

Life Course Socioeconomic Status (LCSES):
Manual of Procedures and Overview of Data Sets Available
Website url: <http://www.lifecourseepi.info>

1.	Introduction.....	2
2.	Data Sources	2
2.1.	LCSES ARIC Study.....	2
2.1.1.	Modifications to Protocol – Parental Education	3
2.2.	ARIC Study.....	3
2.3.	Neighborhood Characteristics Data	4
3.	Quality Assessment of Data Collected in LCSES Study	4
3.1.	SES_B Item-by-Item Checks	5
4.	Preparation of Addresses for Geocoding	5
4.1.	ARIC Visit 1-4 Addresses.....	6
4.2.	Historical Addresses	6
4.2.1.	Childhood Address Correction Summary.....	7
4.2.2.	Age 30 Addresses	7
4.2.3.	Age 40 Addresses	7
4.2.4.	Age 50 Addresses (sesb25a-sesb25e).....	8
5.	Geocoding.....	8
5.1.	Previous Geocoding Efforts in ARIC Study	8
5.2.	Geocoding Vendor and Match Codes	8
5.3.	Geocoding of Contemporary (V1-V4) Addresses.....	9
5.3.1.	Efforts to Increase Match Rates in Current Addresses	10
5.4.	Geocoding of Historical Addresses.....	11
5.4.1.	Assignment of Historic Census Tracts.....	12
5.4.2.	Hand Geocoding	14
5.5.	Quality Assessment of Geocoding.....	15
5.5.1.	Assessment of Reliability of Commercial Geocoding.....	15
5.5.2.	Assessing Short-term Repeatability of Geocodes.....	16
5.5.3.	Repeatability of Geocodes Assigned to Historical Addresses.....	17
5.5.4.	Assessment of Repeatability and Accuracy of Geocodes.....	17
6.	Census Based Contextual Data	17
7.	Data sets.....	18
7.1.	Individual SES Lifecourse Data Set.....	18
7.2.	Neighborhood SES Data Sets Information	18
7.3.	Childhood Neighborhood SES Information**	20

1. Introduction

Life Course SES, Social Context and Cardiovascular Disease (LcSES) is a four-year study sponsored by the National Heart, Lung, and Blood Institute. Its purpose is to study the explanatory mechanisms for the association of historical and contemporary socio-economic status (SES) with cardiovascular disease (CVD) in a large, population-based bi-ethnic cohort of men and women. LCSES is an ancillary study to the Atherosclerosis Risk in Communities (ARIC) study.

Between 2001 and 2002, survivors from the Atherosclerosis Risk in Communities (ARIC) Study were queried about their childhood and earlier adulthood socio-environmental circumstances. Information collected included individual level SES as well as childhood and early adulthood place of residence. Place of residence was linked with census level measures of the social environment from the appropriate time period. The childhood and early adulthood SES information collected as part of the LCSES ARIC Ancillary Study will be used in conjunction with individual and census tract level SES data collected at the ARIC baseline exam and the subsequent ARIC examinations to construct profiles of SES across the life course. This will be used in conjunction with the extant data on behavioral and biomedical risk factors collected over the course of the ARIC study to attempt to elucidate pathways linking SES / the social environment to cardiovascular disease (CVD) across the life course.

Most efforts to date to study the influence of the social context on health have focused only on the current context. Thus, a unique aspect of the LCSES study is the collection of historical residential information including linking this information to historical census-based measures of the social environment. However, linking self-reported previous places of residence with historical census data is not as straightforward as when linking current address with more recent census data. As established procedures for doing this were not identified in the literature, a major purpose of this manual is to provide an overview and document the procedures developed during the LCSES study.

2. Data Sources

There are two primary sources of data for the LCSES Ancillary Study: Individual level information, collected in the LCSES ARIC Annual Followup telephone interview and as part of ARIC Study (Visits 1-4) and neighborhood SES indicators, where the census tract is the unit of observation and pertinent Census years are the source of the data.

2.1. LCSES ARIC Study

The LCSES questionnaire was administered between 2001 and 2002, as part of the ARIC Annual telephone follow-up (AFU). Its official title is the "Residential / Occupational History Form (SES B)". The focus of the questionnaire was on

childhood and early adulthood (pre-ARIC baseline exam) socioeconomic circumstances. Childhood measures included mother's and father's occupation and education (see 2.1.1), as well as home ownership and city and state of residence. Early adult measures included occupation and characteristics of occupation at ages 30, 40, and 50, address and home ownership at ages 30, 40, and 50, and information related to participation in the armed services. A copy of complete SES B Questionnaire can be found in **Appendix 1**.

2.1.1. Modifications to Protocol – Parental Education

Parental education was first queried in the SES-A Questionnaire, as part of ARIC Visit 4. The SES-B questionnaire administered at the LCSES AFU had a skip pattern in place such that items relating to parental education were asked only if participants did not participate in the Visit 4 examination (SES_A).

Frequencies from the 1st SES_B data retrieval indicated that a high proportion of the participants who were asked questions relating to parental education were not able to provide a response (presumably because they did not know their parents' level of educational attainment). Additionally, there were a few reports from the Minnesota field center of complaints about the parental education items making participants feel embarrassed/uncomfortable. Given the combination of the high percentage of non-respondents and the complaints and staffing issues in Minnesota, a decision was made in the fall of 2001 to drop this item from the subsequent AFU interviews.

