2000 National Household Survey on Drug Abuse

Sample Design Report

Contract No. 283-98-9008
RTI Project No. 7190
Phase II, Deliverable No. 10

Authors:

Katherine R. Bowman
James R. Chromy
Dawn M. Odom
Michael A. Penne

Project Director: Thomas G. Virag

Prepared For:

Substance Abuse and Mental Health Services Administration
Rockville, Maryland 20857

Prepared by:
Research Triangle Institute

May 2002

2000 National Household Survey on Drug Abuse

Sample Design Report

Authors:
Katherine R. Bowman
James R. Chromy
Dawn M. Odom
Michael A. Penne

Project Director:
Thomas G. Virag

Prepared for

Substance Abuse and Mental Health Services Administration
Contract No. 283-98-9008
Deliverable No. 10

May 2002

Acknowledgments

This publication was developed for the Substance Abuse and Mental Health Services Administration (SAMHSA), Office of Applied Studies (OAS), by Research Triangle Institute (RTI), Research Triangle Park, North Carolina, under Contract No. 283-98-9008. Significant contributors at RTI include Katherine R. Bowman, James R. Chromy, Dawn M. Odom, Michael A. Penne, R. David Belton, Brenda K. Porter, and Thomas G. Virag (Project Director).

Table of Contents

1. Overview 1

1.1 Target Population 1
1.2 Design Overview 1
1.3 5-Year Design 1
1.4 Stratification and First-Stage Sample Selection 2
1.5 Dwelling Units and Persons 3

2. The Coordinated 5-Year Sample 5

2.1 Formation of and Objectives for Using the Composite Size Measures 5
2.2 Stratification 6
2.3 First-Stage Sample Selection 9
2.4 Survey Year and Quarter Assignment 9
2.5 Creation of Variance Estimation Strata 11

3. General Sample Allocation Procedures for the Main Study 13

3.1 Notation 13
3.2 Determining Person Sample Sizes by State and Age Group 16
3.3 Second-Stage Sample Allocation for Each Segment 20
3.3.1 Dwelling Unit Frame Construction—Counting and Listing 20
3.3.2 Determining Dwelling Unit Sample Size 20
3.4 Determining Third-Stage Sample (Person) Selection Probabilities for Each Segment 22
3.5 Sample Size Constraints: Guaranteeing Sufficient Sample for Additional Studies and Reducing Field Interviewer Burden 22
3.6 Dwelling Unit Selection and Release Partitioning 23
3.7 Half-Open Interval Rule and Procedure for Adding Dwelling Units 24
3.8 Quarter-by-Quarter Deviations 24
3.9 Sample Weighting Procedures 29

4. References 38

Appendix A 1999-2003 NHSDA Field Interviewer Regions

Appendix B 2000 NHSDA Procedure for Adding Missed Dwelling Units

List of Tables

2.1 Number of Segments on Sampling Frame by State 7
2.2 Survey Year and Quarter Assignment Order for 96 Segments within Each FI Region 10
2.3 Segment Identification Number Suffixes for the 1999-2003 NHSDAs 12

3.1 Expected Relative Standard Errors by Race/Ethnicity and Age Group: Main Sample 19
3.2 Expected Main Study Sample Sizes by State and Age Group 21
3.3 Quarterly Sample Sizes and Percent Released 27
3.4 Definitions of Levels for Potential Variables for Dwelling Unit Non-Response Adjustment 32
3.5 Definitions of Levels for Potential Variables for Selected Person Post-Stratification and Non-Response Adjustment 33
3.6 Definitions of Levels for Potential Variables for Dwelling Unit Post-Stratification and Respondent Post-Stratification at the Person Level 35
3.7 Model Group Definitions 36
3.8 Flowchart of Sample Weighting Steps 36
3.9 Sample Weight Components 37

Chapter 1: Overview

1.1 Target Population

The respondent universe for the 2000 National Household Survey on Drug Abuse (NHSDA) was the civilian, noninstitutionalized population aged 12 years or older residing within the United States and the District of Columbia. Consistent with the NHSDA designs since 1991, the 2000 NHSDA universe included residents of noninstitutional group quarters (e.g., shelters, rooming houses, dormitories, and group homes), residents of Alaska and Hawaii, and civilians residing on military bases. Survey coverage before the 1991 NHSDA was limited to residents of the coterminous 48 states and it excluded residents of group quarters and all persons (including civilians) living on military bases. Persons excluded from the 2000 universe included those with no fixed household address (e.g., homeless transients not in shelters) and residents of institutional group quarters, such as jails and hospitals.

1.2 Design Overview

The Substance Abuse and Mental Health Administration (SAMHSA) implemented major changes in the way the NHSDA would be conducted beginning in 1999 and continuing through subsequent years. The 1999 survey was the first conducted using computer-assisted interviewing (CAI) methods. This survey also marked the first year in a transition to improved state estimates based on minimum sample sizes per state. In addition, it was also the first year in which cigarette brand information was obtained for the Centers for Disease Control and Prevention (CDC). To obtain the required precision at the state level and to improve the precision of cigarette brand data for youth at the national level, the total sample size of 67,500 was increased by 2,500 youths aged 12 to 17 to a total of 70,000. This large sample size allowed SAMHSA to continue reporting precise demographic subgroups at the national level without needing to oversample specially targeted demographics, as required in the past. This large sample is referred to as the "main sample" or the "CAI sample." The achieved sample for the 2000 CAI sample was 71,764 persons.

Beginning in 2000, SAMHSA and the University of Delaware conducted the Validity of Self-Reported Drug Use in Population Surveys (VSRDU) to evaluate and establish baseline information on the validity of survey research methods in assessing recent drug use among the general household population. The existing validity research had been conducted on only specific subgroups, resulting in the need to examine the validity on the overall population. In order to obtain the required precision, the sample size consisted of 1,000 per age group (12 to 17 and 18 to 25) per year (2000 and 2001). Although different surveys, the VSRDU was conducted alongside the NHSDA and used the same interviewer staff. The design of the VSRDU will not be covered in this report, but full documentation will be published elsewhere.

1.3 5-Year Design

A coordinated 5-year sample design was developed. The 2000 main sample is a subsample of the 5-year sample. Although there is no overlap with the 1998 sample, a coordinated design for 1999-2003 facilitated 50% overlap in first-stage units (area segments) between each two successive years from 1999 through 2003. This design was intended to increase the precision of estimates in year-to-year trend analyses because of the expected positive correlation resulting from the overlapping sample between successive NHSDA years.

The 1999-2003 design provides for estimates by state in all 50 states plus the District of Columbia. States may therefore be viewed as the first level of stratification as well as a reporting variable. Eight states,

referred to as the "large" states,[1] had a sample designed to yield 3,600 to 4,630 respondents per state for the 2000 survey. This sample size was considered adequate to support direct state estimates. The remaining 43 states[2] had a sample designed to yield 900 to 1,030 respondents per state in the 2000 survey. In these 43 states, adequate data were available to support reliable state estimates based on small area estimation methodology. The youth supplement was allocated to the larger population states to increase precision of smoking-related estimates for youth at the national level.

1.4 Stratification and First-Stage Sample Selection

Within each state, field interviewer (FI) regions were formed. Based on a composite size measure, states were geographically partitioned into roughly equal size regions according to population. In other words, regions were formed such that each area yielded, in expectation, roughly the same number of interviews during each data collection period, thus distributing the workload equally among NHSDA interviewers. The smaller states were partitioned into 12 FI regions, whereas the eight "large" states were divided into 48 regions. Therefore, the partitioning of the United States resulted in the formation of a total of 900 FI regions. FI region maps can be found in Appendix A.

For the first stage of sampling, each of the FI regions was partitioned into noncompact clusters[3] of dwelling units by aggregating adjacent Census blocks. Consistent with the terminology used in previous NHSDAs, these geographic clusters of blocks are referred to as *segments*. A sample *dwelling unit* in the NHSDA refers to either a housing unit or a group-quarters listing unit such as a dormitory room or a shelter bed. To support the overlapping sample design and any special supplemental samples or field tests that SAMHSA may wish to conduct, segments were formed to contain a minimum of 175 dwelling units[4] on average. In prior years, this average minimum segment dwelling unit size was only 90.

Before selecting sample segments, additional implicit stratification was achieved by sorting the first-stage sampling units by an MSA/SES (metropolitan statistical area/socioeconomic status) indicator[5] and by the percent of the population that is non-Hispanic and white. From this well-ordered sample frame, 96[6] segments per FI region were selected with probabilities proportionate to a composite size measure and with

---

[1]For the 1999-2003 NHSDAs, the "large" states are California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas.

[2]For reporting and stratification purposes, the District of Columbia is treated the same as a state and no distinction is made in the discussion.

[3]Noncompact clusters (selection from a list) differ from compact clusters in that not all units within the cluster are included in the sample. While compact cluster designs are less costly and more stable, a noncompact cluster design was used because it provides for greater heterogeneity of dwellings within the sample. Also, social interaction (contagion) among neighboring dwellings is sometimes introduced with compact clusters (Kish, 1965).

[4]Dwelling unit counts were obtained from the 1990 Decennial Census data supplemented with revised population counts from Claritas.

[5]Four categories are defined as: (1) MSA/low SES, (2) MSA/high SES, (3) Non-MSA/low SES, and (4) Non-MSA/high SES. In order to define SES, block group-level median rents and property values were given a rank (1...5) based on state and MSA quintiles. The rent and value ranks were then averaged, weighting by the percent renter and owner occupied dwelling units, respectively. If the resulting score fell in the lower 25th percentile by state and MSA, the area was considered “low SES”; otherwise, it was considered “high SES.”

[6]The 1999-2003 sample was planned such that 48 segments per FI region would be selected. In the implementation, however, an additional 48 segments were added to support any supplemental or field test samples.

minimum replacement (Chromy, 1979). The selected segments were then randomly assigned to a survey year and quarter of data collection as will be described in Section 2.4. Twenty-four of these segments were designated for the coordinated 5-year sample, while the other 72 were designated as "reserve" segments.

1.5 Dwelling Units and Persons

After sample segments for the 2000 NHSDA were selected, specially trained field household listers visited the areas and obtained complete and accurate lists of all eligible dwelling units within the sample segment boundaries. These lists served as the frames for the second stage of sample selection.

The primary objective of the second stage of sample selection (listing units) was to determine the minimum number of dwelling units needed in each segment to meet the targeted sample sizes for all age groups. Thus, listing unit sample sizes for the segment were determined using the age group with the largest sampling rate, which we refer to as the "driving" age group. Using 1990 Census data adjusted to more recent data from Claritas, state- and age-specific sampling rates were computed. These rates were then adjusted by the segment's probability of selection, the subsegmentation inflation factor,[7] if any, the probability of selecting a person in the age group (equal to the maximum or 0.99 for the driving age group), and an adjustment for the "maximum of two" rule.[8] In addition to these factors, historical data from the 1999 and 2000 NHSDAs were used to compute predicted screening and interviewing response rate adjustments. The final adjusted sampling rate was then multiplied by the actual number of dwelling units found in the field during counting and listing activities. The product represents the segment's listing unit sample size.

Some constraints were put on the listing unit sample sizes. For example, to ensure adequate samples for the overlapping design and/or for supplemental studies, the listing unit sample size could not exceed 100 or half of the actual listing unit count. Similarly if five unused listing units remained in the segment, a minimum of five listing units per segment was required for cost efficiency.

Using a random start point and interval-based (systematic) selection, the actual listing units were selected from the segment frame. After dwelling unit selections were made, an interviewer visited each selected dwelling unit to obtain a roster of all persons residing in the dwelling unit. As in previous years, during the data collection period, if an interviewer encountered any new dwelling unit in a segment or found a dwelling unit that was missed during the original counting and listing activities, then the new or missed dwellings were selected into the 2000 NHSDA using the half-open interval selection technique.[9] The selection technique eliminates any frame bias that might be introduced because of errors and/or omissions in the counting and listing activities and also eliminates any bias that might be associated with using "old" segment listings.

---

[7]Segments found to be very large in the field are partitioned into *subsegments*. Then, one subsegment is chosen at random with probability proportional to size to be fielded. The subsegmentation inflation factor accounts for the narrowing down of the segment.

[8]Brewer's Selection Algorithm never allows for greater than two persons per household to be chosen. Thus, sampling rates are adjusted to satisfy this constraint.

[9]In summary, this technique states that, if a dwelling unit is selected for the 2000 study and an interviewer observes any new or missed dwelling units between the selected dwelling unit and the dwelling unit appearing immediately after the selection on the counting and listing form, then all new or missed dwellings falling in this interval will be selected. If a large number of new or missed dwelling units are encountered (generally greater than ten), then a sample of the new or missing dwelling units will be selected.

Using the roster information obtained from an eligible member of the selected dwelling unit, 0, 1, or 2 persons were selected for the survey. Sampling rates were preset by age group and state. Roster information was entered directly into the electronic screening instrument, which automatically implemented this third stage of selection based on the state and age group sampling parameters.

One exciting consequence of using an electronic screening instrument in the NHSDA is the ability to impose a more complicated person-level selection algorithm on the third stage of the NHSDA design. In 1999 and continuing through 2000, one feature that was included in the design was that *any* two survey-eligible people within a dwelling unit had some chance of being selected (i.e., all survey eligible pairs of people had some nonzero chance of being selected). This design feature was of interest to NHSDA researchers because, for example, it allows analysts to examine how the drug use propensity of one individual in a family relates to the drug use propensity of other family members residing in the same dwelling unit (e.g., the relationship of drug use between a parent and his or her child).

Chapter 2: The Coordinated 5-Year Sample

As was previously mentioned, the sample design was simultaneously developed for the 1999-2003 NHSDAs. Starting with a Census block level frame, first stage sampling units or area segments were formed. A sufficient number of segments was then selected to support the 5-year design as well as any supplemental studies SAMHSA may choose to field.

2.1 Formation of and Objectives for Using the Composite Size Measures

The composite size measure procedure is used to obtain self-weighting samples for multiple domains in multistage designs. The NHSDA sample design has employed the composite size measure methodology since 1988. Our goal was to specify size measures for sample areas (segments) and dwelling units that achieve the following objectives:

- Yield the targeted domain sample sizes in expectation (E_s) over repeated samples; that is, if m_{ds} is the domain d sample size achieved by sample s, then

$$E_s(m_{ds}) = m_d \text{ for } d=1,...,D. \quad (1)$$

- Constrain the maximum number of selections per dwelling unit at a specified value; specifically, we limit the total number of within-dwelling unit selections across all age groups to a maximum of two.

- Minimize the number of sample dwelling units that must be screened to achieve the targeted domain sample sizes.

- Eliminate all variation in the sample inclusion probabilities within a domain except for the variation in the within-dwelling unit/within-domain probabilities of selection. The inverse probabilities of selection for each sample segment were used to determine the number of sample lines to select from within each segment. As a consequence, all dwelling units within a specific stratum were selected with approximately the same probability, and therefore, approximately equalized dwelling unit sampling weights. This feature minimizes variance inflation that results from unnecessary variation in sampling weights.

- Equalize the expected number of sample persons per cluster to balance the interviewing workload and to facilitate the assignment of interviewers to regions and segments. This feature also minimizes adverse effects on precision resulting from extreme cluster size variations.

- Simplify the size measure data requirements so that decennial Census data (block level counts) are adequate to implement the method.

Using the 1990 Census data supplemented with revised population projections, a composite size measure was computed for each Census block defined within the United States. The composite size measure began by defining the rate $f_h(d)$ at which we wished to sample each age group domain d (d=1,...,5 for 12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 years or older) from state h.

Let $C_{hijk}(d)$ be the population count from domain d in Census block k of segment j of FI region i within each state h. The composite size measure for block k was defined as

$$S_{hijk} = \sum_{d=1}^{5} f_h(d)\ C_{hijk}(d). \tag{2}$$

The composite size measure for segment j was calculated as

$$S_{hij+} = \sum_{d=1}^{5} f_h(d) \sum_{j=1}^{N_{hij}} C_{hijk}(d), \tag{3}$$

where N_{hij} equals the number of blocks within segment j of FI region i and state h.

2.2 Stratification

Because the 5-year NHSDA design provides for estimates by state in all 50 states plus the District of Columbia, states may be viewed as the first level of stratification. The objective of the next level of stratification was to distribute the number of interviews, in expectation, equally among FIs. Within each state, Census tracts were joined to form mutually exclusive and exhaustive FI regions of approximately equal sizes (aggregate composite size measures of roughly 100). Using desktop computer mapping software, the regions were formed taking into account geographical boundaries, such as mountain ranges and rivers, to the extent possible. Therefore, the resulting regions facilitated ease of access as well as distributing the workload evenly among NHSDA interviewers. Twelve FI regions were formed in each state, except in California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas, where 48 regions were formed.[10]

To form segments within FI regions, adjacent Census blocks were collapsed until the total number of dwelling units within the area was at least 175 and the size measure was at least 9.38 times the maximum of F_1, F_2, F_3, F_4, and F_5, where F_i is the person sampling rate for age group i in the state. The desired number of responding persons in each segment is 9.38. Latitude and longitude and sorting within block groups, tracts, and counties were used to obtain geographic ordering of the blocks. Segments were required to be entirely within FI region and county boundaries; however, they could span Census tracts and block groups. This crossing-over was avoided as much as possible. Table 2.1 summarizes the segment sampling frame by state.

---

[10]The design called for 300 persons in each of three age groups (12 to 17, 18 to 25, and 26 and older) equally allocated to four quarters within each small sample state. Based on an analysis of the cost variance tradeoffs, an average cluster size of 3.125 persons in each of the three age groups (or an average of 9.375 persons over the three age groups combined) was considered near optimal. When applied to the small states, a quarterly sample of 75 persons per quarter per age group could be obtained from 24 clusters or area segments. For unbiased variance estimation purposes, at least two observations are required per stratum (Chromy, 1981); maximum geographic stratification was obtained by defining 12 strata with 2 area segments each per quarter. Two additional segments were selected for each of the other 3 quarters, yielding 8 area segments per stratum or 96 area segments per small sample state. This stratum configuration also corresponded with reasonable average workload for a single FI, leading us to designate the geographic strata within state as FI regions. This approach supported a target sample size for the small states of 300 persons per age group or a total of 900 for the year. In the large sample states, four times as large a sample was required. Optimum cluster size configuration and maximum stratification given the need for unbiased variance estimation were maintained by simply quadrupling the number of FI regions to 48 per large sample state, yielding a sample 300 persons per age group per quarter, 1,200 per age group over four quarters, and 3,600 per year over all three age groups.

Table 2.1 Number of Segments on Sampling Frame by State

State	State Abbreviation	State FIPS Code	Number of Segments on Sampling Frame	Total Number of Segments Selected	Number Selected for Five-Year Sample	Unique Segments in Five-Year Sample
Total U.S.			499,287	86,400		
Northeast						
Connecticut	CT	09	5,978	1,152	288	288
Maine	ME	23	2,573	1,152	288	288
Massachusetts	MA	25	11,413	1,152	288	288
New Hampshire	NH	33	2,246	1,152	288	286
New Jersey	NJ	34	14,343	1,152	288	288
New York	NY	36	30,600	4,608	1,152	1,151
Pennsylvania	PA	42	24,256	4,608	1,152	1,151
Rhode Island	RI	44	1,912	1,152	288	282
Vermont	VT	50	1,248	1,152	288	284
North Central						
Illinois	IL	17	22,549	4,608	1,152	1,151
Indiana	IN	18	11,987	1,152	288	288
Iowa	IA	19	6,210	1,152	288	288
Kansas	KS	20	5,430	1,152	288	288
Michigan	MI	26	18,477	4,608	1,152	1,152
Minnesota	MN	27	9,364	1,152	288	288
Missouri	MO	29	10,871	1,152	288	288
Nebraska	NE	31	3,567	1,152	288	288
North Dakota	ND	38	1,330	1,152	288	286
Ohio	OH	39	21,500	4,608	1,152	1,151
South Dakota	SD	46	1,603	1,152	288	285
Wisconsin	WI	55	10,704	1,152	288	288
South						
Alabama	AL	01	8,702	1,152	288	288
Arkansas	AR	05	5,411	1,152	288	288
Delaware	DE	10	1,346	1,152	288	281
Washington, D.C.	DC	11	943	1,152	288	273
Florida	FL	12	26,545	4,608	1,152	1,152
Georgia	GA	13	13,398	1,152	288	288
Kentucky	KY	21	7,718	1,152	288	287
Louisiana	LA	22	8,216	1,152	288	288
Maryland	MD	24	8,340	1,152	288	288
Mississippi	MS	28	5,473	1,152	288	288
North Carolina	NC	37	14,955	1,152	288	288
Oklahoma	OK	40	6,941	1,152	288	288
South Carolina	SC	45	7,437	1,152	288	287
Tennessee	TN	47	10,764	1,152	288	288
Texas	TX	48	34,367	4,608	1,152	1,151
Virginia	VA	51	11,666	1,152	288	288
West Virginia	WV	54	3,757	1,152	288	288

(continued)

Table 2.1 Number of Segments on Sampling Frame by State (continued)

State	State Abbreviation	State FIPS Code	Number of Segments on Sampling Frame	Number of Segments Selected	Number Selected for Five-Year Sample	Unique Segments in Five-Year Sample
West						
Alaska	AK	02	1,139	1,152	288	273
Arizona	AZ	04	8,212	1,152	288	288
California	CA	06	53,064	4,608	1,152	1,152
Colorado	CO	08	7,977	1,152	288	287
Hawaii	HI	15	1,658	1,152	288	276
Idaho	ID	16	2,611	1,152	288	288
Montana	MT	30	2,028	1,152	288	286
Nevada	NV	32	2,625	1,152	288	276
New Mexico	NM	35	3,369	1,152	288	288
Oregon	OR	41	6,835	1,152	288	288
Utah	UT	49	3,475	1,152	288	288
Washington	WA	53	11,086	1,152	288	287
Wyoming	WY	56	1,068	1,152	288	285

FIPS = Federal Information Processing Standards.

2.3 First-Stage Sample Selection

Once the segments were formed, a probability proportional to size sample of segments was selected with minimum replacement within each FI region. The sampling frame was implicitly stratified by sorting the first-stage sampling units by an MSA/SES indicator[11] and by the percent of the population that is non-Hispanic and white. As Table 2.1 indicates, 96 segments per FI region were chosen for a total of 1,152 segments in each state, except in the large states where a total of 4,608 segments were chosen. Although only 24 segments were needed to support the 5-year study, an additional 72 segments were selected to serve as replacements when segment lines are depleted and/or to support any supplemental studies embedded within the NHSDA.

2.4 Survey Year and Quarter Assignment

Within each FI region, the 96 selected segments were assigned to a survey year and quarter in a random, systematic fashion. Because segments can be selected multiple times, the goal was to avoid putting the same segment in consecutive survey years. Therefore, survey years and quarters were assigned using a random starting point and the order defined in Table 2.2. The notation in the table is as follows:

99A = Segment for the 1999 NHSDA,
99B = Segment for the 1999 NHSDA and used again in the 2000 NHSDA,
00 = Segment for the 2000 NHSDA and used again in the 2001 NHSDA,
01 = Segment for the 2001 NHSDA and used again in the 2002 NHSDA,
02 = Segment for the 2002 NHSDA and used again in the 2003 NHSDA, and
03 = Segment for the 2003 NHSDA.

[11]Four categories are defined as: (1) MSA/low SES, (2) MSA/high SES, (3) Non-MSA/low SES, and (4) Non-MSA/high SES.

Table 2.2 Survey Year and Quarter Assignment Order for 96 Segments within Each FI Region

Order	Survey Year	Quarter	Panel	Variance Replicate
1	99A	1	1	1
2	Y00	1	15	1
3	X99B	1	8	2
4	Z01	1	22	2
5	02	1	5	1
6	Y99A	1	13	1
7	X03	1	12	2
8	Z99B	1	20	2
9	00	1	3	1
10	Y02	1	17	1
11	X01	1	10	2
12	Z03	1	24	2
13	01	1	4	2
14	Y03	1	18	2
15	X02	1	11	1
16	Z99A	1	19	1
17	99B	1	2	2
18	Y01	1	16	2
19	X00	1	9	1
20	Z02	1	23	1
21	03	1	6	2
22	Y99B	1	14	2
23	X99A	1	7	1
24	Z00	1	21	1
25	99A	2	1	1
26	Y00	2	15	1
27	X99B	2	8	2
28	Z01	2	22	2
29	02	2	5	1
30	Y99A	2	13	1
31	X03	2	12	2
32	Z99B	2	20	2
33	00	2	3	1
34	Y02	2	17	1
35	X01	2	10	2
36	Z03	2	24	2
37	01	2	4	2
38	Y03	2	18	2
39	X02	2	11	1
40	Z99A	2	19	1
41	99B	2	2	2
42	Y01	2	16	2
43	X00	2	9	1
44	Z02	2	23	1
45	03	2	6	2
46	Y99B	2	14	2
47	X99A	2	7	1
48	Z00	2	21	1
49	99A	3	1	1
50	Y00	3	15	1
51	X99B	3	8	2
52	Z01	3	22	2
53	02	3	5	1
54	Y99A	3	13	1
55	X03	3	12	2
56	Z99B	3	20	2
57	00	3	3	1
58	Y02	3	17	1
59	X01	3	10	2
60	Z03	3	24	2
61	01	3	4	2
62	Y03	3	18	2
63	X02	3	11	1
64	Z99A	3	19	1
65	99B	3	2	2
73	99A	4	1	1
74	Y00	4	15	1
75	X99B	4	8	2
76	Z01	4	22	2
77	02	4	5	1
78	Y99A	4	13	1
79	X03	4	12	2
80	Z99B	4	20	2
81	00	4	3	1
82	Y02	4	17	1
83	X01	4	10	2
84	Z03	4	24	2
85	01	4	4	2
86	Y03	4	18	2
87	X02	4	11	1
88	Z99A	4	19	1
89	99B	4	2	2

(continued)

Table 2.2 Survey Year and Quarter Assignment Order for 96 Segments within Each FI Region (continued)

Order	Survey Year	Quarter	Panel	Variance Replicate
66	Y01	3	16	2
67	X00	3	9	1
68	Z02	3	23	1
69	03	3	6	2
70	Y99B	3	14	2
71	X99A	3	7	1
72	Z00	3	21	1
90	Y01	4	16	2
91	X00	4	9	1
92	Z02	4	23	1
93	03	4	6	2
94	Y99B	4	14	2
95	X99A	4	7	1
96	Z00	4	21	1

X, Y, and Z denote extra segments for the corresponding NHSDA survey year. The 24 segments assigned to survey years not beginning with X, Y, and Z would then be used to field the 5-year study. Using the survey year and quarter assignments, a sequential segment identification number (SEGID) was then assigned. Table 2.3 describes the relationship between segment identification numbers and quarter assignment. The last two digits in the SEGID are called the "segment suffix" in Table 2.3 and correspond to "panel" in Table 2.2.

2.5 Creation of Variance Estimation Strata

The nature of the stratified clustered sampling design requires that the design structure be taken into consideration when computing variances of survey estimates. Key nesting variables were created to capture explicit stratification and to identify clustering. For the 1999-2003 NHSDAs, each FI region comprised its own stratum.

Two replicates per year were defined within each variance stratum. The first replicate consists of those segments that are "phasing out" or will not be used in the next survey year. The second replicate is made up of those segments that are "phasing in" or will be fielded again the following year, thus constituting the 50% overlap between survey years. Each variance replicate consists of four segments, one for each quarter of data collection. Table 2.2 describes the assignment of segments to variance estimation replicates.

All weighted statistical analyses for which variance estimates are needed should use the stratum and replicate variables to identify nesting. Variance estimates can be computed by using clustered data analysis software packages such as SUDAAN (RTI, 2001). The SUDAAN software package computes variance estimates for nonlinear statistics using procedures such as a first-order Taylor series approximation of the deviations of estimates from their expected values. The approximation is unbiased for sufficiently large samples.

Table 2.3 Segment Identification Number Suffixes for the 1999-2003 NHSDAs

Segment Suffix	1999 NHSDA	2000 NHSDA	2001 NHSDA	2002 NHSDA	2003 NHSDA
01	x (Q1)				
02	x (Q1)	x (Q1)			
03	x (Q2)				
04	x (Q2)	x (Q2)			
05	x (Q3)				
06	x (Q3)	x (Q3)			
07	x (Q4)				
08	x (Q4)	x (Q4)			
09		x (Q1)	x (Q1)		
10		x (Q2)	x (Q2)		
11		x (Q3)	x (Q3)		
12		x (Q4)	x (Q4)		
13			x (Q1)	x (Q1)	
14			x (Q2)	x (Q2)	
15			x (Q3)	x (Q3)	
16			x (Q4)	x (Q4)	
17				x (Q1)	x (Q1)
18				x (Q2)	x (Q2)
19				x (Q3)	x (Q3)
20				x (Q4)	x (Q4)
21					x (Q1)
22					x (Q2)
23					x (Q3)
24					x (Q4)

Note: The segment suffix is defined as the last two digits of the segment identification number.

Chapter 3: General Sample Allocation Procedures for the Main Study

In this chapter, the computational details of the procedural steps used to determine both person and dwelling unit sample sizes will be discussed. The within-dwelling unit age group specific selection probabilities for the 2000 NHSDA main study design are also addressed. This optimization procedure was specifically designed to address SAMHSA's multiple precision and design requirements while simultaneously minimizing the cost of data collection. Costs were minimized by determining the smallest number of interviews and selected dwelling units necessary to achieve the various design requirements. In summary, this three-step optimization procedure proceeded as follows:

1. In the first step, we determined the optimal number of interviews (i.e., responding persons) by domains of interest needed to satisfy the precision requirements for several drug outcome measures. In other words, we initially sought to determine 255 unknown m_{ha} for each state h (51) and age group a (5). A solution to this multiple constraint optimization was achieved utilizing Chromy's Algorithm (Chromy, 1987). This is described in further detail in Section 3.2.

2. Using the m_{ha} determined from Step 1, the next step was to determine the optimal number of selected dwelling (D_{hj}) units (i.e., second-stage sample) necessary. This step was achieved by applying parameter constraints (e.g., probabilities of selection and expected response rates) at the segment level j or the stage at which dwelling units would be selected. This was done on a quarterly basis using 25% of the m_{ha}'s. This step is described in further detail in Section 3.3.

3. The final step in this procedure entails determining age group specific probabilities of selection (S_{hja}) for each segment given m_{ha} and D_{hj} from Steps 1 and 2. This was achieved using a modification of Brewer's Method of Selection (Cochran, 1977, pp. 261-263). The modification was designed to select 0, 1, or 2 persons from each dwelling unit.[12] A detailed discussion of the final step is given in Section 3.4. After calculation of the required dwelling units and the selection probabilities, sample size constraints[13] were applied to ensure adequate sample for overlapping designs and/or supplemental studies and to reduce field interviewer burden. Limits on the total number of expected interviews per segment were also applied. This process became iterative to reallocate the reduction in sample size to other segments not affected by such constraints. Details of this step in the optimization procedure are given in Section 3.5.

3.1 Notation

h = 50 states plus the District of Columbia.

a = Age group. a = 1...5 and represents the following groups: 12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older.

j = Individual segment indicator (total of 7,200; 1,800 per quarter).

s = Design parameter estimated state strata. Utilized for estimating response and eligibility rates from historic NHSDA data. Individual segments are defined into 1 of 3 strata. For the 2000 main study NHSDA, s = 1...3 are defined as follows:

---

[12]Direct application of Brewer's method would require a fixed sample size.

[13]Because of the overlap of the split sample, constraints were applied to the required dwelling unit sample sizes. Specifically, some segments would be revisited in the 2001 survey.

Stratum	Defining Criteria
1: High Response Rate	Overall Response Rate $\geq$ 0.70
2: Medium Response Rate	0.62 $\leq$ Overall Response Rate < 0.70
3: Low Response Rate	Overall Response Rate < 0.62

m_{ha} = Number of completed interviews (person respondents) desired in each state h and age group a. Computation of m_{ha} is discussed in Section 3.2. For quarter computation of selected dwelling unit sample size, 25% of the yearly estimate is used.

y_{ha} = Estimated number of persons in the target population in state h and age group a. The 2000 population is estimated using the 1990 Census data and the 2001 Claritas Population Projections in the compound interest formula, $y = Ae^{Bx}$, where

y = population at time x,
A = initial population,
e = base of the system of natural logarithms,
B = growth rate per unit of time, and
x = period of time over which growth occurs.

