Carnegie-Mellon University.

The foregoing inventory of resources shows what we start with. Most colleges probably have most if not all of them. Many undoubtedly have other resources that facilitate the work we shall describe. If other things work better, we would like to hear about it. At this point we present to you, in detail, the story of how we at Lehman College unlocked the Census storehouse for our beginning undergraduates.

I. Development of Data Files

For 1980 census data going down to the tract and block-group levels there are two files: the complete count STF1 and the sample STF3. Only the latter has information on income and poverty, education, and occupation. Using SPSS-X on the CUNY mainframe computer, we selected the STF1 and STF3 files for New York City and suburban counties, identified and labelled each of the original census variables, created new variables for "non-Hispanic White" (an ethnic category which the Census Bureau should have developed itself), created percentages from most of the census variables, changed the order of variables to make the file look more logical (in our judgment), and finally "Matched" the STF1 and STF3 files to create a single master census file for the Bronx and other metropolitan areas that could be analyzed in SPSS-X. The process involved six successive transformations of data. Each of the six programs can be made available to interested researchers.

For the 1980 PUMS file we first rectangularized the file by "nesting," changed the order of variables to make the file more logical, and recoded ages and ethnic background to simplify for easier analysis. This process involved two successive data transformations, which interested researchers may obtain.

Using the mainframe disks we can quickly generate crosstabulations from the Bronx PUMS files. We have found no PC program that will do so for the largest PUMS, the sample based on 5% of the population (for the Bronx, this sample includes over 58,000 respondents and the datafile on PC would be larger than 5 megs). We have created a PC-usable PUMS subset by randomly selecting one- quarter of the PUMS cases. The file is still over one meg in size, and works best from the v disk of one of the more powerful PC's. For the tract and block group material we have created ABC datafiles on our PC network, so students on their own can do the analyses we shall describe below. This information is also a base for student projects involving maps of The Bronx on our Unix-based PC network. Students can study census items for the 64 Bronx "health areas," the 356 Bronx census tracts, or the 4,132 Bronx city blocks. We must reiterate that all the data displays are based on the transformations we did for STF1 plus STF3, and for PUMS, and the programs for these transformations can be made available to colleagues.

II. Student Projects

In introductory classes, a neighborhood study is the first project that introduces students to our computerized programs. Students describe their Bronx neighborhoods (those who are not Bronx residents "adopt" a local neighborhood). They draw a map of the neighborhood, indicating its boundaries, and describing <u>why</u> they chose the boundaries they did. The reason for their choice is generally socio-economic rather than spatial, so the students already have generated assumptions about their neighborhood. Next, students are presented with tract maps that approximate as closely as possible the neighborhoods they have described (we must be honest: in this project we try to convince students to tailor their neighborhoods to the boundaries of one or more census tracts). Then each student is given a printout of 63 variables, with figures from New York State as a whole and from the Bronx as a whole. These are selected from items in the computer program for each of 338 Bronx census tracts. They include general population figures as well as items on employment and occupation, education, family structure, income and poverty, and housing. Items for 1970 are included as well as 1980 items. Students thus see on the printout certain "norms." They then predict what they will find for the census tract or tracts constituting their neighborhood. Then they are shown how to use the simple "list" command in ABC to retrieve the figures for their local tract, so they see how accurate their guesses were. Great differences will stimulate students to make hypotheses about demographic factors they did not consider - or perhaps about change in their neighborhood since 1980. Throughout this first project, students are encouraged to use their own personal experience in the neighborhood to supplement the statistics they find.

The following items are examples of what the students work with (figures for 1980 unless otherwise specified):

Once they have done the first project, students will have mastered the ABC software package and (we hope) will be interested in exploring their local area in other ways. Using the same dataset just described, we next show students how to aggregate items among all Bronx tracts using weighted means. We soon develop rather complex questions. For example, we can consider all the tracts (there are 55 of them) where the 1980 population was less than half the 1970 population. We can then get weighted means for these tracts to see any peculiarities. We find, for example, that in these 55 tracts that lost so

TABLE 1

NAME	N.Y.STATE	BRONX	LABEL	GUESS FOR YOUR NBRHD
BLPCT	13.68	31.82	% Blacks in Pop, 1980	
BLPCT70		24.30	% Blacks in Pop, 1970	
FEMPLOYD	44.77	37.58	% Females, 16+, who are employed	
WHCOLL	20.40	13.89	N.Hsp.White 25+: % 4 + Yrs College	
KD1PAR	21.68	44.41	Kids -18: % in 1-Parent Homes	
BELOWPOV	13.09	26.98	% Pop, Income Below Poverty Level	
VACANT	5.35	4.80	% Units that are Vacant	

much population, the number over age 65 actually increased by more than a third.