2.2. ARIC Study

The data collected as part of the LCSES Study represents only a portion of the SES data that will be used in the analyses involving life course related topics. Another source of data on earlier life SES is the SES-A questionnaire, which was administered during the Visit 4 examination. It includes information about parents' education as well as own occupation between ages 25-44 and current occupation and family income. A copy of the SES-A questionnaire can be found on the LCSES website, www.lifecourseepi.info, the ARIC website and in **Appendix 2**.

Additional individual-level SES and socio-demographic data is available from ARIC visits 1-4 as well as selected annual telephone follow-ups. These variables, as well as relevant items from the SES-A questionnaire have been incorporated into the LCSES dataset entitled 'lcdsc05a'. See Sec. 7.1 for further description of this data set.

2.3. Neighborhood Characteristics Data

One of the goals of the LCSES Study is to determine the extent to which current and historical neighborhood context modify the association of individual-level life course SES exposures and CVD events. Historical addresses were geocoded at the level of the appropriate (1960, 1970 or 1980) census tract, allowing us to link census-based area-level (neighborhood) measures to individual records and to examine the impact of the long-term social environment. For a further discussion of the geocoding process, see section 4 and 5. Census based data will be discussed in section 7.

Addresses from the ARIC visits 1-4 were also geocoded and linked to 1990 census data. Addresses from visit 4 were also linked to 2000 data, as the time period covered corresponded most closely to 2000. A more detailed description of this geocoding procedure will be presented in Section 4 and 5 of this manual .

3. Quality Assessment of Data Collected in LCSES Study

The LCSES investigators developed a series of steps to edit the SES-B data prior to the creation of the final analysis files. The general steps are outlined below:

- Data collection took place over a period of approximately one year. The Collaborative Studies Coordinating Center at UNC (CSCC) received the AFU-SES_B telephone interview results from the four field centers at intervals over the data collection period and released restricted-use partial datasets to the LCSES investigators. These files included participants' addresses. In order to protect participant confidentiality, the ARIC study IDs were replaced with encrypted IDs before distribution to the LCSES investigators. These intermediate datasets were used to develop programs for editing and "cleaning" the data.
- The editing process occasionally resulted in feedback to the ARIC interviewers, specifically if repeated errors were noticed (e.g., consistent misspelling of a city name).
- After the final data retrieval there were a total of 12,716 observations. The final dataset released by the CSCC was a 'clean' file with participant addresses but encrypted ARIC study ID's. The historical addresses from data set were sent to a commercial geocoding firm. When the geocoding process (to be described in Sections 4 and 5) was complete, final data sets were prepared; these data sets include the participants' actual study IDs and codes corresponding to the geographical areas in which they resided, ie state, county, tract. However, all address information was removed prior to replacing the encrypted with the participants' real IDs.

3.1. SES_B Item-by-Item Checks

The primary purpose of the quality control assessment was to evaluate the extent of missing data. As a first step, a SAS program was written to differentiate true missing values from those missing as a result of skip patterns inherent in the questionnaire. Major reasons for the skip patterns included:

- Participants who attended visit 4 and responded to the SES A questionnaire were not re-queried about their parents' education in the SES B interview.
- All participants were asked about place of residence at age 30; participants were queried selectively about place of residence at age 40 and 50, depending on their age at the ARIC baseline exam. For example, if a participant was between aged 50 and 59 at baseline examination, he/she was not queried about address at age 50.
- Skip patterns were built into the SES-B questionnaire reflecting the logic of the questions. For example, if a participant answered 'yes' to Question 1, they were directed to skip Questions 2-4 and answer Question 5; if they answered 'no' they were asked Question 2a. See **Appendix 3** for SAS code identifying skips.

Once we were able to distinguish true missings from those resulting from skip patterns, patterns of true missing data and 'unknown' values were examined. The questions with the most missing data were items related to parental / caretaker's education (Questions 2a, 2b, 4a, 4b), which were discontinued midway through the study, as described previously under 'Modifications to Protocol'. Due to this decision, of the 14% of the LCSES participants who were eligible to be asked these questions, (i.e., they had not participated the visit 4 clinic exam) 46% were missing the information.

Appendix 3 tabulates the number and proportion of missing answers to the various questionnaire items. Counts are given for true missing and unknown for questions which include a category of "I don't know" as a possible value. Only a small percentage of participants were either unable to answer the question or responded with "I don't know".

4. Preparation of Addresses for Geocoding

The process of geocoding involves assigning geographic reference coordinates to non-geographic data. Addresses which include street name and house or structure number are linked to geographic databases which contain this same information. Geocoding software attempts to match the address from each observation in the non-geographic file with a street segment in the geographic database which includes street names, street addresses ranges on each side of the street, and street latitude/longitude coordinates. Commercial address geocoding services use data derived from the U.S. Census TIGER/Line files. Because the geographic extent and identifying codes for census units (tracts, etc.) are also stored in the TIGER/Line files, the individual address information can be tagged with appropriate census codes. (Ref: 'Linking Survey Respondents' Residential Addresses to Contextual Data: Address Geocoding Issues, Udry, J.R., Carolina Population Center, 1995). Because missing street name, multiple and/or local

street names, and inaccurate address information can substantially reduce the success of the geocode matching, it is essential to edit residential address for obvious problems before submitting the addresses to a commercial geocoding service.