First, B is computed as $\{ln(y/A)\}/x$, where y = the population in 2001, A = the population in 1990 and x = 11. Then, the 2000 population (y^*_{ha}) is computed using the original formula and this time allowing x to be 10. Finally, the 2000 population is adjusted by the ratio of estimated eligible listed dwelling units to the Claritas dwelling unit counts (U_{hj}). This adjustment factor considers the number of added dwelling units expected to be obtained through the half-open interval rule (1.01) and the probability of a dwelling unit being eligible (ε_s), both determined via historic data. The coefficient adjustment of 1.01 is estimated using historical data and is the proportion of all screened dwelling units (includes added) over the original total of selected dwelling units (excluding added dwelling units). So, $y_{ha} = [(1.01 * \varepsilon_s * L_{hj} * (1/I_{hj})) / U_{hj}] * y^*_{ha}$, where ε_s , L_{hj} , and I_{hj} are defined further below. This adjustment is computed at the Census block level then aggregated to the state level.

f_{ha} = m_{ha} / y_{ha}. State-specific age group sampling fraction.

F_h = $Max\{f_{ha} / (\varphi_s * \lambda_{sa} * \delta_{sa})$, a=1-5$\}$.

P_{hj} = Inverse of the segment selection probability. Dwelling unit sample sizes are computed on a quarterly basis and segments are selected on a yearly basis. Since each quarter only contains a fourth of the selected segments, these probabilities are adjusted by a factor of 4 so that weights will add to the yearly totals.

I_{hj} = Subsegmentation inflation factor. For segments too large to count and list efficiently in both time and cost, field listing personnel are allowed to subsegment the segment into roughly equal size subdivisions. They perform a quick count (best guess: L^*_{hj}) of the entire segment and then subdivide (taking also a best guess estimate of the number of dwelling units in each subsegment: B^*_{hj}). Using a selection algorithm provided by RTI , one subsegment is selected for regular counting and listing. For the subsegment to represent the entire segment, the weights are adjusted up to reflect the unused portion of the segment.

$= (B^*_{hj} / L^*_{hj})$.

= 1, if no subsegmenting was done.

D_{hj} = Minimum number of dwelling units to select for screening in segment *j* to meet the targeted sample sizes for all age groups.

L_{hj} = Final segment count of dwelling units available for screening.

S_{hja} = State, segment-specific probability of selecting a person in age group *a*. A design constraint implemented is that no single age group selection probability could exceed 1. The maximum allowable probability was then set to .99.

S_{sa} = Stratum-specific probability of selecting a person in age group *a*. Only used in calculation of Max of 2 rule (δ_{sa}) described below. As with S_{hja}, the maximum allowable probability is .99.

ε_s = Stratum-specific, dwelling unit eligibility rate. Derived from 1999 NHSDA Quarter 3 data by taking the average eligibility rate among all states in the same stratum.

φ_s = Stratum-specific, screening response rates. Calculated using the same methodology as described for the dwelling unit eligibility rate (ε_s).

λ_{sa} = Stratum and age group-specific interview response rate. Using data from Quarter 3 of the 1999 NHSDA, the additive effects of stratum and age group on interviewer response were determined by taking the average interview response rate among all states in the same stratum.

γ_a = Expected number of persons within an age group per dwelling unit. Calculated using 1998 NHSDA data by dividing the weighted total number of rostered persons in an age group by the weighted total number of complete screened dwelling units.

δ_{sa} = Stratum and age group-specific maximum-of-two rule adjustment. The survey design restricts the number of interviews per dwelling unit to a total of two. This is achieved through a modified Brewer's method of selection. This results in a loss of potential interviews in dwelling units where selection probabilities sum greater than two. The adjustment is designed to inflate the number of required dwelling units to compensate for this loss. This procedure is iterative and utilizes 1998 NHSDA data as described below. (Note that, since prior NHSDA data are unavailable for each segment, maximum-of-two rule adjustments are computed at the stratum level.)

1. Determine the number of required dwelling units (R_{sa}) necessary to obtain desired person sample sizes under the assumption that age group sample sizes are the same across the strata (use overall national sample sizes).

$$R_{sa} = \frac{m_a}{(\varepsilon_s * \varphi_s * \lambda_{sa} * \gamma_a * \delta_{sa})} \quad \text{where } m_a = \sum_h m_{ha}, \quad \delta_{sa} = 1 \text{ for first iteration} \tag{4}$$

2. Set S_{sa} = .99 for the age group with the largest R_{sa}. All other age group probabilities are set in proportion to the largest:

$$S_{sa} = \frac{R_{sa}}{Max(R_{sa})} \tag{5}$$

3. Assign S_{sa} to respective person record in 1998 NHSDA data. With the modified Brewer's method, selection probabilities are now adjusted to reflect the total household composition. In short, if selection probabilities for all eligible dwelling unit members sum greater than two, then probabilities are ratio adjusted to sum to two. This will be denoted as S^*_{sa}. However, sums less than two are unadjusted.

4. Sum S_{sa} and S^*_{sa} within stratum. The maximum-of-two rule (δ_{sa}) is then calculated as the ratio of the summed S^*_{sa} / S_{sa}.

5. Insert new calculated δ_{sa} into Step 1 and repeat Steps 1 through 5. Continue until the absolute difference between δ_{sa} of the current cycle and the previous cycle is less than .001, usually about three to four iterations.

3.2 Determining Person Sample Sizes by State and Age Group

The first step in the design of the third stage of selection was to determine the optimal number of respondents for each of the 255 domains that would be needed to minimize costs associated with data collection, subject to multiple precision requirements established by SAMHSA. In summary, the precision requirements on the relative standard error (RSE) of an estimate of 10% for SAMHSA's 17 subpopulations of interest are:

- RSE = 3.40% for the total, national population.
- RSE = 5.00% for the national population in each of the four age groups: 12 to 17, 18 to 25, 26 to 34, 35 or older.
- RSE = 5.00% for the population within each of the four age groups for white (i.e., nonblack, non-Hispanic).
- RSE = 11.00% for the population within each of the four age groups for blacks (i.e., black, non-Hispanic).
- RSE = 11.00% for the population within each of the four age groups for Hispanics.

Note, one stratification feature that we used in previous NHSDA designs and was worth including in the design of the current NHSDA is the expansion of the age group domain to 12 to 17, 18 to 25, 26 to 34, 35 to 49, and 50 or older age groups. This age group stratification parallels SAMHSA's NHSDA subpopulation of interest, as implied by the precision constraints, except for the age group 35 or older. As we have done with the NHSDA designs since 1992, we have chosen to further stratify this important age group by 35 to 49 and 50 or older to decrease the total number of 35 or older respondents needed to meet precision requirements. Since substance abuse is more prevalent among the 35 to 49 year olds compared to the 50 or older age group, oversampling this younger age group will increase the precision of the estimates generated for the 35 or older age group, while minimizing the total number of respondents aged 35 years or older needed in the sample.

To form precision constraints that reflect the above standard error requirements, we have set up a preliminary Step-1 Optimization using (1) design effects estimated from the 1994-1996 NHSDA data, (2) population counts obtained from Claritas, Inc., and (3) various outcome measures that were estimated for each block group in the United States from our recently completed 1991-1993 NHSDA small area estimation (SAE) project. Appropriate variance constraints were defined for nine outcome measures of interest. These outcome measures of interest were included to address not only the NHSDA recency-of-use estimates but

also such related generic substance abuse measures as treatment received for alcohol and illicit drug use and dependency on alcohol and illicit drug use.

Specifically, the nine classes of NHSDA outcomes we considered were:

Use of Legal (Licit) Substances

1. *Cigarette Use in the Past Month.* Smoked cigarettes at least once within past month.

2. *Alcohol Use in the Past Month.* Had at least one drink of an alcoholic beverage (beer, wine, liquor, or a mixed alcohol drink) within the past month.

Use of Illicit Substances

3. *Any Illicit Drug Use in the Past Month.* Includes hallucinogens, heroin, marijuana, cocaine, inhalants, opiates or nonmedical use of sedatives, tranquilizers, stimulants, or analgesics.

4. *Any Illicit Drug Use Other than Marijuana in the Past Month.* Past month use of any illicit drug excluding those whose only illicit drug use was marijuana.

5. *Cocaine Use in the Past Month.* Use within the past month of cocaine in any form, including crack.

Note that current use of any illicit drug provides a broad measure of illicit drug use; however, it is dominated by marijuana and cocaine use. Therefore, estimates of marijuana and cocaine are included since these two measures reflect different types of drug abuse.

Drug or Alcohol Dependence

6. *Dependent on Illicit Drugs in the Past Year.* Dependent on the same drugs listed in *3. Any Illicit Drug Use in the Past Month* above. Those who are dependent on both alcohol and another illicit substance are included, but those who are dependent on alcohol only are not.

7. *Dependent on Alcohol and Not Illicit Drugs in the Past Year.* Dependent on alcohol and not dependent on any illicit drug.

Treatment for Drugs and Alcohol Problems

8. *Received Treatment for Illicit Drugs in the Past Year.* Received treatment in the past 12 months at any location (including hospitals, clinics, self-help groups, doctors) for any illicit drugs.

9. *Received Treatment for Alcohol Use but Not Illicit Drugs in the Past Year.* Received treatment in the past 12 months for drinking (including hospitals, clinics, self-help groups, doctors). These estimates exclude those who received treatment in the past 12 months for both drinking and illicit drugs.

These outcome measures considered, as well as the precision that is expected from this 2000 NHSDA design, are presented in Table 3.1. RSEs were based on an average prevalence rate of 10% for each measure.

Additionally, initial sample size requirements were implemented:

- Minimum sample size of 3,600 persons per state in the eight large states and 900 persons in the remaining 43 states.

- Equal allocation of the sample across the three age groups: 12 to 17, 18 to 25, and 26 or older within each state.

A tobacco brand interview supplement and an additional sample of 2,500 youths aged 12 to 17 were added to the NHSDA to allow for estimation of tobacco brand usage by youth. The 2,500 additional youth were allocated to states in the following manner.

1. Calculate the expected 12 to 17 sample for each state based on population.[14]

$$m^{Pop}_{h(12\text{-}17)} = (22{,}500 - (1.85 * 2{,}500)) * (N_{h(12\text{-}17)} / N_{(12\text{-}17)}) \; . \qquad (6)$$

2. Find the difference between original sample allocation ($m_{h(12\text{-}17)}$) and allocation based on population ($m^{Pop}_{h(12\text{-}17)}$):

$$diff = m_{h(12\text{-}17)} - m^{Pop}_{h(12\text{-}17)}. \qquad (7)$$

   - If *diff* is negative, the original sample requires over-sampling and for cost purposes no additional 12 to 17 sample is allocated to these states.

   - If *diff* is positive, no oversampling of the 12 to 17 is necessary for the original sample allocation. The additional 2,500 sample is allocated to these states as the value of *diff*.

Furthermore, race/ethnicity groups are not oversampled for the 2000 main study. However, consistent with previous NHSDAs, the 2000 NHSDA is designed to over-sample the younger age groups.

Among the 51 states, a required total sample size of 70,000 respondents is necessary to meet all precision and sample size requirements. Table 3.2 shows expected state by age group sample sizes which were equally allocated to each of the four quarters.

[14]Sample size determination based on population alone would have required a reduction of the originally allocated sample in some states. Since this was not feasible for state-level precision and estimates purposes, a reduction in total sample size was necessary to compensate for not reducing the sample below the original sample size. 1.85 was iteratively computed as the value necessary for correct sample allocation calculation.

Table 3.1 Expected Relative Standard Errors By Race/Ethnicity and Age Group: Main Sample

	Total Respondents					Hispanic Respondents				
Outcome Measure	**12-17**	**18-25**	**26-34**	**35+**	**Total**	**12-17**	**18-25**	**26-34**	**35+**	**Total**
Expected Relative Standard Error for Classes of Outcome Measures										
Past Year, Dependence on Alcohol (not Illicit Drugs)	2.49	2.70	4.29	3.56	2.27	6.16	7.54	10.72	11.65	6.04
Past Month Alcohol Use	2.57	2.71	4.23	3.58	2.47	6.42	7.47	10.62	11.40	6.42
Past Month Cigarette Use	2.31	2.62	4.13	3.30	2.22	6.92	7.11	10.31	11.99	6.90
Past Month Cocaine Use	2.29	2.50	3.57	2.30	1.55	6.32	7.42	10.21	9.95	5.18
Past Year Received Treatment For Illicit Drug Use	2.44	2.57	3.58	2.97	1.87	6.53	7.17	10.44	10.72	5.65
Past Year Received Treatment For Alcohol Use	2.43	2.51	3.52	3.05	2.02	6.47	7.24	10.04	10.67	5.82
Past Month Use of Any Illicit Drug But Marijuana	2.31	2.49	3.60	3.03	1.82	6.43	7.57	10.40	11.08	5.14
Dependence on Illicit Drugs	2.43	2.63	3.61	2.93	1.77	6.49	7.42	10.43	10.61	4.93
Past Month Illicit Drug Use	2.44	2.57	3.60	3.15	1.80	6.49	7.13	10.31	10.94	5.19
Average Relative Standard Error	2.41	2.59	3.79	3.10	1.98	6.47	7.34	10.39	11.00	5.69
Target Relative Standard Error	5.00	5.00	5.00	5.00	3.40	11.00	11.00	11.00	11.00	

	Black Respondents					White Respondents				
Outcome Measure	**12-17**	**18-25**	**26-34**	**35+**	**Total**	**12-17**	**18-25**	**26-34**	**35+**	**Total**
Expected Relative Standard Error for Classes of Outcome Measures										
Past Year, Dependence on Alcohol (not Illicit Drugs)	6.40	7.14	10.13	10.14	6.28	2.79	3.10	4.33	3.70	2.51
Past Month Alcohol Use	6.65	7.19	10.03	10.28	6.23	2.88	3.11	4.33	3.73	2.78
Past Month Cigarette Use	6.29	7.31	10.17	10.11	6.42	2.70	3.02	4.50	3.61	2.48
Past Month Cocaine Use	6.36	6.48	9.23	8.87	5.55	2.75	2.85	4.15	2.62	1.64
Past Year Received Treatment For Illicit Drug Use	6.08	6.98	10.23	9.15	5.77	2.82	3.07	4.14	3.22	2.05
Past Year Received Treatment For Alcohol Use	6.09	6.52	10.18	9.43	6.11	2.79	3.00	4.08	3.21	2.26
Past Month Use of Any Illicit Drug But Marijuana	6.33	6.84	9.96	9.31	5.24	2.68	2.87	4.18	3.35	1.96
Dependence on Illicit Drugs	6.12	7.01	10.14	9.38	5.78	2.78	3.15	4.24	3.14	1.96
Past Month Illicit Drug Use	6.10	6.85	10.15	9.56	5.26	2.78	3.07	4.20	3.35	2.00
Average Relative Standard Error	6.27	6.92	10.02	9.58	5.85	2.78	3.03	4.24	3.33	2.18
Target Relative Standard Error	11.00	11.00	11.00	11.00		5.00	5.00	5.00	5.00	

Note: Relative Standard Errors are based on a prevalence rate of 10%.

3.3 Second-Stage Sample Allocation for Each Segment

Given the desired respondent sample size for each state and age group (m_{ha}) needed to meet the design parameters established by SAMHSA, the next step is to determine the minimal number of dwelling units to select for each segment to meet the targeted sample sizes. In short, this step involves determining the sample size of the second-stage of selection. This sample size determination is performed on a quarterly basis to take advantage of both segment differences and, if necessary, make adjustments to design parameters. Procedures described below were originally developed for initial implementation in Quarter 1 of the survey. The description below is specific to Quarter 1. Any modifications/corrections were made in subsequent quarters and are explained in detail in Section 3.7.

3.3.1 Dwelling Unit Frame Construction—Counting and Listing

The process by which the dwelling unit frame is constructed is called counting and listing. In summary, a certified lister visits the selected area and lists a detailed and accurate address (or description if no address is available) for each dwelling unit within the segment boundaries. The lister is given a series of maps on which to make note of the location of these dwelling units. The resulting list of dwelling units is entered into a database and serves as the frame from which the second-stage sample is drawn.

In some situations, the number of dwelling units within the segment boundaries is much larger than the specified maximum. To obtain a reasonable number of dwelling units for the frame, the lister will first count the dwelling units in such an area. The sampling staff at RTI will then partition the segment into smaller pieces or subsegments and randomly select one to be listed. For more information on the subsegmenting procedures, see the Counting and Listing Supplement for Subsegmenting (RTI, 1996).

During counting and listing, the lister moves about the segment in a prescribed fashion called the "continuous path of travel." In short, the lister attempts to move in a clockwise fashion, makes each possible right turn, makes U-turns at segment boundaries, and doesn't break street sections. Following these defined rules and always looking for dwelling units on the right hand side of the street, the lister minimizes the chance of not listing a dwelling unit within the segment. Also, using a defined path of travel makes it easier for the FI assigned to the segment to locate the sampled dwelling units. Finally, the continuous path of travel lays the groundwork for the half-open interval procedure for recovering missed dwelling units as is described in Section 3.7 of this report. A detailed description of the counting and listing procedures is provided in the 1999 NHSDA: Counting and Listing General Manual (RTI, 1999).

3.3.2 Determining Dwelling Unit Sample Size

For the main study, the optimization formula is as follows:

$$f_{ha} = P_{hj} * I_{hj} * \left(\frac{D_{hj}}{L_{hj}}\right) * S_{hja} * \varphi_s * \lambda_{sa} * \delta_{sa}. \quad (8)$$

At this point in the procedure, only two components in the formula are unknown: D_{hj} and S_{hja}. Selection probabilities are segment- and age-group specific, and to maximize the number of selected persons

Table 3.2 Expected Main Study Sample Sizes by State and Age Group

State	State FIPS	FI Regions	Total Segments	Total Respondents					
				12-17	18-25	26-34	35-49	50+	Total
Total Population		900	7,200	25,000	22,500	9,352	6,900	6,248	70,000
Northeast									
Connecticut	09	12	96	300	300	130	90	80	900
Maine	23	12	96	300	300	78	95	127	900
Massachusetts	25	12	96	350	300	139	87	74	950
New Hampshire	33	12	96	300	300	81	100	119	900
New Jersey	34	12	96	472	300	135	91	74	1,072
New York	36	48	384	1,200	1,200	572	350	278	3,600
Pennsylvania	42	48	384	1,200	1,200	556	341	303	3,600
Rhode Island	44	12	96	300	300	79	95	126	900
Vermont	09	12	96	300	300	79	99	122	900
North Central									
Illinois	17	48	384	1,200	1,200	571	376	253	3,600
Indiana	18	12	96	415	300	128	94	78	1,015
Iowa	19	12	96	300	300	120	95	86	900
Kansas	20	12	96	300	300	124	93	83	900
Michigan	26	48	384	1,200	1,200	532	372	296	3,600
Minnesota	27	12	96	319	300	127	94	80	919
Missouri	29	12	96	358	300	133	90	77	958
Nebraska	31	12	96	300	300	80	93	127	900
North Dakota	38	12	96	300	300	80	94	126	900
Ohio	39	48	384	1,200	1,200	508	381	311	3,600
South Dakota	46	12	96	300	300	80	92	128	900
Wisconsin	55	12	96	356	300	131	91	78	956
South									
Alabama	01	12	96	301	300	145	84	70	901
Arkansas	05	12	96	300	300	79	87	134	900
Delaware	10	12	96	300	300	82	96	123	900
District of Columbia	11	12	96	300	300	87	98	115	900
Florida	12	48	384	1,200	1,200	535	353	312	3,600
Georgia	13	12	96	522	300	145	91	65	1,122
Kentucky	21	12	96	300	300	136	87	77	900
Louisiana	22	12	96	315	300	142	89	69	915
Maryland	24	12	96	317	300	146	89	65	917
Mississippi	28	12	96	300	300	147	84	69	900
North Carolina	37	12	96	464	300	142	89	70	1,064
Oklahoma	40	12	96	300	300	140	85	75	900
South Carolina	45	12	96	300	300	141	89	70	900
Tennessee	47	12	96	360	300	138	89	73	960
Texas	48	48	384	1,484	1,200	629	348	223	3,884
Virginia	51	12	96	428	300	141	92	67	1,028
West Virginia	54	12	96	300	300	76	88	136	900

(continued)

Table 3.2 Expected Main Study Sample Sizes by State and Age Group (continued)

State	State FIPS	FI Regions	Total Segments	Total Respondents					
				12-17	18-25	26-34	35-49	50+	Total
West									
Alaska	02	12	96	300	300	89	117	94	900
Arizona	04	12	96	319	300	145	89	66	919
California	06	48	384	2,231	1,200	599	383	218	4,631
Colorado	08	12	96	300	300	139	96	65	900
Hawaii	15	12	96	300	300	83	95	122	900
Idaho	16	12	96	300	300	80	97	123	900
Montana	30	12	96	300	300	77	95	128	900
Nevada	32	12	96	300	300	83	101	116	900
New Mexico	35	12	96	300	300	85	102	113	900
Oregon	41	12	96	300	300	144	85	71	900
Utah	49	12	96	300	300	85	105	110	900
Washington	53	12	96	389	300	154	82	64	989
Wyoming	56	12	96	300	300	80	101	119	900

FIPS = Federal Information Processing Standards.

within a dwelling unit, the age group whose adjusted sampling fraction ($f_{ha} / (\varphi_s * \lambda_{sa} * \delta_{sa})) = F_h$, known now as the driving age group, is set to the largest allowable selection probability (S_{hja}) of .99. D_{hj} is then computed as:

$$D_{hj} = \frac{f_{ha}}{(P_{hj} * I_{hj} * S_{hja} * \varphi_s * \lambda_{sa} * \delta_{sa})} * L_{hj} . \quad (9)$$

3.4 Determining Third-Stage Sample (Person) Selection Probabilities for Each Segment

$$S_{hja} = \frac{f_{ha}}{P_{hj} * I_{hj} * (\frac{D_{hj}}{L_{hj}}) * \varphi_s * \lambda_{sa} * \delta_{sa}} \quad (10)$$

Having solved for D_{hj}, solve the selection probabilities for the remaining age groups. If L_{hj} equals 0 and subsequently D_{hj} equals 0, then all S_{hja} equals 0.

3.5 Sample Size Constraints: Guaranteeing Sufficient Sample for Additional Studies and Reducing Field Interviewer Burden

A major area of interest for the survey is to ensure that an adequate sample of eligible dwelling units remain within each segment. This sample surplus is needed to provide for the yearly 50% overlap across segments and to allow SAMHSA to implement supplemental studies. An adequate remaining sample has two advantages: (1) for the 50% overlap design, this will provide better precision in year-to-year trend estimates because of the expected positive correlation between successive NHSDA years; and (2) it will reduce the amount of counting and listing costs.

In addition, concern was noted about guaranteeing that FIs would be able to complete the amount of work assigned to them within the quarterly time frame. These concerns prompted adjustments to the D_{hj} sample size:

1. Number of selected dwelling units for screening: < 100 or < ½*L_{hj}. Adjustments were made by adjusting the D_{hj} counts to equal the minimum of 100 or ½*L_{hj}.

2. Number of selected dwelling units: > 5. For cost purposes, if at least five dwelling units remain in the segment, the minimum number of selected dwelling units was set to five.

3. Expected number of interviews: < 40.

This expected number of interviews ($m^*_{hja(\text{main})}$) was computed for the main study as follows:

$$m^*_{hja(\text{main})} = D^*_{hj} * \varepsilon_s * \varphi_s * \gamma_{sa} * S_{hja} * \lambda_{sa} * \delta_{\text{sa}}, \tag{11}$$

where D^*_{hj} has been adjusted for constraint 1. This value is the total number of interviews expected within each segment. The calculation of the first adjustment, the screening adjustment, is:

$$5 / D^*_{hj} . \tag{12}$$

Similarly, the interview adjustment is computed as:

$$40 / m^*_{hja(\text{main})} . \tag{13}$$

This second adjustment is applied to D_{hj} under the assumption of an equal number of screened dwelling units for each completed interview.

Both constraints 1 and 3 reduce the second-stage sample. This in turn could potentially reduce the expected third-stage sample size. Therefore, the reduction in second-stage sample is reallocated back to the segments by applying a marginal adjustment to the third-stage sample size (m_{ha}) at the state and age group level. As a result, segments that were not subject to these constraints could be affected. This adjustment to reallocate the dwelling unit sample is iterative until the expected person sample sizes are met.

Note: The optimization procedures implemented for the derivation of D_{hj} assign the larger dwelling unit samples to segments with better response rates. Often such segments are the first to be affected by the sample size constraints. Hence, when forced to reallocate the reduction in dwelling unit sample size to segments with poorer response rates, the overall dwelling unit sample size will increase in nonlinear amounts. In short, segments with worse response rates require more screened dwelling units per completed interview.

3.6 Dwelling Unit Selection and Release Partitioning

After derivation of the required dwelling unit sample size (D_{hj}), the sample is selected from the frame of counted and listed dwelling units for each segment (L_{hj}). The frame is ordered in the same manner as described in Section 3.3.1 and selection is completed using systematic sampling with a random start value.

In order to compensate for quarterly variations in response rates and yields, a sample partitioning procedure was implemented in all quarters. The entire sample (D_{hj}) would still be selected, but only certain percentages of the total would be released into the field. An initial percentage would be released to all segments at the beginning of the quarter and based on interquarter work projections, additional percentages would be released if field staff could handle the added workload. Each partitioning of the sample is a valid sample and helps to control the amount of nonresponse without jeopardizing the validity of the study.

Incidentally, a reserve sample of 10% was also selected, over and above the required D_{hj} sample, to allow for supplemental releases based on state experiences within each quarter. In Quarter 1, the D_{hj} sample was allocated out to states in the following release percentages:

Release 1: 75% of main sample
Release 2: 25% of main sample
Release 3: 100% of reserve sample (10% of main sample)

A summary of the quarterly sample sizes and percents released is provided in Table 3.3.

3.7 Half-Open Interval Rule and Procedure for Adding Dwelling Units

To guarantee that every dwelling unit has a chance of selection and to eliminate any bias associated with incomplete frames, the NHSDA implements a procedure called the half-open interval rule. This procedure requires that the interviewer look both on the property of each selected dwelling unit and between that dwelling unit and the next listed dwelling unit for any unlisted units. When found in these specific locations, the unlisted units become part of the sample (added dwelling units). If the number of added dwelling units linked to any particular sample dwelling unit did not exceed six or if the number for the entire segment was less than or equal to ten, the FI was instructed to consider these dwelling units as part of their assignment. If either of these limits was exceeded, special subsampling procedures were implemented, as described in Appendix B.

3.8 Quarter-by-Quarter Deviations

The following section describes corrections and/or modifications that were implemented in the process of design optimization. *Design* refers to deviations from the original proposed plan of design. *Procedural* refers to changes made in the calculation methodologies. Finally, *Dwelling Unit Selection* will address changes that occurred after sample size derivations. Specifically, corrections implemented during fielding of the sample (i.e., sample partitioning as described in Section 3.6). Quarter 1 deviations are not included since the methods and procedures described above were all implemented in Quarter 1. Subsequently, any changes would have been made after Quarter 1.

Quarter 2

Design: An additional 10% sample was included to allow for supplemental releases where needed.

Procedural: In order to predict state response rates more accurately, an additional quarter of 1999 NHSDA data was used in the computation of response rates by strata. Thus, data from Quarters 3 and 4 were used to compute average dwelling unit eligibility, screening response, and interviewer response rates for the high, medium, and low response rate strata. In addition to incorporating our experience for the computation of response rates, state yield experience was incorporated in Quarter 2 by forming low, medium, and high 12 to 17 year old yield strata and computing average yields within these groupings. The 12 to 17 year old yield strata were defined using data from Quarters 3 and 4 of the 1999 NHSDA as follows:

Stratum	Defining Criteria
1: High 12 to 17 Yield	Yield > 0.242041
2: Medium 12 to 17 Yield	0.197924 < Yield ≤ 0.242041
3: Low 12 to 17 Yield	Yield ≤ 0.197924

Dwelling Unit Selection: Quarter 2 D_{hj} sample was partitioned into the following release percentages:

Release 1: 72% of entire sample (79/110, main sample + 10% reserve)
Release 2: 23% of entire sample (25/110, main sample + 10% reserve)
Release 3: 5% of entire sample (6/110, main sample + 10% reserve)

Unlike Quarter 1, the sample releases were allocated out at the FI Region level.

Quarter 3

Design: Using the completed cases from Quarter 1 and the projected number of completes from Quarter 2, each state's surplus/shortfall was computed. The remaining state sample was then distributed to Quarters 3 and 4 in the ratio of 104% and 92%. An additional 10% sample was also included, therefore bringing the total Quarter 3 sample to the 114.4% level.

Procedural: Similar to Quarter 2, the accumulated state yield and response rate experience was used to predict design parameters. However, an additional quarter of data was added to the pool. Therefore, average yields and response rates were defined by averaging across strata defined by high, medium, and low yield and response rate experience, respectively, using data from Quarters 3 and 4 of 1999 and Quarter 1 of 2000.

Dwelling Unit Selection: The Quarter 3 D_{hj} sample was partitioned into the following release percentages:

Release 1: 100% of main sample
Release 2: 100% of reserve sample (10% of main sample)

Similar to Quarter 1, the sample releases were allocated at the state level.

Quarter 4

Design: The state and age sample sizes were adjusted in order to meet the yearly targets. In order to minimize unequal weighting effects, we required that the Quarter 4 person sample size be enough to achieve an effective sample for the year of 95.5% of each state's target. The new Quarter 4 sample was then equally distributed across the three age groups. Finally, a few age group samples were incremented by 25% where the age-specific effective sample size was less than 80% of the target. An additional 10% sample was also included.

Procedural: As in previous quarters, we used the accumulated state yield and response rate experience to predict design parameters. In Quarter 4, however, enough quality data was available to use state-specific rates rather than averaging across groups of states.

Thus, data from Quarters 3 and 4 of 1999 and Quarters 1 and 2 of 2000 were used to predict state-specific yield and response rates.

Dwelling Unit Selection:

The Quarter 4 D_{hj} sample was partitioned into the following release percentages:

Release 1: 80% of main sample
Release 2: 20% of main sample
Release 3: 100% of reserve sample (10% of main sample).

Similar to Quarter 1, the sample releases were allocated at the state level.

