How to Draw a Nationally-Representative Sample: Updating and Reassessing Monitoring the Future's Sampling Procedures

Richard Miech
Research Professor
Principal Investigator, Monitoring the Future
University of Michigan

How to Draw a Nationally-Representative Sample??
Updating and Reassessing Monitoring the Future's Sampling Procedures

Richard Miech
Research Professor
Principal Investigator, Monitoring the Future
University of Michigan

Introduction

- The goal of Monitoring the Future (MTF) is to provide nationally-representative estimates of adolescent drug use.
    - Our bread and butter is current ***trends***
        - Is adolescent marijuana use increasing as more state legalize?
        - Increase in adolescent nicotine vaping from 2017-2018 was largest MTF ever recorded for any substance, prompting the FDA to act.
            - E.g., FDA cites MTF results in its policy banning flavors in cartridge-based vaping devices
        - Will adolescent drug use levels return to pre-pandemic levels after large drop in 2021?

Introduction

- The goal of Monitoring the Future (MTF) is to provide nationally-representative estimates of adolescent drug use.
    - A philosophical question: How would we know?
    - Do we need to survey all ~12 million U.S. adolescents in order to get accurate national estimates?
    - Fortunately, there is another way.
        - (in fact, surveying all ~12 million adolescents would be a misuse of resources)

Assessing National Trends

- We draw a ***sample*** of all U.S. adolescents.
  - Our frame is a list of all middle schools and high schools in the U.S., from which we draw a random sample.
- Results from the sample are what we would find if we surveyed all 12 million U.S. 8^{th}, 10^{th}, and 12^{th} graders U.S, + or - a percentage point or two.
- Much of the value-added of MTF comes from its scientifically-based ability to generalize to the U.S. as a whole.
  - Federal government considers MTF one of few studies that can inform federal policy.

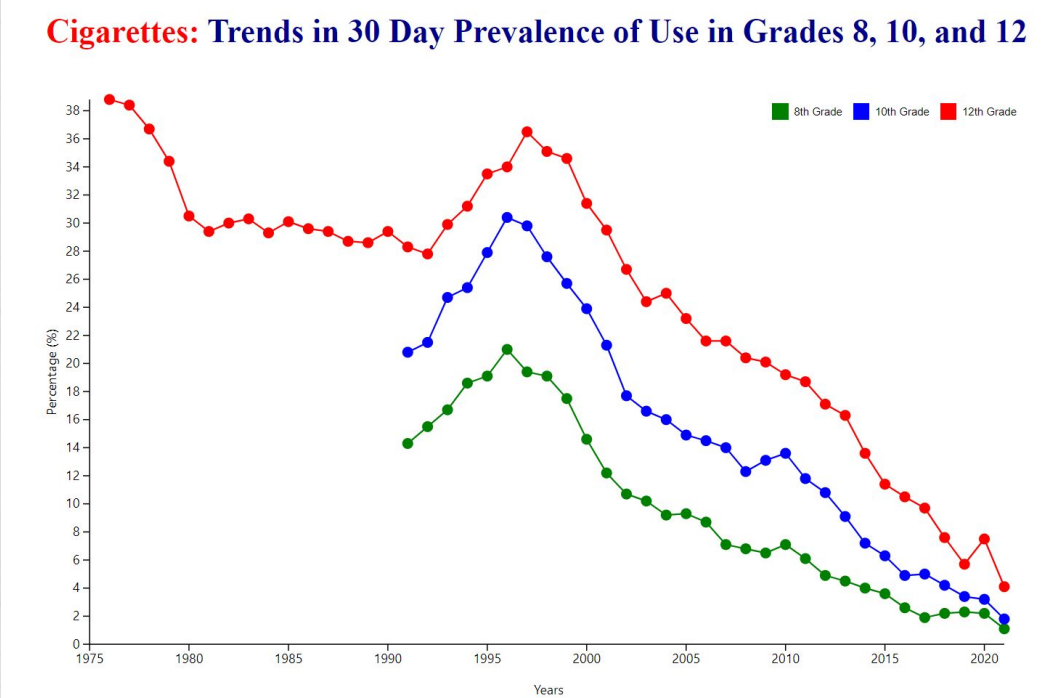
Research Design: Monitoring the Future

- Separate, nationally representative, annual surveys of 8th, 10th, and 12th grade students.
  - Sampled from a list of all public and private schools in the U.S. (under guidance of professional sampling statistician John Haeussler).
  - Approximately 40,000 students total in about 410 public and private secondary schools per year.
  - As of 2021, more than 1.5 million completed surveys.
    - That being said, it is the representativeness of the data and not the large sample size that is the key characteristic
    - Other sampling methods may or may not provide a representative sample – difficult to tell. And representativeness may change over time.