Or we can select areas where few Blacks are below poverty and compare them to areas where many are below poverty, and see the differences in family structure. We can do the same for Hispanics and have a couple of dimensions for interesting speculation (at this point we must remind students that they are working with areas, not with individuals). Illustrative tables are given below.

Those students who are particularly interested in the preceding studies are introduced to a second dataset, based on the Public Use Microdata Sample (PUMS file) from the 1980 census. We use the PUMS file for the Bronx, which, when tailored for the PC, includes over 14,000 individuals, around 1.3% of the Bronx population. Though we cannot look at areas within a county, the PUMS file uses individuals as its cases, so we can define specific characteristics of a population and make comparisons through crosstabulations without fear of confusing people with census tracts.

With PUMS, we can find unexpected differences among populations, which may well call for reconsideration of

TABLE 2

Procedure:	Univariate
Datafile:	BXCOR78
Partition:	popratio lt 50
Number of cases passing partition:	55
Number of cases not passing partition:	283
Variable:	OLDPCRAT (TotalPop, % 65+: 1980 Compared to 1970)
Weight:	TOTALPOP (Total Population)
N total:	55
N included:	55
N weighted:	110,580
Minimum code:	0.00
Maximum code:	274.75
Num. unique codes:	44
Mean:	138.902
Mode:	183
Median:	140.
Sum:	15,359,836.00
Standard deviation:	61.981
Variance:	3,841.698

TABLE 3

Procedure:	Univariate
Datafile:	BXCOR78
Partition:	blblopov lt 10
Number of cases passing partition:	120
Number of cases not passing partition:	218
Variable:	BLMAKDS (Black: % Fams, No Hsbnd, Own Kids)
Weight:	BLACKPOP (Black Population)
N total:	111
N included:	38
N weighted:	60,138
Minimum code:	4
Maximum code:	28
Num. unique codes:	15
Mean:	12.6
Mode:	15
Median:	13.2
Sum:	756,937
Standard deviation:	3.8
Variance:	14.4

TABLE 4

Procedure: Univariate
Datafile: BXCOR78
Partition: blblopov ge 40

Number of cases passing partition: 98
Number of cases not passing partition: 240

Variable:	BLMAKDS (Black: % Fams, No Hsbnd, Own Kids)
Weight:	BLACKPOP (Black Population)
N total:	98
N included:	94
N weighted:	147,976
Minimum code:	9
Maximum code:	68
Num. unique codes:	34
Mean:	32.2
Mode:	40
Median:	32.0
Sum:	4,758,267
Standard deviation:	7.6
Variance:	57.5

TABLE 5

Procedure: Univariate
Datafile: BXCOR78
Partition: hsblopov lt 10

Number of cases passing partition: 82
Number of cases not passing partition: 256

Variable:	HSMAKDS (Hisp: % Fams, No Hsbnd, Own Kids)
Weight:	HISPPOP (Hispanic Population)
N total:	82
N included:	34
N weighted:	16,143
Minimum code:	4
Maximum code:	46
Num. unique codes:	17
Mean:	10.5
Mode:	6
Median:	8.0
Sum:	169,706
Standard deviation:	5.9
Variance:	34.9

TABLE 6

Procedure: Univariate
Datafile: BXCOR78
Partition: hsblopov ge 40

Number of cases passing partition: 130
Number of cases not passing partition: 208