Table 3.3 Quarterly Sample Sizes and Percent Released

State	Quarter 1			Quarter 2		
	# Selected	# Released	Percent	# Selected	# Released	Percent
Total	57,982	52,631	91%	59,686	57,287	96%
Northeast						
Connecticut	750	681	91%	896	896	100%
Maine	695	631	91%	699	699	100%
Massachusetts	878	799	91%	902	902	100%
New Hampshire	697	632	91%	616	616	100%
New Jersey	1179	1071	91%	1407	1292	92%
New York	3002	2727	91%	3583	3383	94%
Pennsylvania	2550	2317	91%	2566	2549	99%
Rhode Island	752	686	91%	896	847	95%
Vermont	698	635	91%	614	614	100%
North Central						
Illinois	3004	2723	91%	3090	3025	98%
Indiana	964	878	91%	1121	1060	95%
Iowa	695	633	91%	700	656	94%
Kansas	750	681	91%	772	730	95%
Michigan	2790	2492	89%	2795	2647	95%
Minnesota	675	616	91%	602	567	94%
Missouri	833	758	91%	735	695	95%
Nebraska	698	633	91%	697	688	99%
North Dakota	638	580	91%	643	605	94%
Ohio	2790	2535	91%	3238	3069	95%
South Dakota	638	578	91%	565	535	95%
Wisconsin	756	685	91%	672	671	100%
South						
Alabama	753	686	91%	899	847	94%
Arkansas	636	576	91%	643	643	100%
Delaware	752	683	91%	775	732	94%
District of Columbia	638	582	91%	746	721	97%
Florida	2797	2538	91%	3228	3190	99%
Georgia	1208	1100	91%	1211	1143	94%
Kentucky	697	636	91%	704	668	95%
Louisiana	733	667	91%	737	696	94%
Maryland	738	673	91%	738	696	94%
Mississippi	631	574	91%	642	605	94%
North Carolina	986	895	91%	993	939	95%
Oklahoma	750	678	90%	776	733	94%
South Carolina	699	635	91%	691	691	100%
Tennessee	903	821	91%	927	874	94%
Texas	3449	3129	91%	3045	2878	95%
Virginia	991	901	91%	1155	1093	95%
West Virginia	637	579	91%	744	742	100%

(continued)

Table 3.3 Quarterly Sample Sizes and Percent Released (continued)

State	Quarter 1			Quarter 2		
	# Selected	# Released	Percent	# Selected	# Released	Percent
West						
Alaska	640	585	91%	564	561	99%
Arizona	742	672	91%	743	703	95%
California	5174	4700	91%	5164	4889	95%
Colorado	698	632	91%	806	800	99%
Hawaii	698	635	91%	695	658	95%
Idaho	697	631	91%	616	582	94%
Montana	637	578	91%	643	611	95%
Nevada	749	683	91%	690	648	94%
New Mexico	638	579	91%	565	565	100%
Oregon	702	637	91%	699	662	95%
Utah	633	574	91%	565	532	94%
Washington	908	825	91%	907	907	100%
Wyoming	636	576	91%	566	532	94%

State	Quarter 3			Quarter 4		
	# Selected	# Released	Percent	# Selected	# Released	Percent
Total Population	62,459	57,975	93%	49,102	45,969	94%
Northeast						
Connecticut	802	730	91%	811	736	91%
Maine	1059	1059	100%	881	799	91%
Massachusetts	1174	1064	91%	913	827	91%
New Hampshire	817	741	91%	621	564	91%
New Jersey	1549	1408	91%	906	825	91%
New York	3052	2773	91%	2416	2415	100%
Pennsylvania	4209	4209	100%	3708	3364	91%
Rhode Island	753	687	91%	491	491	100%
Vermont	771	700	91%	1140	1039	91%
North Central						
Illinois	3149	2861	91%	2567	2328	91%
Indiana	1389	1261	91%	609	609	100%
Iowa	722	656	91%	560	506	90%
Kansas	553	501	91%	394	358	91%
Michigan	2918	2918	100%	2501	2500	100%
Minnesota	686	621	91%	495	447	90%
Missouri	900	899	100%	895	814	91%
Nebraska	655	594	91%	451	408	90%
North Dakota	667	667	100%	613	555	91%
Ohio	2986	2709	91%	2092	1899	91%
South Dakota	636	577	91%	571	519	91%
Wisconsin	1145	1138	99%	1384	1256	91%

(continued)

Table 3.3 Quarterly Sample Sizes and Percent Released (continued)

State	Quarter 3			Quarter 4		
	# Selected	# Released	Percent	# Selected	# Released	Percent
South						
Alabama	867	783	90%	434	434	100%
Arkansas	924	837	91%	825	749	91%
Delaware	642	586	91%	432	432	100%
District of Columbia	1006	910	90%	1926	1748	91%
Florida	2990	2717	91%	2416	2416	100%
Georgia	1281	1164	91%	793	722	91%
Kentucky	977	888	91%	586	586	100%
Louisiana	635	579	91%	594	540	91%
Maryland	745	674	90%	387	353	91%
Mississippi	672	672	100%	595	543	91%
North Carolina	1049	1049	100%	933	846	91%
Oklahoma	688	628	91%	413	373	90%
South Carolina	629	571	91%	547	547	100%
Tennessee	821	745	91%	530	530	100%
Texas	2543	2316	91%	1557	1557	100%
Virginia	1045	950	91%	736	668	91%
West Virginia	987	899	91%	1006	913	91%
West						
Alaska	577	527	91%	568	568	100%
Arizona	581	581	100%	585	583	100%
California	5368	4881	91%	3399	3088	91%
Colorado	626	567	91%	445	445	100%
Hawaii	736	670	91%	417	378	91%
Idaho	626	555	89%	414	414	100%
Montana	557	557	100%	793	793	100%
Nevada	582	530	91%	427	388	91%
New Mexico	927	840	91%	547	498	91%
Oregon	725	657	91%	484	440	91%
Utah	502	453	90%	193	173	90%
Washington	960	871	91%	702	620	88%
Wyoming	599	545	91%	399	365	91%

3.9 Sample Weighting Procedures

At the conclusion of data collection for the last quarter, sample weights were constructed for each quarter of the state-level study that reflected the various stages of sampling described earlier in Section 1.2.2. The calculation of the sampling weights was based on the stratified, three-stage design of the study. Specifically, the person-level sampling weights were the product of the three stagewise sampling weights, each equal to the inverse of the selection probability for that stage. In review, the stages are as follows:

Stage 1: Selection of segment.

Stage 2: Selection of dwelling unit.

Three possible adjustments exist with this stage of selection:
(1) Subsegmentation inflation – by-product of counting and listing,
(2) Added dwelling unit – results from the half-open interval rule, and
(3) Release adjustment.

Stage 3: Selection of person within a dwelling unit.

A total of seven nondesign-based adjustments were necessary for the calculation of the final analysis sample weight. All nondesign-based adjustments were implemented using a generalized exponential modeling technique. These are listed in the order in which they were implemented:

1. *Nonresponse Adjustment at the Dwelling Unit Level.* This is to account for the failure to complete the within-dwelling unit roster. The potential list of variables for the 51 state main study dwelling unit nonresponse modeling is presented in Table 3.4.

2. *Dwelling Unit Level Post-Stratification.* This involves using screener data of demographic information (e.g., age, race, gender, etc.). Dwelling unit weights were adjusted to the intercensal population estimates obtained from the U.S. Bureau of the Census' National Estimates and Projections Branch. In short, explanatory variables used during modeling consisted of counts of eligible persons within each dwelling unit that fell into the various demographic categories. Subsequently, these counts multiplied by the newly adjusted dwelling unit weight and summed across all dwelling units for various domains add to the Census control totals. This adjustment is necessary for the proper calculation of pairwise weights and allows us to achieve greater precision in subsequent adjustments. Screener level potential variables are listed in Table 3.6.

3. *Extreme Weight Treatment at the Dwelling Unit Level.* If it was determined that design-based weights (Stages 1 and 2) along with any of their respective adjustments result in an unsatisfactory unequal weighting effect (i.e., variance between the dwelling unit level weights is too high as well as high frequency of extreme weights), then high weights were properly adjusted. This was implemented by doing another weighting calibration. The control totals are the dwelling unit level post-stratified weights, and the same explanatory variables as in dwelling unit level post-stratification were used so that the extreme weights were controlled and all the distributions in various demographic groups were preserved.

4. *Selected Person Weight Adjustment for Post-Stratification to Roster Data.* This step utilizes control totals derived from the dwelling unit roster that are equal to the Census estimates. This assists in bias reduction and improves precision by taking advantage of the properties of a larger sample size. Selected person sample weights (i.e., those that have been adjusted at the dwelling unit level and account for third stage sampling) were adjusted to the dwelling unit weight sums of all eligible rostered persons. Any demographic information used in modeling is based solely on screener information since this is the only information available for all rostered persons. Potential variables for this adjustment are a combination of the variables presented in Table 3.5.

5. *Person Level Nonresponse Adjustment.* This adjustment allows for the correction of weights resulting from the failure of selected sample persons to complete the interview. Respondent sample weights were adjusted to the total weight sum of all selected persons, adjusted for post-stratification to the eligible roster of persons. Again, demographic information used in modeling is based solely on screener information. Potential variables for this adjustment are a combination of the variables presented in Table 3.5.

6. *Person Level Post-Stratification.* This step is to adjust the final person sample weights to the Census Bureau's estimates. These are the same outside control totals used in the third adjustment. However, explanatory variables for this adjustment are based on questionnaire data, not screener data as in adjustment 3. Data can differ between the two sources. Variables used in modeling are presented in Table 3.6.

7. *Extreme Weight Treatment at the Person Level.* This was implemented in the same manner as described above in adjustment 3 except the weights reflect the third stage of selection.

Table 3.4 Definitions of Levels for Potential Variables for Dwelling Unit Non-Response Adjustment

Group Quarter Indicator
- 1: College Dorm
- 2: Other Group Quarter
- 3: Non-Group Quarter

Percent of Owner-Occupied Dwelling Units in Segment (% Owner)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Percent of Segments That are Black (% Black)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Percent of Segments That are Hispanic (% Hispanic)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Population Density
- 1: MSA $\geq$ 1,000,000
- 2: MSA less than 1,000,000
- 3: Non-MSA urban
- 4: Non-MSA rural

Quarter
- 1: Quarter 1
- 2: Quarter 2
- 3: Quarter 3
- 4: Quarter 4

Segment Combined Median Rent and Housing Value (Rent/Housing)
- 1: First Quintile
- 2: Second Quintile
- 3: Third Quintile
- 4: Fourth Quintile
- 5: Fifth Quintile

States

Interactions among the main effect variables are also considered.

Table 3.5 Definitions of Levels for Potential Variables for Selected Person Post-Stratification and Non-Response Adjustment

Group Quarter Indicator
- 1: College Dorm
- 2: Other Group Quarter
- 3: Non-Group Quarter

Percent of Owner-Occupied Dwelling Units in Segment (% Owner)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Percent of Segments That are Black (% Black)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Percent of Segments That are Hispanic (% Hispanic)
- 1: 0 - <10%
- 2: 10% - <50%
- 3: 50% - 100%

Population Density
- 1: MSA $\geq$ 1,000,000
- 2: MSA less than 1,000,000
- 3: Non-MSA urban
- 4: Non-MSA rural

Quarter
- 1: Quarter 1
- 2: Quarter 2
- 3: Quarter 3
- 4: Quarter 4

Segment Combined Median Rent and Housing Value (Rent/Housing)
- 1: First Quintile
- 2: Second Quintile
- 3: Third Quintile
- 4: Fourth Quintile
- 5: Fifth Quintile

States

Age
- 1: 12-17
- 2: 18-25
- 3: 26-34
- 4: 35-49
- 5: 50+

Gender
- 1: Male
- 2: Female

Table 3.5 Definitions of Levels for Potential Variables for Selected Person Post-Stratification and Non-Response Adjustment (continued)

Hispanicity
1: Hispanic
2: Non-Hispanic
Race
1: White
2: Black
3: Indian / Native American
4: Asian
Relation to Householder
1: Householder or Spouse
2: Child
3: Other Relative
4: Non-Relative

Interactions among the main effect variables are also considered.

Table 3.6 Definitions of Levels for Potential Variables for Dwelling Unit Post-Stratification and Respondent Post-Stratification at the Person Level

Age
- 1: 12-17
- 2: 18-25
- 3: 26-34
- 4: 35-49
- 5: 50+

Gender
- 1: Male
- 2: Female

Hispanicity
- 1: Hispanic
- 2: Non-Hispanic

Quarter
- 1: Quarter 1
- 2: Quarter 2
- 3: Quarter 3
- 4: Quarter 4

Race
- 1: White
- 2: Black
- 3: Indian / Native American
- 4: Asian

State

Interactions among the main effect variables are also considered.

All adjustments for the 2000 main study final analysis weights were done with exponential adjustment factors derived from modeling the response variable with a generalized exponential regression formula. To help reduce computational burden at all adjustment steps, separate models were fit for clusters of states, based on Census Region Division definitions as shown in Table 3.7. Furthermore, model variable selection at each adjustment was done using a forward insertion process. The final adjusted weight, which is the product of weight components 1 through 14, is the analysis weight used in estimation. Table 3.8 presents a flowchart of steps used in the weighting process and Table 3.9 displays all individual weight components.

Table 3.7 Model Group Definitions

Model	Defined State
1	Connecticut, Maine, New Hampshire, Rhode Island, Vermont, Massachusetts
2	New Jersey, New York, Pennsylvania
3	Illinois, Indiana, Michigan, Wisconsin, Ohio
4	Indiana, Kansas, Minnesota, Missouri, Nebraska, South Dakota, North Dakota
5	Delaware, District of Columbia, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia
6	Alabama, Kentucky, Mississippi, Tennessee
7	Arkansas, Louisiana, Oklahoma, Texas
8	Colorado, Iowa, Montana, Nevada, New Mexico, Utah, Wyoming, Arizona
9	Alaska, Hawaii, Oregon, Washington, California

Table 3.8 Flowchart of Sample Weighting Steps

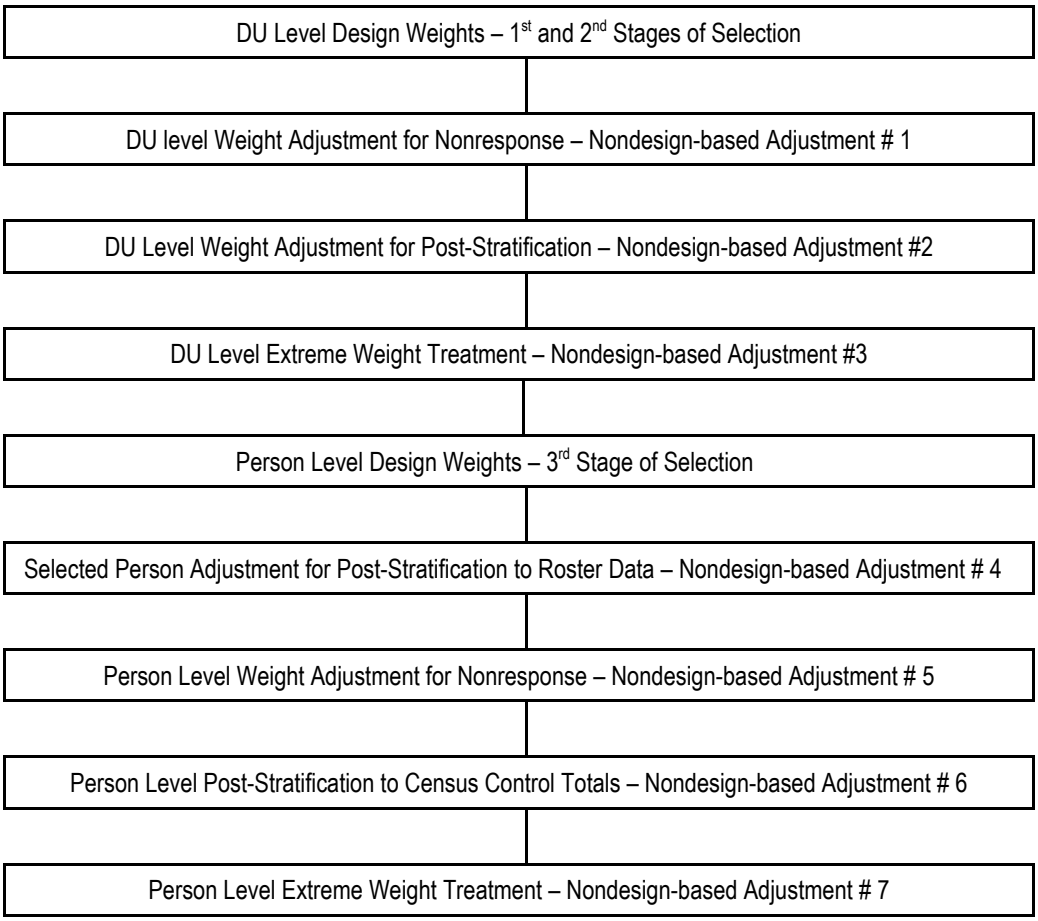


Table 3.9 Sample Weight Components

DU Level Design Weight Components
1 Inverse Probability of Selecting Segment
2 Quarter Segment Weight Adjustment
3 Subsegmentation Inflation Adjustment
4 Inverse Probability of Selecting Dwelling Unit
5 Inverse Probability of Added Dwelling Unit
6 Dwelling Unit Release Adjustment
7 Dwelling Unit Nonresponse Adjustment
8 Dwelling Unit Post-Stratification
9 Dwelling Unit Extreme Weight Treatment
Person Level Design Weight Components
10 Inverse Probability of Selecting a Person Within a Dwelling Unit
11 Roster Adjustment
12 Person Level Nonresponse Adjustment
13 Person Level Post-Stratification Adjustment
14 Person Level Extreme Weight Treatment

Full details of the finalized modeling procedures, as well as final variables used in each adjustment step, can be found in the Sampling Weight Calibration for the 2000 NHSDA (Chen, 2002).

References

Chen, P., Emrich, S., Gordek, H., Singh, A., & Westlake, M. (2002). Sampling weight calibration for the 2000 NHSDA. In *2000 National Household Survey on Drug Abuse: Methodological resource book* (prepared for the Substance Abuse and Mental Health Services Administration, Contract No. 283-98-9008). Research Triangle Park, NC: Research Triangle Institute.

Chromy, J.R. (1987). Design optimization with multiple objectives. In *Proceedings of the American Statistical Association, Survey Research Methods Sections*, pp. 194-199. Washington, DC: American Statistical Association.

Chromy, J. R. (1981). Variance estimates for a sequential sample selection procedure. In D. Krewski, R. Platek, & J.N.K. Rao (Eds.), *Current topics in survey sampling*. *Proceedings of he symposium held May 7-9, 1980, sponsored by the Ottawa Chapter and the Survey Research Methods Section of the American Statistical Association* (pp. 329-347). New York: Academic Press.

Chromy, J. R. (1979). Sequential sample selection methods. In *Proceedings of the American Statistical Association, Survey Research Methods Section* (pp. 401-406). Alexandria, VA: American Statistical Association.

Cochran, W. (1977). *Sampling techniques* (3rd ed.) (pp. 261-263). New York: John Wiley & Sons.

Kish, L. (1965). *Survey sampling* (pp. 313-315). New York: John Wiley & Sons.

RTI (1996). *Counting and listing supplement for subsegmenting*. Research Triangle Park, NC: Author.

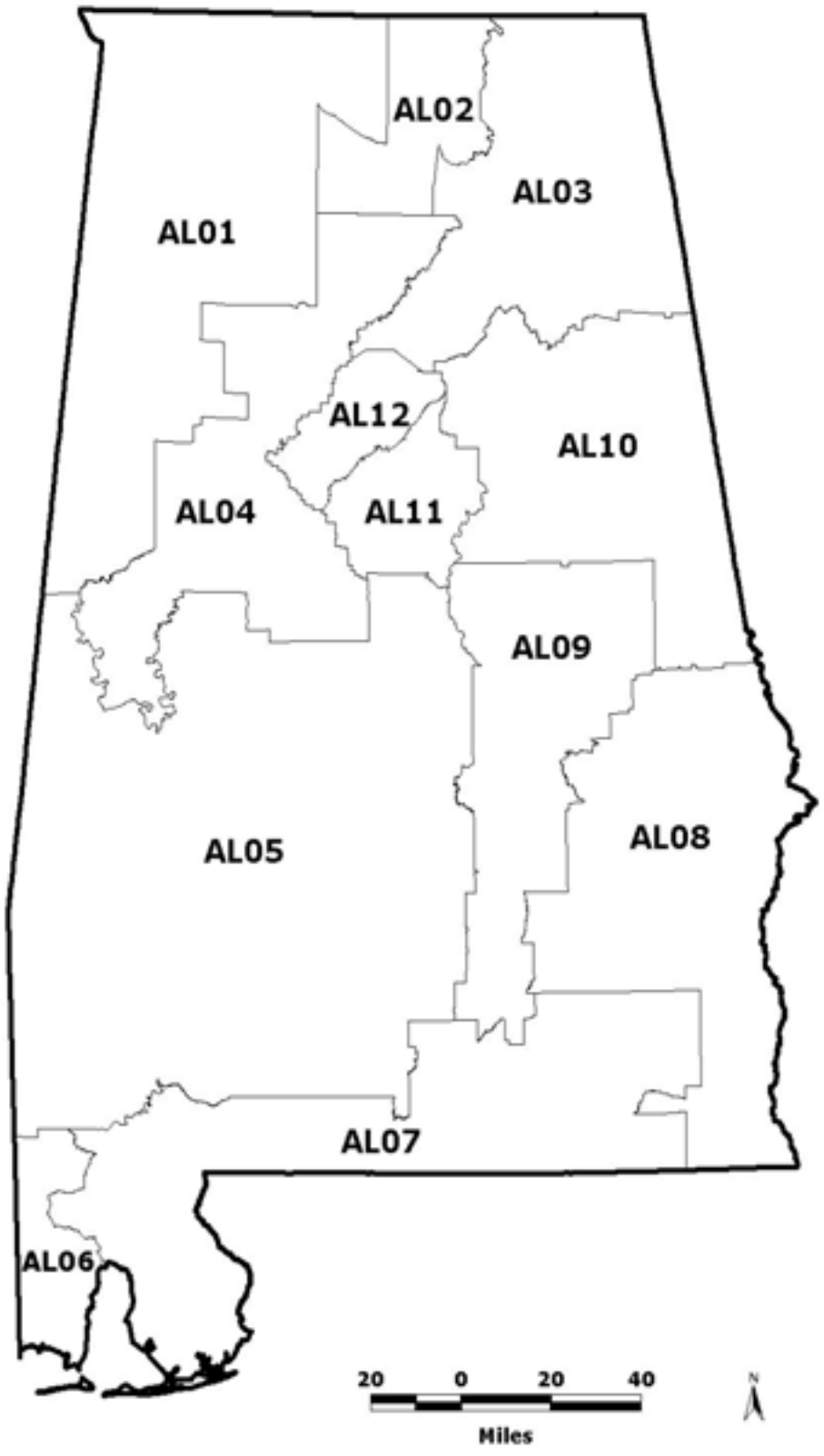
RTI (1999). *Counting and listing general manual*. Research Triangle Park, NC: Author.

RTI (2001). *SUDAAN user's manual, release 8.0*. Research Triangle Park, NC: Author.

Appendix A

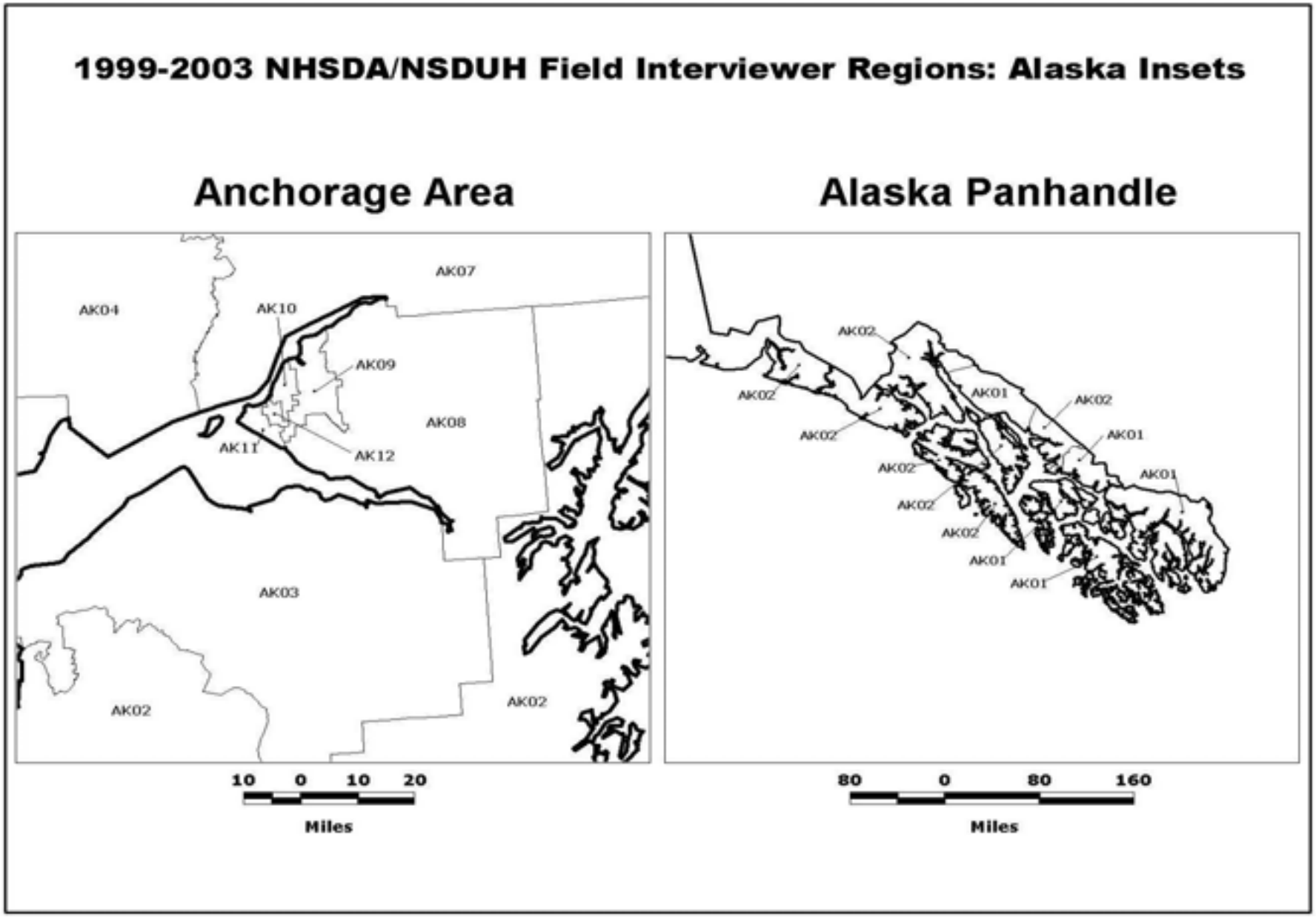
1999-2003 NHSDA Field Interviewer Regions

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Alabama

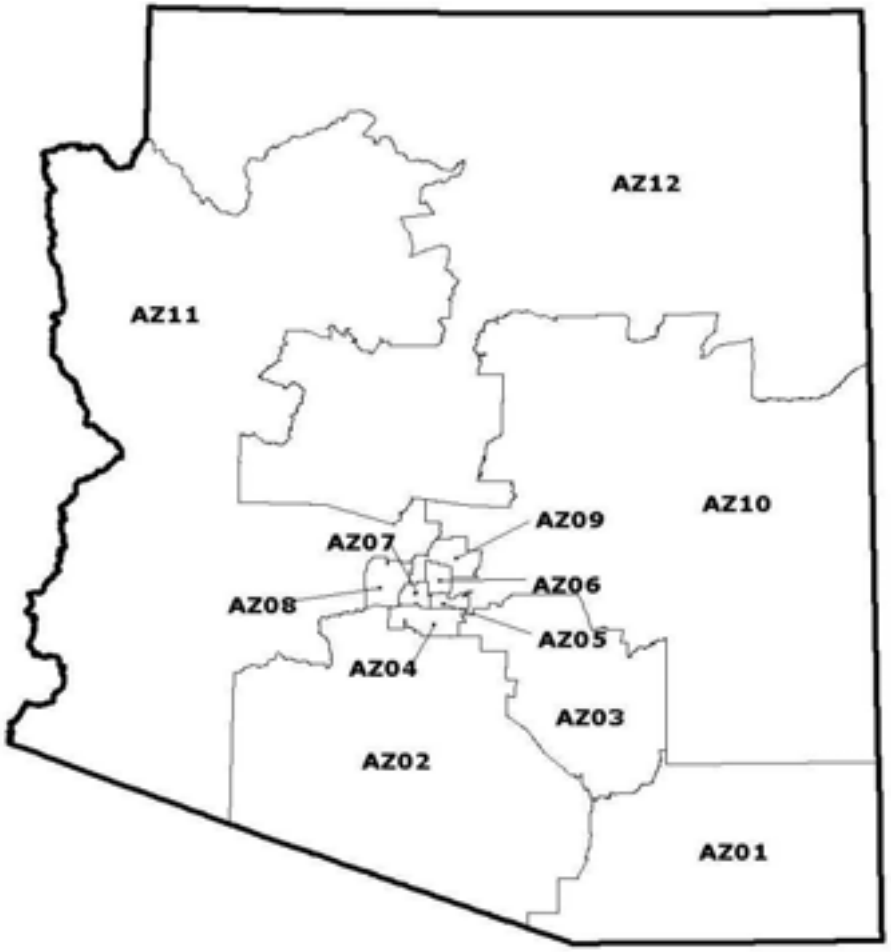


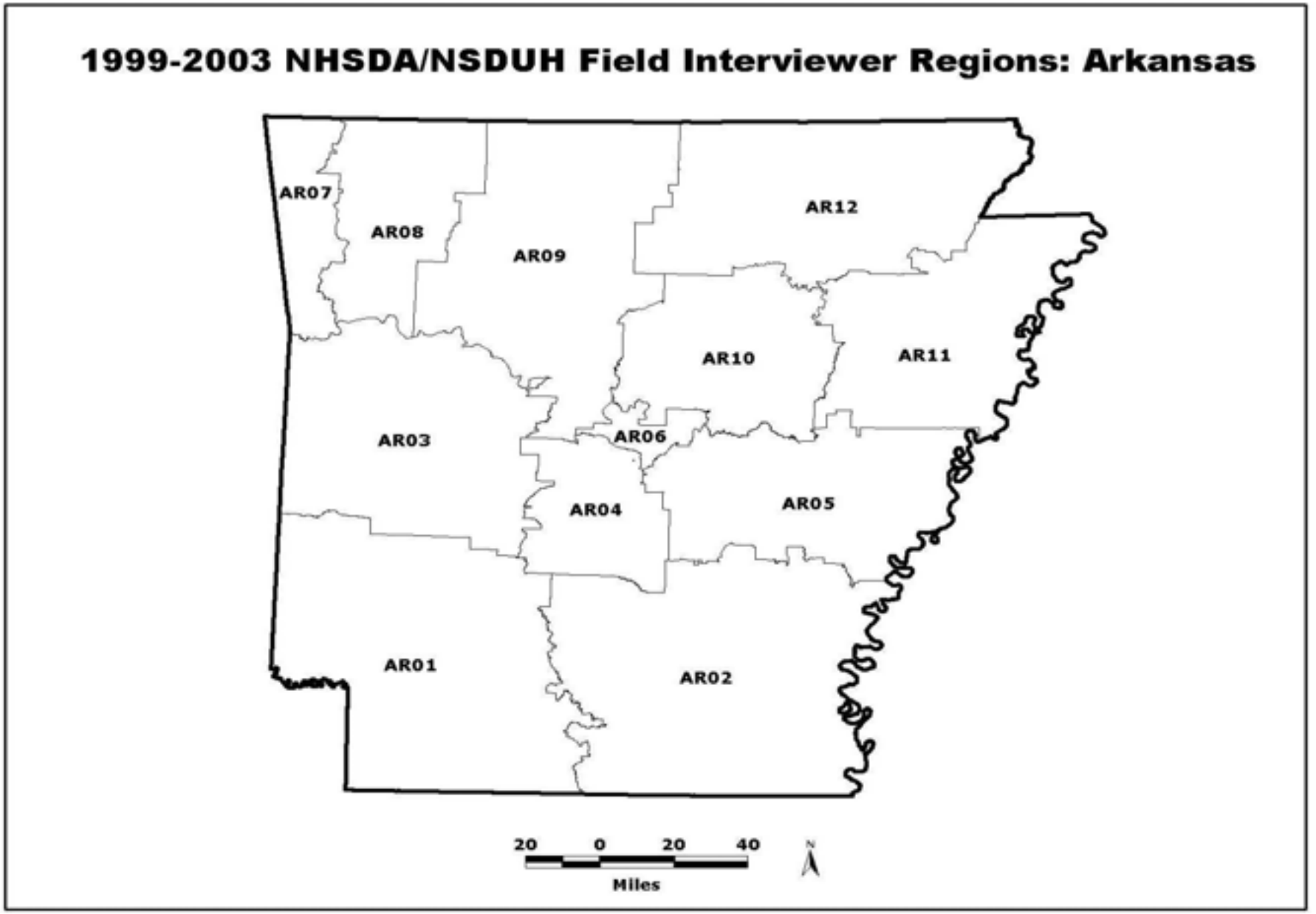

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Alaska
AK04
AK04
AK05
AK06
AK07
AK03
AK02
AK04
See inset next page
AK03
AK03
AK02
See inset next page
AK03
AK03
AK03
100 0 100 200
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Alaska Insets

