Documentation for the 1970, 1980, and 1990 Census PSID Geocode Match Files

Sixth Edition
March 15, 2005

http://www.isr.umich.edu/src/psid/
Psidhelp@isr.umich.edu
734-763-5166

Panel Study of Income Dynamics
Economic Behavior Program
Survey Research Center
Institute for Social Research
University of Michigan
PO Box 1248
Ann Arbor, MI 48106-1248

Table of Contents

PSID Address Data

The Geocoding process was done in two phases. The first geocoded 1968-1985 addresses to 1970 and 1980 census identifiers; the second geocoded 1968 and 1970-1999 addresses to the 1990 census identifiers.

Address Data Used for 1970 and 1980 Geocode Match Files

The list of addresses that we geocoded came from the PSID Address files, a confidential set of addresses of PSID respondents kept separately from the main PSID data. The addresses on these files are used to generate labels for mailings to respondents, including annual reports and payments for completed interviews. As a basis for geocoding the residential addresses of respondents, the address data on these files are problematic in several respects:

a) the purpose of collecting the addresses is to get mail to respondents, not to specify where they lived. Significant numbers of respondents use Post Office Boxes, Rural Route numbers, General Delivery, or, in a few cases, a relative's address as their mail address, and we cannot determine their actual street address;

b) for many of the early interview years, we had retained only one Address file; in later years, we had several from which to choose. In both cases, we typically chose the one used for final payments to respondents, dated several months after completion of the interview (which normally is in the spring or summer of the interview year). Respondents could easily have moved during those several months, and the file address would be the new one. To a large extent, analysts can determine the seriousness of this problem through use of variables in each wave of the PSID, including questions noted in the PSID documentation volumes as "if moved", "month moved", "month of interview", "state now" [as of interview], and "county now" [as of interview]. The dates of the Address files are indicated in the table below.

c) prior to the beginning of the geocoding project, the Address files were used only as described above, and there was no particular reason to keep them on hand after the final respondent payments had been made. Among the files for the 1968 through 1985 interviewing years, four had in fact been overwritten or otherwise destroyed: 1969, 1975, 1977, and 1978. We therefore had to impute an address, and associated geocodes, based on prior and subsequent addresses. Our imputation is problematic for families that moved during the period covered by the missing address files, but not for those who did not move.

Table 1 shows the number of addresses available to use for each interviewing year. On the 14 Address files we had available, there were 83,081 non-unique addresses, each associated with a unique set of family and year identification codes. We aggregated these addresses to combine exact duplicates, collapsing them into 37,327 exactly unique addresses (including exact matches on spelling and punctuation on street number, street name, street type, unit number, city, state, and ZIP code). Because the computer consolidation treated only exact matches as identical (e.g., "123 Jones Av" is different from "123 Jones" and "123 Jones Ave"), we believe there were in fact approximately 20,000 genuinely unique addresses.

Table 1: PSID Address Files Used for 1970 and 1980 Geocode Match Files

Interview Year	Date of file	Number of addresses
1968	07-25-69	4,802
1969	Missing	[4,460]
1970	02-04-71	4,645
1971	11-22-71	4,840
1972	01-16-73	5,060
1973	10-31-73	5,285
1974	02-18-75	5,517
1975	Missing	[5,725]
1976	03-18-77	5,862
1977	Missing	[6,007]
1978	Missing	[6,154]
1979	03-03-80	6,373
1980	03-13-81	6,533
1981	02-22-82	6,620
1982	02-14-83	6,742
1983	01-04-84	6,852
1984	01-10-85	6,918
1985	02-12-86	7,032

The problem of the missing years of PSID Address files was addressed by imputing to the missing year the geocodes for the previous year, if there was a previous year, and the geocodes for the following year if there was no valid data for the previous year:

* for 1969 addresses, 1968 codes were used if available, and 1970 if not

* for 1975 addresses, 1974 codes were used if available, and 1976 if not

* for 1977 and 1978 addresses, 1976 codes were used if available, and 1979 if not

Analysts who wish to try a more sophisticated imputation scheme based on PSID interview date and responses to the "whether moved since last spring" and "date moved" questions, and on the dates of the address files, can, of course, do so. We found that the assumptions we needed to make were too complex to be easily justified.

Address Data Used for 1990 Geocode Match File

The addresses used in this project were the mailing addresses to which SRC had mailed nominal payments for participation during the years 1969 through 1999. Whenever possible, we used the addresses that we had for respondents at the end of the interviewing season (late fall to early winter, depending on the year, but immediately prior to creation of the address file for the upcoming year). This may mean a few changes of address subsequent to the interview may have been entered based on information received from the respondents or the US Postal Service, but that all the address changes provided during the interview have been entered. The 1968 addresses are an exception to this general rule, since they appear to be the addresses we knew at the beginning of the 1969 interviewing season, and probably include a significant number of address

changes of which we learned in the course of our pre-1969 mailings to 1968 respondents. Overall, we estimate that over 99% of the addresses at the time of our selection, except for 1968, were the addresses at the time of the interview. Table 2 shows the numbers of addresses that were used for the 1990 Geocode Match file.

Table 2: PSID Address Files Used for 1990 Geocode Match File

Interview year	Number of addresses
1968	4,802
1969	Missing
1970	4,645
1971	4,840
1972	5,060
1973	5,285
1974	5,517
1975	5,724
1976	5,861
1977	6,006
1978	6,153
1979	6,372
1980	6,533
1981	6,619
1982	6,741
1983	6,851
1984	6,917
1985	7,031
1986	7,017
1987	7,060
1988	7,113
1989	7,113
1990	9,371
1991	9,363
1992	9,829
1993	10,195
1994	11,003
1995	10,401
1996	8,511
1997	6,748
1999	6,997

Because the Public Release final versions of the 1994-1999 family files were not available when we processed the addresses, some addresses that we geocoded in those years will not be associated with a family record in the final version of the Public Release data of the 1994-1999 family files. The numbers of records for 1994-1999 in the 1990 Geocode Match file may not

match the number of records in corresponding Public Release[1] family files because of retroactive family composition changes in the first version Public Release data.

When we did the geocoding of 1968-1985 addresses for 1980 and 1970 Census codes, we had been unable to locate address files for 1969, 1975, 1977, and 1978; this time, we found the 1975, 1977, and 1978 address files in archival tapes in a bank safe deposit box, and thus were able to code a more complete set of addresses. Unfortunately, we were not able to locate any dataset with final 1969 addresses. In addition, it should be noted that the 4,802 "1968" addresses were the addresses of the 1968 respondents as we knew them at the beginning of the 1969 interviewing, not at the end of 1968 interviewing, and reflected some changes of address of which we learned between the 1968 and 1969 interview seasons.

For the years 1975-1989 our address data contained one fewer case in each year than the PSID final-release family files. At some point in the address standardization process, a case must have been deleted in each of these years, but we did not discover this until it was too late to look up and restore it. In the 1990 Geocode Match File, described below, records for these missing families have been included with missing data for the identifiers. Table 4 shows the year and the PSID interview number of these records. Note that other records as well may also have missing data for the identifiers.

Table 4: Families with Missing Data for All Census Identifiers in the 1990 Geocode Match File

V1 Year	V2 Interview Number
1975	5173
1976	5842
1977	1331
1978	6090
1979	2945
1980	3370
1981	1540
1982	6647
1983	5297
1984	3730
1985	6484
1986	6446
1987	4081
1988	7072
1989	5214

[1] See http://psidonline.isr.umich.edu/Guide/FAQ.aspx#2 for more information about Public Release data.

The Geocoding Process

As noted earlier, the Geocoding process was done in two phases: the first geocoding 1968-1985 addresses to 1970 and 1980 census identifiers, and the second geocoding 1968 and 1970-1999 addresses to the 1990 census identifiers. These two geocoding efforts are described separately below.

1970 and 1980 Geocoding Process

Our primary objective was to characterize the "neighborhoods" in which respondents lived. We had decided in advance that we would use census tract as the approximation of neighborhood in tracted areas; Block Numbering Areas (BNAs) in blocked but untracted areas; and Enumeration Districts (EDs) in areas with neither tracts nor blocks. We also coded Minor Civil Division/Census County Division (MCD/ CCD) and census Place, and retained the ZIP code from the address, so that these could be used as substitutes in instances where we could not find tract, BNA, or ED codes. In addition, we used county information to assign codes for broader areas that might be used to represent "labor markets" – PMSA/SMSAs, CMSA/SCSAs, SEAs, ESRs, and our own specially-created Labor Market Areas (LMAs). See the Glossary at the end of this section for a brief overview of the various geographic levels coded.