Variable:	HSMAKDS (Hisp: % Fams, No Hsbnd, Own Kids)
Weight:	HISPPOP (Hispanic Population)
N total:	130
N included:	130
N weighted:	242,455
Minimum code:	5
Maximum code:	51
Num. unique codes:	34
Mean:	33.2
Mode:	31
Median:	32.4
Sum:	8,058,798
Standard deviation:	6.9
Variance:	47.3

certain social policies. For example, if we concentrate on the three major ethnic groups in the Bronx (non-Hispanic Whites, Blacks, and Hispanics), we note very significant age differences. We can further divided these groups into those who are native born and those who are not (in the Bronx, the Blacks who are not native born are in their majority Jamaicans. Dominicans are the largest non-native Hispanic group, while almost all native born Hispanics are Puerto Ricans). If we do this, we find that age differences are even more magnified: Kids under 17 are twice as large a constituent of the native born Black and Hispanic groups than of the non-native groups, while those over age 65 actually constitute a majority of the non native White group!

The PUMS tables presented below show only one striking aspect of Bronx population groups. With the PUMS file we can spend hours examining other characteristics and relationships. Does income increase with education in the same way for male Black heads of household as for male White heads of household? How does the specific ancestry of non-native born Whites differ from the ancestry of native born Whites? Does a larger percentage of Bronx residents of Albanian origin have air conditioners in their apartments than Bronx residents of Irish background? If age and marital status are held constant, do Hispanics of Dominican origin still have higher household incomes than Hispanics of Puerto Rican origin? Do Bronx residents who spend over an

TABLE 6

Procedure: Xtables
Datafile: BXPUMS
Partition: citizen eq 0
Number of cases passing partition: 11820
Number of cases not passing partition: 2736

Row: AGEGROUP (Age in 1980, Categorized)
Column: SIMPLRAC (Racial/Ethnic Categories)
N total: 11820 N included: 11696

Col %	NH White	NH Black	His panic	Total	N's
4 AndUnder:	4.5	10.5	12.4:	9.3	1085
5-12 :	8.0	16.1	17.0:	13.9	1621
13-17 :	6.9	11.8	13.0:	10.7	1251
18-24 :	12.8	13.0	13.1:	13.0	1518
25-29 :	7.3	8.3	8.7:	8.1	950
30-39 :	10.4	14.7	13.7:	13.0	1519
40-49 :	8.4	9.0	10.3:	9.3	1086
50-64 :	22.5	11.3	8.5:	13.8	1619
65 And Up :	19.3	5.3	3.2:	9.0	1047
Total	100.0	100.0	100.0	100.0	
N's	3682	3914	4100		11696

TABLE 7

Procedure: Xtables
Datafile: BXPUMS
Partition: citizen ne 0
Number of cases passing partition: 2736
Number of cases not passing partition: 11280

Row: AGEGROUP (Age in 1980, Categorized)
Column: SIMPLRAC (Racial/Ethnic Categories)
N total: 2736 N included: 2522

Col %	NH White	NH Black	His panic	Total	N's
4 AndUnder:	0.3	1.3	1.6:	.9	22
5-12 :	2.1	5.3	5.8:	3.9	98
13-17 :	1.8	9.0	8.3:	5.4	137
18-24 :	3.4	14.1	13.0:	8.8	221
25-29 :	2.5	11.0	15.5:	8.1	204
30-39 :	8.3	19.1	19.2:	14.0	353
40-49 :	8.4	14.7	15.4:	11.9	300
50-64 :	19.9	16.2	14.2:	17.5	441
65 And Up :	19.3	9.4	7.0:	29.6	746
Total	100.0	100.0	100.0	100.0	
N's	1195	702	625		2522

hour commuting to work have fewer bedrooms than Bronx residents who walk to work? From the ridiculous to the sublime, one cannot predict which of these questions will stimulate an undergraduate.