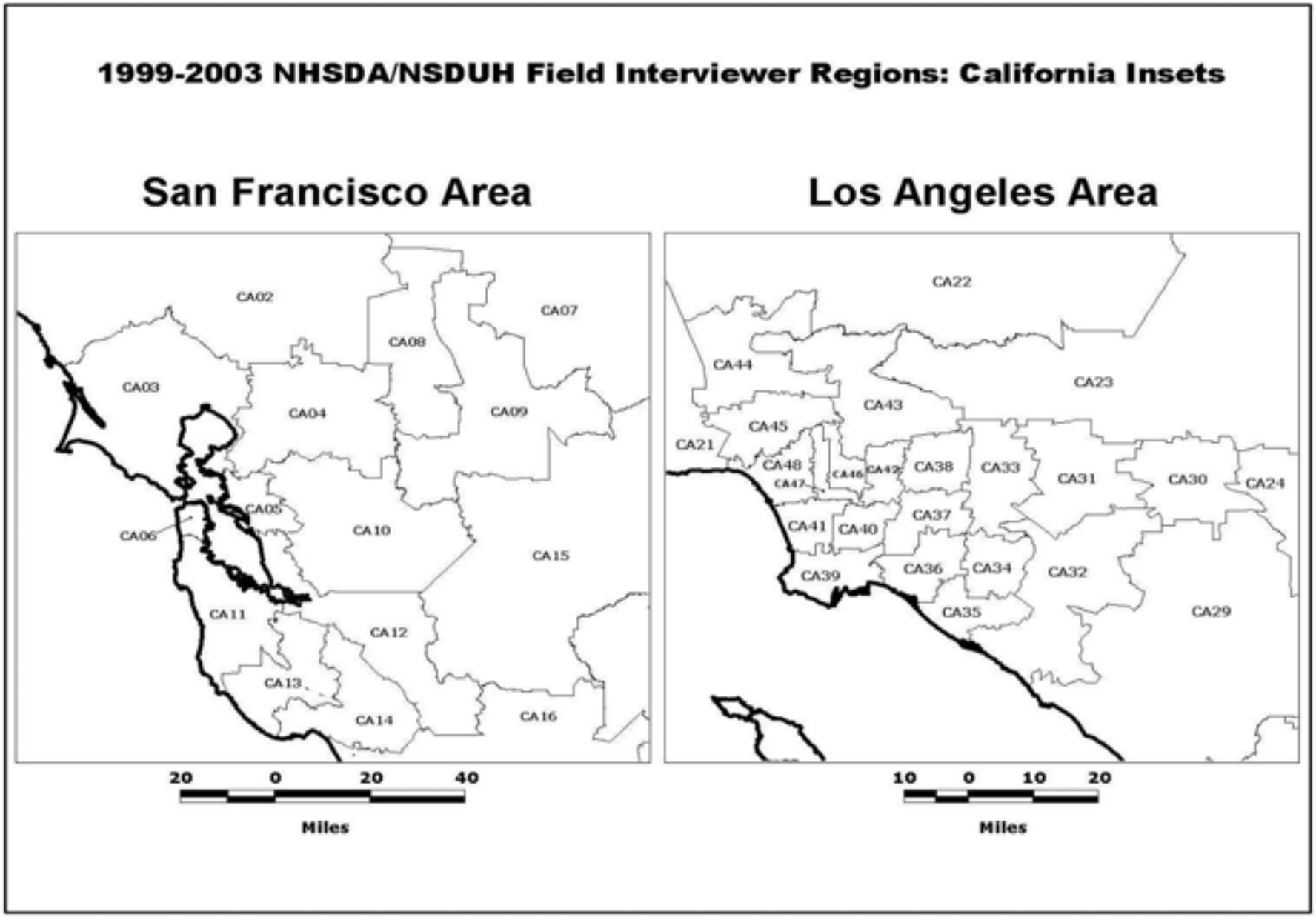


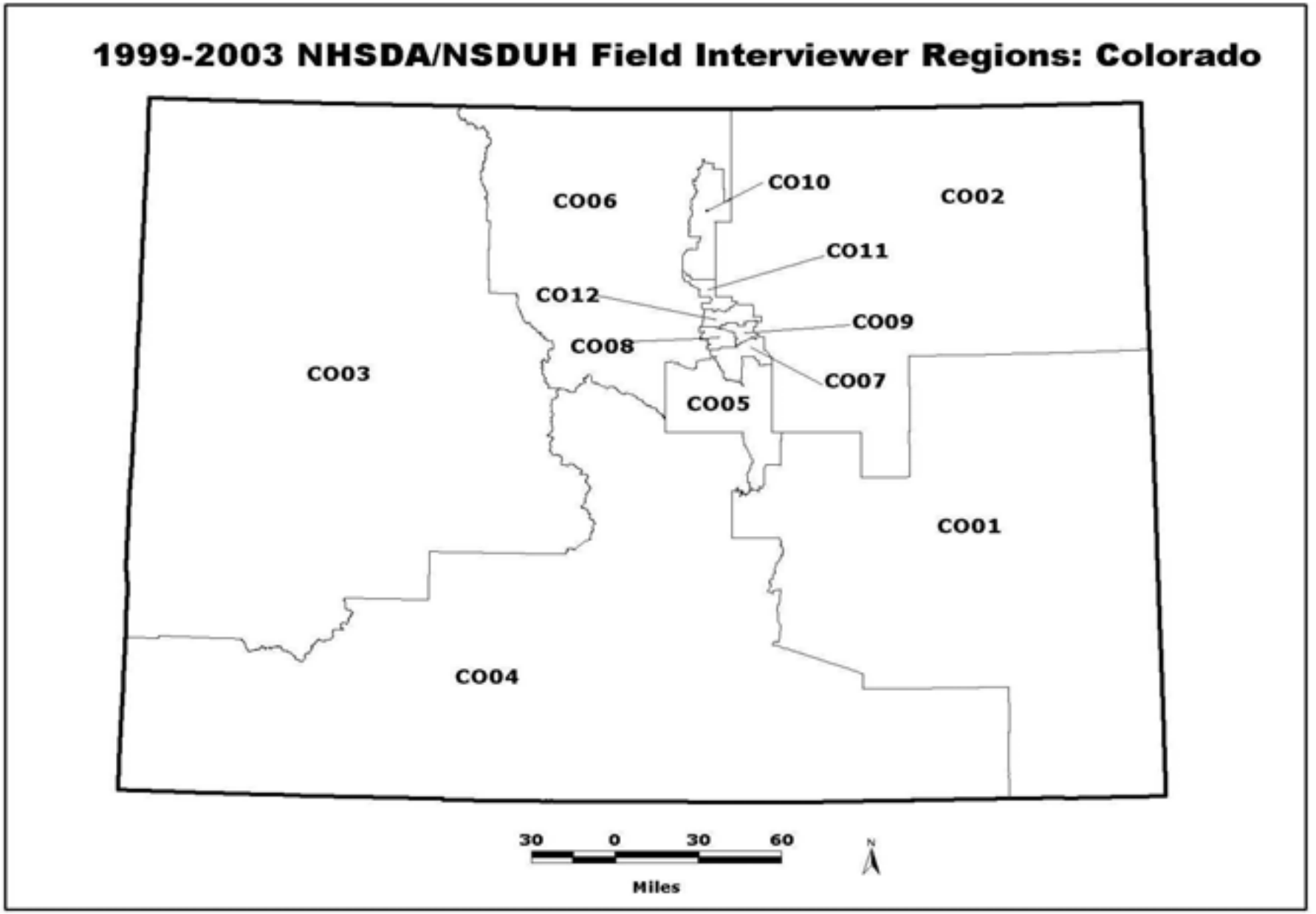
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Arizona

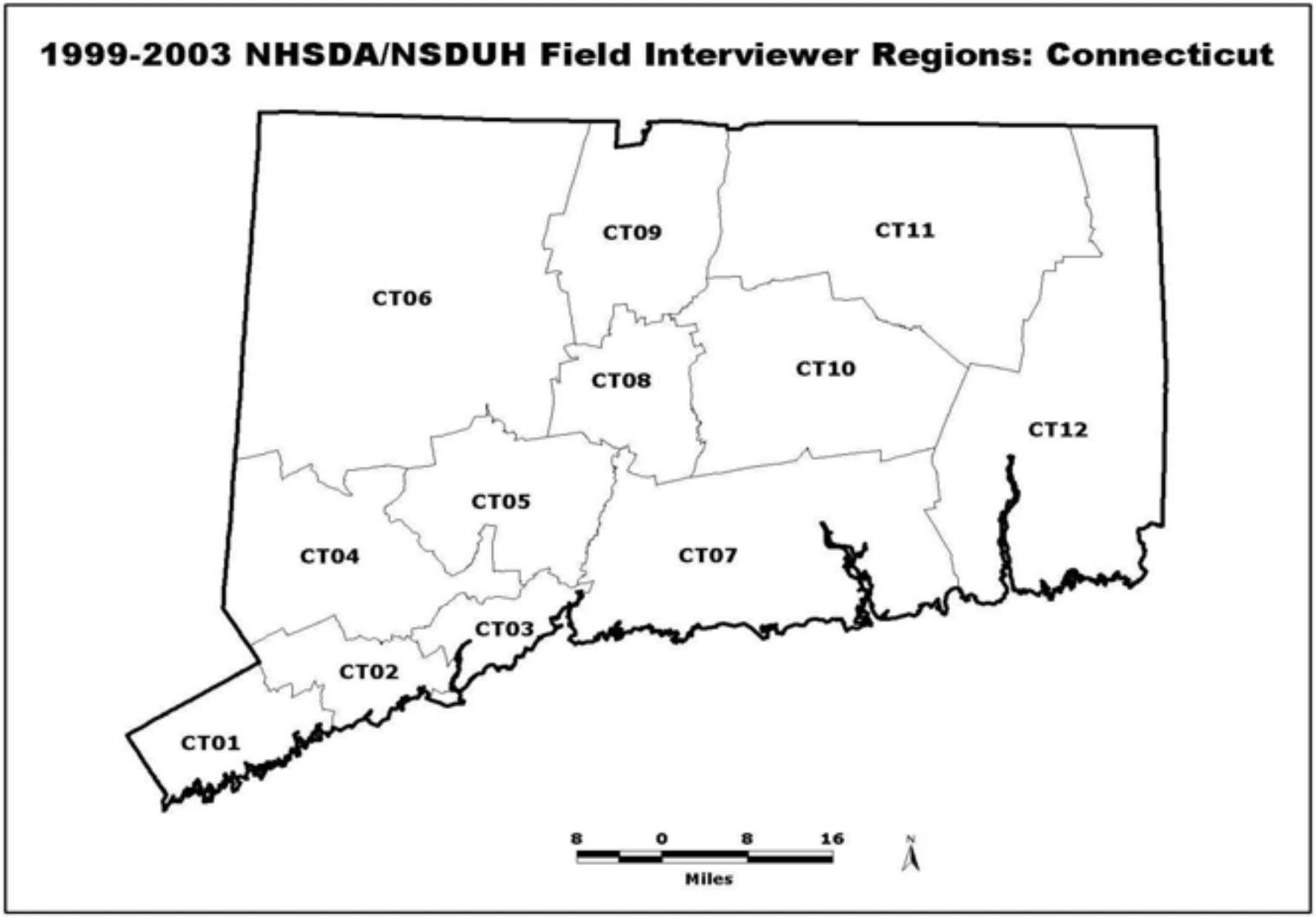



1999-2003 NHSDA/NSDUH Field Interviewer Regions: Arkansas
AR07
AR08
AR09
AR12
AR10
AR11
AR03
AR06
AR04
AR05
AR01
AR02
20
0
20
40
Miles
N

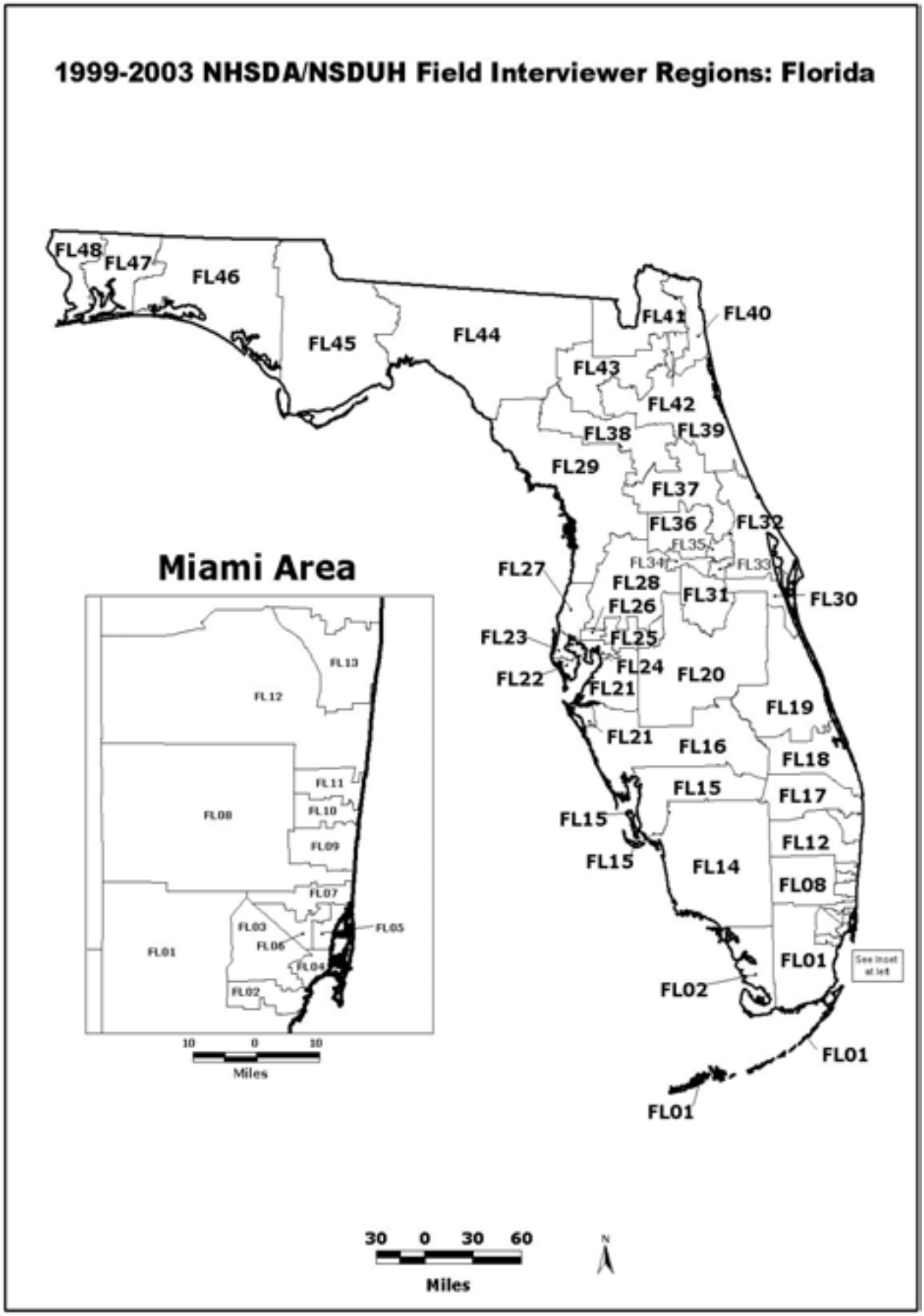
1999-2003 NHSDA/NSDUH Field Interviewer Regions: California


1999-2003 NHSDA/NSDUH Field Interviewer Regions: California Insets
San Francisco Area
Los Angeles Area
CA02
CA07
CA08
CA03
CA04
CA09
CA05
CA06
CA10
CA15
CA11
CA12
CA13
CA14
CA16
20 0 20 40
Miles
CA22
CA44
CA23
CA43
CA45
CA21
CA48
CA46
CA42
CA38
CA33
CA31
CA30
CA24
CA47
CA37
CA41
CA40
CA39
CA36
CA34
CA32
CA35
CA29
10 0 10 20
Miles

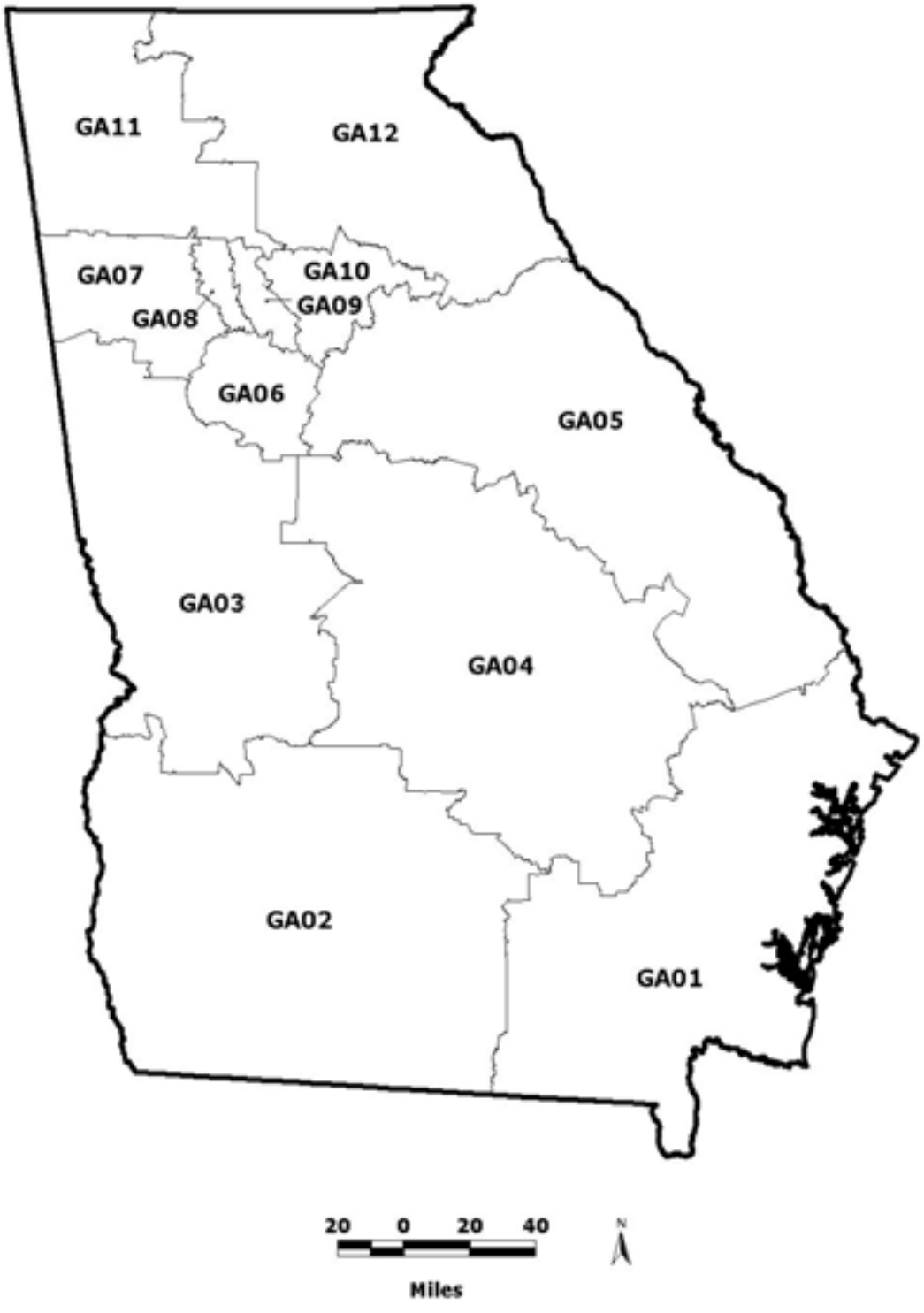

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Colorado
CO10
CO06
CO02
CO11
CO12
CO09
CO08
CO03
CO07
CO05
CO01
CO04
30
0
30
60
Miles
N

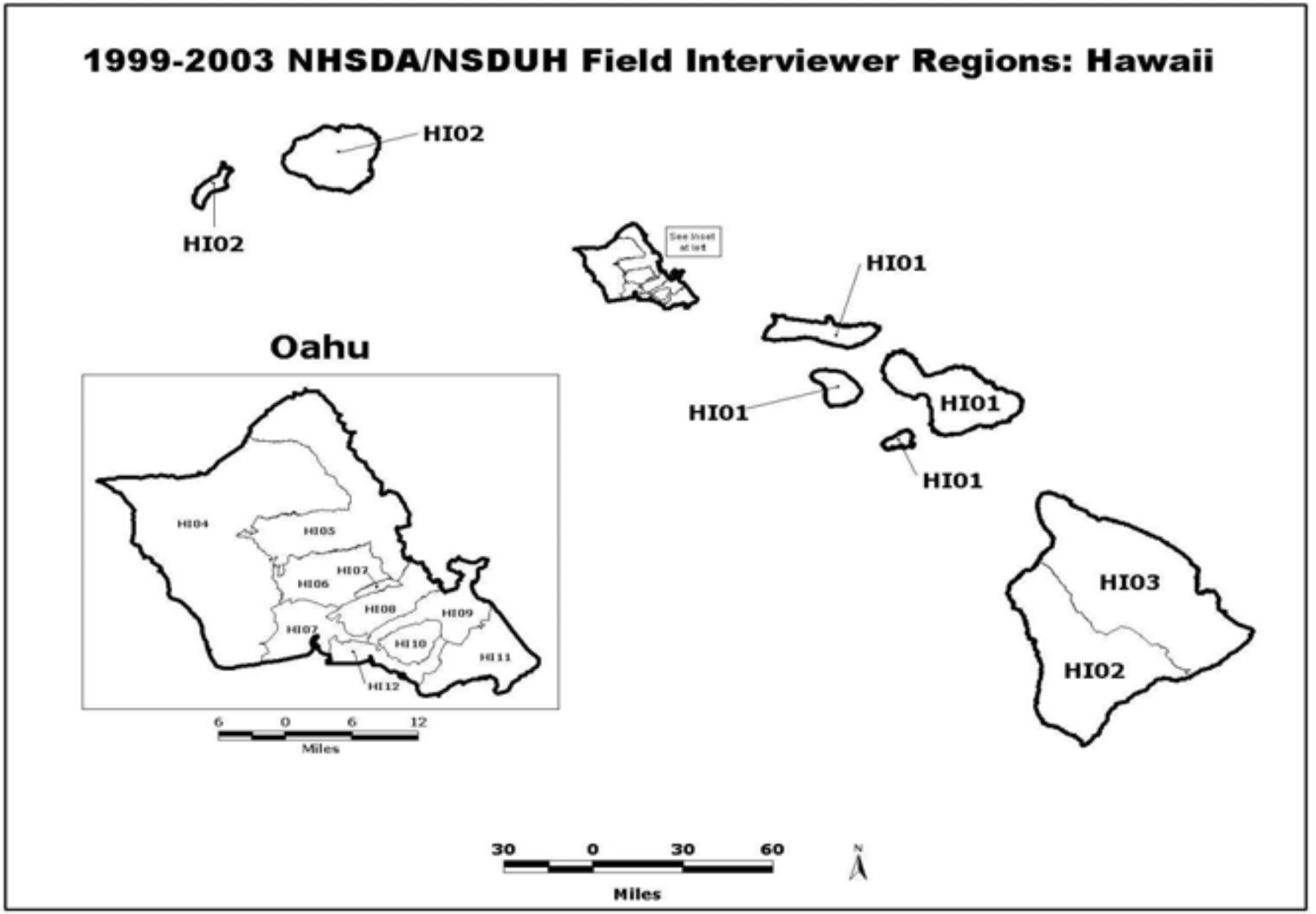

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Connecticut
CT09
CT11
CT06
CT08
CT10
CT12
CT05
CT04
CT07
CT03
CT02
CT01
8 0 8 16
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Delaware

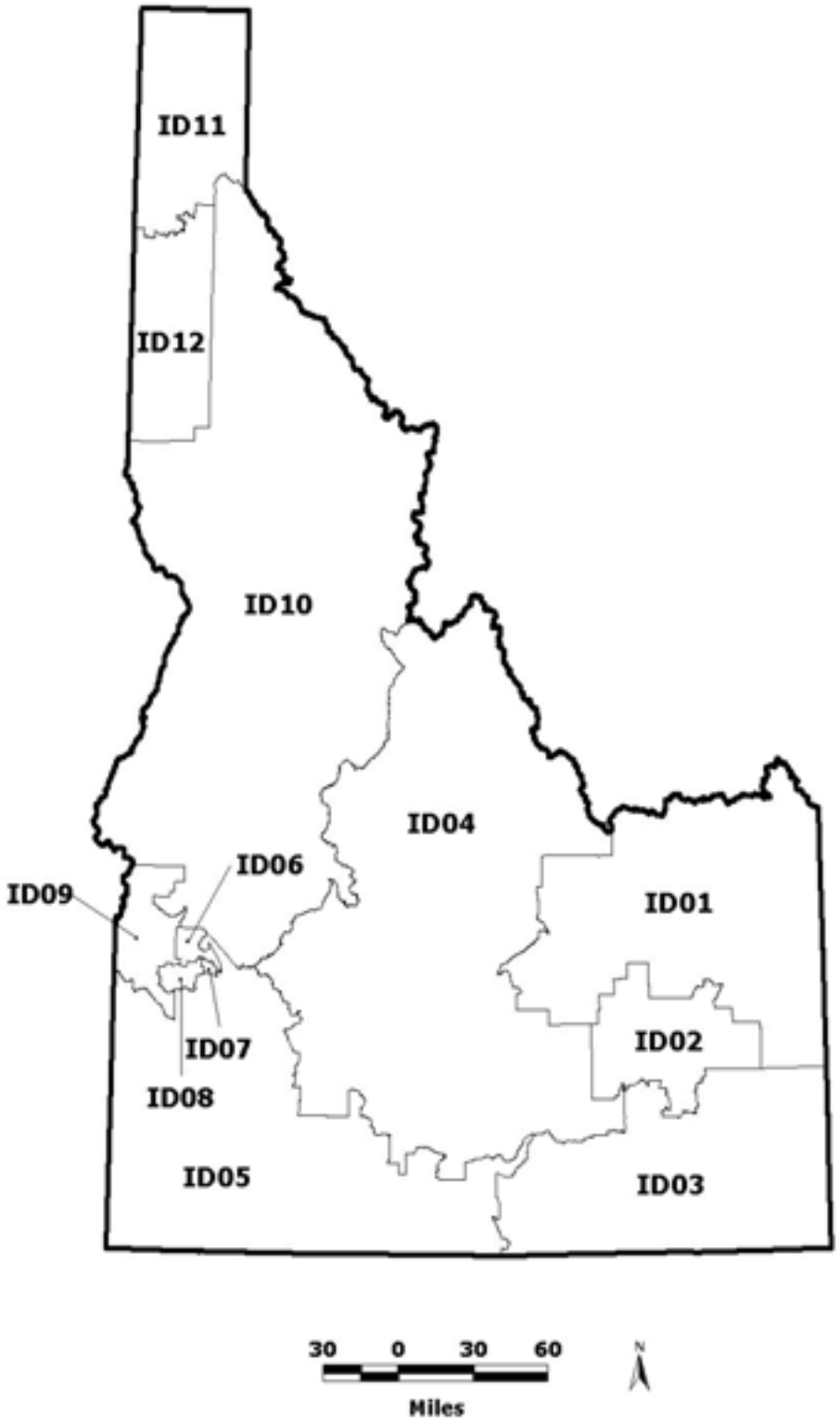

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Florida
FL48
FL47
FL46
FL45
FL44
FL41
FL40
FL43
FL42
FL38
FL39
FL29
FL37
FL36
FL32
FL35
FL34
FL33
FL27
FL28
FL31
FL30
FL26
FL23
FL25
FL24
FL20
FL22
FL21
FL19
FL21
FL16
FL18
FL15
FL17
FL15
FL12
FL15
FL14
FL08
FL01
FL02
FL01
FL01
See inset at left
30 0 30 60
Miles
N
Miami Area
FL13
FL12
FL11
FL08
FL10
FL09
FL07
FL03
FL05
FL06
FL01
FL04
FL02
10 0 10
Miles

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Georgia

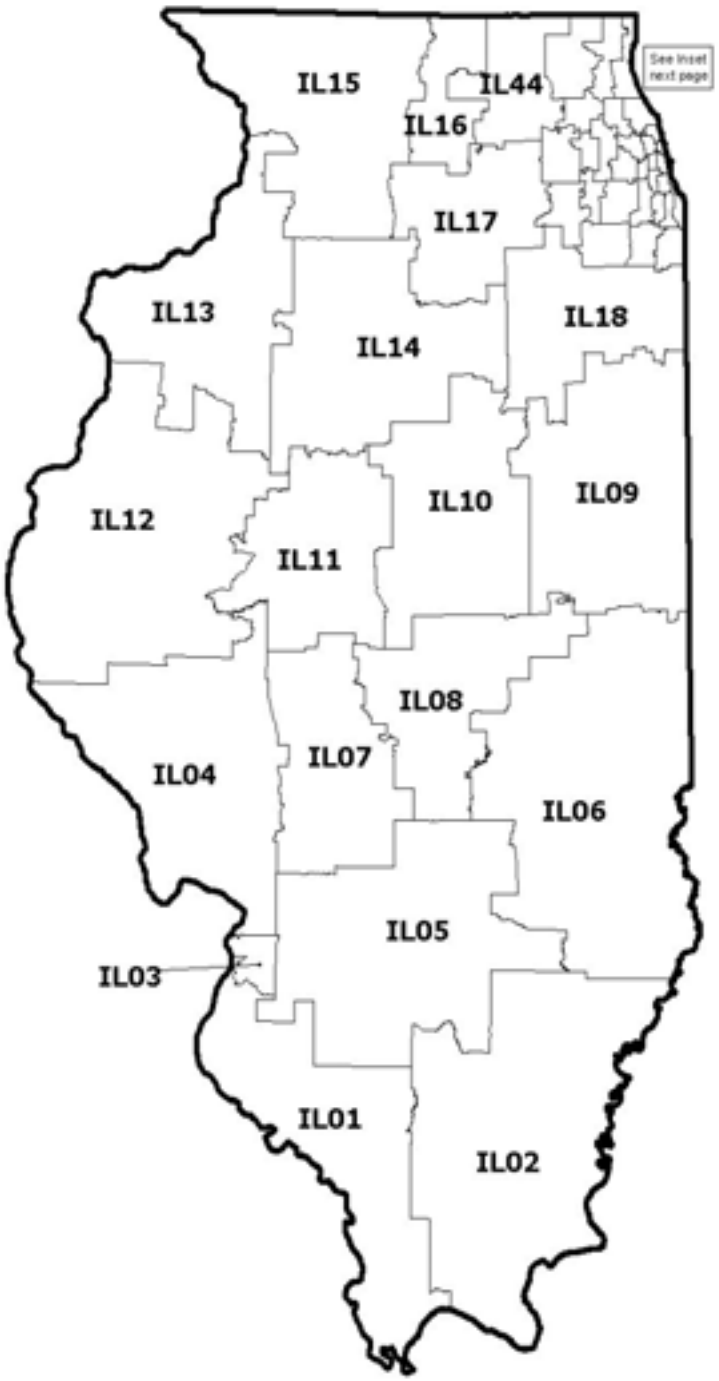



1999-2003 NHSDA/NSDUH Field Interviewer Regions: Hawaii
HI02
HI02
See Inset at left
HI01
HI01
HI01
HI01
Oahu
HI04
HI05
HI07
HI06
HI08
HI09
HI07
HI10
HI11
HI12
6 0 6 12
Miles
HI03
HI02
30 0 30 60
Miles
N

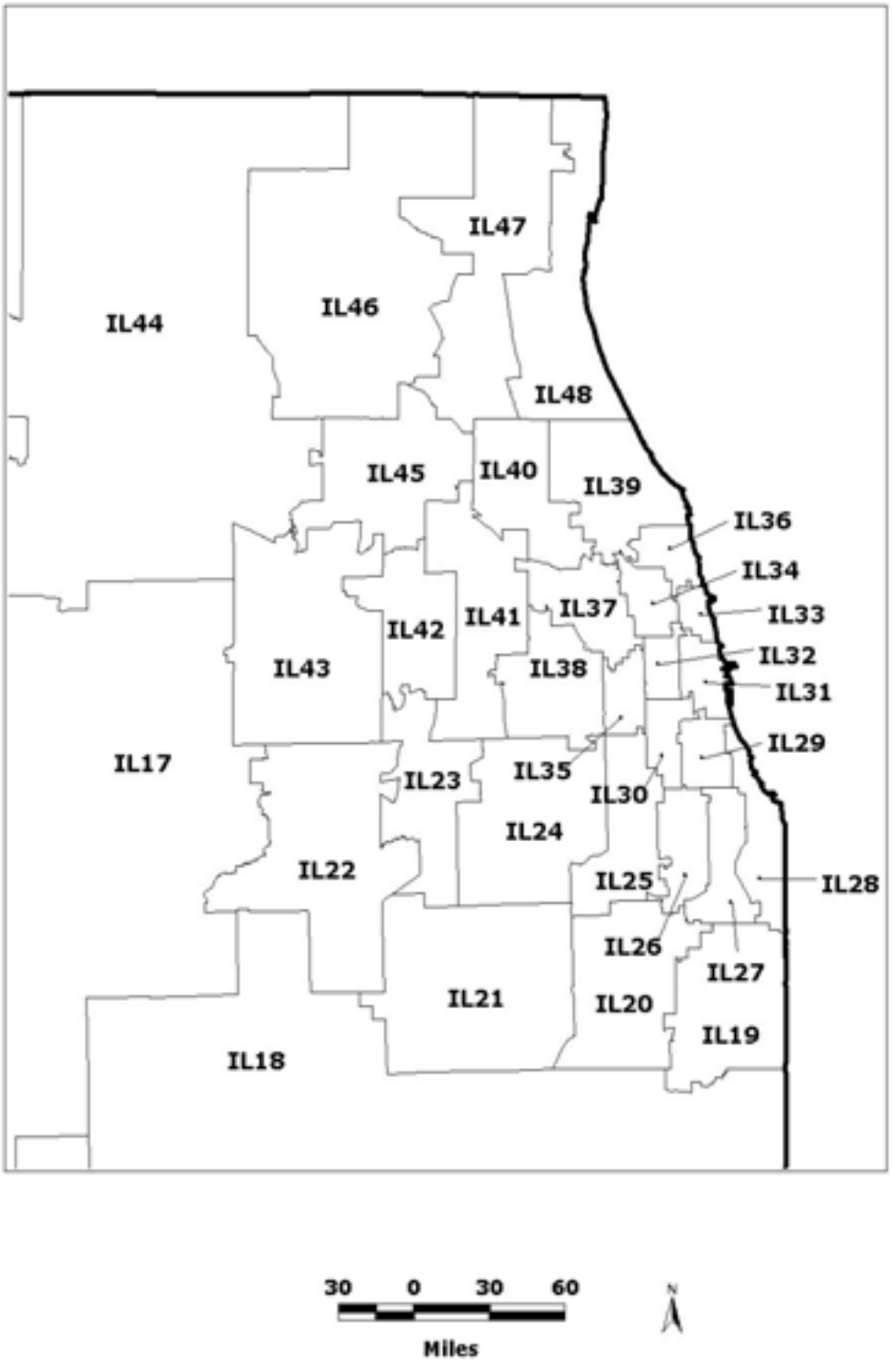
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Idaho