4.1. ARIC Visit 1-4 Addresses

In addition to reviewing and correcting the self-reported city, county and state data for Visits 1-4, we reviewed street addresses to see if elements necessary for geocoding were present. If they were deemed adequate, the address was sent to the commercial geocoder. Major reasons that addresses were determined to be not adequate include:

- State missing (most often military addresses)
- All address fields missing
- C/O (in care of)
- Temporary address
- Address not in the U.S
- Apartment name without address or where no address could be located
- Only address is PO Box

Table 1 summarizes characteristics of the addresses by whether or not they were sent to the geocoder .

Table 1: Visit 1 – Visit 4 Addresses by Type of Information Provided

	Visit 1 (n=15792)	Visit 2 (n=14348)	Visit 3 (n=12885)	Visit 4 (n=11656)
SENT (N)		(14327)	(13887)	
Apparently complete address	14260	13769	12280	11166
Partialaddress	25	34	29	59
Zip code only	3	3		4
Street name only	125	18	32	7
PO box & address	0	0	72	0
Rural Route number	1052	209	101	74
TOTAL	15465	14033	12514	11310
NOT SENT (N)				
Address not adequate	13	16	10	7
PO only-no street address	314	278	363	339
TOTAL	327	294	373	346

4.2. Historical Addresses

The SES-B questionnaire queried participants’ address at various life epochs: childhood and ages 30, 40, and 50 years. As a first step in preparing these addresses for further processing, the spellings of all cities, counties, and states were checked and, when possible, corrected. States were corrected for misspellings and standardized to the two digit abbreviations used by the US Postal Service. Within each state, the cities and counties were hand-checked against look-up tables available through the internet (<http://www.accesschecks.com/xref.htm>). When a city was

provided but no county listed we used either Map Quest (www.mapquest.com) or the National Association of Counties (www.naco.org) to identify the corresponding county, if extant. For the major study states (Maryland, Mississippi, Minnesota and North Carolina), we also verified that changes in counties had not occurred during the years of interest. When mistakes were found, they were corrected and ‘flag variables’ were created to track the number and types of errors. Flag variables were created for each set of address questions (childhood, age 30, age 40, and age 50). A value was added to a ‘city flag’ variable whenever there was an error associated with the city field of the address. Similarly, ‘county flags’ and ‘state flags’ were also created to track similar errors. The codebook and explanation for these variables are listed in the Address Variable Codebook (**Appendix 4**). If no errors were detected in the participants’ response to the address questions, the value for the flag variables was ‘OK’.

4.2.1. Childhood Address Correction Summary

Fewer than 15% of the childhood addresses required any correction for the city, county, or state. Of the errors identified, more than 70% were misspellings. Misspellings and incorrect counties accounted for over 90% of the errors for childhood counties. Finally, nearly 100% of the corrections to childhood state were misspellings or simple substitutions of the US Postal Code state abbreviation. More detailed information can be found in the Childhood Address Correction Results document (**Appendix 5**).

4.2.2. Age 30 Addresses

Based on their age of entry into the ARIC study, all 12,724 LCSES participants were asked their residence at age 30 years. The Age 30 Address Correction Results (**Appendix 6**) tabulates the proportion and type of errors in the cities (fgcity33), counties (fgcnty33), and states (fgst33) reported by ARIC participants for their residence at age 30 years. Approximately 80% of the reported cities, 85% of the reported counties, and 75% of the reported states appeared correct.

4.2.3. Age 40 Addresses

Based on their age of entry into the ARIC study, 8,947 participants – those age 50 and older - were asked their residence at age 40 years. The Age 40 Address Correction Results document (**Appendix 7**) lists the proportion and type of errors in the cities (fgcity29), counties (fgcnty29), and states (fgst29) reported by ARIC participants for their residence at age 40 years. Approximately 81% of the reported cities, 92% of the counties, and 78% of the states did not require correction.

4.2.4. Age 50 Addresses (sesb25a-sesb25e)

If age of entry into the ARIC study was 60 or older participants were asked their residence at age 50 years (n=2483). The Age 50 Address Correction Results (**Appendix 8**) document lists the proportion and type of errors in the cities (fgcity25), counties (fgcnty28), and states (fgst25) reported by ARIC participants for their residence at age 50 years. Approximately 76% of the reported cities, 95% of the counties, and 80% of the states did not require correction.

5. Geocoding

This section describes the procedures used to assign appropriate Census tracts contemporary (V1-4) and historical addresses and also reports on success rates and procedures used to increase match rates when commercial geocoding was not successful.