1980 Geocodes

We began the geocoding by searching for the codes associated with the 1980 Census because we had far more 1980 than 1970 Census information on hand. Our first step was to use the Census Bureau's data tapes of the Geographic Base File - Dual Independent Map Encoding (GBF-DIME) dataset. GBF-DIME includes a file for each metropolitan area (as defined in 1978), and consists of listings of street names and number ranges in tracted areas; each record includes, for the left and right side of the street (as viewed looking from lower to higher numbers) of each census block segment, codes for the state, county, metropolitan area, Place, ZIP code, tract, and block. We created listings of GBF-DIME for each metro area, sorted by street name and number range. Because census Place names and postal city names rarely have the same boundaries, we did not want to sort by Place – a "Jonesville" postal address may well not be inside the city limits of the census Place named "Jonesville". When there were two or more streets with the same name within the metropolitan area, a common occurrence, we used ZIP codes, street number ranges, and Place codes to find the correct record. We were able to assign 1980 neighborhood geocodes to over half of our addresses by comparing the PSID address to the GBF-DIME address listing for the relevant metropolitan area.

We also had at our disposal a set of Block Statistics Maps from the Census Bureau, also compiled in 1978 for metropolitan areas identified in that year (plus a few urbanized "selected areas" not in metropolitan areas). These maps were difficult to use because of the lack of street name indices, but did allow us in some cases to find the apparent locations of addresses with higher street numbers than those appearing in GBF-DIME. The Block Statistics Maps were also very helpful in locating addresses in the urbanized "selected areas" which, because they were not in metropolitan areas, did not appear in GBF-DIME.

For addresses in areas not covered by GBF-DIME or the Block Statistics Maps, we used Census Bureau Place-and-Vicinity and County Maps. These are usually local planning department

maps adapted by the Census Bureau, are of substantially lower print quality than the Block Statistics Maps, and also lack street name indices.

When we had difficulty locating a street on one of the Census Bureau maps (including a substantial number of cases where the street did not yet exist in 1978), we tried to locate the address on a recent commercial map of the city and then transpose the location to the Census Bureau map and record the corresponding geocodes. This was not possible for many smaller cities, for which no commercial maps appear to exist.

As a final check on the validity of our geocodes, we matched them to the corresponding level in our Census Extracts datasets for 1980. This allowed us to locate, for example, tracts in the geocode file that were not in the Census Extract file and thus presumably not valid. We located several dozen transcription and key-entry errors in this way. The only remaining non-matches in the Geocode file are approximately 50 ZIP codes which are valid according to the US Postal Service's ZIP code Directory (or at least were in 1980), but which do not appear in the Census Extract ZIP code dataset because they had no residential population (e.g., ZIP codes associated exclusively with commercial districts or with city Post Office Box addresses), or were otherwise excluded from the ZIP code version of Summary Tape File 3 prepared for the Census Bureau by National Planning Data Corporation.

1970 Geocodes

Our first step in geocoding the PSID addresses for 1970 was to electronically translate as many as possible of 1980 tract numbers to their 1970 equivalents. For this purpose, we combined two Census Bureau Tract Comparability Files: a pre-1980 Census version with correspondences of all 44,550 tracts that existed in either 1970 or 1980, and a post-1980 Census version that included only the 18,979 1980 tracts that had changed in some way from 1970. Both files indicated the number of 1970 tracts associated with a given 1980 tract, and the number of 1980 tracts associated with a given 1970 tract. The post-Census version also included a characterization of the effects of the boundary changes with an indicator that the change was "minor" (i.e., involved less than 100 persons). In combining the files, we gave precedence to the post-Census version in the case of conflicts or duplication. We then characterized the nature of the change in the resulting file of 49,337 records:

1	No apparent change in boundaries	25,326	(51%)
2	1970 tract split exactly into several 1980 tracts	9,097	(18%)
3	"Minor" changes in tract boundaries	4,402	(9%)
4	Untracted in 1970, tracted in 1980	4,504	(9%)
7	Other changes	6,008	(11%)

Categories 1, 2, and 3 were then used to make the translation of 1980 tract codes to 1970 tract codes. Even when tract boundaries did not change, their numbers could. We were able to directly translate 76% of the 1980 tract codes to their 1970 equivalents using this technique. In addition, a listing of the combined Tract Comparability Files was useful in distinguishing the "untracted in 1970" areas from the "other changes" tracts, and in providing a listing of the possible 1970 tracts for the latter.

We had hoped to use the 1970 version of the GBF-DIME or its predecessor, the Address Coding Guide, but it turned out that the Census Bureau could not locate any copies and had not provided any to the National Archives. Our inquiries with numerous other possible sources proved fruitless.

That left us with the 1970 Block Statistics Maps (for 1968 urbanized portions of metropolitan and selected non-metropolitan areas), and 1970 Census Place and County Maps to attempt geocoding of tracts, BNAs, and EDs for 1970.

We translated Census Place and MCD/CCD codes from 1980 to 1970 electronically as well, using information on changes in codes from the Census Bureau's 1980 *Geographic Identification Code Scheme* (publication PHC80-R5), and found over 250 additional codes in the 1970 version of GICS for Places that had apparently ceased to be considered Places by 1980 (either through annexation or loss of population). It should be noted again here that the boundaries of Places and MCDs change as cities annex surrounding territory, and much more substantially for county legislative districts that are MCDs, even though the code numbers may remain the same. An address that was in a given MCD or Place in 1980 may therefore not have been in it in 1970, although our translation scheme assumes that it was.

Results of 1970 and 1980 Geocoding to Date (August 1991)

Among the 37,324 semi-unique addresses we attempted to geocode, there were 4,136 that clearly could not be geocoded at all – 96 where our only information was that the respondent was deceased, 3,854 where all we knew was that there was a non-response to the questionnaire, 181 with foreign addresses, and 5 with a response but no address of any kind. For the other 33,188 addresses, the geocode results as of August 1991 are summarized in Table 5.

Table 5: Valid Geocodes Located

Matches to Census Extract data, if different, in parentheses

Level	**1970**	**1980**
Tract	22,357	23,605
BNA	190	1,239
ED	657	870
MCD/CCD	33,179	33,179
Place	31,873	31,619
	(27,122)	
County	33,188	33,188
LMA	33,188	33,188
ZIPCode5	33,175	33,175
	(23,515)	(33,045)
ZIPCode3	33,175	N/A
	(33,170)	
PMSA/SMSA	25,145	25,747
CMSA/SCSA	9,248	10,786
SEA	33,188	33,188
ESR	33,188	33,188

Table 5: Valid Geocodes Located

Matches to Census Extract data, if different, in parentheses

Level	1970	1980
State	33,188	33,188

Some of the geocodes we found did not match the geocodes in our Census Extract [2] datasets. In both 1970 and 1980, the Census Bureau roughly approximated Level-level areas by consolidating tracts in tracted areas, and MCD/CCDs in untracted areas, despite the fact that these areas often had crosscutting boundaries. Another problem was that many of our addresses were Post Office Boxes, and their associated ZIP codes did not contain any residential population, so there was no corresponding Level-level Census Extract data. A third problem was that some of our ZIP codes came into existence for the first time after 1968 or 1978, and others ceased to exist before 1978. Finally, in 1970, the Census Bureau created 5-digit Level-level datasets only in metropolitan areas. All these problems resulted in our having valid geocodes for numerous addresses for which we had no matching Census Extract records – 9,660 for 1970 5-digit ZIP codes, 5 for 1970 3-digit ZIP codes, and 130 for 1980 5-digit ZIP codes.

There was a similar problem for 1970 Place-level data – Census Extract records were available only for Places of 2500 or greater population. This meant we had 4,751 addresses with apparently valid Place geocodes but no matching Census Extract records.