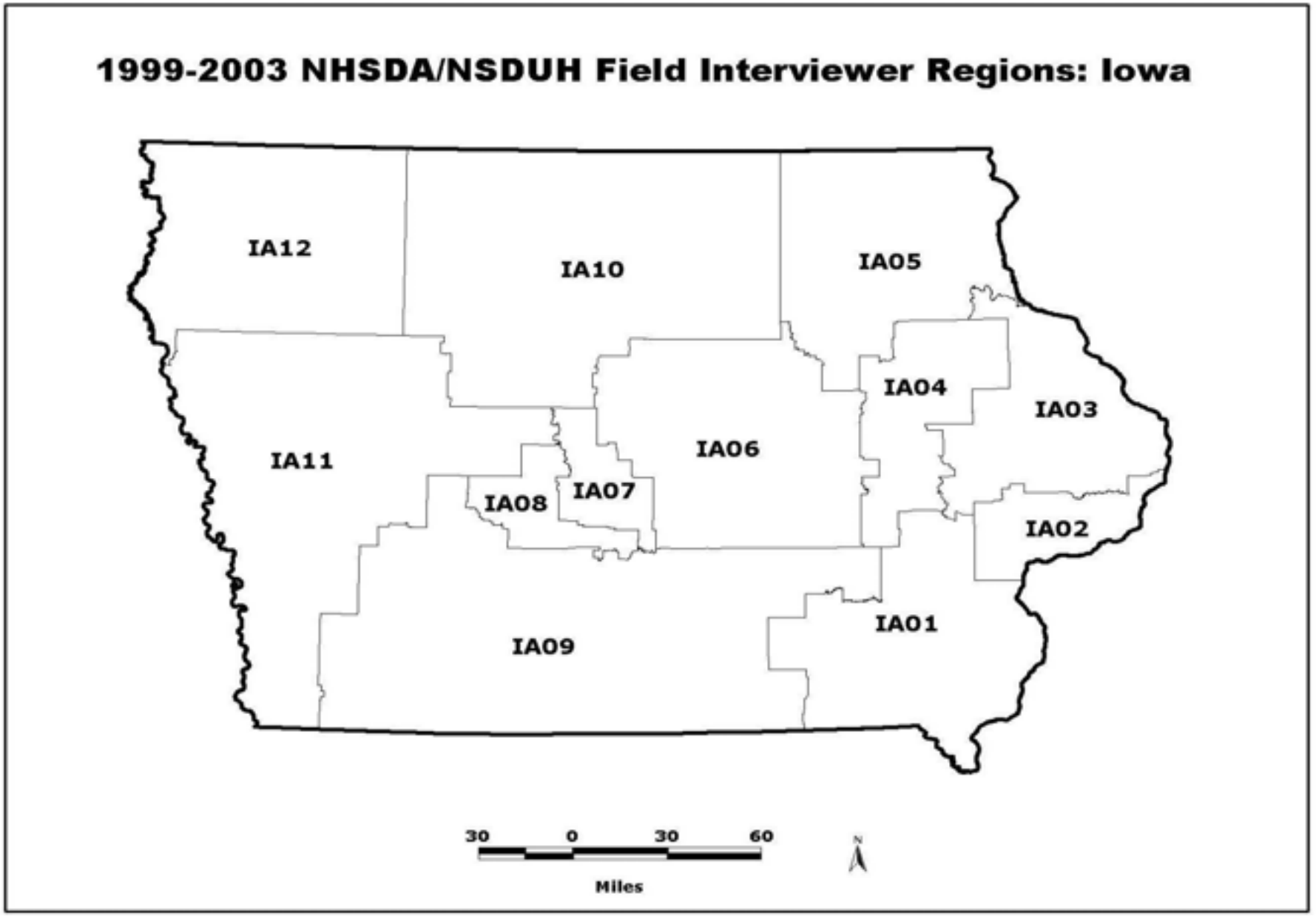
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Illinois



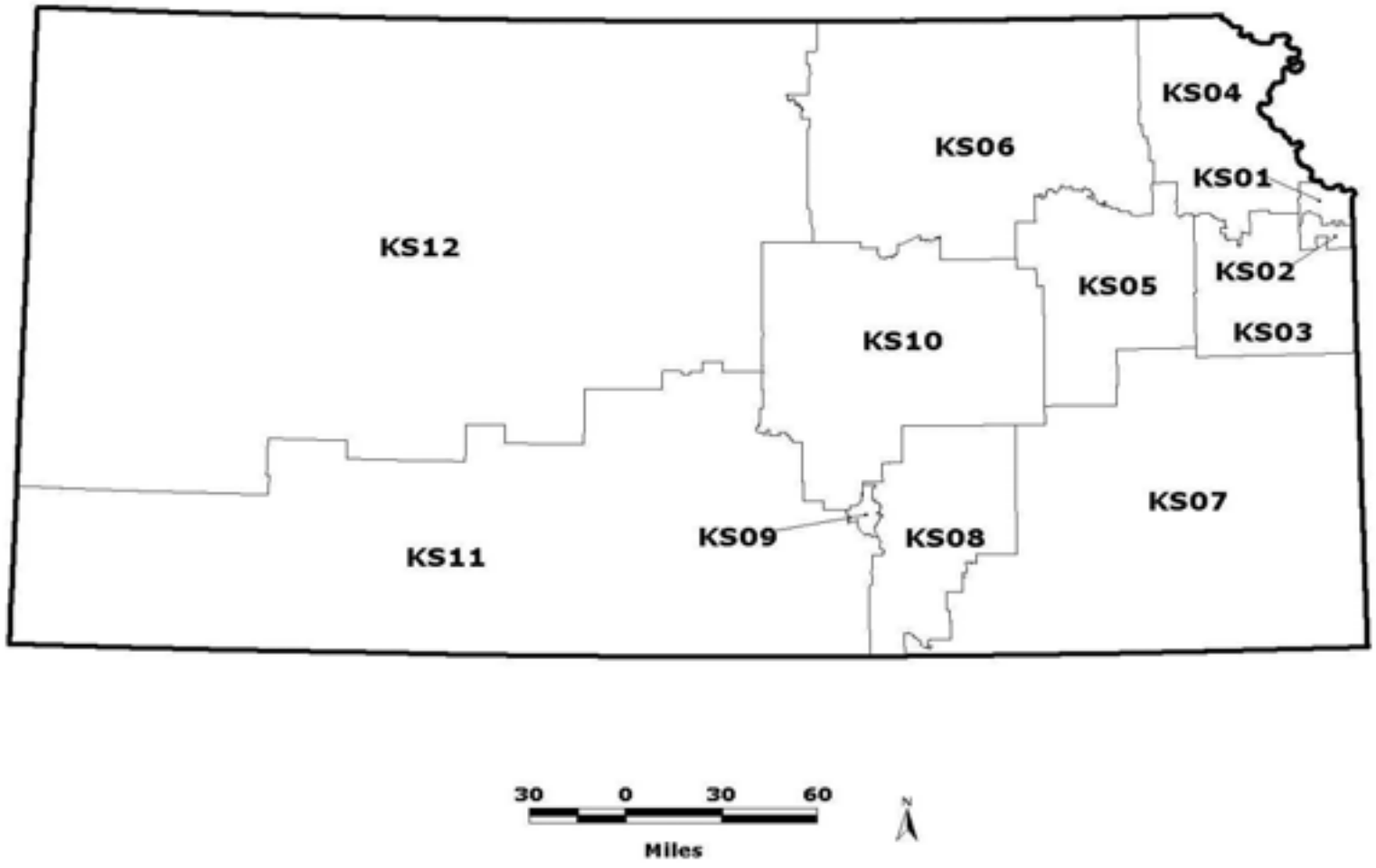
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Chicago Area

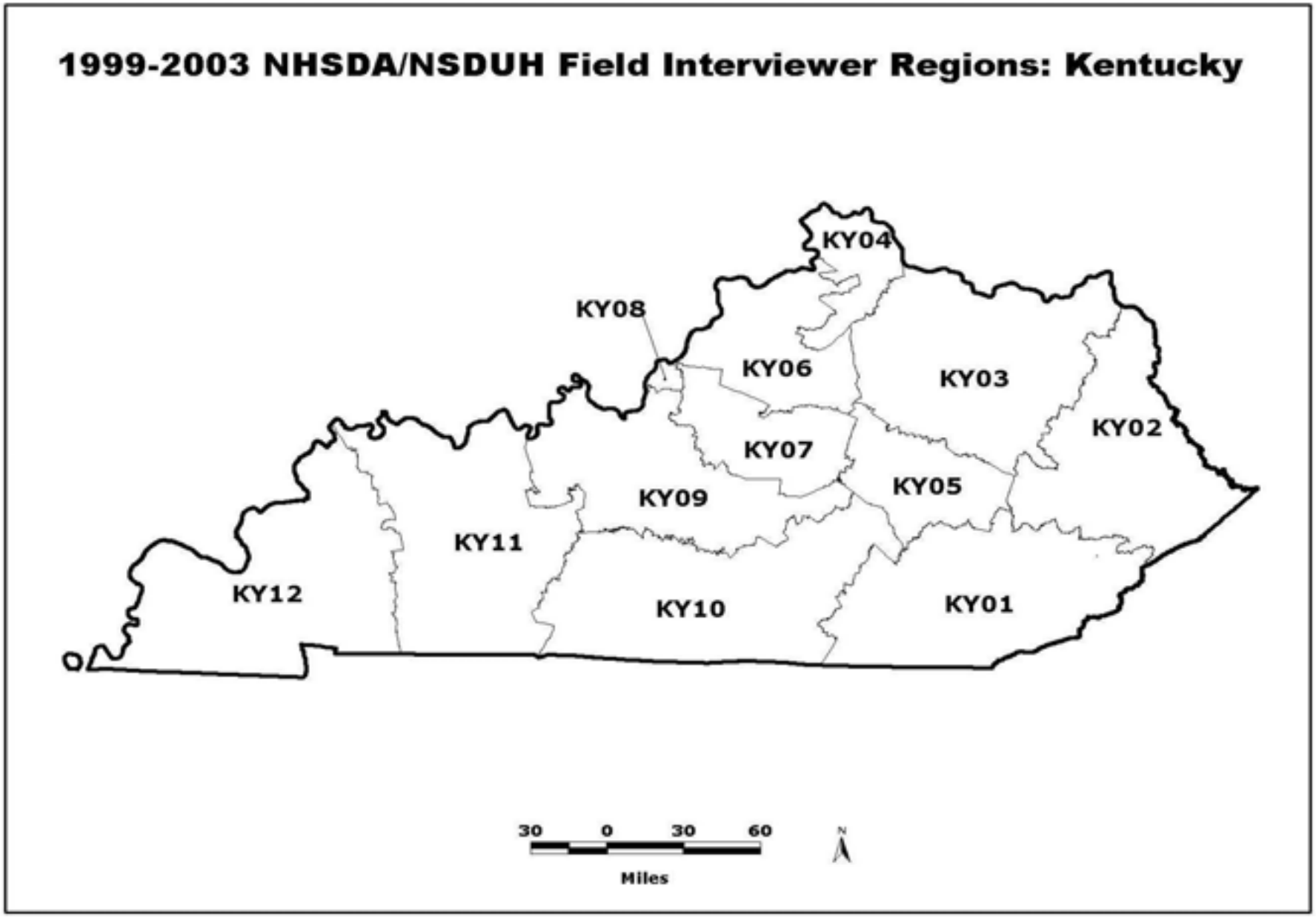


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Indiana

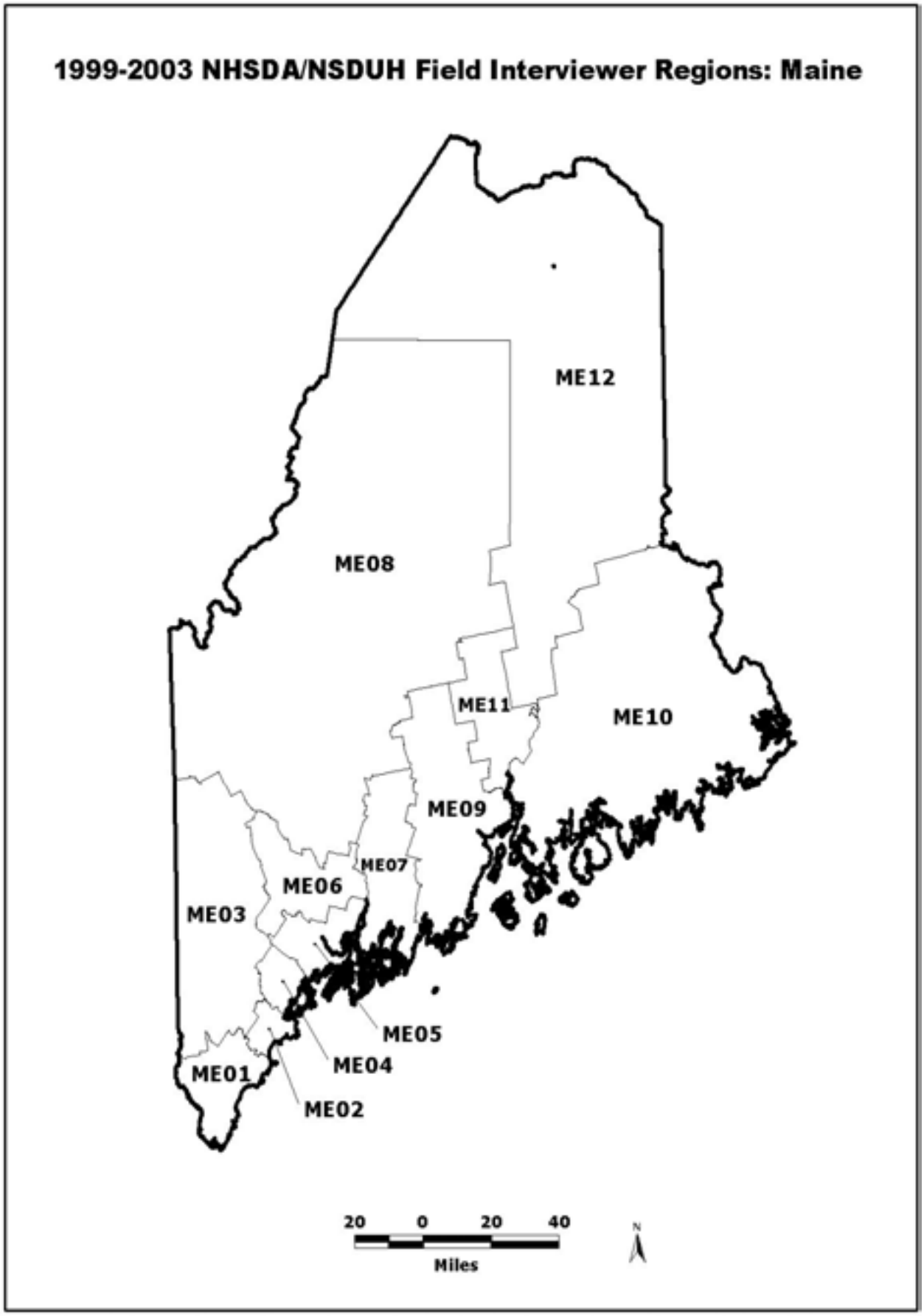

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Iowa
IA12
IA10
IA05
IA04
IA03
IA11
IA06
IA08
IA07
IA02
IA01
IA09
30
0
30
60
Miles
N

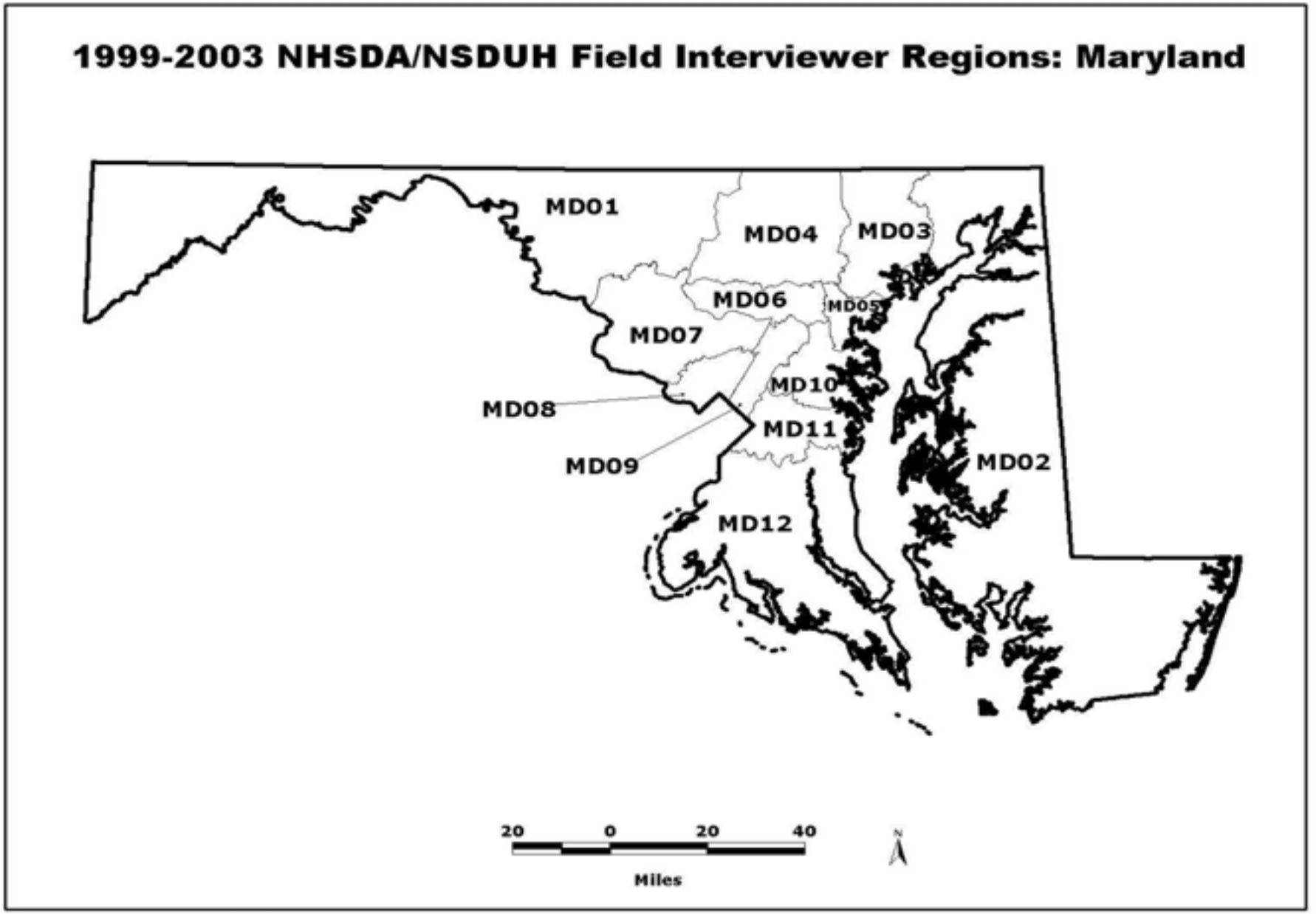
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Kansas

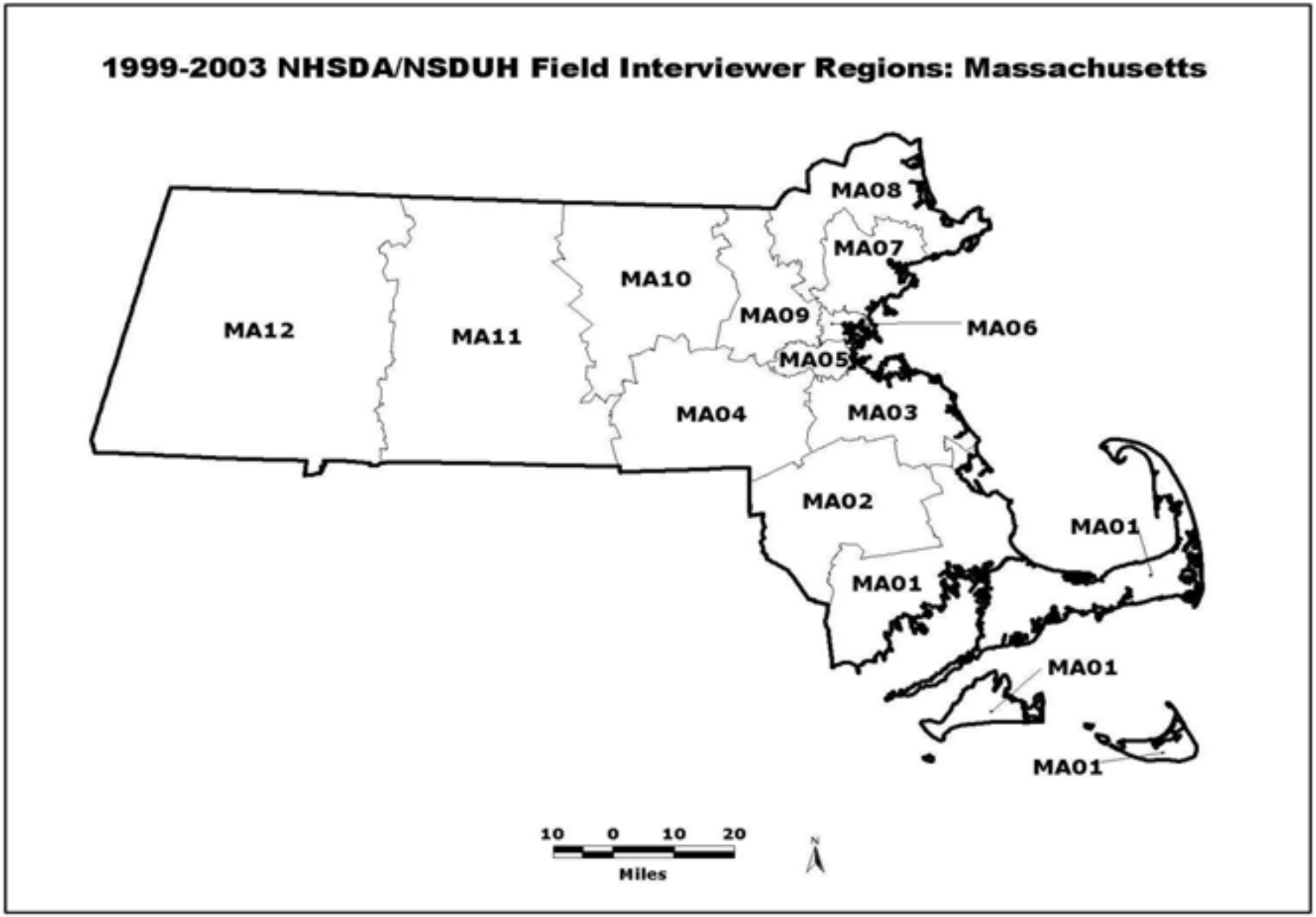



1999-2003 NHSDA/NSDUH Field Interviewer Regions: Kentucky
KY04
KY08
KY06
KY03
KY02
KY07
KY05
KY09
KY11
KY12
KY10
KY01
30 0 30 60
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Louisiana

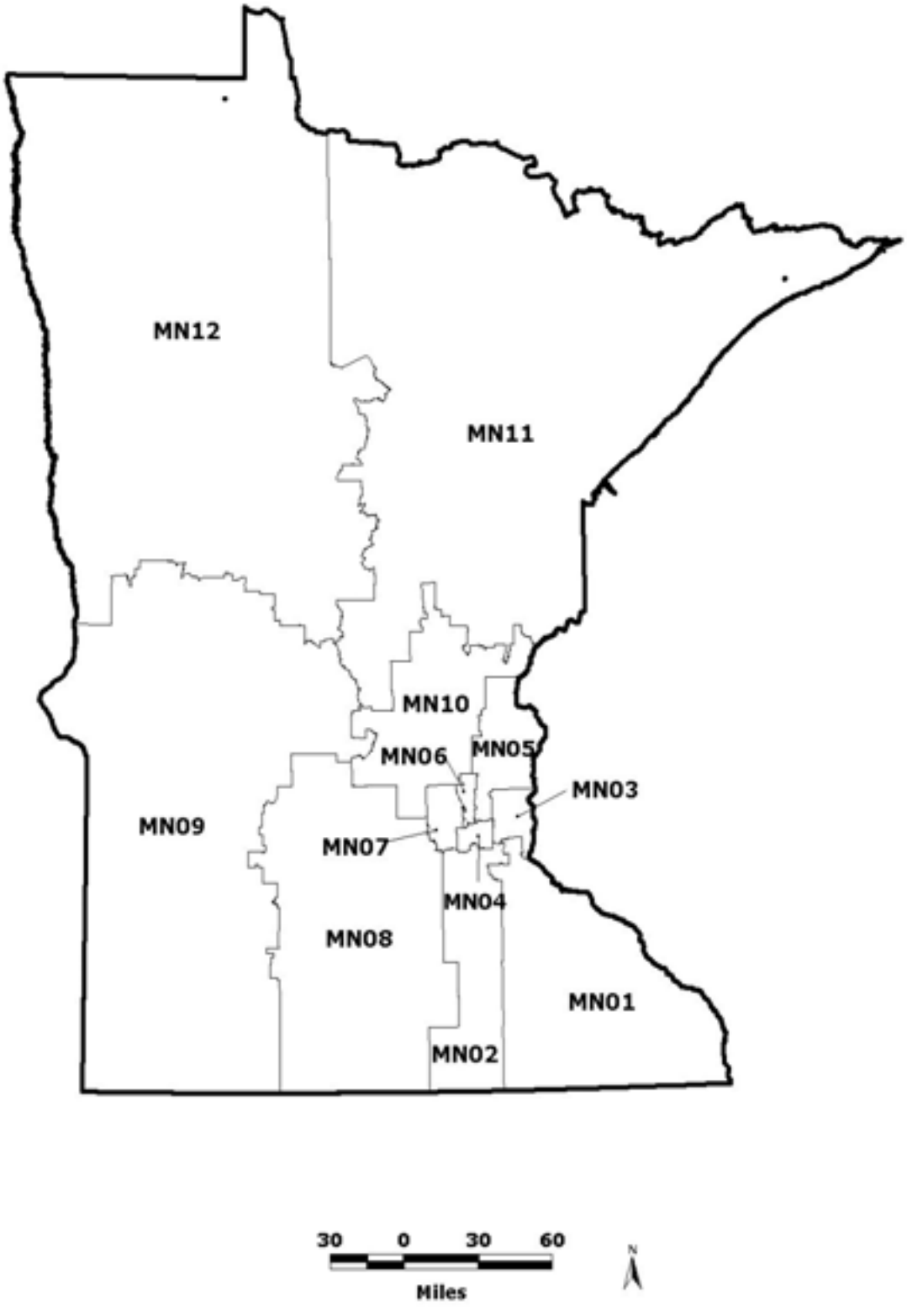

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Maine
ME12
ME08
ME11
ME10
ME09
ME07
ME06
ME03
ME05
ME04
ME01
ME02
20
0
20
40
Miles
N

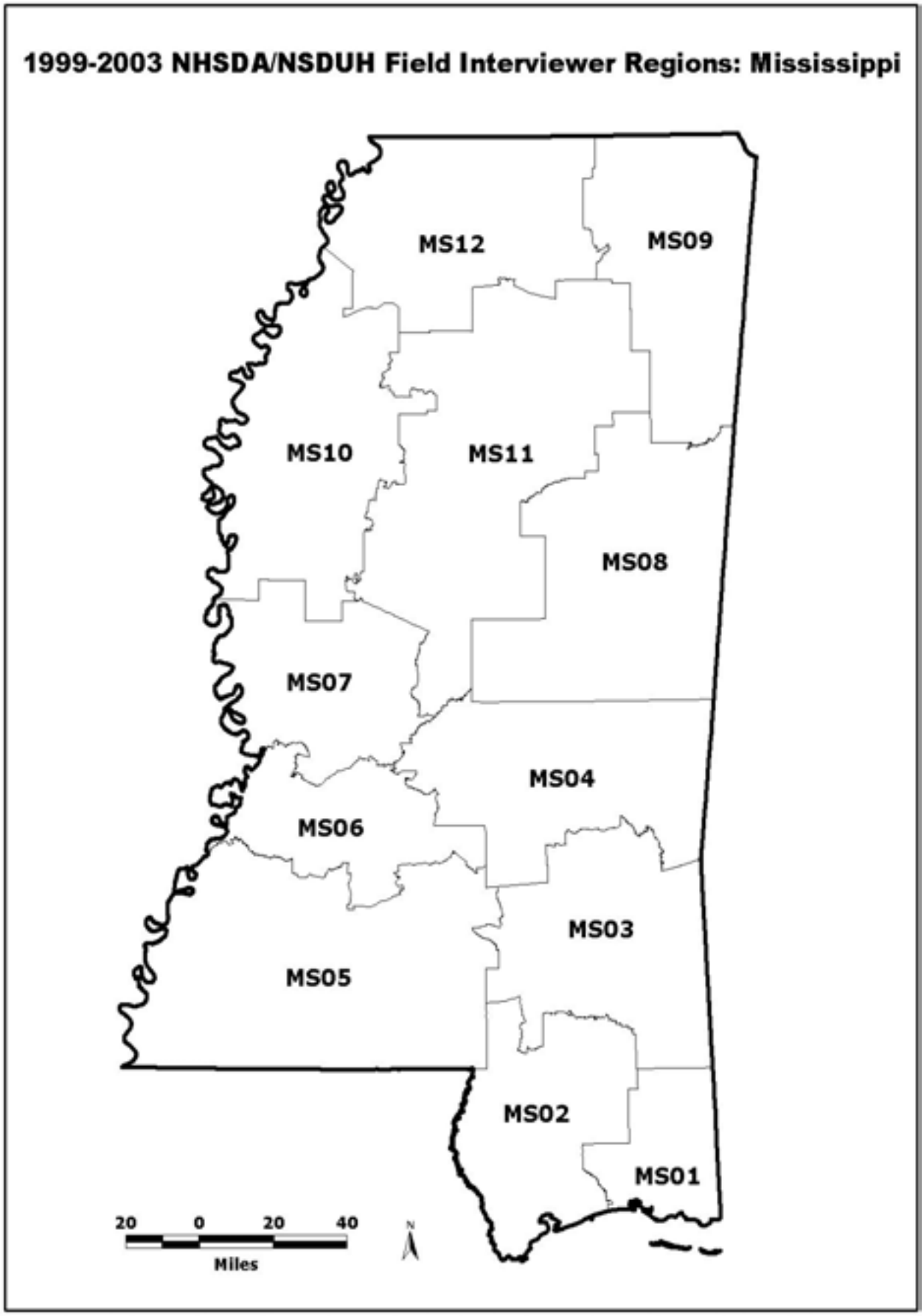

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Maryland
MD01
MD04
MD03
MD06
MD05
MD07
MD10
MD08
MD11
MD09
MD02
MD12
20
0
20
40
Miles
N