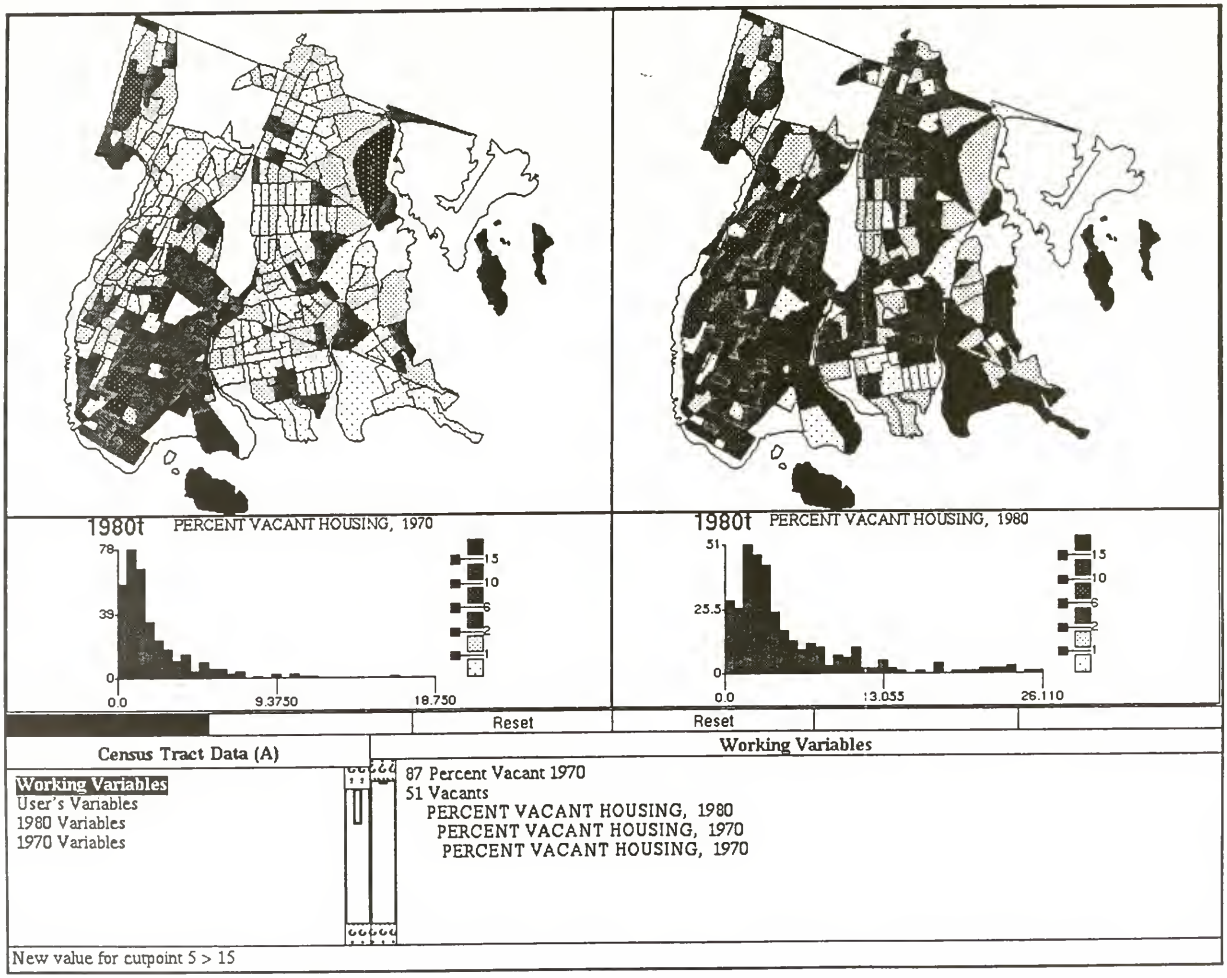
The accompanying maps indicate the final stage in our process of introducing undergraduates to census information. Our Unix-based PC network includes Bronx maps showing three sorts of geographic units: the 64 Bronx health areas, the 356 Bronx census tracts, and the 4,132 city blocks into which the Bronx is divided. Available data is displayed for each unit (note that income, education, and occupation figures are available only down to the census tract level; for city blocks our data shows age and ethnic divisions, family structure, and housing). The mapping system is extremely flexible. Students can change the cutpoint values of an item and see the changes instantly on a new map. New variables can be created from two or more existing ones. The screen can be split so that two maps (showing the same or different geographic units) can be displayed simultaneously. For greater detail there is a powerful zoom feature. Most important, all the manipulations are shown instantly and can be printed. The History Machine program, mouse-based and insulating students from the horrors of Unix, is also very easy to learn. We include here three maps to illustrate each geographic unit we are able to present: the health area and census tract maps have a split screen to show changes in variables through time. The block map is just too detailed to reproduce perfectly without zooming in on one part of The Bronx; nonetheless, the complete map reproduced here will give an idea of what we can do with our map program.