Table 6 gives a summary of what we think are our best approximations of neighborhoods and the reasons we were unable to do better than MCD/CCDs in many cases:

Table 6: Best Geocode Located

Level	1970		1980	
Tract	22,357	67%	23,605	71%
BNA	190	1%	1,239	4%
ED	657	2%	870	3%
MCD/CCD	9,975	30%	7,465	22%
Rural Route	3,778	11%	3,778	11%
P O Box	1,925	6%	1,925	6%
General Delivery	254	1%	254	1%
No street address	338	1%	287	1%
Metro, not in GBF-DIME	920	3%	838	3%
Nonmetro, no map available	2,759	8%	383	1%
None	9	0%	9	0%

Our largest problem in locating tract, BNA, and ED was respondents' use of postal rather than street addresses. This was the situation for nearly half our non-metropolitan addresses, and for about 5% of metropolitan addresses as well. Since we had no specific street addresses, we had to

[2] The Census Extract datasets are no longer distributed by the PSID. See the *Introduction* section of this document for more information.

guess at both the MCD/CCD and Place codes of the residential address. For the former, we coded the MCD/CCD with the largest population in the county. For the latter, we coded the Place that had a name corresponding to the postal city name, if there was one. These approximations are obviously rough (as is the ZIP code for Post Office Box addresses), since the respondent may well live outside the MCD/CCD or the Place we chose. However, since the MCD/CCD and Place we chose had the largest populations in the areas involved, they also necessarily had the highest probability of containing the actual residence of our respondent of any MCD/CCD or Place in the area.

A second major problem for our geocoding efforts was the absence of many metropolitan addresses from the GBF-DIME listings and Block Statistics Maps, either because the street name was not present on those sources or the street number was greater than the highest number there. We believe that most of these cases involve housing that was constructed after 1978, when GBF-DIME and the Block Statistics Maps for the 1980 Census were fixed. We were able to reduce the original number of cases in this category by half through the use of newer commercial maps of large cities and transposing to the Block Statistics Maps, but could not locate commercial maps for many smaller cities. In those cases where we could locate the street on a Block Statistics Map, we believe we could make better guesses as to MCD/CCD and Place than was possible for the postal address cases.

Our third major geocoding problem was with addresses consisting of the names of apartment buildings or trailer parks. For some, we were able to find street names and numbers in telephone books or the *ZIP code Directory*, but for most we had to guess at MCD/CCD and Place in the same manner as we did for postal address cases.

Finally, we had nine cases where the mailing address we had could not be located on a map, or in any Census Bureau materials, despite having a valid ZIP code, and we couldn't make any intelligent guess as to MCD/CCD or Place. These few cases were left without a neighborhood geocode.

Where these problems prevented assigning a tract, BNA, or ED code, we indicated the nature of the problem in the "neighborhood/problem code" variable. Analysts may use that variable to inform their own decisions about the adequacy of our use of MCD/CCD as the primary substitute for neighborhood information.

Analysts should feel free not to accept our characterization of MCD/CCD as the "second-best" approximation of neighborhood. We suggest that, rather than decide on an abstract basis that one level is always better than alternatives, analysts compare the possibilities empirically. For example, when tract, BNA, and ED are not available, analysts might attach to a record the population totals for the relevant MCD/CCD, Place, and 5-digit ZIP code, and choose as the neighborhood level the one among them that has the smallest population above some minimum number (at least 30, since that is when suppression of substantive data becomes very likely); the designated level can become a new variable that analysts use when selecting cases for attachment of the substantive Census Extract data.

1990 Geocoding Process for 1968-1995 Addresses

The process of geocoding addresses to 1990 census identifiers was somewhat different than the process, described above, of geocoding addresses to the 1970 and 1980 census identifiers. This was largely due to differences in the data available from the US Census and other

sources. For the 1980 decennial Census, the Census Bureau made available an electronic file of street segments in tracted portions of metropolitan areas (the GBF-DIME files) and high quality paper maps for streets and Census geographic boundaries in all parts of the country. There was also an electronic translation file for 1970 and 1980 tracts. For the 1990 Census, however, the Census Bureau had decided to rely heavily on the availability to users of proprietary Geographic Information System (GIS) software, and designed the Census electronic files of street segments (TIGER/Line files) with GIS users in mind. Paper maps were available only on a customized basis, and did not include the layouts of streets. The translation file for 1980 and 1990 tract boundaries was never completed or released for budgetary reasons.

All these changes meant we had to rely on GIS systems for geocoding addresses to the 1990 Census boundaries. This meant we had to be certain that addresses were spelled correctly, and had the correct ZIP codes (as of the time of our geocoding). Primitive GIS software available in 1995 lacked the ability to make an accurate decision about misspelled addresses or incorrect ZIP codes, however, the current (as of 1999) GIS software does have this ability. Our process was therefore divided into two stages, first, cleaning and standardizing the addresses and second, geocoding the cleaned and standardized addresses. These efforts are described below.

Postal Standardization of Addresses

As noted previously, the PSID address files contain mailing addresses and therefore did not always reflect the residential address of the respondents. In particular, there were a significant number of Rural Route, General Delivery, and Post Office Box addresses that required special treatment, as well as some other addresses that included no street number. All PSID addresses were categorized to make address standardization easier and to allow the exclusion of certain types of addresses. The address typology for 1968-1995 was as follows:

Street name and number	154,403
Street name only	1,486
Rural route	18,513
General delivery	1379
PO Box	11,792
Building/complex name only	713
Foreign	661
No address	39

The foreign and no address cases were not processed further. "Foreign" does not include areas outside the US with US ZIP codes such as Puerto Rico, US Virgin Islands, Guam, etc.

Standardization of addresses according to US Postal Service conventions served a number of important purposes. First, it ensures that addresses are recognizable to the US Postal Service; addresses that are not recognizable in the USPS database may well be misspelled or otherwise defective (e.g., missing street type or direction or have an incorrect ZIP code), and need hand lookup and correction. Second, it results in standard alphabetic and numeric versions of the address that can be used for visual and computerized matching. A USPS-standardized address not only includes uniform spellings of street names and types (converting "Street" and

"St" to "ST", for example), but also results in uniquely identifying code numbers (5-digit ZIP code, ZIP+4, delivery point, and carrier route numbers) which are much easier to use in computerized matching that alphabetic strings.

Delivery points are two-digit numbers basically representing mailboxes, an individual mailbox on single-unit dwellings or a group mailbox in multi-unit dwellings. In the overwhelming majority of cases, there should be only one delivery point with a given number in a given ZIP+4 area; the only exceptions appear to be group mailboxes in common areas of newer apartment complexes with more than 100 units in a single division of the complex. Delivery points should have substantial historical continuity, with new ones added as new units are constructed and old ones disappearing when units are demolished, but no change in the numbers of existing units.

ZIP+4 areas are basically a side of city block, such as the even numbers on the 100 block of Elm Street; the USPS has created analogous areas for cul-de-sacs and other non-rectangular "blocks". It is also the case that USPS uses ZIP+4s to represent rural carrier routes in ways that don't necessarily have a nice geographic coherence like a side of a city block. As the name implies, ZIP+4 codes have four digits. ZIP+4s should have substantial historical continuity, with new ones added as new streets are constructed but no changes in numbers on existing blocks.

Carrier routes include groups of ZIP+4 areas assigned to a single carrier. Historically, since carrier routes came first, it would be more accurate to say that ZIP+4s are the segments within a carrier route. Carrier route identifiers consist of four digits, one letter and three numbers, with the letters representing ordinary routes (C), rural routes (R), private contractor routes (H), and Post Office Boxes (B). The shape and content of carrier routes change fairly frequently, as the local Post Office attempts to balance the workload of individual carriers in the face of new construction in the area.

Five-digit ZIP code areas include an average of about 8500 persons, but the population range is quite wide – from 1 person to 112,000 persons in 1990. ZIP code areas change with some frequency, mainly when old areas subdivide (within exactly the same outer borders) to reflect growth within the area, or merge to reflect depopulation and the closing of a small post office. Five-digit ZIP code numbers sometimes change even when the area involved has not, in anticipation of future subdivision, as occurred in central Florida in the 1980s.

For this project, we standardized PSID addresses using the commercial software package AccuMail from Group One Software. This matches addresses in a database to the USPS database of valid addresses, makes spelling corrections, and assigns corrected 5-digit ZIP codes, and added ZIP+4, delivery point, carrier route, and county codes, as well as an error code for those addresses not recognized. The error codes might indicate, for example, that the 5-digit ZIP code was not valid, or there was no such street name in the ZIP code, or the street number was not in a valid range, or the street direction or street type is missing but required to distinguish addresses with the same street name.