5.1. Previous Geocoding Efforts in ARIC Study

Visit 1 addresses, collected in 1987-89, were originally geocoded by a commercial geocoder (vendor unknown) to 1990 census boundaries, under the direction of Ana Diez Roux as part of her dissertation work. When the geocode was successful, an address was associated with a Federal Information Processing Standards (FIPs) code, which includes all census boundary information for that location (state, county, tract, block group). An unexpectedly high proportion of Washington County, MD. addresses did not geocode, apparently due to a change in addresses in the early 1990's as part of the state of Maryland's effort to change to a grid address system to improve emergency response. In order to minimize missing data for the participants with these addresses changes, Dr. Diez-Roux solicited the help of Washington County ARIC field center personnel who were familiar with the area. With the use of a street map and 1990 census tract boundaries, they attempted to place the affected addresses into their correct geographic location so that tract numbers could be manually assigned.

Visit 3 addresses were originally geocoded in 1990 by a commercial company, GDT, to the 1990 census boundaries as part of an ARIC ancillary study.

5.2. Geocoding Vendor and Match Codes

For our work in the LCSES study we chose the firm Mapping Analytics (MA). This decision was based on the recommendation of Dr. Nancy Krieger, who found this vendor to be most accurate of four vendors evaluated as part of a geocoding accuracy study (Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. 'On the wrong side of the tracts? Evaluating accuracy of geocoding in public health research', *AJPH*: 2001, 91, 1114-1116). Additional efforts aimed at evaluating the repeatability and accuracy of the geocodes provided by MA over the course of our study will be described later in this section of the manual.

All addresses sent to MA were returned with geographic coordinates (latitude and longitude), geographic identifiers (FIPS code) and a match code indicating the level of precision of each match. Match codes range from A (exact address) to E (no match). For our purposes, we decided to accept only addresses which matched at least at the level of the census tract (as this will be the smallest geographical unit at which we will aggregate census data). This includes selected Z (Centroid) code matching with tract level accuracy. Table 2 illustrates the codes we considered acceptable; a full list of codes can be found in **Appendix (9)**.

Table 2: Definition of MAPPING ANALYTICS Codes

A codes – Address match

ASn Indicates a house range address geocode. This is the most accurate geocode available. The digit at the end indicates the following:

- 0: Best location
 - 1: Street side is unknown. The Census FIPS Block ID from the left side is assigned, however no offset is assigned, the point is placed directly on the intersection.
 - 2: Address was interpolated on to a TIGER segment that did not contain address ranges initially.
 - 3: Both 1 and 2.
- AX3 Indicates an intersection geocode. The street side can not be determined, and no address ranges can be assigned, hence the "i3" is returned to be consistent with AS codes.

Z codes – ZIP+4 Centroid Match Code Descriptions

2nd Character (Census ID Accuracy):

- B Indicates Block Group accuracy (most accurate).
- T Indicates Census Tract accuracy.
- C Indicates unclassified Census accuracy. Normally accurate to at least the County level.

Detail description of ZC codes

Acceptable

ZC7N – Location derived from a ZIP Code Centroid – unclassified Census accuracy – normally accurate to the county level – indicates a location based upon a ZIP+2 centroid where the street segment is matched and the ZIP+2 cluster centroid is the same for the entire street. **The BG, tract and County are all the same for this street.** Location is a ZIP+2 centroid.

Probably good

ZC5W – Location derived from a ZIP Code Centroid – unclassified Census accuracy – normally accurate to the county level – based on Post Office location (that delivers the mail to this address) where MORE than 80% of addresses in this ZIP Code are in a single census tract. **REASONABLE Census ID accuracy.** Location is assigned to the ZIP code centroid.

5.3. Geocoding of Contemporary (V1-V4) Addresses

Databases and programs used to generate geocodes vary and databases used by the same vendor are regularly updated. Therefore, we decided that we would resubmit Visit 1 and Visit 3 address with Visit 2 and Visit 4 address so that all geocodes assigned would be obtained from a common source. The match rate for Washington County for Visit 1 was lower than that of other sites, as expected, due to the changes in addresses that occurred in the early 1990s. The address changes were related to a change in the grid address system by the state of Maryland made to improve 911 emergency responses by the ambulance, fire, and police department. Table 3 presents data on the yield of V1 to V4 addresses from the 1st submission to MA.

Table 3: Results from Submission of V1-V4 ARIC Addresses to Mapping Analytics

CODE from MA	visit 1	visit 2	visit 3	visit 4	Total
Nothing	19	17	1	13	50
A/AS	0	8	17	10	35
AS0	12826	13473	11665	10446	48410
AS1	174	161	138	131	604
AS2	42	48	48	44	182
AS3	2	0	0	0	2
To address: total	13044	13690	11868	10631	49233
To Block: total	75	295	323	379	1072
To Tract: total	117	92	48	41	298
ZC5W	227	32	20	18	297
ZC7N	1	4	1	0	6
ZC5X,5Y,5Z,9D,9G,9H (Not acceptable)	1982	495	253	228	2958
To County: total	2210	531	274	246	3261
Total	15465	14625	12514	11310	53914

Major reasons why addresses did not lead to a match include:

- Streets were renumbered between the time the participant provided the address and the time of geocoding
- Streets were renamed
- Route/street was moved or eliminated (e.g., US 29 is moved to a different route through town)
- The mailing address provided was not necessarily the place of residence (e.g., PO Box)

5.3.1. Efforts to Increase Match Rates in Current Addresses

Participants with Washington County visit 1 addresses which did not geocode (n=998) were compared with the file originally prepared by A. Diez Roux. . In cases where a geocode was assigned in her file (presumably due to her supplemental manual geocoding, using maps and the help of a Washington County field center staff person), the geocodes assigned were used to replace the missing values in the more current data set. In the file (lcdsc05a) containing the neighborhood SES variables, these participants were assigned the code 'A' for the variable TRTCODED. We also compared visit 1 and 2 addresses for these participants and found that approximately 80% have the same street but different number in visit 1 and visit 2.