- - - - - - - - -

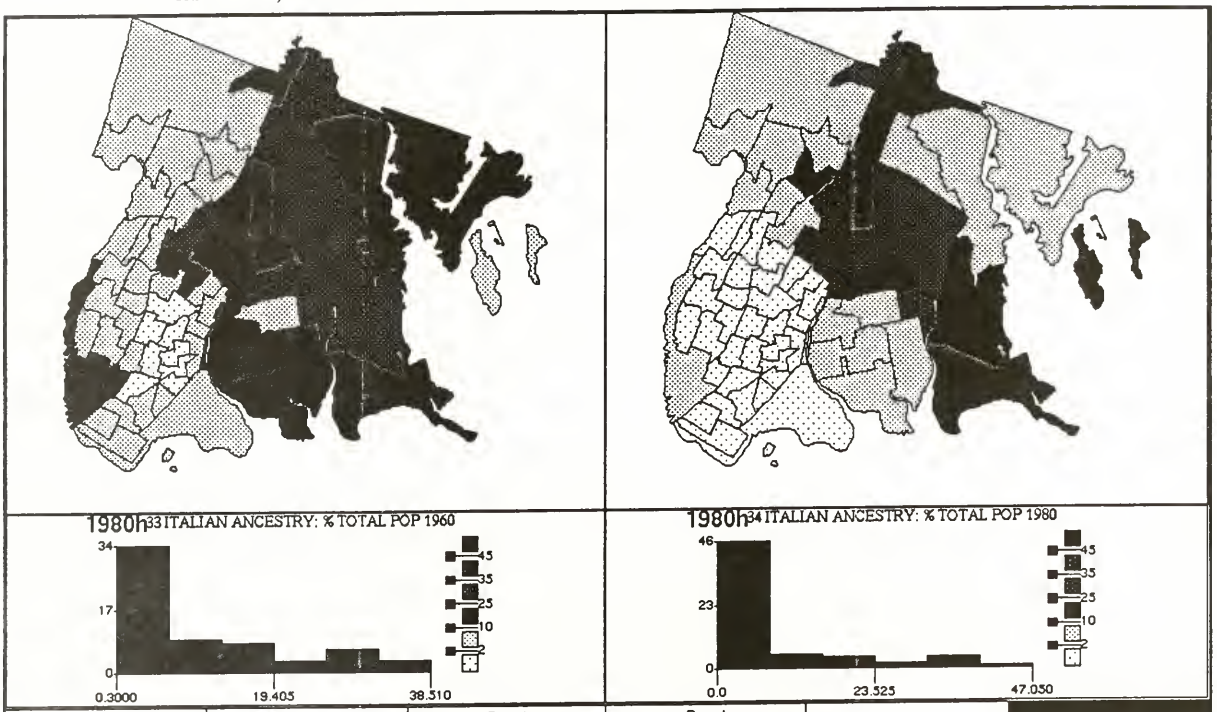
We continue to enlarge the kinds of data manipulation as well as the census information available to students. We are just beginning to incorporate the items of the 1980 STF4A file into our systems. We shall soon create new units from the existing city block map (police precincts; community school districts, for example) so that the data associated with these units can be compared visually with our census data. And, of course, we are preparing to plug in the 1990 census data as soon as it becomes available. Whatever the research potential for all this material, we shall not forget that in the first instance it was designed for use in undergraduate teaching. We stand ready to share the materials we have developed with others who think like us

[1] Presented at the IASSIST 90 Conference held in Poughkeepsie, N.Y. May 30 - June 2, 1990.