Addresses not recognized and standardized using AccuMail were submitted to a commercial firm certified by the USPS, Lorton Data of Minneapolis, for additional address correction and standardization. Lorton's software had somewhat more flexible parameters than AccuMail. It could, for example, estimate the carrier route for street numbers not in the USPS database but just outside the USPS address range, including addresses that were no longer deliverable because they had been torn down. Lorton had the disadvantage of always presuming that the 5-digit ZIP code was correct when there was a conflict between the ZIP code and the city name, and altered

the city name to fit the ZIP code; in our experience, it was more likely that there was a transcription error in the ZIP code, so we took care to examine each case in which the city supplied by Lorton was different from the input city.

After both the AccuMail and the Lorton processing stages, there was extensive examination of the addresses that were not recognized by the USPS database, to do additional checking and corrections of spelling mistakes and ZIP code transcription errors, and those addresses which had been recognized by AccuMail or Lorton but which seemed to have a level of precision greater than warranted by the address type (e.g., when a street-name-only address had a delivery point assigned). Use was made of the USPS ZIP code Directory, of the USPS database in AccuMail, of an electronic street map of the United States (StreetAtlas USA from DeLorme mapping), and checks against addresses for the same family in other interview years. After a wave of hand corrections was made, another stage of AccuMail and Lorton processing was done on the residual of non-non-standardized addresses.

After several iterations of the standardization process, we had the following levels of precision in our coding of addresses for 1968-1995:

Delivery point (with 5-digit ZIP code and ZIP+4)	144,798
ZIP+4 (with 5-digit ZIP code)	1,865
Carrier route (with 5-digit ZIP code)	6173
5-digit ZIP code	35,187
Subtotal usable for matching	*188,023*
Not usable (no ZIP code) or no address	963
Total	**188,986**

Our standardization process resulted in 93% of the PSID addresses being recognized and fully coded using US Postal Service databases. The great majority of the addresses to which we could not assign delivery points, ZIP+4s, and carrier routes, were ordinary-looking addresses, with street names and numbers. We apparently had no difficulty getting mail delivered to these addresses, so it seems likely that the great majority are valid and deliverable, but for some reason are not in the USPS database. The USPS database does systematically exclude addresses in small towns where there is only one carrier, or there is no home delivery, since coding for carrier sorting was the main impetus for developing the database. However, many of the addresses for which we found no USPS match were in large cities, which should be entirely in the USPS database. The error codes cited by Lorton indicate that "no matching street name" is the basis for about half the errors, and "invalid address number range" is the basis for about a third. Our best guess is that these addresses include transcription errors in the street address that are sufficient to confuse a computer matching system like AccuMail and Lorton, but which a human mail carrier can translate and deliver to the proper address. The unusable PSID addresses were primarily foreign (about 70%) or 1995 non-interviews (25%).

Because MapInfo and other GIS software packages have difficulty with multiple datapoints at the same geographic coordinates, we attempted to eliminate duplicate address records (such as when the same family lives at the same address for several years). For addresses recognized by the post office, we used ZIP code, ZIP+4, and delivery point to identify duplicates; for non-standardized addresses, we used ZIP code and street address. This gave us approximately

42,600 apparently unique addresses to use in geocoding. There are probably a few duplicates in this number, due to different apartment numbers and other slight differences in spelling among the addresses not recognized by the post office.

1990 Geocodes for 1968,1970-1995 Addresses

As noted above, the geocoding of 1968-1985 addresses to 1980 and 1970 geocodes used paper listings and paper maps, which were not available for the 1990 Census. Instead, we attempted to use Geographic Information System (GIS) software and other electronic aids to assign the geocodes. Our first attempt was to use the Census Bureau's TIGER/Line files in combination with the MapInfo GIS software. We used software called TMT to extract from the 75 CD-ROMs for the 1992 version of the Census Bureau's TIGER/Line files boundary data for each county in the United States, containing street maps and geographic boundaries for 1990 states, 1990 counties, 1990 Places, 1990 tracts and BNAs, 1990 metropolitan statistical areas. These county boundary files were consolidated into state-level boundary files. The boundary files were then read into MapInfo to create GIS files. At each stage, the files for a state took from 200Mb to 2 GB of space, and data storage was primarily on backup tape.

Geocoding in MapInfo is done by converting addresses to points defined by latitude and longitude, joining the point map to a boundary map of the tracts, etc., and then exporting the point-level data (with associated IDs and geographic areas) to a database. We were able to create point maps for about 65% of the input addresses using an automated routine that compared the input addresses with the street name listings derived from the TIGER/Line files. Another 10% of addresses could be matched one at a time by comparing the input addresses with the TIGER/Line street listings. The remaining 25% of addresses (more than half of which were rural route, general delivery, PO Box, or building name addresses) had to be processed separately to determine latitude and longitude. We used a map package called StreetAtlas USA from DeLorme mapping. We were able to obtain exact latitude and longitude for about half of the ordinary street addresses remaining, and used the centroid of the street (essentially, the ZIP+2 line) for instances of street names without numbers. For rural route addresses, we assigned the latitude and longitude of the centroid of the 5-digit ZIP code area. For general delivery, PO Box, building name, and the remainder of the ordinary street addresses, the latitude and longitude of the geographic centroid of the Place (city) was assigned. The resulting latitudes and longitudes were added to those created by the automated and manual address matching to for street names and numbers, the point maps created and joined to the boundary maps, and the output databases produced.

Our intention was to go through this process for all 50 states, DC, and eight colonies. We completed the entire process for 15 states in the Northeast and Southeast regions, areas that contained about 24% of all the addresses we were attempting to geocode. We completed the automated and manual address matching, and the supplementary latitude/longitude lookup portions of the process, for another 27 states in the Southeast, Midwest, and Mountain regions, containing another 44% of all our addresses. Unfortunately, at that point, we had a hard disk crash that destroyed the file allocation table on the hard drive, and were unable to recover the files created from the automated and manual address matching. We did not have usable tape backups of these huge files. Lack of time, staff, and funding to redo that portion of the process meant that we had to devise a shortcut solution. One possibility was to use StreetAtlas USA to give us latitude and longitude for all the addresses, make point maps from those, and join them to the boundary layers. However, that also involved time, staffing, and funding beyond the constraints of the budget.

The course taken was to submit the full set of unique addresses (approximately 45,600) to two commercial geocoding firms, Geographic Data Technologies and DecisionMark. They were able to provide us with a limited set of 1990 geocodes: state, county, Place, Minor Civil Division, and tract/BNA. We used two firms since we wanted to compare the results and choose the better of the two. The results were very similar for the two firms. The latitudes and longitudes they provided were identical in 88% of the cases, as were the associated geocodes. Most of the discrepancies were in the treatment of rural route, general delivery, PO Box, and building name addresses. For those cases and others for which neither company seemed to have provided a latitude and longitude close to what we thought was appropriate (n=3,206), we used StreetAtlas USA to generate our own latitudes and longitudes and resubmitted them along with the addresses to GDT for geocoding based on the latitude/longitude points. Unfortunately, in the two-month interval between the original quote and our submission of the second wave of data, GDT had eliminated its capacity to geocode based on latitude/longitude points. Nonetheless, the returned geocoded data seemed considerably improved. We combined the two waves of GDT data and the one wave of DM data, and selected what appeared to be the best information from the three datasets, using two criteria: (1) whether the level of precision of the geocoding (street number, ZIP+4 [side of block] centroid, ZIP+2 [street length] centroid, or 5-digit ZIP code centroid) was appropriate to the type of input address (street name and number, street name only, rural route, PO Box, building name); and (2) our determination that GDT had come closer to the StreetAtlas latitudes and longitudes in a plurality of the sample of cases we checked.

Finally, we recombined the best geocodes from the two firms with our original PSID identifiers, year and interview ID. The file includes also includes the address type, the geocode precision level and the string of six 1990 geocodes.

1990 Geocodes for 1996-1999 Addresses

The geocoding for 1996-1999 is significantly different from the previous works. It was done in three phases. In the first stage, we developed highly organized and automated procedures for 1996-1999, having a possible application of future geocoding tasks in mind. Under the new system, SAS/GIS first recognizes each address and assigns basic geographical variables (such as block, tract, county, state, and ZIP code) by opening a corresponding TIGER/LINE map automatically. By minimizing human involvement in this way, we tried to avoid possible human errors in the geocoding process. For the cases that SAS/GIS was unable to geocode, the system utilizes previous address matching files or uses the *Federal Financial Institutions Examination Council's*[3] Geocoding Web site to obtain census tracts. Finally, for the remaining cases with a valid ZIP code, an imputation was made such that geographic variables were assigned based on the most populated area within each ZIP code. Using the computerized sequence of steps, at the conclusion, we were able to obtain tract-level information for 90-95% of all addresses. Again, our goal has been to have a seamlessly integrated and accurate procedure for geocoding.