During a review of these manually assigned address we found that 60 visit 1 Washington County participants with manually assigned a tracts (trtcoded=A) had been placed in a tract in which the only building seemed to be a prison. Mapping Analytics did not assign any of these participants' visit 2 addresses this tract; most were placed in 2 tracts adjoining the visit 1 "prison" tract. We randomly selected 10 of these participants and asked Pat Crowley, Project Coordinator in the Hagerstown ARIC field center to check reported address changes and moves using non-electronic files kept for this purpose.. She

determined that although the addresses had “changed” the participants had not moved.

We tabulated the number of reported addresses changes in Washington County between subsequent ARIC exams. Changes between visits 1 and 2 were substantially greater at 19% than between visits 2 and 3 (7%) and visits 3 and 4 (7%). This difference and the questionable assignment of V1 addresses to the “prison” tract has raised some concern about the accuracy of the manual geocoding. In order to investigate this problem, we identified tracts where clusters of 5 or more address changes occurred between V1 and V2. We grouped these clusters by visit 1 tract and used ARCVIEW to overlay the corresponding visit 2 addresses on a tract map of Washington County. Using this empiric approach, we identified 274 people who appeared to move from a visit 1 tract to an adjacent visit 2 tract, usually mapping to locations on or near the shared tract boundary. We then enlisted the help of the Washington County ARIC field center staff who reviewed written records and found that 255 of the 274 people had not actually moved, although their address (eg street name or number) had changed. As a consequence of this review, we changed the visit 1 tracts (which had been manually assigned) to the tract associated with the visit 2 address. These people are identified in the data set with a `trtcoded='c'`.

Other addresses in the data set that did not geocode were reviewed for obvious errors that might have resulted in a failure to geocode. Examples of common mistakes include spelling errors and a missing space between two words (MainST instead of Main ST). The obvious errors were corrected and these addresses were re-submitted to MA.

5.4. Geocoding of Historical Addresses

Approximately 24,000 addresses representing places of residence at ages 30, 40 and 50 were edited and sent to Mapping Analytics for geocoding. Table 4 presents data on the yield.

Table 3: Geocoding yield of final submission of historical addresses to MA

Addresses expected		24225
Sent for Geocoding		22176 (91.54%)
Returned with:	A code	16471 (74.27%)
	Usable Z code	145
	E or non usable Z code	5560
Addresses without usable geocoding:	address with street number	1693
	cross streets	925
	street name (or institution name)	2942

Major reasons for lack of geocodes include:

- Participant could not recall the complete address
- Streets were renamed or addresses numbering changed across time. (This problem was most common in Washington County. No efforts have been made to date to work with these addresses).
- Street or route moved or eliminated
- Zip codes not usable
- Street information misspelled or inaccurate (road instead of street, etc)

5.4.1. Assignment of Historic Census Tracts

Given that census tract boundaries change across time, the 1990 census tract assignment provided by the geocoder could not be used to assign tracts to historic addresses. In order to abstract context-appropriate census information, we followed the procedure below:

- We first determined the census year (1960, 1970, 1980) that corresponded most closely to the year in which a participant was age 30, 40 or 50 (Table 5).

Table 5: Assignment of Historical Census to Participants by Birth Cohort and Age Closest to Census

Birth cohort	Age 30	Age 40	Age 50
'22-'25 1469	'52-'55 (1960 C**)	'62-'65 (1960 C)	'72-'75 (1970 C)
'26-'30 3349	'55-'60 (1960 C)	'65-'70 (1970 C)	'75-'80 (1980 C)
'31-'35 3708	'60-'65 (1960 C)	'70-'75 (1970 C)	'80-'85 (1980 C)
'36-'40 3978	'65-'70 (1970 C)	'75-'80 (1980 C)	'85-'90 (1990 C)
'41-'44 2230	'70-'74 (1970 C)	'80-'84 (1980 C)	'90-'94 (1990 C)

**Note: 1469 people turned 30 before 1956 and should have used 1950 census data: however, 1950 data is not available by tract so 1960 tract data is used.

- We next determined the appropriate method of assigning historic addresses, depending on this Census year. There are two methods available:
 - Overlay method – the longitude and latitude points of an address are overlaid onto digitized files of tract lines. The advantage of this method is that any address that is geocoded (has a lat/long) can be placed in a tract.
 - Comparability file method – The US Census Bureau maintains files which indicate how tract boundaries change from one census to the next. These are called comparability files. The advantage of this method is that accuracy is as good as the comparability files (we

found only one error). The disadvantage is that when a more recent tract is made up of 2 or more tracts, or parts of 2 or more tracts from the previous decade, it is not possible to determine the correct historical tract placement. A check of our data shows that as many as 23% of addresses could not be placed in tracts when attempting to use take 1990 tracts back to 1970.