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Massachusetts
MA08
MA07
MA10
MA09
MA06
MA12
MA11
MA05
MA04
MA03
MA02
MA01
MA01
MA01
MA01
10 0 10 20
Miles
N

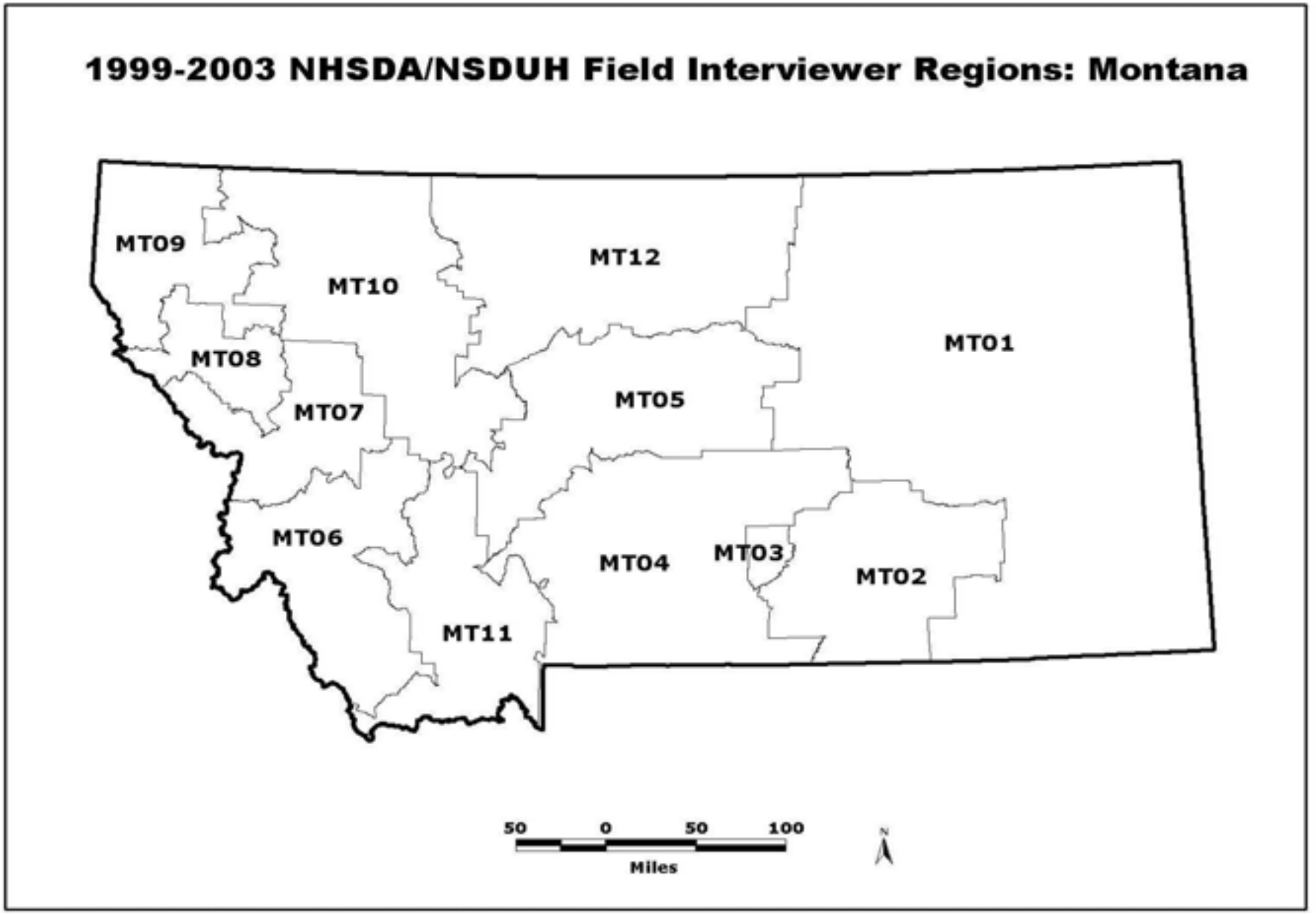

1999 -2003 NHSDA/NSDUH Field Interviewer Regions: Michigan
MI01
MI01
MI02
Detroit Area
MI03
MI04
MI05
MI06
MI14
MI15
MI07
MI08
MI13
MI16
MI17
MI18
MI19
MI21
MI20
MI12
MI11
MI10
MI09
MI22
MI48
MI43
MI45
MI41
MI44
MI47
MI46
MI42
See Inset at left
10 0 10 20
Miles
30 0 30 60
Miles

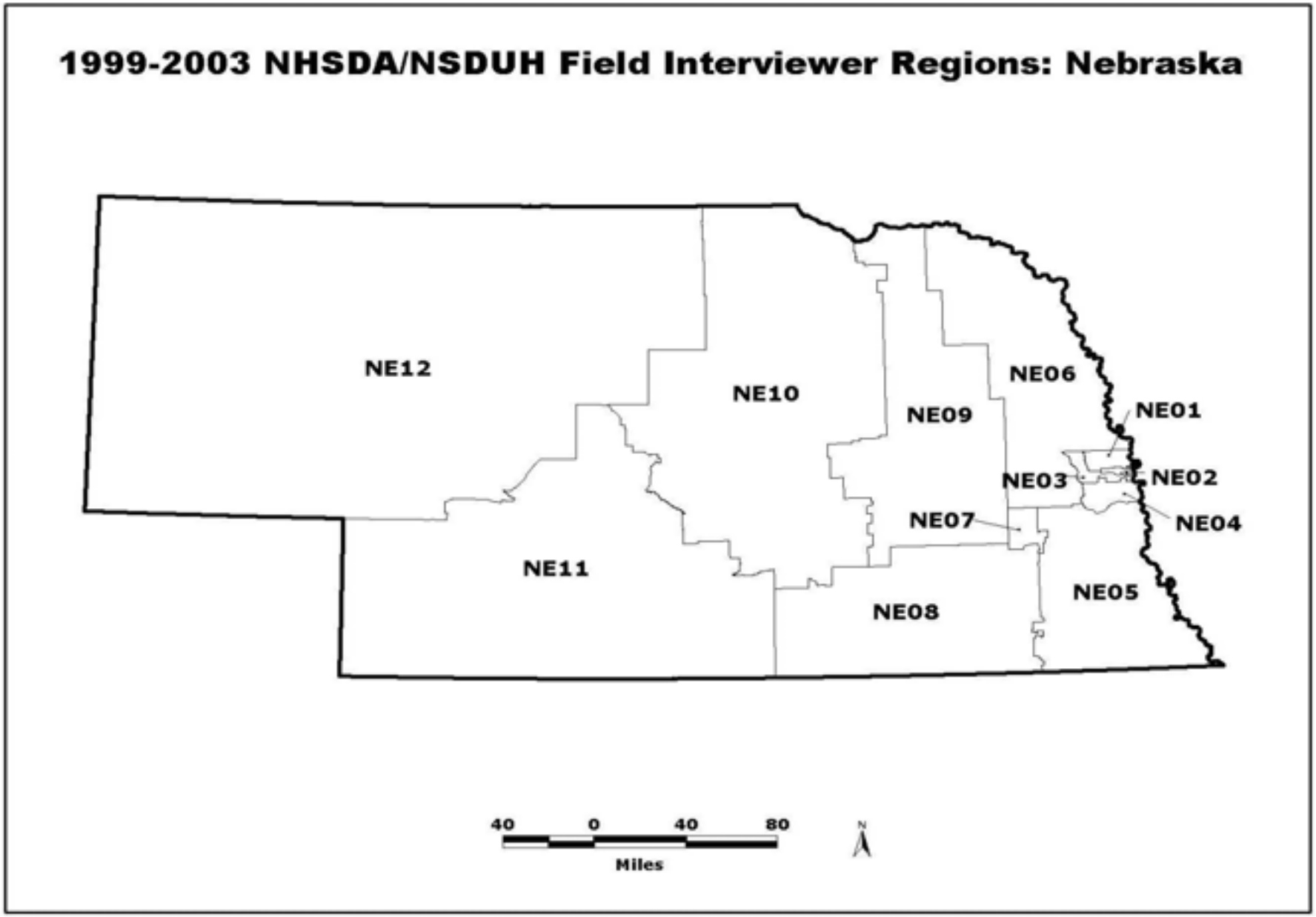
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Minnesota




1999-2003 NHSDA/NSDUH Field Interviewer Regions: Mississippi
MS12
MS09
MS10
MS11
MS08
MS07
MS04
MS06
MS03
MS05
MS02
MS01
20 0 20 40
Miles
N


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Missouri
MO11
MO09
MO12
MO06
MO05
MO04
MO03
MO10
MO02
MO07
MO01
MO08
30 0 30 60
Miles

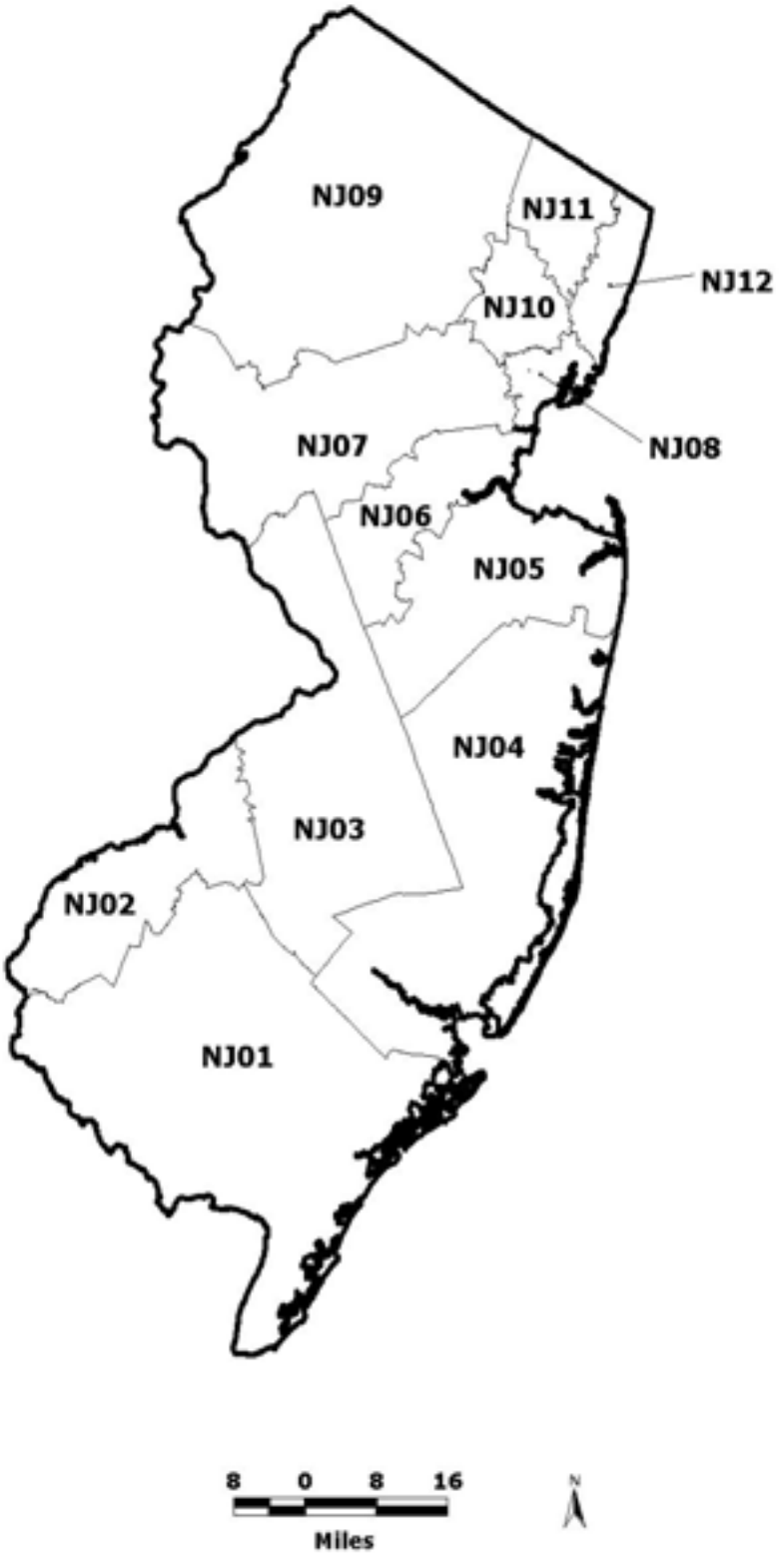

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Montana
MT09
MT10
MT12
MT01
MT08
MT07
MT05
MT06
MT04
MT03
MT02
MT11
50
0
50
100
Miles
N


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Nebraska
NE12
NE10
NE09
NE06
NE01
NE03
NE02
NE07
NE04
NE11
NE08
NE05
40
0
40
80
Miles
N

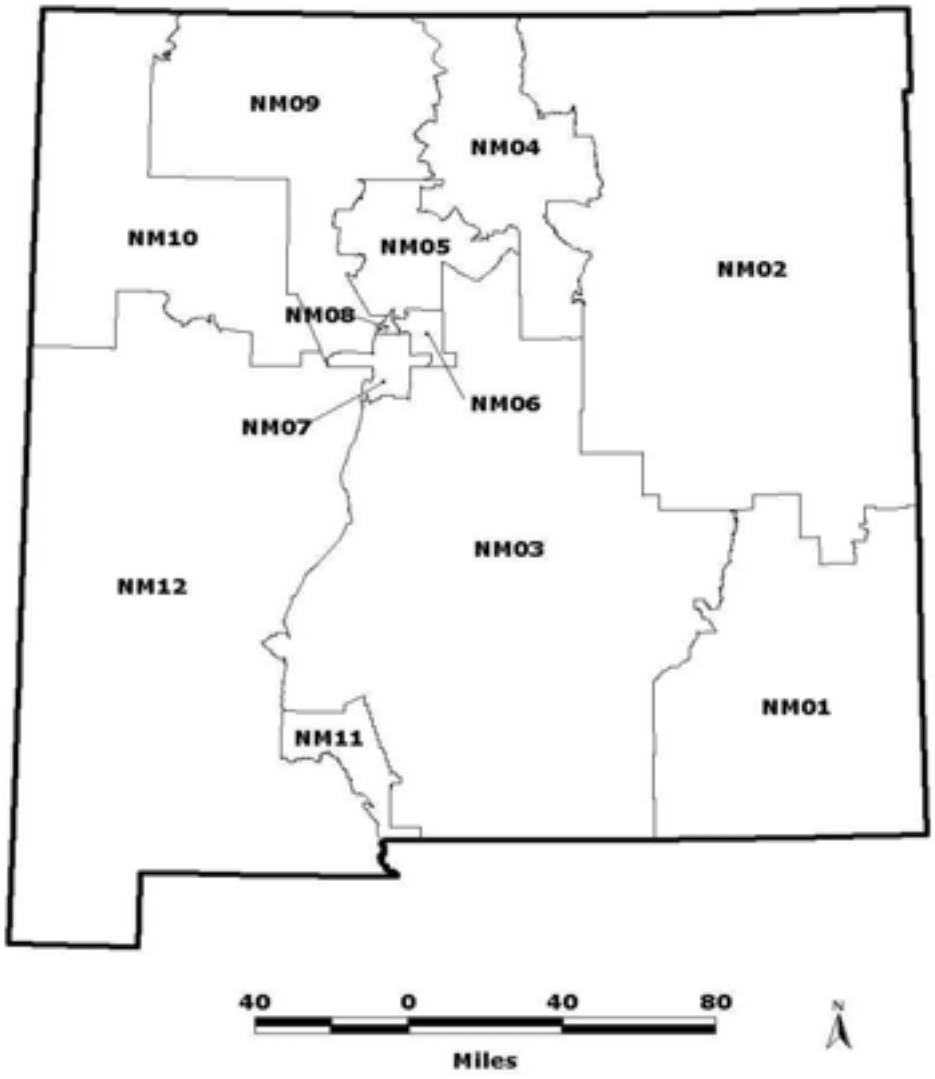
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Nevada

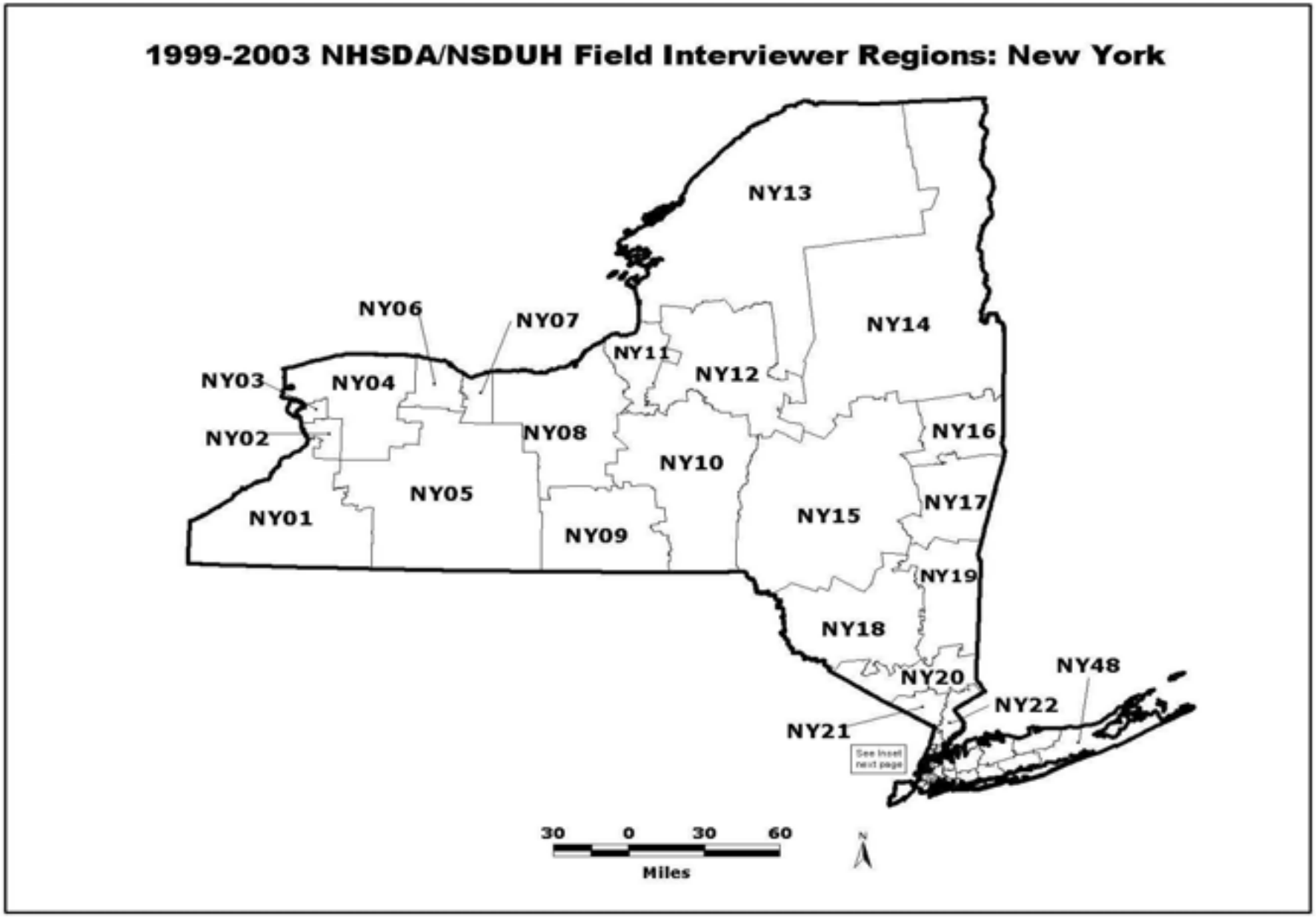
1999-2003 NHSDA/NSDUH Field Interviewer Regions: New Hampshire

1999-2003 NHSDA/NSDUH Field Interviewer Regions: New Jersey

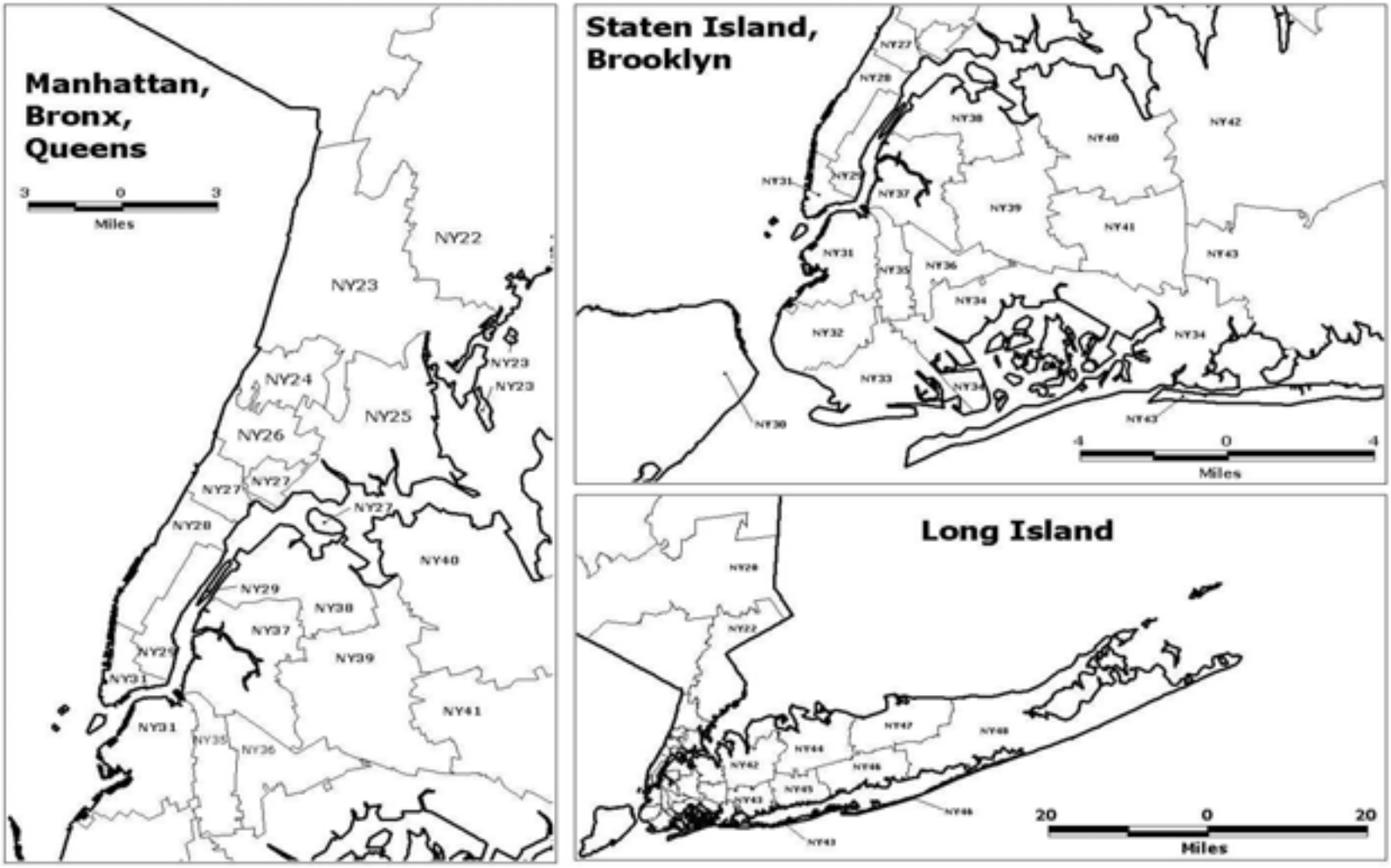


1999-2003 NHSDA/NSDUH Field Interviewer Regions: New Mexico

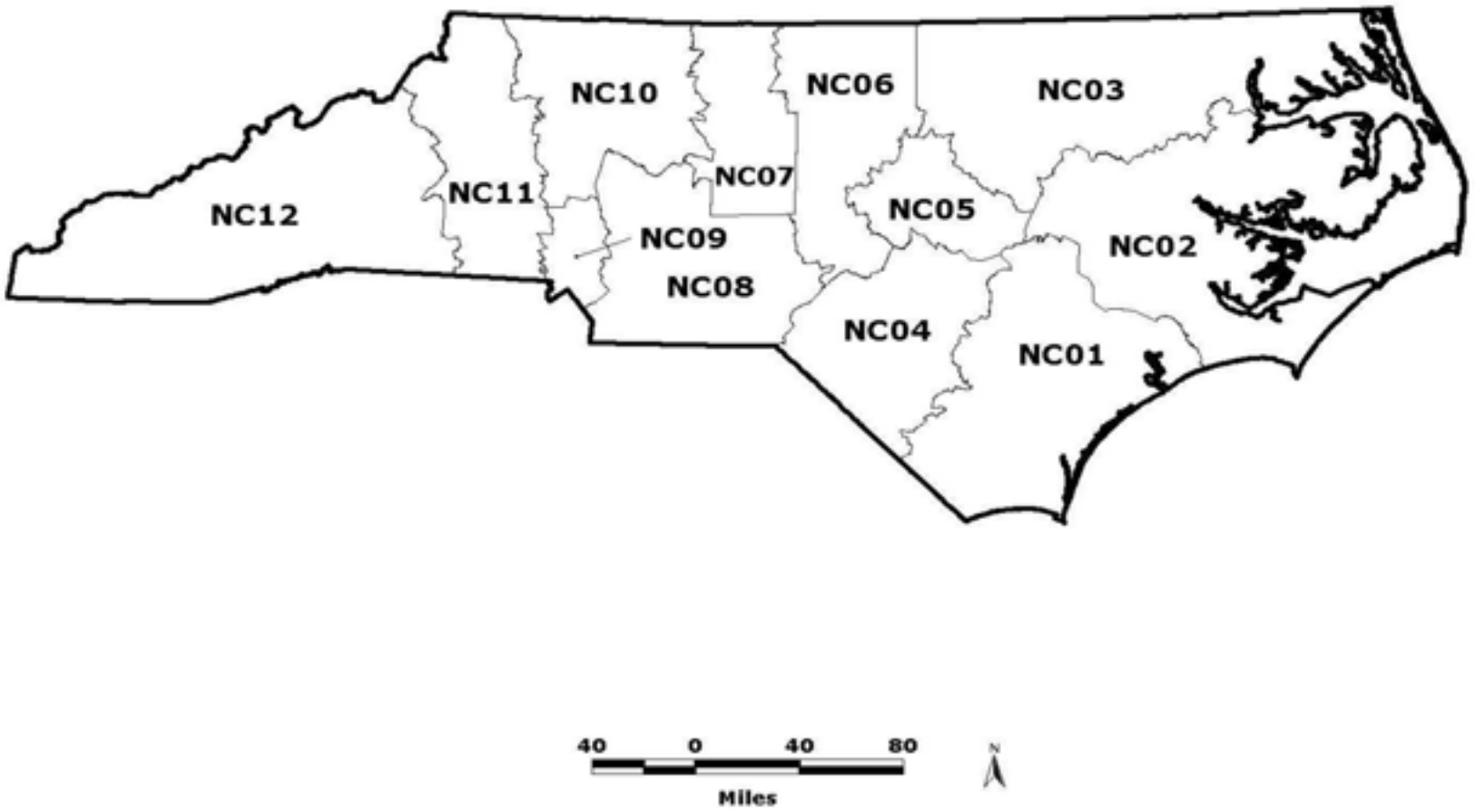


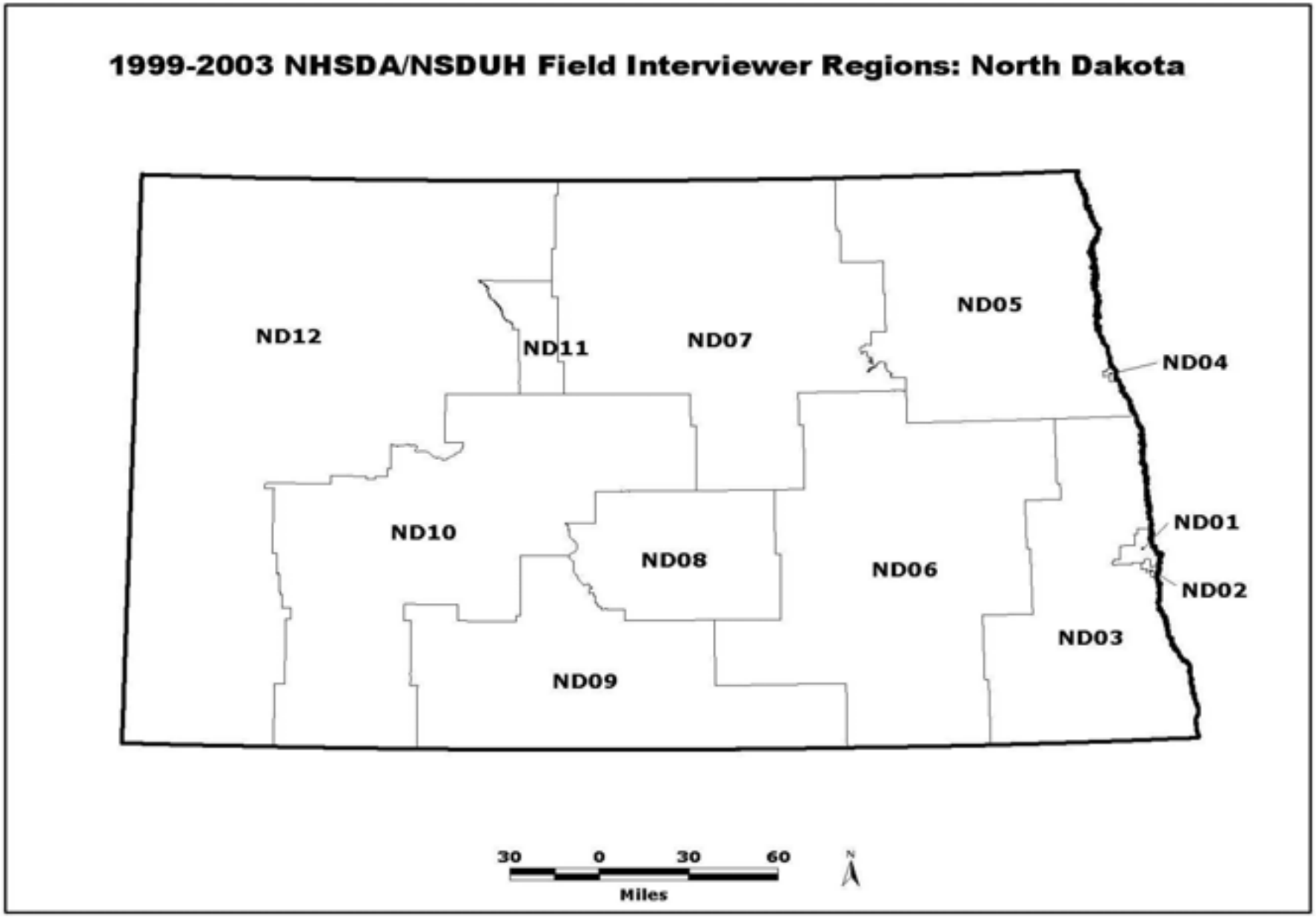

1999-2003 NHSDA/NSDUH Field Interviewer Regions: New York
NY13
NY14
NY06
NY07
NY11
NY12
NY03
NY04
NY02
NY08
NY16
NY10
NY05
NY17
NY01
NY15
NY09
NY19
NY18
NY48
NY20
NY22
NY21
See inset next page
30 0 30 60
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: New York City Area Insets