In the second stage, we adopted the idea of a 'correlation list' approach of the MABLE/Geocorr (http://www.census.gov/plue/geocorr) in our geocoding procedure. A correlation list is a table that connects a geographic area with a 'source geocode' to corresponding geographic coverage

[3] The FFIEC (http://www.ffiec.gov) website utilizes the high quality map from the Geographic Data Technology (GDT – http://www.geographic.com)

specified by a 'target code'. The advantages of this approach are twofold. First, by separating a source code from a target code, we were able to focus on the basic 4-level hierarchy of geographic entities (i.e., block-tract-county-state) at the first stage. Second, by keeping target codes separately, geographic variables can be easily aggregated at several levels. For this time, we included 1990 Places, 1990 (primary) metropolitan statistical areas (or NECMSA for New England states) as target codes.

Release 2

Finally, a third stage of geocoding was implemented in June, 2005. Advancements in GIS technology made it possible to geocode addresses from 1996, 1997, and 1999 that were previously imputed during release 1. Geocoding software available for release 1 did not return geocodes with enough confidence for these addresses to be released, hence they were imputed. During this third stage, we were able to find geocodes with almost 100% confidence for 2,866 addresses that were imputed in release 1. The table below illustrates.

Year of Data Collection	**Percent of Imputed Addresses in release 1**	**Percent of Imputed Addresses in release 2**
1996	18%	10%
1997	6%	3%
1999	38%	10%

Glossary of Geographic Area Terms and Abbreviations

It may take analysts of this collection of datasets a while to get used to some of the geographic concepts represented in the text of the individual variable descriptions, so we offer the following brief (and incomplete) glossary for temporary guidance. However, no analyst should make choices among geographic levels for analysis without thoroughly studying the full descriptions of the geographic identifier variables below.

BNA: Block Numbering Area, a "neighborhood"-like area analogous to a tract in an area (typically a small city) that is blocked but not tracted.

CCD: Census County Division, a Census Bureau-created approximation of a township in counties without township-like subdivisions; a possible substitute "neighborhood" if tract, BNA, and ED are not available.

CMSA/SCSA: Consolidated Metropolitan Statistical Area, formerly called Standard Consolidated Statistical Area (SCSA), a group of bordering PMSA/SMSAs with substantial cross commuting of workers; a possible "economic area", with the disadvantage of including only a small part of the land area of the US.

ED: Enumeration District, the basic work area for a single Census enumerator; a possible "neighborhood" approximation in rural (untracted and unblocked) areas.

ESR: Economic Sub-Region, a group of two or more topographically and economically similar counties, often crossing state lines, comprised of two or more SEAs; a possible "economic area", with the advantage of being geographically comprehensive.

LMA: Labor Market Area, one or more counties with close economic ties defined by patterns of commuting to work; specially created for this dataset as a geographically comprehensive "economic area".

MCD: Minor Civil Division, a legal subdivision of a county, typically a township or a city; a possible substitute for a "neighborhood" in areas where tract, BNA, and ED are not available.

NECMA: New England County Metropolitan Area, an alternative form of metropolitan areas in New England states, with the advantage of being comprised of whole counties, not of portions of counties as is the case for PMSA/SMSAs in the region.

PLACE: Census Place, typically, a city or other municipality, sometimes crossing county lines; a possible substitute for "neighborhood" if tract, BNA, and ED are unavailable, with the disadvantage of including only a minority of land area in the US.

PMSA/SMSA: Primary Metropolitan Statistical Area, formerly called Standard Metropolitan Statistical Area (SMSA), a group of one or more counties defined by large urban populations and patterns of commuting to work; a possible "economic area" with the disadvantage of including only the large urban areas of the US.

SEA: State Economic Area, a group of counties within a state, defined by topographic and economic similarities; a subdivision of an ESR; a possible "economic area", with the advantage of being geographically comprehensive.

TRACT: Census tract, a "neighborhood"-like area in larger urban settings.

ZIP code: US Postal Service Zoning Improvement Plan area, a possible substitute for "neighborhood" if tract, BNA, and ED are not available.

Variable Descriptions (N, Mean, Minimum and Maximum for All Variables)

Note: In the tables below missing data, typically coded as a field of nines, has been removed from the mean, minimum and maximum calculations. The reduced N indicates the number of records affected for each variable.

1970 Geocode Match data statistics

Variable	Label	N	Minimum	Maximum
V1	Year	105427	1968	1985
V2	Interview Number	105427	1	7032
V701	70 STATE FIPS ID	104981	1	56
V702	70 ECONOMIC SUBREGION	104981	1	121
V703	70 STATE ECONOMIC AREA	104981	1	36
V704	70 PMSA	78507	40	9320
V705	70 COUNTY FIPS ID	104981	1	840
V706	70 MCD/CCD	104955	1	645
V707	70 ED	2272	300	120900
V708	70 TRACT 6-DIGIT	70357	100	950800
V709	70 PLACE	99829	3	9052
V710	70 CONSOL METRO STAT ARE	29856	7	91
V711	70 ZIPCODE5	104960	1002	99801
V712	70 LABOR MARKET AREA	104981	100007	455008
V713	70 NBRHD/PROB TYPE	104992	1	13
V714	WH 70 ST-CO MATCH CNX	105427	0	1
V715	WH 70 TRACT MATCH CNX	105427	0	1
V716	WH 70 ED MATCH CNX	105427	0	1
V717	WH 70 MCD/CCD MATCH CNX	105427	0	1
V718	WH 70 PLACE MATCH CNX	105427	0	1
V719	WH 70 ZIP5 MATCH CNX	105427	0	1
V720	70 ZIPCODE3	104960	10	998
V721	WH 70 BNA MATCH CNX	105427	0	1
V722	WH 70 ZIP3 MATCH CNX	105427	0	1

1980 Geocode Match data statistics

Variable	Label	N	Minimum	Maximum
V1	Year	105427	1968	1985
V2	Interview Number	105427	1	7032
V801	80 STATE FIPS	104981	1	56
V802	80 ECONOMIC SUBREGION	104981	1	121
V803	80 STATE ECONOMIC AREA	104981	1	36
V804	80 PRIM METRO STAT AREA	79924	40	9340
V805	80 COUNTY FIPS	104981	1	840
V806	80 MCD/CCD	104955	1	663
V807	80 ED	2728	100	176000
V808	80 TRACT/BNA	77882	100	993600
V809	80 CENSUS PLACE	98802	3	9052
V810	80 CONSOL METRO STAT ARE	34309	7	91
V811	80 ZIPCODE5	104960	1002	99801
V812	80 LABOR MARKET AREA	104981	100007	455008
V813	80 NBRHD/PROB TYPE	104992	1	13
V814	WH 80 ST-CO MATCH CNX	105427	0	1
V815	WH 80 TR/BNA MATCH CNX	105427	0	1
V816	WH 80 ED MATCH CNX	105427	0	1
V817	WH 80 MCD/CCD MATCH CNX	105427	0	1
V818	WH 80 PLACE MATCH CNX	105427	0	1
V819	WH 80 ZIP5 MATCH CNX	105427	0	1

1990 Geocode Match data statistics (see codebook for frequencies)

Variable	Label	N	Min	Max
YEAR90	Year	211242	1968	1999
RLS90	Release Number	211242	2	2
FAMID90	Interview Number	211242	1	16970
IMPUTE90	Whether tract was imputed	211242	1	5
ADDRTYPE90	Address type	211162	1	8
GEOPRC90	Geocode precision level	211169	1	8
STATE90	State FIPS 90	210165	1	72
CNTY90	County FIPS 90	210162	1	840
MSA90	MSA 90	210083	40	9360
TRACT90	Tract/BNA 90	167842	100	9983.00
ZIP590	5-digit ZIP code (year of data collection)	187963	657	99921
PLACE90	Place FIPS 90	156441	00100	89140

Notes and Problems on Variables

INTERVIEW YEAR

OTHER NOTES AND PROBLEMS

This variable, in combination with the Interview Number variable uniquely identifies records in the Geocode Match files.

OTHER NOTES AND PROBLEMS

This variable, in combination with the Interview Year variable uniquely identifies records in the Geocode Match files.