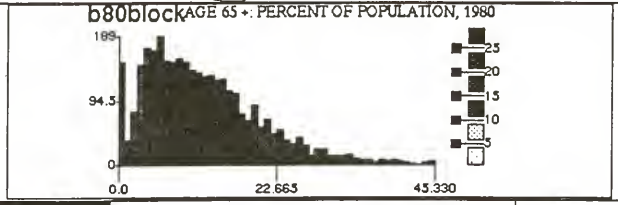
THE BRONX, NEW YORK. THE 356 CENSUS TRACTS



THE BRONX, NEW YORK. THE 64 HEALTH AREAS



THE BRONX, NEW YORK. THE 4,132 BLOCKS



Reset

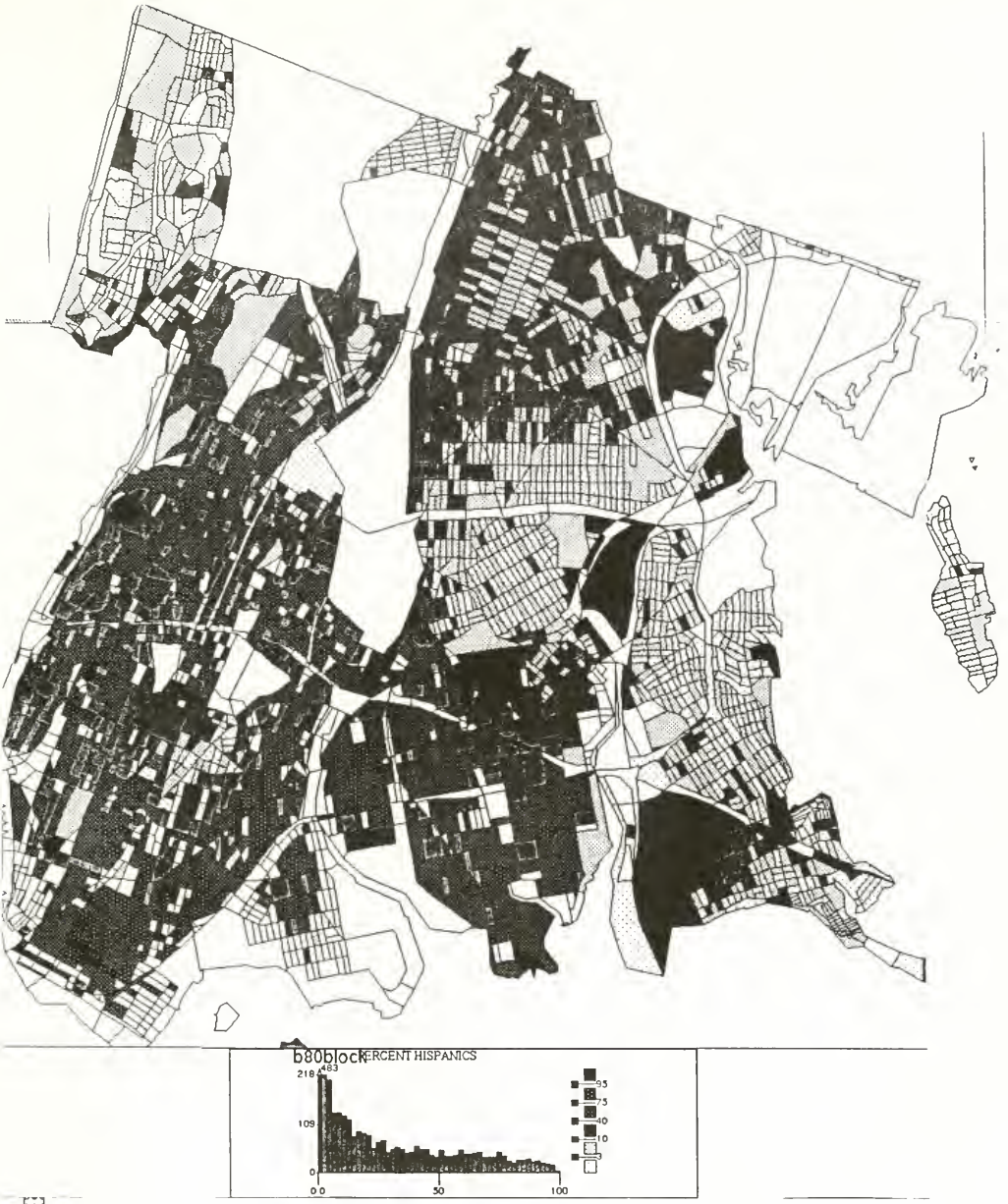
test

Working Variables
User's Variables
Bronx Blocks: 1980 Variables
Bronx Blocks: 1970 Variables