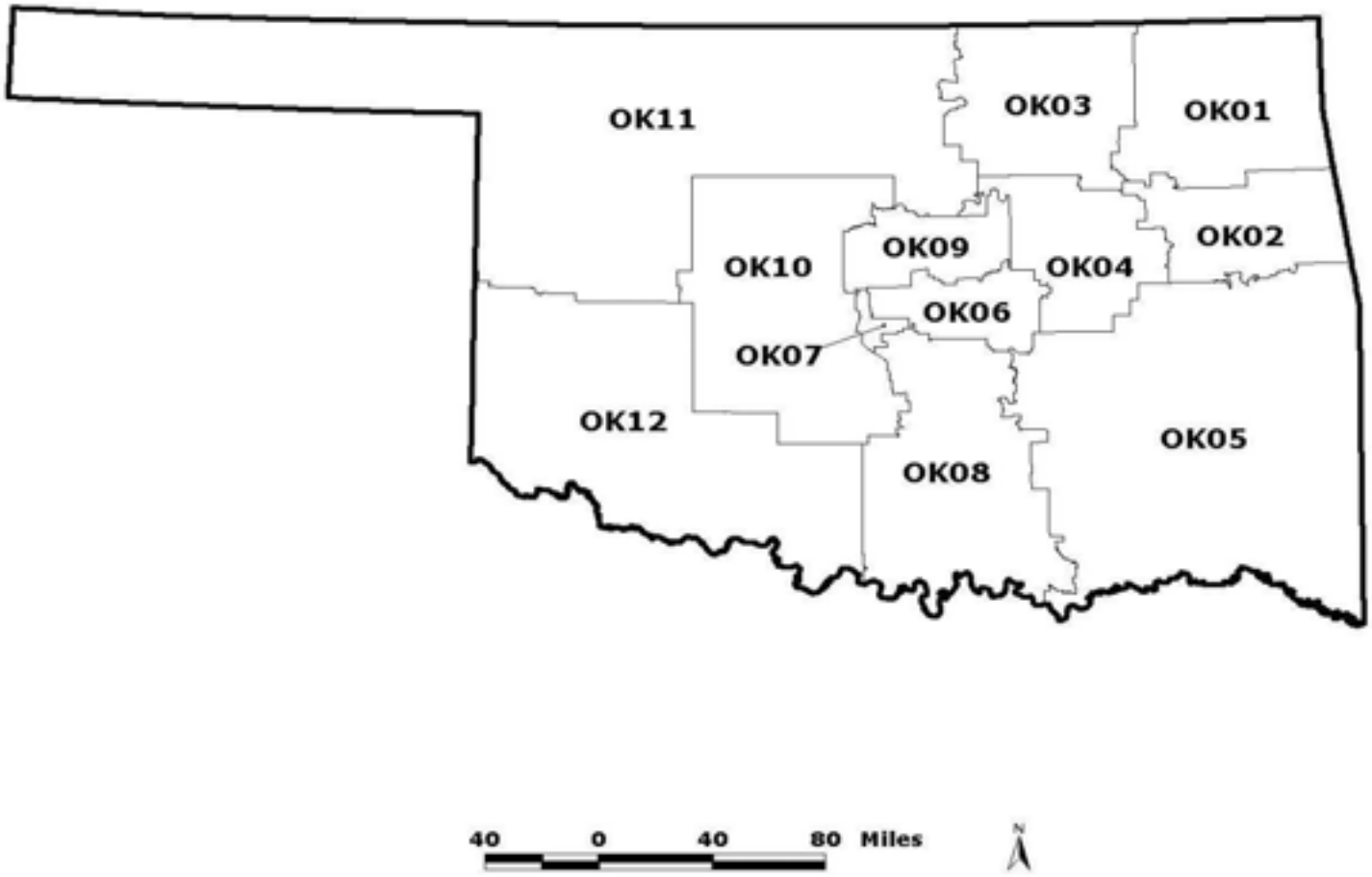
1999-2003 NHSDA/NSDUH Field Interviewer Regions: North Carolina

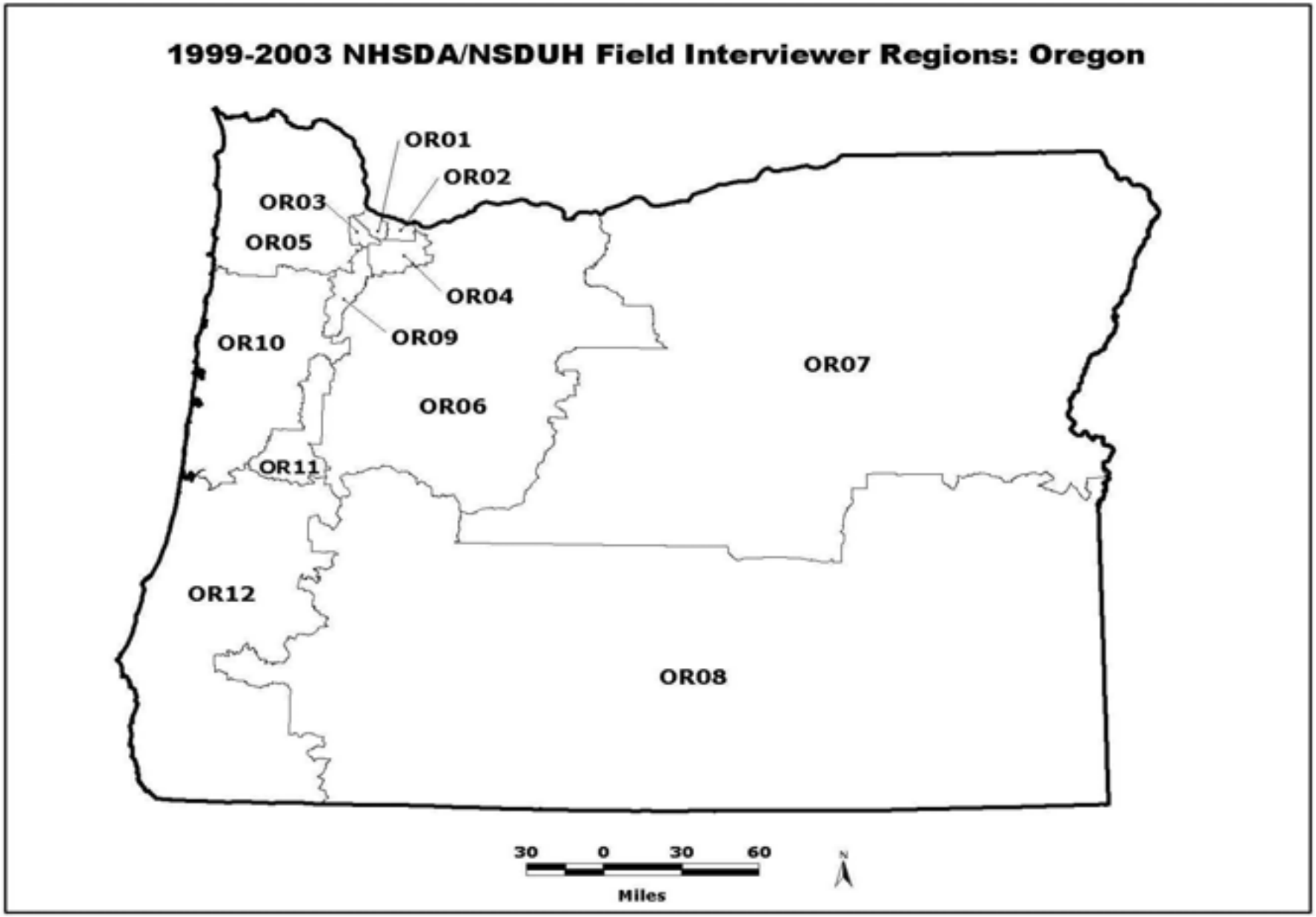



1999-2003 NHSDA/NSDUH Field Interviewer Regions: North Dakota
ND12
ND11
ND07
ND05
ND04
ND10
ND08
ND06
ND01
ND02
ND03
ND09
30 0 30 60
Miles
N

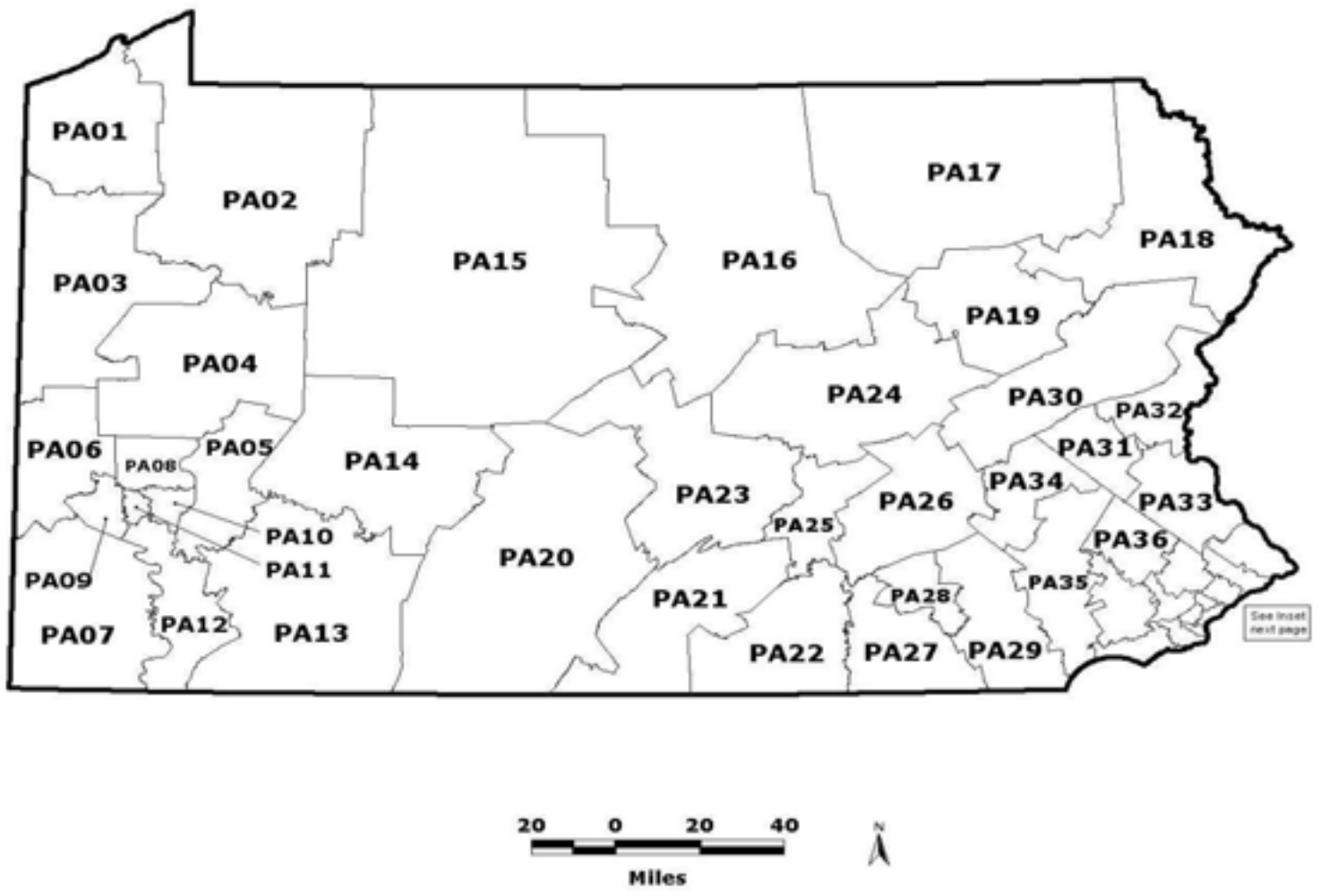
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Ohio

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Oklahoma

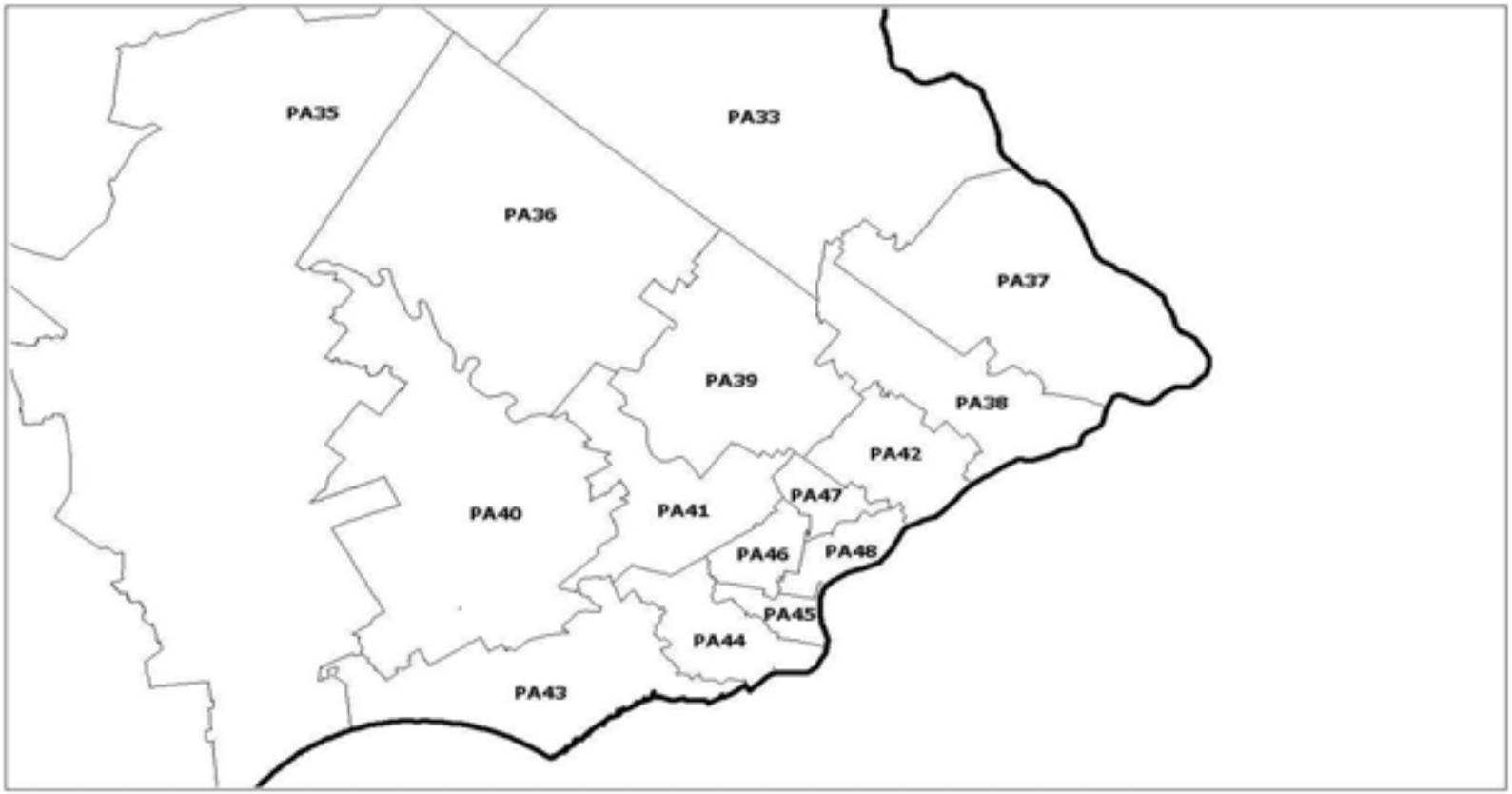



1999-2003 NHSDA/NSDUH Field Interviewer Regions: Oregon
OR01
OR02
OR03
OR05
OR04
OR09
OR10
OR07
OR06
OR11
OR12
OR08
30 0 30 60
Miles
N

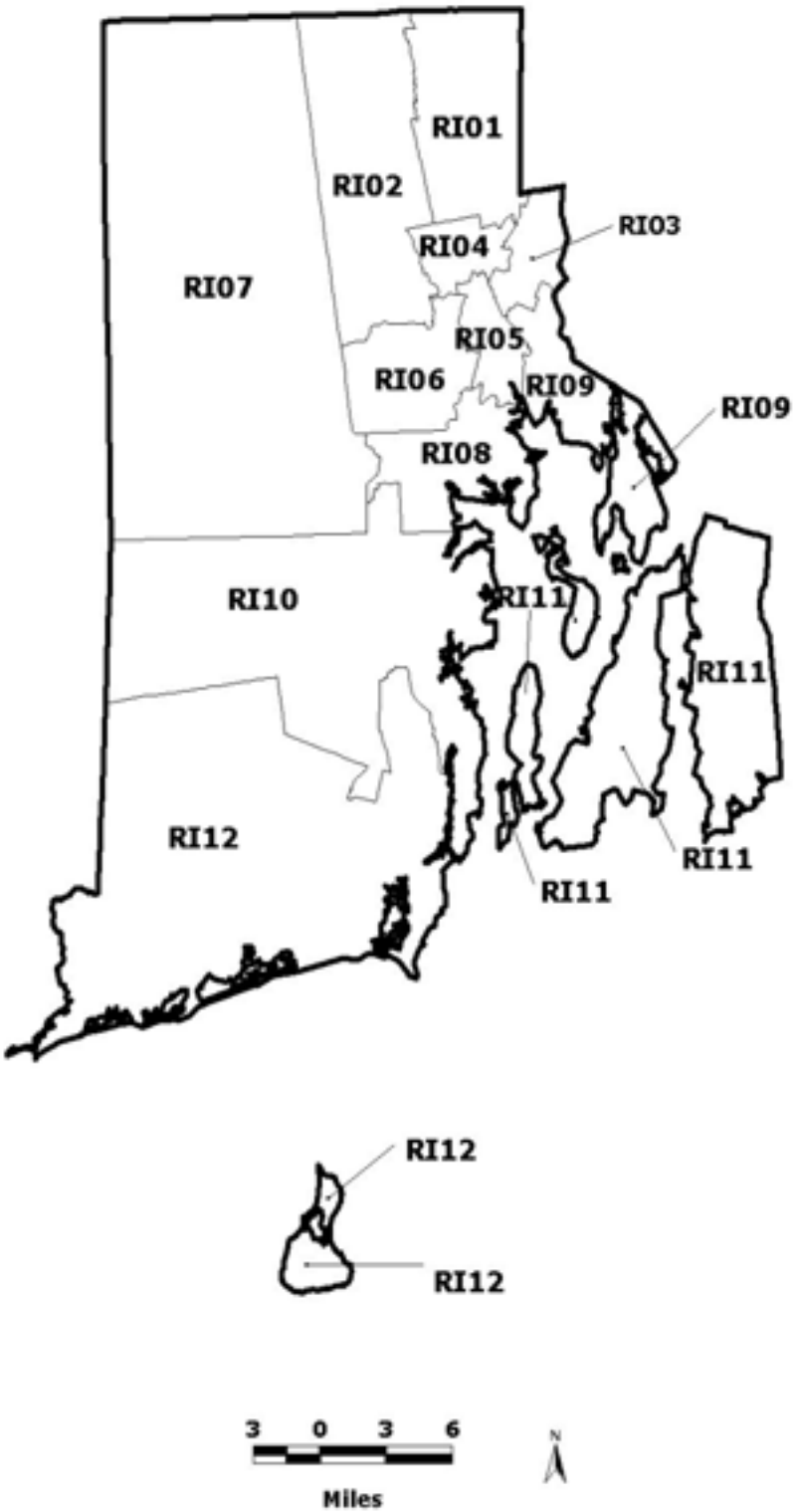
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Pennsylvania

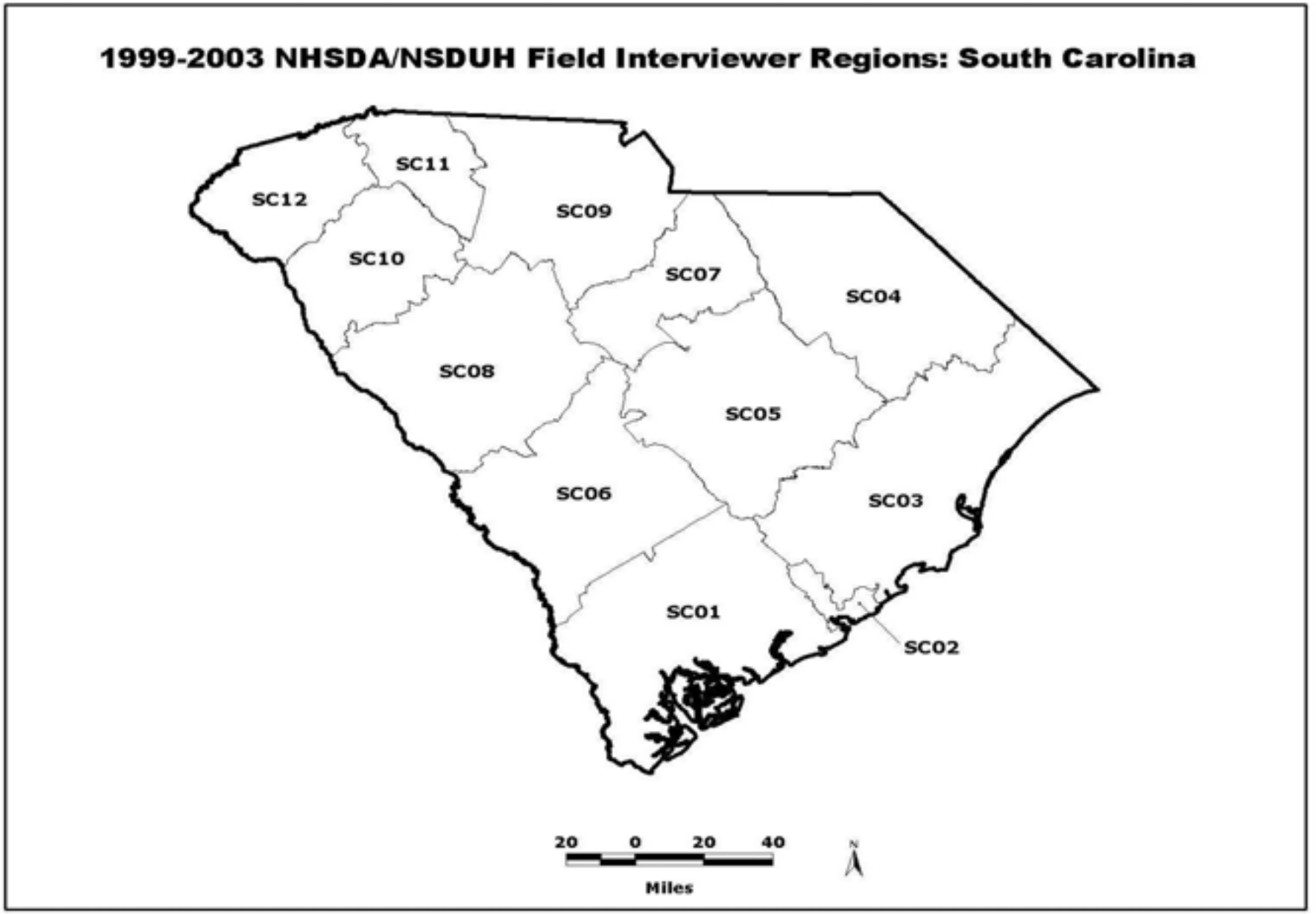


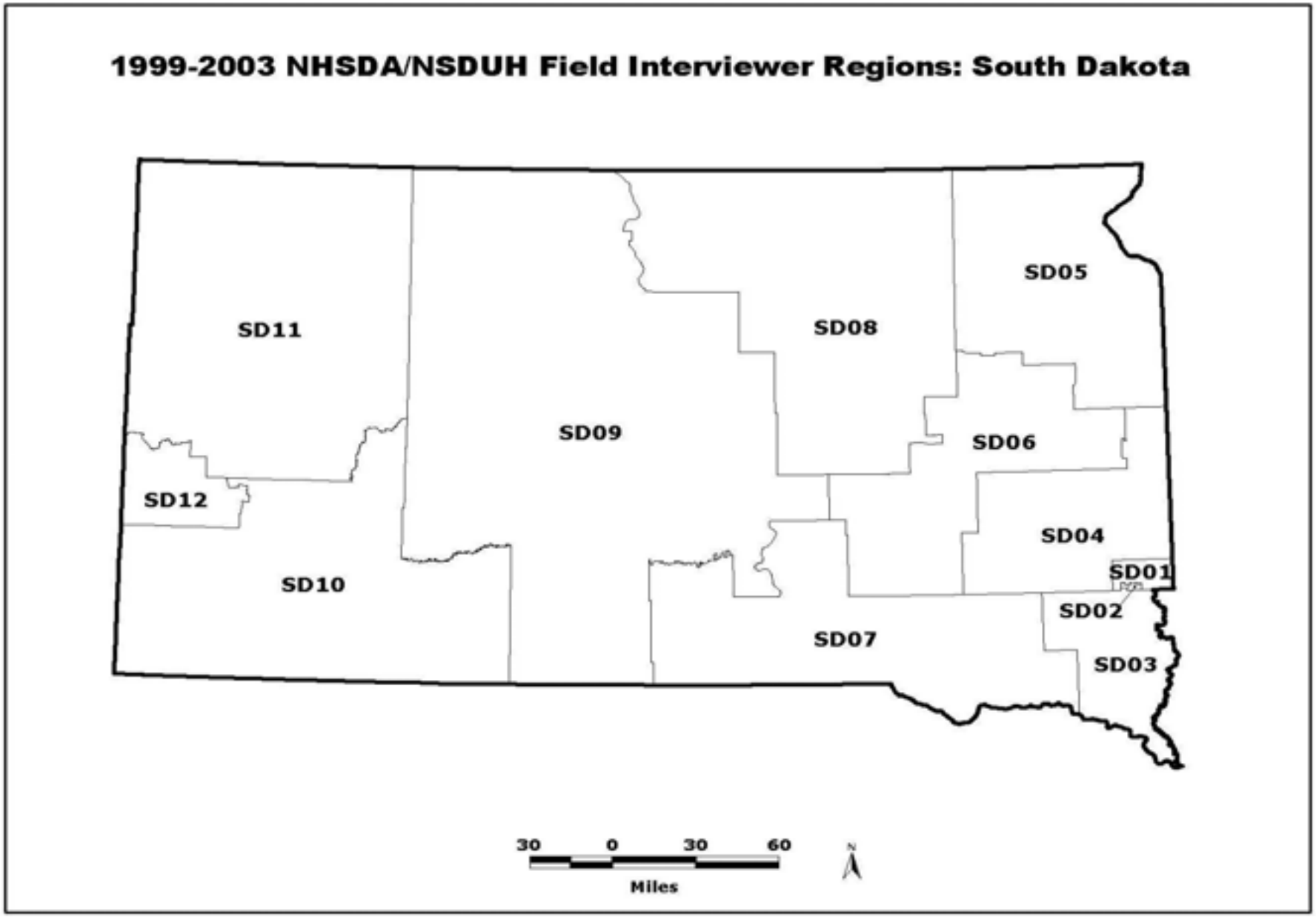
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Philadelphia Area



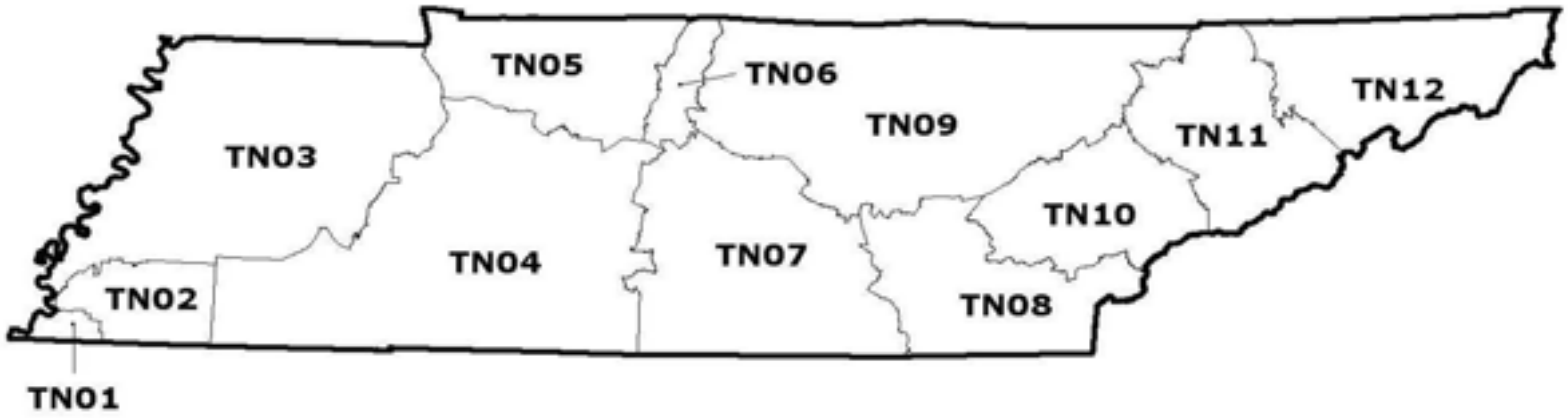
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Rhode Island

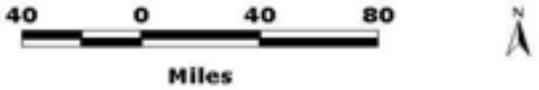


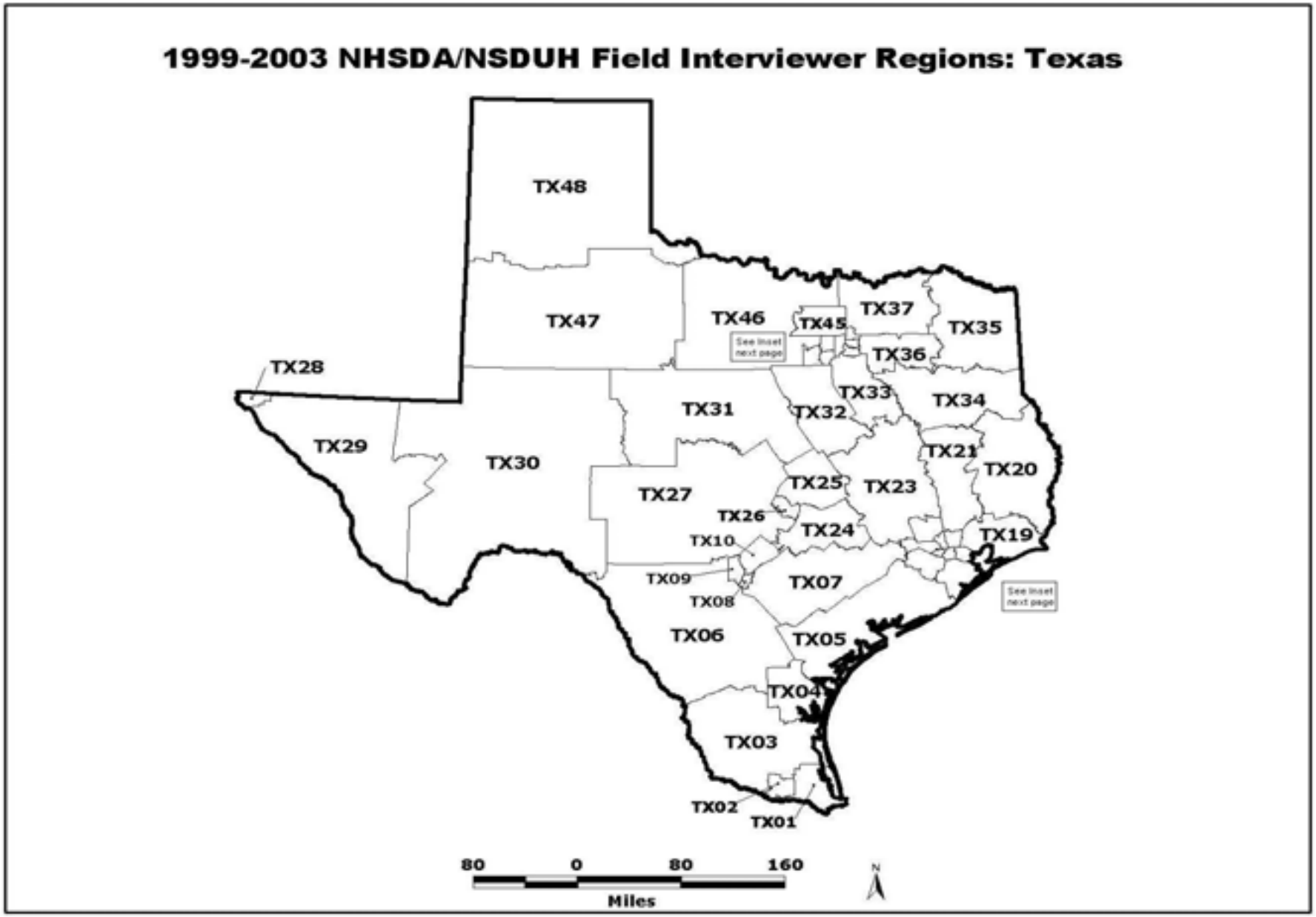

1999-2003 NHSDA/NSDUH Field Interviewer Regions: South Carolina
SC11
SC12
SC09
SC10
SC07
SC04
SC08
SC05
SC06
SC03
SC01
SC02
20 0 20 40
Miles
N