As a test, we compared the 1970 tract assigned by each method, using 13,044 addresses that were originally assigned latitude and longitude and 1990 tracts. Of these, 36% could not be assigned a 1970 tract using the comparability files because of identifiable tract splits between the periods. Of the remaining addresses (n=8348), 97% were assigned identical tracts by both methods.

- Latitudes and longitudes of the addresses that were successfully matched were linked with the appropriate census tract by historic census year. This was done by using the 1970 and 1980 digitized census tract maps produced on CD ROM by Geolytics, Inc. These map files, which include both map boundaries at the several geographic levels and census data from population and housing summary files, were available through Davis Library UNC Chapel Hill. See **Appendix 12** for a description of the process of overlaying address locations on tract maps.
- These electronic polygon files were not available for 1960 so comparability file method was used. 1960 addresses were assigned 1970 tracts as described above then placed in the 1960 tracts using information on the correspondence between 1970 and 1960 tracts.

Problems:

1. If the 1970 tract was made up of 2 or more 1960 tracts a person could not be placed in a 1960 tract without going back to actual addresses and attempting to geocode 'by hand'. Hand geocoding will be described in more detail below.
2. Jackson, MS & Washington Co. MD were not tracked in 1960. Some socioeconomic housing data was available in print volumes for Jackson and Hagerstown at the level of city block groups. Data from these areas were aggregated at the level of 1970 tract boundaries and used.
3. For other nontraced areas in 1960 we used the 1970 census data as the next best thing. A variable, 'gapyr', gives the gap between the year of the census data and the year the participant was a specified age. Normally gapyr should have a value less than 5. There were 1,469 participants for whom the 1950 census would have been more

appropriate. This census information was not available, so the 1960 census was substituted, making gapyr as large as 8 years. See **Table 5** above. In those instances when 1970 census data must be substituted for 1960 data, the gap can be as great as 15 years. In one instance, a person was age 30 in 1952 and lived in a non-tract area in 1960; therefore, the 1970 census was the only available source of data and the gapyr value is 18.

4. In order to use the overlay method described above, it was necessary to have longitude and latitude coordinates for an address obtained from an exact address (match code starts with A). If the address is one of those with a match code starting with Z, coded to tract level, it is necessary to use the comparability file method to find the tract for the historic year.
5. When the historic addresses were ready for geocoding, Mapping Analytics had already begun using the 2000 Census sources. We requested that they geocode to 1990 tracts and they agreed. Approximately 2 years after the coding was completed, we noticed an anomaly which led us to question the coding: for several streets which were identified tract boundaries, addresses with both odd and even dwelling numbers had been assigned to the same tract. Discussion with MA disclosed that they had used an overlay method to assign 1990 tracts to historic addresses previously coded to the 2000 census. We also discovered that MA did not process the files inhouse; all major geocoders sent addresses to the same company for batch geocoding.

5.4.2. Hand Geocoding

There were several situations in which we attempted to hand geocode participant addresses. This was most commonly done when the participant provided partial address information (street name but not number, cross streets, etc) and was also used for 1960 addresses when a 1970 tract mapped to 2 or more 1960 tracts. Detailed street maps of the four main study areas were obtained and census tract boundaries from the three historical censuses (1960, 1970 and 1980) were drawn by hand on these maps. Using standardized procedures (**Appendix 10**) we attempted to locate each address on the map. If the street could be located and was contained within the boundary of a tract, it was assigned the appropriate tract number. If a street was located but in multiple tracts or served as the boundary for two or more tracts, a census tract was not assigned.

A large number of Washington County, Md. historical addresses were obsolete, because a major renumbering/naming of streets occurred in the early 1990's. We obtained detailed historical street maps from the Hagerstown Public Library and

attempted to locate the original street names. We then attempted to hand geocode those found using the methods described above.

Table 6 summarizes our success rate at assigning geocodes and ultimately assigning a tract from the 1960-1980 censuses. After including those addresses that were geocoded by hand, 50% of the addresses for 1960, 76% of those for 1970 and 85% of the 1980 addresses were assigned a tract. The yield for 1960 increases considerably (66%) when the 1970 tract is substituted for missing information.

Table 6: Geocoding and Tract Assignment Success Rate

Census Year	1960 N=8223	1970 N=12706	1980 N=3283
Percentage of addresses sent to geocoder	89	93	94
Percentage successfully coded to 1990 tract	61	71	81
Percentage with good tract in historic year	42	69	80
Adjustments			
Geocoded by Hand	8	7	5
Substituted 1970 tract	16		
Percentage total yield	66	76	85

5.5. Quality Assessment of Geocoding

Assessment of the reliability of commercial geocoding included three investigations: a comparison of the geocodes from two different commercial vendors; a comparison of assigned codes from a single vendor from submissions at two different time periods; and, a review of the process of assigning geocodes to historical addresses. Each of these investigations is reviewed below.