This variable provides a match to the PSID main annual family and individual files. See table below for corresponding variables in the main PSID files.

Family Interview Numbers in Single-year Family Files and in Cross-year Individual File

	Interview Number	
Year	**Family File**	**Individual File**
1968	V3	V30001
1969	V442	V30020
1970	V1102	V30043
1971	V1802	V30067
1972	V2402	V30091
1973	V3002	V30117
1974	V3402	V30138
1975	V3802	V30160
1976	V4302	V30188
1977	V5202	V30217
1978	V5702	V30246
1979	V6302	V30283
1980	V6902	V30313
1981	V7502	V30343
1982	V8202	V30373
1983	V8802	V30399
1984	V10002	V30429
1985	V11102	V30463
1986	V12502	V30498
1987	V13702	V30535
1988	V14802	V30570
1989	V16302	V30606
1990	V17702	V30642
1991	V19002	V30689
1992	V20302	V30733

Family Interview Numbers in Single-year Family Files and in Cross-year Individual File

	Interview Number	
Year	**Family File**	**Individual File**
1993	V21602	V30806
1994	ER2002	ER33101
1995	ER5002	ER33201
1996	ER7002	ER33301
1997	ER10002	ER33401
1999	ER13002	ER33501

FIPS STATE IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

FIPS is an acronym for Federal Information Processing Standard.

The Census Bureau treats the 50 states proper, plus the District of Columbia, as "states" for statistical reporting purposes. (The colonies American Samoa, Guam, Puerto Rico, and the US Virgin Islands are also assigned state identification numbers, but, except for Puerto Rico, little statistical data is available about them, and they are excluded from our datasets).

ECONOMIC SUB-REGION IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

Economic Sub-Regions (ESRs) were devised in 1950, and slightly revised in 1960, by the Census Bureau and the Department of Agriculture (Economic Geography Division), and provide rough nationwide analogues to Consolidated Metropolitan Statistical Areas. ESRs consist of groupings of State Economic Areas and frequently cross-state lines. They include from 3 to 93 counties and vary greatly in land area and population. Similarities of topography and natural resources seem to have been more important than work commuting patterns in defining ESRs; the shapes are often irregular, but nearly all are contiguous (the exceptions being a few coastal areas broken by bays.) Although ESRs in metropolitan areas were originally designed to include PMSA/SMSAs and CMSA/SCSAs, the boundaries of PMSA/SMSAs and CMSA/SCSAs have expanded and otherwise changed, while those of ESRs have not; ESR lines therefore now split some PMSA/SMSAs and CMSA/SCSAs.

STATE ECONOMIC AREA IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

State Economic Areas (SEAs) were devised in 1950, and slightly modified in 1960, by the Census Bureau and the Department of Agriculture (Economic Geography Division) in an attempt to provide nationwide analogues to Standard Metropolitan Statistical Areas. SEAs were conceived as subdivisions of ESRs, with boundaries based in part on state lines and in part on the combination of topographic and natural resource considerations used to define ESRs. Metropolitan areas (as of 1960) are treated as separate SEAs, and this sometimes results in the surrounding area being treated as a single non-contiguous SEA or as an oddly shaped contiguous area fully or partially "ringing" the PMSA/SMSA or CMSA/SCSA. Because the boundaries of PMSA/SMSAs have changed since 1960, and those of SEAs have not, the two areas frequently have crosscutting boundaries.

The SEA code in the original Census data is of a mixed type: a 2-digit numeric code ranging from 1 to 10 for non-metropolitan SEAs, and a 1-character alphabetic code ranging from A to P for metropolitan SEAs. In this collection of datasets, the alphabetic codes have been translated into 2-digit numeric codes, with "A" becoming "21", "B" becoming "22", ... and "P" becoming "36".

This variable must be used in conjunction with the FIPS State code to uniquely identify an SEA.

PRIMARY METROPOLITAN STATISTICAL AREA/STANDARD METROPOLITAN STATISTICAL AREA IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

The Census Bureau currently uses two slightly different names for metropolitan areas; Metropolitan Statistical Areas (MSAs) for those that are not considered part of a Consolidated Metropolitan Statistical Area (CMSA/SCSA), and Primary Metropolitan Statistical Areas (PMSA/SMSAs) for those that are a part of a CMSA/ SCSA. We found this dichotomy unnecessarily confusing, and use the PMSA designation to cover both MSAs and PMSA/SMSAs proper. Previously, both MSAs and PMSA/SMSAs were called Standard Metropolitan Statistical Areas (SMSAs).

PMSA/SMSAs consist of one or more counties (in New England, one or more towns) that a) include a city of 50,000 or greater population, **or** b) have an urbanized area of 50,000 or greater population **and** a total population of 100,000 or more. In addition to the "central" county (or town) containing the city or urbanized area, adjacent ("fringe") counties are included in the metropolitan area if 10% or more of the employed residents of the "fringe" area work in the "central" area. PMSA/SMSA boundaries change with some frequency, typically two years before a decennial census (anticipating what population and commuting patterns will be at the time of the census) and again three years after the decennial census (based on actual results). At these points, it is common for new PMSA/SMSAs to be formed when a county passes a population threshold, for adjacent counties to be added to an existing PMSA/SMSA as commuting increases, for adjacent PMSA/SMSAs to be combined into one as cross-commuting increases, and for a multi-county PMSA/SMSA to split into two as cross-commuting decreases. In addition, counties previously part of a PMSA/SMSA may cease to be part of any PMSA/SMSA due to decreases in commuting to the central county, as happened with 11 counties between 1970 and 1980; or may shift from one PMSA/SMSA to another due to changing commuting patterns, as happened with six counties between 1970 and 1980. The increase from 247 SMSAs just before the 1980 Census to 328 PMSAs after the 1980 Census included 69 newly created PMSAs (in previously non-metropolitan areas) and 101 other boundary changes.

The Census Bureau has used counties as the basic constituents of PMSA/SMSAs in all parts of the nation except the New England region. In New England, the basic constituents of PMSA/SMSAs are towns (called townships in most of the rest of the country). This use of town constituents was somewhat problematic for us, because we created our PMSA/SMSA datasets by aggregation, and there is considerably more suppressed and otherwise missing data at the township (MCD/CCD) level than at the county level. Therefore, we chose to use the Census Bureau's alternate form of PMSA/SMSAs for New England, called New England County Metropolitan Areas, or NECMAs. NECMAs, like PMSA/SMSAs in the rest of the nation, consist of whole counties, and allowed us to aggregate with many fewer suppression and missing data problems. There are fewer NECMAs than PMSA/SMSAs in New England, and they cover a larger geographic area.

Our definitions of SMSAs for the 1970 Census data come from the Federal Committee on Standard Metropolitan Statistical Areas, Standard Metropolitan Statistical Areas 1975 Revised Edition. We used the definitions of areas in Parts II and V for areas outside of New England to create 237 SMSAs, and those in Part VII to create 13 NECMAs (rather than the 27 SMSAs) in the

region, for our total of 250 SMSAs. (We used the historical data in Part IX to reflect the definitions as of April 1973).

We adopted the 1983 definitions of PMSAs from Appendix A of the Census Bureau's 1983 County and City Data Book, which had 298 non-New England PMSAs and 16 NECMAs (rather than the 30 PMSAs in the region), to give us 314 unique PMSAs.

We used the 1973 and 1983 definitions of PMSA/SMSAs from the above sources, rather than the 1969 and 1979 definitions that appeared in the county-level data in the Census Bureau's datasets, because we wanted to have areas defined with the benefit of the Census information itself, not the Census Bureau's advance guess as to the outcomes. This meant we had to create our own PMSA/SMSA-level datasets, because those in Count 4C for 1970 and STF3A for 1980 were based on the pre-Census definitions. Analysts should consult the sources cited above for keys to the areas associated with each PMSA/SMSA ID number.

FIPS COUNTY IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

FIPS is an acronym for Federal Information Processing Standards.

The Census Bureau defines "counties" as the primary political and administrative subdivisions of states. In most states these sub-divisions are called "counties", but in Louisiana they are "parishes" and in Alaska "boroughs". In addition, the Census Bureau treats as "county equivalents" the District of Columbia (which is also treated as a state equivalent), the several "Census Areas" drawn by the Census in areas not included in Alaska's boroughs, and numerous cities in four states that are politically independent of the county in which they are located. Except for Alaska Census Areas, county boundaries are relatively stable – between 1970 and 1980 there were a handful of border areas that shifted from the jurisdiction of one county to another and three merger/annexations that caused previously separate counties to end their separate existence. There were 3141 areas treated as counties in 1970, and 3137 in 1980. The numbering of "independent cities" is distinctive in the states where they exist (Maryland [Baltimore], Missouri [St. Louis], Nevada [Carson City], and Virginia [38 cities in 1970, 41 in 1980]), in the range 500 and above.