1999-2003 NHSDA/NSDUH Field Interviewer Regions: South Dakota
SD11
SD08
SD05
SD09
SD06
SD12
SD04
SD01
SD10
SD02
SD07
SD03
30
0
30
60
Miles
N

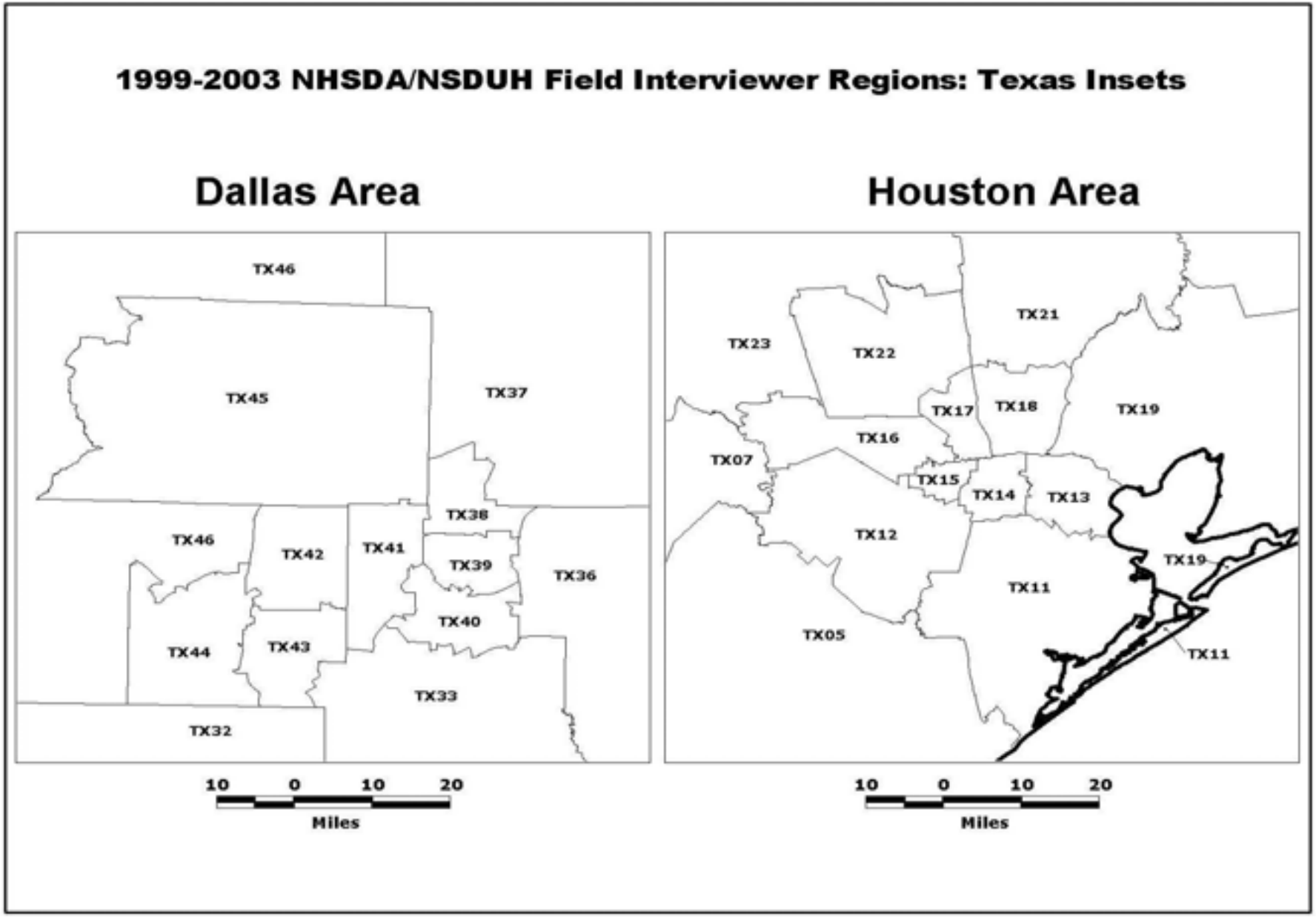
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Tennessee



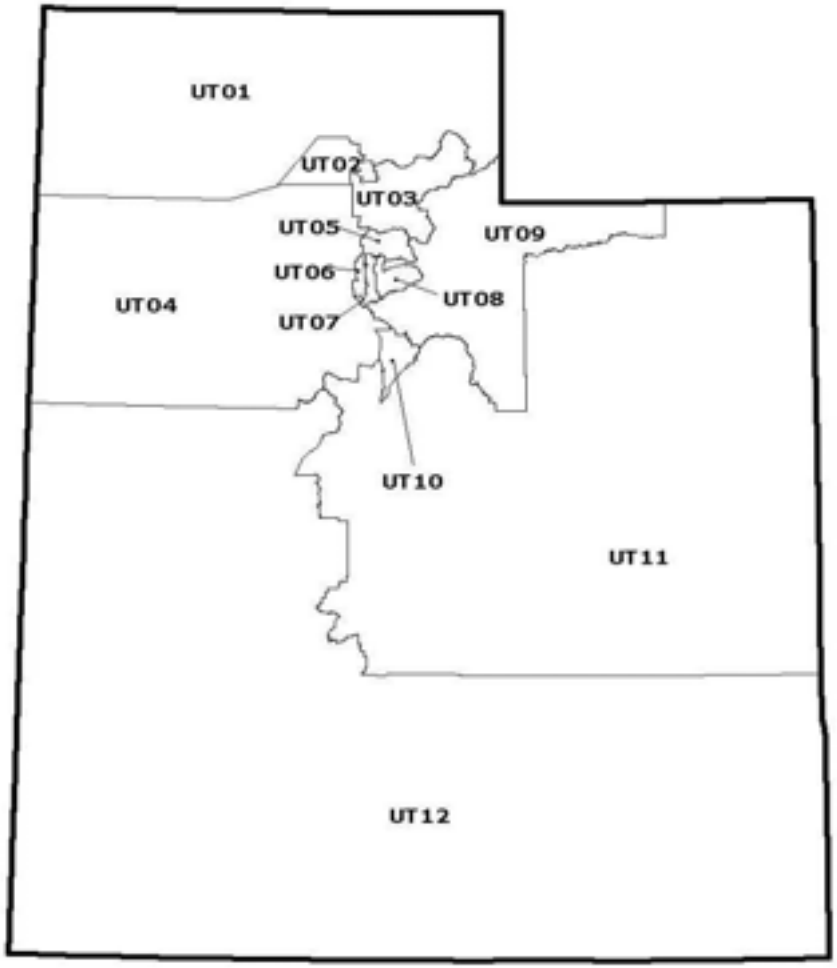



1999-2003 NHSDA/NSDUH Field Interviewer Regions: Texas
TX48
TX47
TX46
TX45
TX37
TX35
TX36
TX28
TX31
TX32
TX33
TX34
TX29
TX30
TX21
TX20
TX27
TX25
TX23
TX26
TX24
TX19
TX10
TX09
TX07
TX08
TX06
TX05
TX04
TX03
TX02
TX01
See Inset next page
80 0 80 160
Miles
N

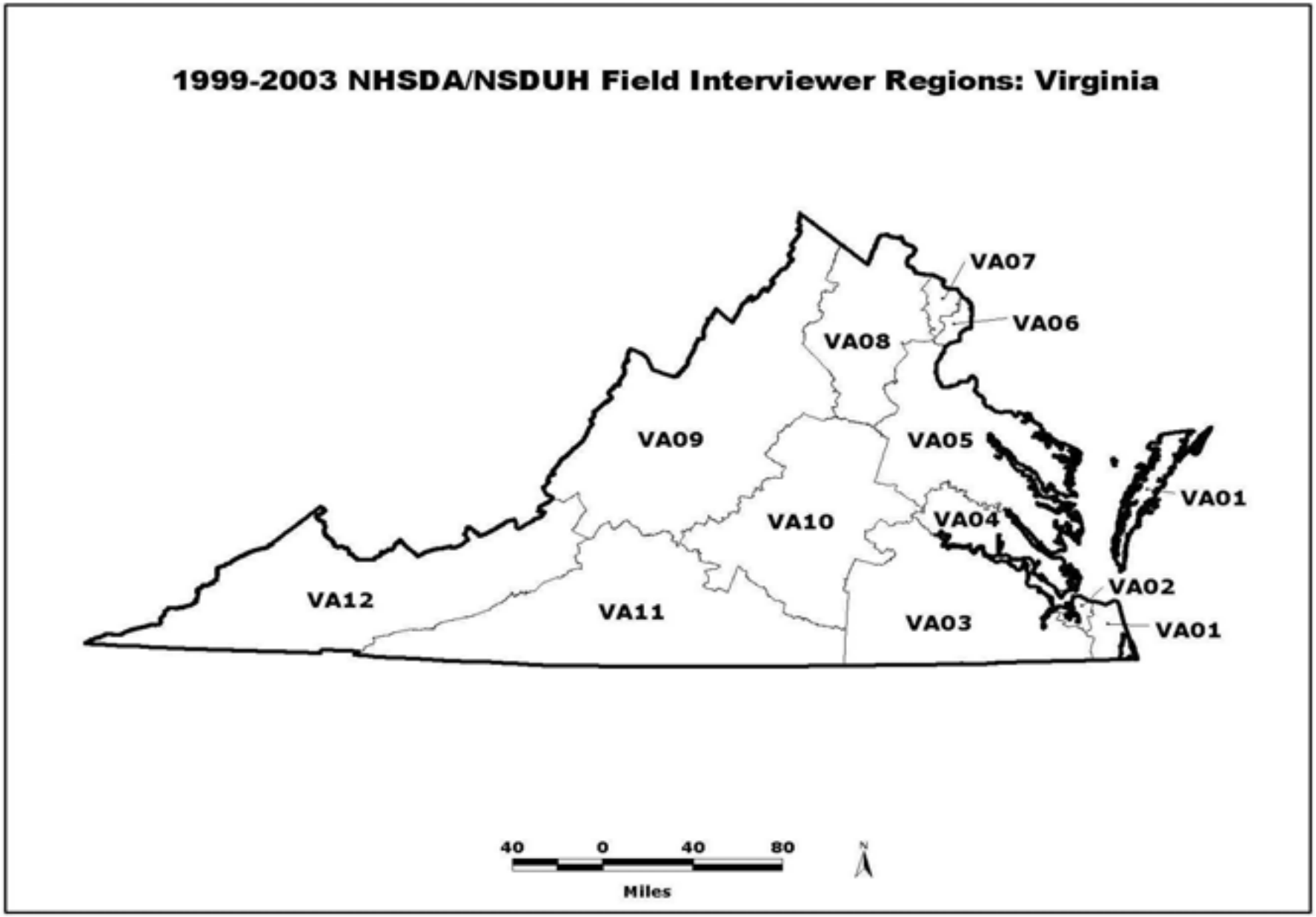
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Texas Insets

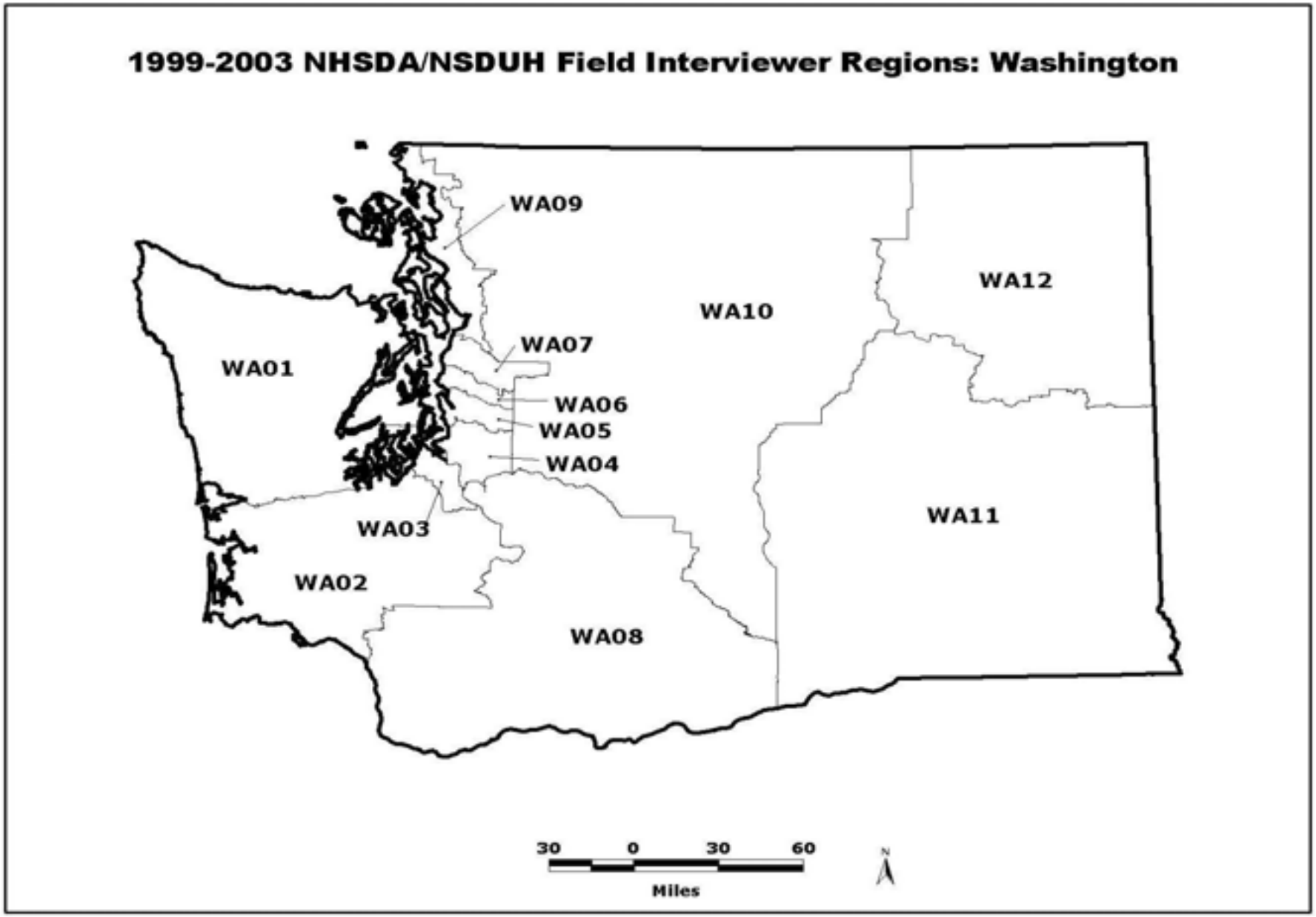


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Utah

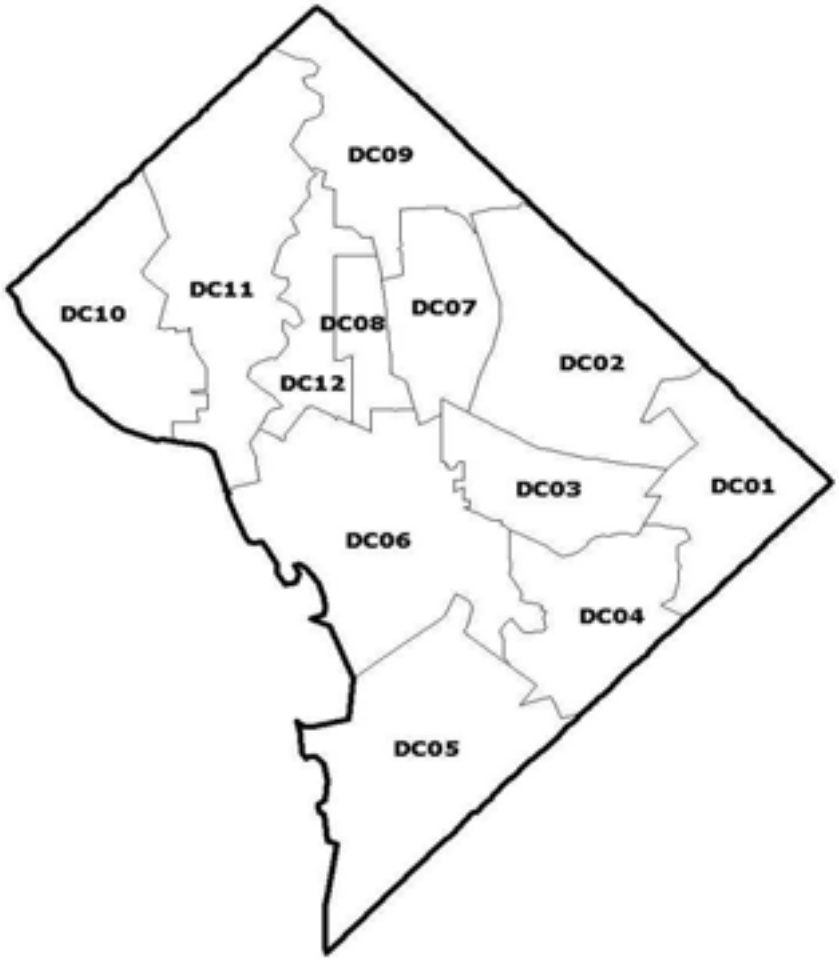


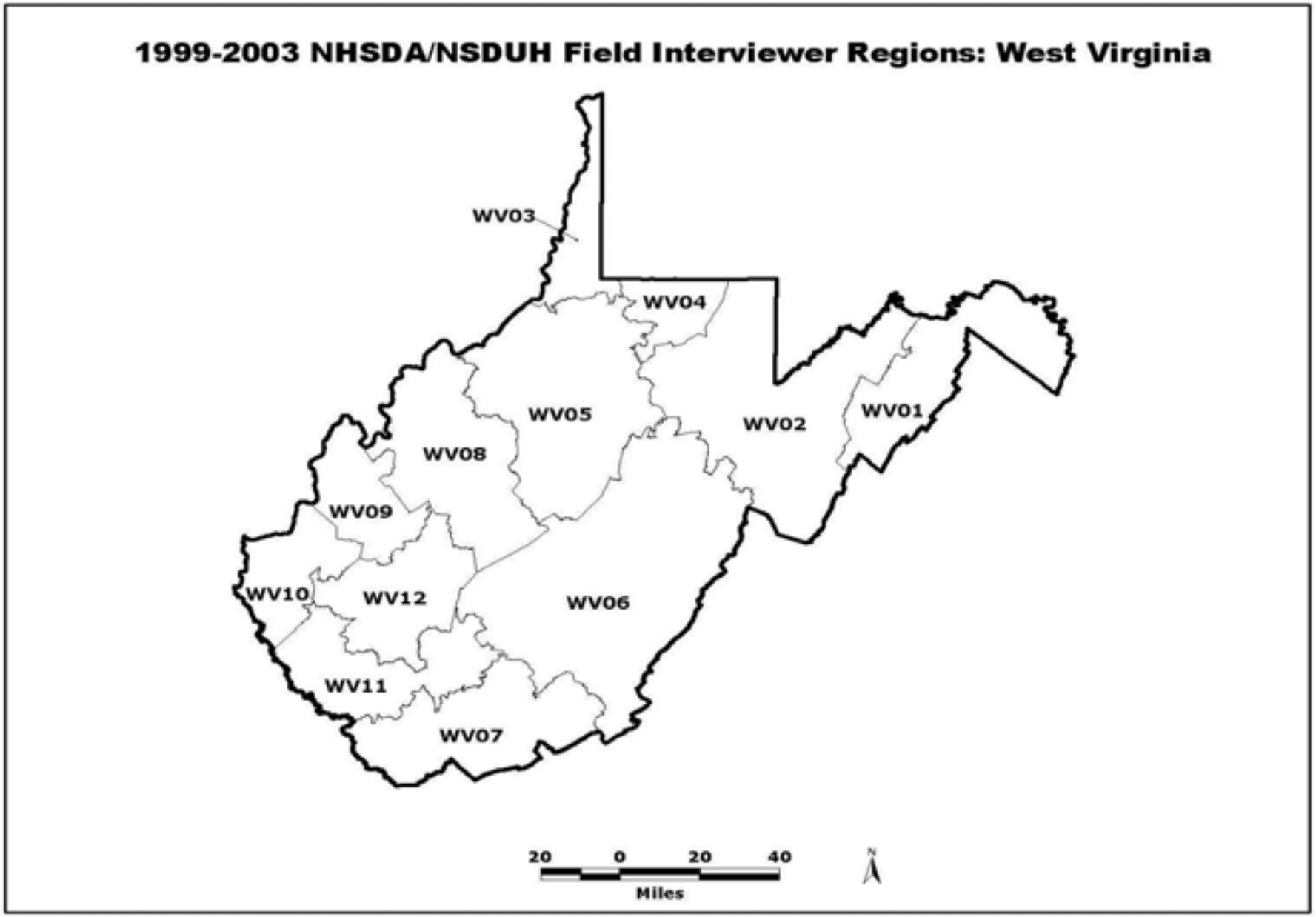
1999-2003 NHSDA/NSDUH Field Interviewer Regions: Vermont


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Virginia
VA07
VA06
VA08
VA09
VA05
VA01
VA10
VA04
VA02
VA12
VA11
VA03
VA01
40
0
40
80
Miles
N

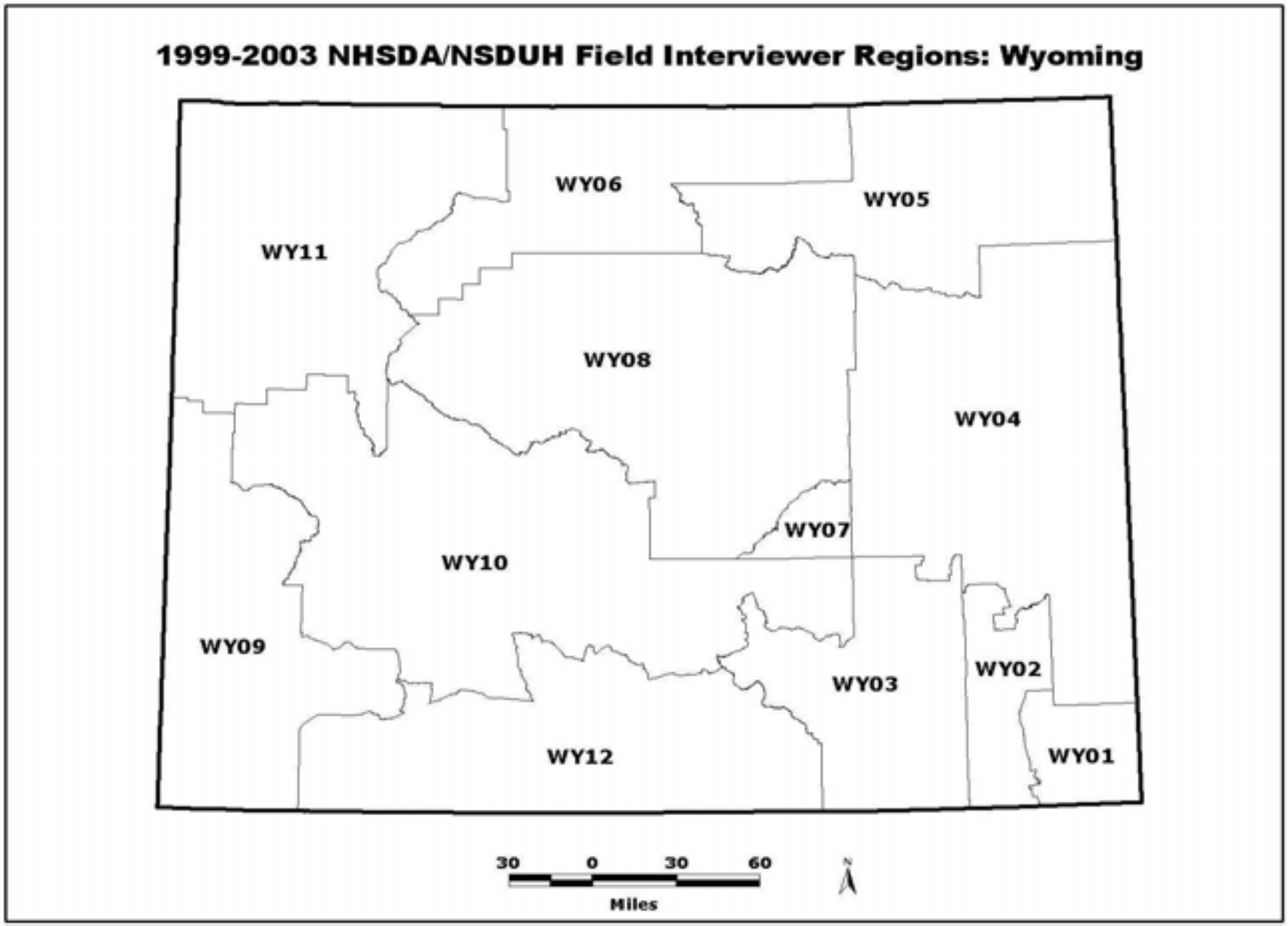

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Washington
WA09
WA10
WA12
WA07
WA01
WA06
WA05
WA04
WA11
WA03
WA02
WA08
30
0
30
60
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Washington, D.C.




1999-2003 NHSDA/NSDUH Field Interviewer Regions: West Virginia
WV03
WV04
WV05
WV02
WV01
WV08
WV09
WV10
WV12
WV06
WV11
WV07
20
0
20
40
Miles
N

1999-2003 NHSDA/NSDUH Field Interviewer Regions: Wisconsin


1999-2003 NHSDA/NSDUH Field Interviewer Regions: Wyoming
WY06
WY05
WY11
WY08
WY04
WY07
WY10
WY09
WY03
WY02
WY12
WY01
30
0
30
60
Miles
N

Appendix B

2000 NHSDA Procedure for Adding Missed Dwelling Units

1. Introduction

The 2000 National Household Survey on Drug Abuse (NHSDA) requires field interviewers (FIs) to visit sample segments and screen and interview dwelling units (DUs) that were selected from an ordered list. The list of DUs, which includes housing units and group quarters, was constructed by the counting and listing staff during the summer and fall of 1998 for the overlapping segments and the summer and fall of 1999 for the replacement segments. Because the listing was done a short time before the 2000 screening and interviewing activities began, no major discrepancies were expected. However, factors such as new construction, demolition, and inaccurate listing may be present in some cases. More commonly, DUs may have been "hidden" and therefore overlooked by the counter and lister.

In order for all DUs to be given a chance of being selected, the NHSDA has a procedure for locating and adding missed DUs. It requires FIs to look on the property of selected DUs and between that DU and the next listed DU (half-open interval rule). In 2000, the rule was modified such that the half-open interval is closed on each map page. Therefore, if the selected DU is the last on a page, the "next listed DU" will be the first one on listed on the same page. If the number of added DUs linked to any particular DU does not exceed five or if the number for the entire segment is less than or equal to ten, the FI is instructed to consider these DUs as part of their assignment. However, if either of these limits is exceeded, the FI will contact RTI for subsampling to be considered.

This document outlines the proposed procedures for RTI to use when discrepant segments are found in the field. For this document, procedures for adding missed DUs will be classified into three categories: adding housing units (HUs), adding group quarter units, and "busts."

2. Motivation

Prior to the 1999 NHSDA, if the number of added DUs exceeded the defined limits, the added DUs were subsampled at the same rate of the original selection for the segment. To maintain unequal weighting effect and to control costs associated with adding DUs, a new subsampling procedure was implemented:

Number of Added DUs	Sampling Rate
0	No action
1 to 10	Automatic (all DUs added to the sample)
11 to 25	1/2
26 to 40	1/3
41 to 50	1/4
50 or more	1/5

3. Procedure for Adding Housing Units

This section refers to housing units that are obtained through the half-open interval rule. This method of dealing with added HUs is preferable to all others because it is probability-based

and maintains the integrity of the sample. When possible, this methodology will be used to resolve added DU problems.

1. Once the limit of five (or ten) rule is exceeded, the FI should stop screening and interviewing activities on added HUs and contact RTI. The FI will be instructed to do a quick check of the segment to see if any other listing problems might arise. At this time, the FI will complete a paper list of added HUs for the entire segment.

2. Once the final list of added HUs has been received by RTI:

    a) Sampling will examine the added HUs and determine whether they are linked to a sample dwelling unit (SDU);
    b) If the number of added HUs linked to any *one* SDU exceeds 50, these units will be treated as a "bust" (see Section 6);
    c) If the number of added HUs linked to any *one* non-sampled DU exceeds 50, these units will also be treated using the procedure for "busts" (see Section 6);
    d) Sampling will calculate the total number of added DUs by adding the number of sampling units obtained through the "bust" procedure to the number of added DUs obtained through the half-open interval rule;
    e) If the total number of added DUs exceeds 10, a subsampling rate will be determined using the criteria above.

3. RTI will add the HUs to the system and subsample if necessary:

    a) Data entry of the added HUs will be done. Lines will be entered for all units that collectively qualify as a "bust" and units obtained through the half-open interval rule—not for all missed DUs found in the segment. The link number will then be entered and a line number will be assigned. For lines obtained through the "bust" procedure, the sampling link number (SLN) will also be recorded. Finally, it will be necessary to check that none of the lines have already been entered in the Newton so that lines don't appear in the system twice.
    b) Select lines from the added HUs at the rate defined above. Record the subsampling rate in a data field.
    c) Bring over probabilities of selection as appropriate for the segment.
    d) Add a random number for the Newton selection algorithm.

4. Selected lines will be added to the FI's assignment during the next transmission.

4. Procedure for Adding Group Quarter Structures

In the case of an entire group quarter (GQ) structure not being listed (or erroneously being listed as a HU), the half-open interval rule will be applied. For example, if the DU preceding the GQ was selected, or if the HU that is really a GQ was selected, the entire GQ structure will be added to the sample. The exception to this rule will be if the number of GQ units in the missed GQ structure exceeds 50. In this last case, the "bust" procedure will be applied (see Section 6).

5. Procedure for Adding Group Quarter Units

In the case of discrepant GQ listings, we will know in advance the number of sampling units (rooms, persons, or beds) and the number of selected units. If the actual number of sampling units equals the amount listed in advance, the Newton will only need to be notified of the new unit type in order to function properly. However, if the actual units do not equal the advance units, two approaches will be taken.

5.1. Number of Actual GQ Units Less Than Number of Advance GQ Units

In the case that there are extra GQ units listed, the units at the end of the list will be assigned an ineligible code such as "Not A DU." All other units will remain eligible.

5.2. Number of Actual GQ Units Greater Than Number of Advance GQ Units

If there are more GQ units in the structure than were previously listed, a complete list will be made and the units will be consecutively numbered. Assume, for example, that 11 units were listed and 45 were actually found. Also, assume that units 1, 5, and 10 were selected for Screening and Interviewing (indicated in bold).

Original list:	**1**
	2
	3
	4
	5
	6
	7
	8
	9
	10
	11

Then, the additional units will be numbered consecutively and a SLN corresponding to each of the originally listed units will be assigned. Next, the added GQ units with SLNs corresponding to the original selected units will be added to the sample.

Unit Number	SLN
12	**1**
13	2
14	3
15	4
16	**5**
17	6
18	7
19	8
20	9
21	**10**

22	11
23	**1**
24	2
25	3
26	4
27	**5**
28	6
29	7
30	8
31	9
32	**10**
33	11
34	**1**
35	2
36	3
37	4
38	**5**
39	6
40	7
41	8
42	9
43	**10**
44	11
45	**1**

6. "Busts"

Any segment listing with a major discrepancy (defined by 150 or more total unlisted units or 50 or more added DUs linked to any one SDU) or that is completely unrepresentative of what is actually found is called a "bust." In the case of a fictitious listing, RTI will relist the segment as quickly as possible. Otherwise, the following approach will be employed. First, if any DUs have disappeared since the time of the listing, all selected "disappears" will be assigned an "ineligible" final screening code. Then, any new DUs will be listed consecutively, assigned a SLN, and added to the sample if the SLN corresponds to the line number of an originally selected DU. Note that if the DU was coded as ineligible in the first step, the new DUs having its line number as the SLN will still be added. This procedure is identical to the procedure for adding extra GQ units, however the list can contain any combination of HUs and GQ units in this case. Again, if the number of DUs added is greater than 10, then resampling will occur from all non-finalized DUs as in Section 3.

7. Quality Control

In order to ensure quality, RTI will employ several quality control checks:

- Mapping will ensure that the correct information has been keyed by data entry,
- Checks within the computing division will be performed; and

- Sampling will check the number of selected lines and the person probabilities of selection assigned to each DU selected in the subsampling routine.