5.5.1. Assessment of Reliability of Commercial Geocoding

We compared the geocodes obtained for the 12,289 V3 addresses sent to two different vendors (GDT and MA) approximately five years apart. Of these, 500, or approximately four percent were not assigned the same geocodes by the two companies.

In a subset of 84 of these addresses from Winston Salem, we tried to determine which geocode was correct. First, we manually geocoded these addresses. We found that:

- 32 (38%) were in tracts assigned by MA
- 5 (6%) were in tracts assigned by GDT
- 47 (56%) were on roads that divided the two tracts.

We followed up on the 47 IDs that were on roads that served as the border of tracts by using Yahoo files & Tiger files so that we could attempt to determine on which side of the road the addresses were located. On roads that fall on tract borders, side of the road is often a determinant of tract assignment. According to an official of the US Bureau of the Census, the center of the road is considered the tract boundary; houses on one side of the street are assigned one tract and those on the other side are assigned an adjacent tract. Typically, this corresponds to even numbered addresses being assigned one tract while odd-numbered addresses are assigned the other tract. Our results are as follows:

- 36 (77%) fell in the tract assigned by MA
- 3 (6%) fell in the tract assigned by GDT
- 7 (15%) could not be located

Our results on albeit a limited sample, favor the geocodes provided by MA. However, given that the Tiger files used by the commercial geocoding companies are regularly updated it may be that the more favorable results seen for MA are related to the files being more accurate at a later date and not the accuracy of the vendor per se.

5.5.2. Assessing Short-term Repeatability of Geocodes (Discuss with JOY. See Kathy note)

A set of 481 “complicated” V3 addresses originally sent to MA were resubmitted a few months later. The results are summarized in the table below:

Table 7: Comparison of longitudes & latitudes Assigned to 481 Addresses on two Submissions to MA, Two Months Apart

Code	Same place	Geocoded to address	Located near	Located >.25 mile	Frequency/%
Not assigned on either submission					5 (1%)
1 st submission only					110* (23%)
Same code on both submissions	Yes				366 (76%)
Different code on two submissions					9 (2%)
9 that came back different	No	No			1
		Yes	Yes	No	6
			No	Yes	2

* The resubmission took place after MA had begun using 2000 files. We believe that these 110 were not geocoded because they were using an overlay method to assign 1990 tracts. These 110 addresses had originally been geocoded with z-code (zip + centroid) level accuracy which would not be specific enough for an overlay method.

5.5.3. Repeatability of Geocodes Assigned to Historical Addresses

We resubmitted a subset of 10362 historical addresses to MA 8 months after the original submission to the same geocoder. The results on repeatability across submissions is summarized in the table below:

Table 8: Correspondence between FIPs code

Match state	Match county	Match tract	Frequency	Percent
no	No	no	1*	0.01
yes	No	no	3	0.03
yes	Yes	no	205	1.99
yes	Yes	yes	10110	97.97

* address was on the state line

5.5.4. Assessment of Repeatability and Accuracy of Geocodes

A more formal assesment of repeatability and accuracy was conducted by Dr. Eric Whitsel. In addition to examining the repeatability of coordinates and tracts assigned by Mapping Analytics (MA), accuracy was assessed by comparing spacial coordinates associated with ‘gold standard’ addresses of air pollution monitors in study sites to those obtained by MA. The 9-month repeatability of geocodes assigned by MA to 1,032 participant addresses was uniformly high. Match rates for addresses of EPA monitors were lower for MA versus a 2nd commercial geocoder (Geocoder B) (76% vs. 88%). In contrast, discordance at the block group, tract and county level was greater for Geocoder B vs. MA. Coordinates assigned by Geocoder B vs. A also were further from those in the EPA database. A manuscript based on this work has been submitted for publication. The complete abstract can be found in **Appendix 11**.

6. Census Based Contextual Data

Socioeconomic census data for 1970 and 1980 was obtained from Geolytics at the same time as the polygon tract files were obtained. A description of this process can be found in **Appendix 12 (Marilyn to write)**. The 1960 census data was not available through Geolytics, and was obtained from electronic files available through the Odum Institute at UNC. Much of the country was not tracted in 1960, but there are print files of socioeconomic housing data of towns and cities with a population greater then (10,000) by city block. This data was keyed and compiled by tract as described above (#2) for Jackson and Hagerstown. See **Appendix 13** for specific information about the sources of Census data.

7. Data sets

7.1. Individual SES Lifecourse Data Set

Latest version: lcdsc05a

Date: 3/2003

A dataset of containing information from the ARIC visits

- Data set description: One record per person with variables selected from the following sources:
 - Visit 1-4 data files
 - Annual Followup (AFU):
 - marital status, when changed - variable 40a & 40 b
 - health compared to others - variable 6
 - SES AFU – All the questions from the SES AFU are included except the questions about residence that were used to locate census tracts lived in at ages 30, 40, 50 and county at the age of 10 years

7.2. Neighborhood SES Data Sets Information

Latest version: Lcdsc35a

Date: 11/2003

A dataset of historic neighborhood census data for ages 30, 40 and 50 was combined with a dataset of contemporary neighborhood data to create one of continuous age from 30-70.