Counties are uniquely identified by the combination of state and county codes.

OTHER NOTES AND PROBLEMS

Minor Civil Divisions are primary political/administrative subdivisions of counties (or county-equivalents). Most MCDs are called "townships", but there are other names – "towns" in New England, New York, and Wisconsin; "magisterial districts" in Virginia and West Virginia; "supervisor districts" in Mississippi, "election districts" in Maryland, and "police jury wards" in Louisiana. In the District of Columbia, the four "quadrants" are treated as MCDs. In Alaska, the Census Bureau draws its own "Census Sub-Areas" as MCDs. There are two substantial complications to this scheme. First, municipalities can be treated as MCDs. The "independent cities" treated as county equivalents are also treated as MCDs in all states, but states themselves can decide which other municipalities should be treated as MCDs separate from their surrounding townships, and there is great variation among states in the criteria they have applied and in proportion of municipalities designated as MCDs.

A second major complication is that MCD boundaries change frequently and substantially. Municipalities annex surrounding township land in all states. In addition, those MCDs that are county electoral districts change boundaries due to population shifts shown by the Census itself. In some states, sub-county divisions are drawn for administrative convenience and change frequently and/or are of quite small size. And some states have no sub-county divisions.

To address some of these problems, the Census Bureau has devised Census County Divisions (CCDs) that it uses for statistical reporting purposes instead of MCDs in most states where MCDs are small or have frequently changing boundaries. The Census Bureau in cooperation with local planning authorities so as to be bounded by relatively unchanging roads and natural features defines Census County Divisions. The 20 CCD states in 1980 were Alabama, Arizona, California, Colorado, Delaware, Florida, Georgia, Hawaii, Idaho, Kentucky, Montana, New Mexico, Oklahoma, Oregon, South Carolina, Tennessee, Texas, Utah, Washington, and Wyoming. In 1970, North Dakota was a CCD state, but was an MCD state in 1980.

Most MCD states had **relatively** unchanging MCD boundaries, with changes occurring primarily due to annexation. The 23 "stable MCD" states in 1980 (and 1970) were Arkansas, Connecticut, District of Columbia, Illinois, Indiana, Iowa, Kansas, Maine, Massachusetts, Michigan, Minnesota, Missouri, Nevada, New Hampshire, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Rhode Island, South Dakota, Vermont, and Wisconsin.

The most problematic category is those states which have unstable MCD boundaries but which had **not** been designated as CCD states by the Census Bureau despite that instability. There were eight "unstable MCD" states in 1980: Alaska, Louisiana, Maryland, Mississippi, Nebraska, North Dakota, Virginia, and West Virginia; all these (except North Dakota, which was a CCD) were treated as MCDs in 1970 as well.

MCDs and CCDs are geographically comprehensive – all the land in the United States is located in such an area.

MCDs/CCDs are identified by a three-digit numeric code assigned by the Census; the codes are assigned in alphabetical order within county, and change primarily when MCD/CCD names

change – there were about 550 such changes between 1970 and 1980. An MCD can be uniquely identified with a combination of state code, county code, and MCD/CCD code.

In those states in which MCDs are county legislative districts (Virginia, West Virginia, Mississippi, Maryland, and Louisiana), the boundaries of districts may change substantially between (and because of) decennial censuses. However, the names of the districts (e.g., "District 2") may not change even if they refer to very different pieces of land. Since the MCD numbering system is based on names, the same MCD number may refer to very different areas in different censuses. In these five states, more than most others, cross-time comparisons of MCD characteristics should be undertaken with extreme caution.

ENUMERATION DISTRICT IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

Enumeration Districts (EDs) are the most basic work units for the Census Bureau, the area assigned to a single enumerator. EDs do not cross the boundaries of legal areas (counties, MCDs) or of statistical areas (tracts, CCDs), but are otherwise drawn so as to be bounded by roads and other natural features. EDs may be redrawn and renumbered for each decennial census, and there is no convenient way to translate ED identification numbers from one decennial census to the next.

While there were over 250,000 EDs defined in 1970 and 1980, covering the entire land area of the United States, the Census Bureau released statistical data only for those roughly 100,000 EDs that are in counties that were not fully tracted (1970) or fully blocked (1980).

The Census Bureau's ED code is comprised of two parts – a four-digit numeric prefix, and an one-character alphabetic suffix (which may be a blank). We have transformed this into a six-digit numeric code by changing the one-character alphabetic suffix into a two-digit numeric suffix – Blank=00, A=01, B=02,....Z=26.

EDs are uniquely identified by a combination of state, county, and ED codes.

ED-level data is available only in untracted, unblocked areas, i.e., those non-urbanized areas for which tract- and BNA-level data is not available. ED-level data should therefore be used in combination with tract and BNA data to obtain a national perspective on both urban and rural areas.

OTHER NOTES AND PROBLEMS

Census tracts are the basic statistical reporting unit in metropolitan areas; block-numbering areas (BNAs) serve the same function in untracted urbanized areas, and the Census Bureau in most respects treats tract and BNA data as a single level of aggregation. Tracts and BNAs are designed to be bounded by roads and natural features, and relatively homogeneous as to population characteristics, economic status, and living conditions; the local committees that establish tract and BNA boundaries typically intend them to represent subjective "neighborhoods".

Tract- and BNA-level data are available primarily for urbanized areas of the US (the primary exceptions being the fully-tracted states noted below), and should be used in combination with ED-level data to obtain a national perspective on both urban and rural areas.

Tract boundaries are relatively stable from one decennial census to the next. For example the Census Bureau's 1970-1980 tract comparability file indicates the following relationships:

Tracted in both 1970 and 1980	
No change in boundaries	25,800
Minor Change in Boundaries (affecting less than 100 persons)	4,402
Exact Split of 1970 Tract into several 1980 tracts	9,097
Other more complex change (including consolidation of several 1970 Tracts into one 1980 Tract)	5,712
Untracted in one year	
Untracted in 1970	4,504
Untracted in 1980	5
Tracts in 1970	44,980
Tracts in 1980	49,519

(This analysis is based on a dataset we created from two Census Bureau files – a pre-1980 Census file that included all 1970 tracts, and a post-1980 census file that included only those 1980 tracts that had different boundaries and/or tract numbers than they had in 1980. BNAs are **not** included in these files.)

Census tract numbers include a four-digit numeric prefix, and a two-digit numeric suffix (which is often blank). In map and some other Census Bureau representations of tract numbers, a decimal point appears between the prefix and suffix. In all of our work, we have eliminated the decimal point, and substituted "00" for blank suffixes.

Block numbering areas (BNAs) are numbered in the same manner as tracts, differing only in the range – 940100 and up in 1970 and 990100 and up in 1980.

Tract and BNA numbers are unique within counties, and can be uniquely identified by use of the state, county, and tract/BNA codes.

Most Census tracts (about 95%) are located in metropolitan areas. But over 260 non-metro counties contain over 3000 tracts. The states of Connecticut, Delaware, District of Columbia, Hawaii, New Jersey, and Rhode Island were fully tracted in 1980, as were DC, Hawaii, and Rhode Island in 1970.

BNAs typically appear in untracted urbanized areas, but four states were fully blocked in 1980 – Georgia, Mississippi, New York, and Virginia – and therefore have BNA data available for all untracted areas. (DC and Rhode Island are both fully tracted and blocked, so there is no BNA data for them.)

CENSUS PLACE IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

Census Places are of two types – incorporated Places, such as cities, villages, or towns, which have legally prescribed powers and functions; and Census Designated Places, (CDPs, previously "unincorporated areas") which are densely settled areas (at least 1000 persons per square mile) with a locally-used distinctive name. The Census Bureau makes data available for all incorporated Places and for CDPs with a minimum population (5000 in urbanized areas with a central city of 50,000 or greater population; 1000 in other areas). It is not unusual for a CDP to coincide exactly with an MCD. Places frequently cross county and MCD lines; in 1980, more than 4000 of the over 22,000 Places crossed county lines. Census Places, although including 73% of the US population in 1980, include only about 15% of the land area. Place-level data, therefore includes only "urbanized" portions of the US population, and should not be used as the sole geographic level if the objective of the analysis is to represent the entire population. The Census Bureau assigns a four-digit numeric code to each Place it recognizes. Places are unique within states, and can be uniquely identified with state and Place codes. Place boundaries change often as cities annex portions of surrounding townships, but the changes tend to be small compared to the original area. Over two-thirds of 1970 incorporated Places had boundary changes by 1980, and nearly half of 1970 CDPs changed boundaries by 1980.