Bronx Blocks: 1980 Variables

12 VACANT UNITS: % TOTAL HOUSING,1980
13 NUMBER OF VACANT UNITS, 1980
14 10+ UNITS AT ADDRESS: % HOUSING, 1980
15 NUMBER OF HOUSING UNITS, 1980
16 POPULATION AGE 65 AND OVER, 1980
17 AGE 65 AND UP: % OF POP., 1980
18 NUMBER OF KIDS UNDER AGE 18, 1980
19 NUMBER OF HOUSEHOLDS, 1980

New value for cutpoint 5 > 25

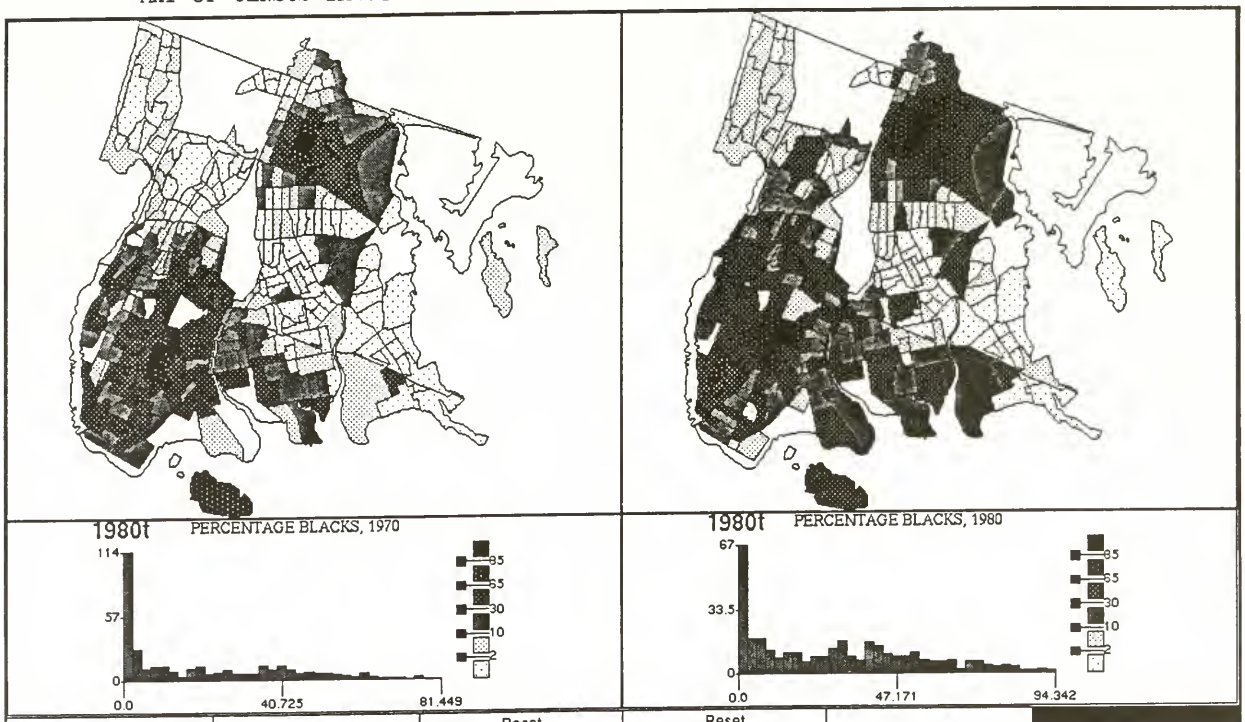


MAP OF THE 4,132 BLOCKS IN THE BRONX, SHOWING PERCENT HISPANICS: lightest to darkest cross-hatches:

Population 0 - 2% 3 - 9% 10 - 39% 40-74% 75 to 94% 95% and above

(Blank areas contain no population)

MAP OF CENSUS TRACTS IN THE BRONX



MAP OF THE 64 BRONX HEALTH AREAS