- Data set description: One record per person with census tract information for participants address when they were 30-70 years old with up to 5 age points (age decades 30, 40, 50 ,60 70) over those years. Census range is 1960-2000.
- Source of Address Data: Age 30 address was asked of everyone at the SES AFU, age 40 & 50 were asked of some depending on their age at the interview. The tract information for addresses of ages not asked at the interview was filled in using the persons Visit 1 – visit 4 tract information. For some of the older participants we were able to fill in age 60 & 70 tract information also.
- Address data year gaps: The ideal would be to have address of residence at each age decade, and it fall on the calendar decade year.
 - Problem 1(addresses from visits only): the date of the visits being used to fill in addresses did not necessarily fall on a participants age decade, the visit with the age closest to the age decade being filled in was used & an age variable included with the actual age for the age decade be represented.

- Problem 2(all): the year of the address* did not usually fall on the census year so there is a variable (gapyr) that is the number of years between the date of the address and the census year.

*year of address from AFU form was year participant was 30, 40, 50, for addresses being filled in it was year of visit being used

- Naming Convention & description Because participants in the LCSES study were born in different birth cohorts, the census most appropriate to a given age will not be uniform across the participants. Table 9 describes the distribution of census by age epoch.

Table 9: Census Closest to Year of Age

Census	Age 30	Age 40	Age 50	Total
1960	7,123	1,122		8,245
1970	5,629	5,996	1,117	12,742
1980		1,903	1,390	3,293
Total	12,752	9,021	2,507	

The Neighborhood SES data has 23 variables (see **Table 10**) repeated for each of the 5 age categories. Because of the variation in source of characteristics by age, it was necessary to devise a meaningful convention for naming variables. The variable names of information for age 30, 40, 50, 60, 70 are prefaced with a3, a4, a5, a6 and a7 respectively, eg. a3unempl, a4unempl, and a7unempl are ‘Percentage unemployed’ at age 30, 40 & 70. Also each of the 5 points also has a variable CENSUS named a3census, a4census, etc. and gives the census year of the data used for that age. For the childhood data the variables are prefaced with a1. Childhood data is county based data and participants were asked ‘where did you mostly live as a child’. The census closest to their age 10 was used. Census range is 1930-1950. Adult age points with data taken from visit 1 to 4 data also have a nonmissing age (a4age, a5age, ..).

- Quality of data? It would be expected that gapyr would be 5 or less, ie address is within 5 years of the census data being used. However, in order to reduce the amount of missing data we have assigned census data from 1970 tracts to people in areas that in 1960 had no available census data. Because we were using 1960 census data (the first year with data compiled by tracts) for people turning 30 in the early 50s, it is possible to have gapyr as large as 18 years. There is one person who turned 30 in 1952 and lived in an untraced area that has 1970 data for age 30 (a3gapyr=18), age 40 (a4gapyr=8) and age 50 (a5gapyr=2). We have left it to the investigator to decide what data to use. Usually agegap would be only a few years, that is the address is within a few years of the age being represented.

Table 10: Data available for adult neighborhood SES

			1960	'60*	'70*	'80*	'90*	2000
			tracts					#
Income	Median household income	incmdh				X	X	X
	Mean household income	incmnh				X	X	X
	Median family income	incmdf				X	X	X
	Mean family income	incmnf		X	X	X	X	X
	% Individuals below poverty	povind			X	X	X	X
	% Families below poverty	povfam						X
	% of households with income interest, dividends, rent % housing units that are owner occupied	incoth owno	X	X	X	X	X	X
Education	% adults 25+ yrs who have high school education	edhs		X	X	X	X	X
	% adults 25 + with college degree	edcol		X	X	X	X	X
Occupation	% ages 16+ in professional, managerial, and executive	profe		X	X	X	X	X
	Percentage unemployed	unempl		X	X	X	X	X
housing	% housing units occupied	unito		X	X	X	X	X
	% housing units dilapidated/deteriorating	delat/deter	X					
	% occupied housing with > one person per room	pproom	X	X	X	X	X	X
	% H units - Living in same house for last five years	Hu5yr		X	X	X	X	X
	Number of individuals living in household	nhhold	X	X	X	X	X	X
	Median value of owner occupied house	Hmdval					X	X
	Mean value of owner occupied house	Hmval	X	X	X	X	X	X
other	% households headed by a female (<i>with minor children</i>)	shhfem			X	X	X	X
	% households headed by a male (<i>with minor children</i>)	shhmal				X		
	% single parent homes (<i>men or women</i>)	snglhh					X	X
	% people in Urban area	urban		x		X	X	X

* US Census tract level data, 1960

** censusCD70 - Geolytics

***censusCD80 – Geolytics

**** Census of population and housing, 1990 summary tape file 3 (STF3)

Census of population and housing, 2000 summary tract files (ICPSR)

7.3. Childhood Neighborhood SES Information

Latest version: lcdsc53a

Date: 11/2003

Neighborhood characteristic data for early childhood.

- Data set description: There is one record per person containing census information based on reported address during childhood (linked to census closest to when the participant was age ten). All variables are county level (rather than tract level) and include: average sales for retail stores, average wages for worker in manufacturing industry, and predicted income using these 2 variables