CONSOLIDATED METROPOLITAN STATISTICAL AREA/STANDARD CONSOLIDATED STATISTICAL AREA IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

Consolidated Metropolitan Statistical Areas (CMSAs, formerly called Standard Consolidated Statistical Areas, or SCSAs) are groupings of two or more contiguous Primary Metropolitan Statistical Areas (see Variable 7 in the Census Extract file codebook). A CMSA/SCSA is designated when two or more contiguous PMSA/SMSAs meet all three of the following conditions:

a) there is substantial commuting of employed persons living in the smaller PMSA/SMSA to work in the larger – either 15% of workers, or 10% where central urbanized areas are contiguous or shared;

b) at least 60% of the population of each PMSA/SMSA is urban; and

c) the combined population for the PMSA/SMSAs is at least one million.

We used the definitions of SCSAs found in Standard Metropolitan Statistical Areas 1975 Revised Edition Parts III and IV to establish SCSAs as defined in April 1973; and 1983 County and City Data Book Appendix A to establish CMSAs as of June 1980 using the 1980 data. See the citations and discussion of these sources above under Variable 7 in the Census Extract file codebook. Because we used New England County Metropolitan Areas (NECMAs) instead of PMSA/SMSAs proper in New England, we had to somewhat redefine CMSA/SCSAs differently in that region as well. This resulted in the following CMSA/SCSAs in New England:

1970 SCSA	1970 NECMA Constituents
Boston (07)	Boston (1123)

1980 CMSA	1980 NECMA Constituents
Boston (07)	Boston (1123), Manchester-Nashua (4763)
Hartford (41)	Hartford (3283)
Providence (80)	New Bedford (5403), Providence (6483)

In addition, the Bridgeport NECMA (1163) is treated as part of the New York CMSA/ SCSA (70) in 1980.

POSTAL ZIPCODE IDENTIFICATION NUMBER (3-DIGIT)

OTHER NOTES AND PROBLEMS

The US Postal Service introduced the Zoning Improvement Plan in 1963 as a means of improving the routing of mail. Each postal address, including non-residential addresses such as office buildings and post office boxes, is assigned a 5-digit numeric code. ZIP codes include roughly equal postal delivery workloads. ZIP code boundaries are relatively stable, changing primarily by subdivisions in which new codes are added within the boundaries of previous codes; however, codes also can disappear as postal delivery areas are merged.

There is a very rough correspondence of the names associated with ZIP codes and those of the census Places on which they are centered. However, there are many census Places that do not have their own ZIP code, and many ZIP code names that do not exist (or no longer exist) as separate census Places. In addition, the boundaries of a ZIP code associated with a postal city name frequently extend well beyond the city (Place) limits. For 1970, the Census Bureau created and released data at the 5-digit ZIP code level only for metropolitan areas. For non-metropolitan areas in 1970, this variable is the closest approximation to the 5-digit ZIP code data described below.

Analysts should also note that although the first three digits of a ZIP code nearly always uniquely identify the state in which the post office serving a given area is located, some offices service addresses in another state; therefore, a postal address in one state may in fact represent a residence in another state very nearby.

POSTAL ZIPCODE IDENTIFICATION NUMBER (5-digit)

OTHER NOTES AND PROBLEMS

The US Postal Service introduced the Zoning Improvement Plan in 1963 as a means of improving the routing of mail. Each postal address, including non-residential addresses such as office buildings and post office boxes, is assigned a 5-digit numeric code. ZIP codes include roughly equal postal delivery workloads. ZIP code boundaries are relatively stable, changing primarily by subdivisions in which new codes are added within the boundaries of previous codes; however, codes also can disappear as postal delivery areas are merged.

There is a very rough correspondence of the names associated with ZIP codes and those of the census Places on which they are centered. However, there are many census Places that do not have their own ZIP code, and many ZIP code names that do not exist (or no longer exist) as separate census Places. In addition, the boundaries of a ZIP code associated with a postal city name frequently extend well beyond the city (Place) limits.

For 1970, the Census Bureau created and released data at the 5-digit ZIP code level only for metropolitan areas. For non-metropolitan areas in 1970, the closest approximation is the 3-digit ZIP code data described above.

Analysts should also note that although the first three digits of a ZIP code nearly always uniquely identify the state in which the post office serving a given area is located, some offices service addresses in another state; therefore, a postal address in one state may in fact represent a residence in another state very nearby.

LABOR MARKET AREA IDENTIFICATION NUMBER

OTHER NOTES AND PROBLEMS

This six-digit numeric variable is our best approximation of what constitutes a "labor market", and is based on the Census Extract dataset Labor Market Area Type Code, which forms its first digit, followed by five digits of identifiers specific to type:

l: CMSA/SCSA: 1000AA, where AA is the 2-digit numeric CMSA/SCSA identification code; this is used for all counties located in CMSA/SCSAs.

2: PMSA/SMSA: 20BBBB, where BBBB is the 4-digit numeric PMSA/SMSA identification code ; this is used for all counties located in PMSA/SMSAs but not in CMSA/SCSAs.

3: County: 3CCDDD, where CC is the 2-digit numeric state identification code and DDD is the 3-digit numeric county identification code ; this is used for all non-metropolitan counties in which less than 20% of the employed population commutes to work outside the county.

4: Pseudo-SEA: 4CC0EE, where CC is the 2-digit numeric state identification code and EE is the 2-digit numeric SEA identification code; this is used for all non-metropolitan counties in which 20% or more of the employed population commutes to work outside the county.

It should be noted that, because the process of creating variables was hierarchical (e.g., a county located in a CMSA/SCSA was assigned a CMSA/SCSA code regardless of which PMSA/SMSA it was in; a county in a PMSA/SMSA was assigned a PMSA/SMSA code regardless of which SEA it was in), aggregation of type 4 counties by SEA number creates a grouping that **ex**cludes counties that are part of metropolitan areas. Hence, our labeling type 4 as "Pseudo-SEAs".

NEIGHBORHOOD/PROBLEM CODE

OTHER NOTES AND PROBLEMS

This variable indicates the "best" neighborhood available for an address, assuming that tract is best for tracted areas, BNA for blocked but untracted areas, and ED for areas neither tracted nor blocked. We assumed the MCD/CCD was second best, and have indicated why we were unable to obtain tract, BNA, or ED.

Our use of MCD/CCD as the second-best address should not constrain analysts from using Place or ZIP code as an alternative for neighborhood. The problem codes associated with MCD/CCD codes should serve as an indication of how good such matches will be.

WHETHER PLACE CODE MATCHES CENSUS EXTRACT RECORD

OTHER NOTES AND PROBLEMS

This variable indicates whether there is a matching Place record in the Census Extract dataset. Analysts should note that there are a significant number of non-matches in 1970, because the Census Bureau released electronic Place-level data for 1970 only for Places with populations greater than 2500 persons.

WHETHER 5-DIGIT ZIPCODE MATCHES CENSUS EXTRACT RECORD

OTHER NOTES AND PROBLEMS

This variable indicates whether there is a matching 5-digit ZIP code in the Census Extract dataset. Analysts should note that there are frequent non-matches for 1970 because the Census Bureau created and released 5-digit Level-level data only for metropolitan areas and some non-matches for both 1970 and 1980 because some ZIP codes were not included in the aggregation.

WHETHER BLOCK NUMBERING AREA MATCHES CENSUS EXTRACT RECORD

OTHER NOTES AND PROBLEMS

This variable indicates whether there is a matching Block Numbering Area in the 1970 Census Extract dataset. There is no corresponding variable for 1980 because BNAs are treated as tracts in that year.

WHETHER 3-DIGIT ZIPCODE MATCHES CENSUS EXTRACT RECORD

OTHER NOTES AND PROBLEMS

This variable indicates whether there is a matching 3-digit ZIP code in the 1970 Census Extract dataset. There is no corresponding variable for 1980 because the Census Bureau did not aggregate to the 3-digit ZIP code level in that year.