Research Design: Monitoring the Future

- Why the large samples?
  - Many drugs we monitor have low prevalence and require large samples for sufficient sample size
  - Large samples allow detection of small year-to-year changes
  - We can and do run randomized-controlled experiments when updating survey measures
  - Large samples allow analyses of smaller demographic subgroups

Example: Trends in Adolescent Cigarette Smoking



Source: Monitoring the Future

Monitoring the Future: Sampling

- Outline of remaining presentation
  - I will describe MTF's current sampling procedures
  - In doing so, will point to the need to update these procedures
  - Would very much like any update to be seamless – do not want to have a discontinuity in MTF's long term trends that results from a methodological artifact

Monitoring the Future: Sampling

- Three stage design; Stage 1, Geographic Areas:
  - The geographic areas used in this study are based on the geographic strata used in the 1990 Survey Research Center's National Sample of the contiguous United States
    - Divides the U.S. into 110 strata
  - In addition to the 28 largest metropolitan areas containing about one third of the nation's population, there are 80 other strata.
  - MTF selects the eight largest U.S. cities "with certainty."

Monitoring the Future: Sampling

- ISSUES: Three stage design; Stage 1, Geographic areas
  - Cities with largest population not necessarily cities with most students, and city sizes have changed since 1990

Table: Cities That Rank in the Top 8 in 1990 and 2021

1990 "Super 8" Stratum By 1990 Population Size	1990 Stratum by 2021 Student Size
New York	Los Angeles
Los Angeles	New York
Chicago	Dallas
Dallas	Houston
Philadelphia	Chicago
Boston-Lawrence-Salem-Lowell-Brockton	Atlanta
Detroit	Sacramento
Washington DC	Washington DC

Monitoring the Future: Sampling

- ISSUES: Three stage design; Stage 1, Geographic areas
  - Top cities have changed in population size since 1990
    - E.g. Detroit now much smaller, and has fallen out of the top eight.
      - But for sampling purposes MTF still considers Detroit in the top eight and selects it with 100% certainty for all grades
    - E.g. Houston now much larger, and it is now in the top eight.
      - But for sampling purposes MTF still considers Houston smaller and does not select it with 100% certainty for all grades.
  - Household size of a city not a perfect correlate with student size
    - E.g. Sacramento

Monitoring the Future: Sampling

- ISSUES: Three stage design; Stage 1, Geographic Areas
  - Ideally, strata other than largest cities should have similar population size.
    - To explore this I looked at # of students in each strata outside of the largest cities using the MTF frame of all students in the U.S.
      - Smallest strata: 41,244
      - Largest strata: 130,253 (3 times larger)
      - Maybe strata mostly cluster around a common size? Let's take a look

Monitoring the Future: Sampling

- ISSUES: Three stage design; Stage 1, Geographic Areas

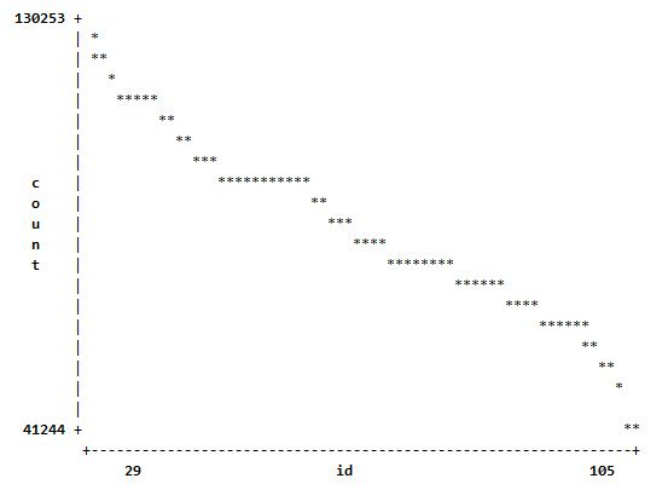


Figure: Number of Students in Each Stratum, in Descending Order
(Top 28 Cities Removed from Analysis Pool)

Monitoring the Future: Sampling

- Three stage design; Stage 2, Schools
  - In the major metropolitan areas, two or more high schools often are included in the sampling design; in most other sampling areas, a single high school is sampled.
  - All schools are selected with probability proportionate to the size of the class (pps). The larger the class, the higher the selection probability.
  - For practical reasons, 12^{th} grade schools with classes <25 are usually excluded from the sample, which this omits about 3% of all seniors from the sampling frame.

Monitoring the Future: Sampling

- Three stage design; Stage 2, Schools continued
  - If a sampled school is unwilling to participate, a replacement school is selected from the same geographic area.

Monitoring the Future: Sampling

- ISSUES: Three stage design; Stage 2, Schools
  - Replacements are becoming more and more common
    - Requires much time and energy from our sampling statistician to examine each school and select a replacement per his judgement
    - Our sampling statistician will be difficult to replace when he retires
    - Would be nice if we could turn our sampling statistician into a bot

Monitoring the Future: Sampling

- Three stage design, Stage 3, Students
  - When possible, survey all of the school's students in a target grade
    - Used to have a maximum of 350, but with transition to web no longer a need for a maximum
  - In some cases school only allows sampling of a subset of students, such as half of the English classes.

Monitoring the Future: Sampling

- ISSUES: Three stage design, Stage 3, Students
  - No issues!

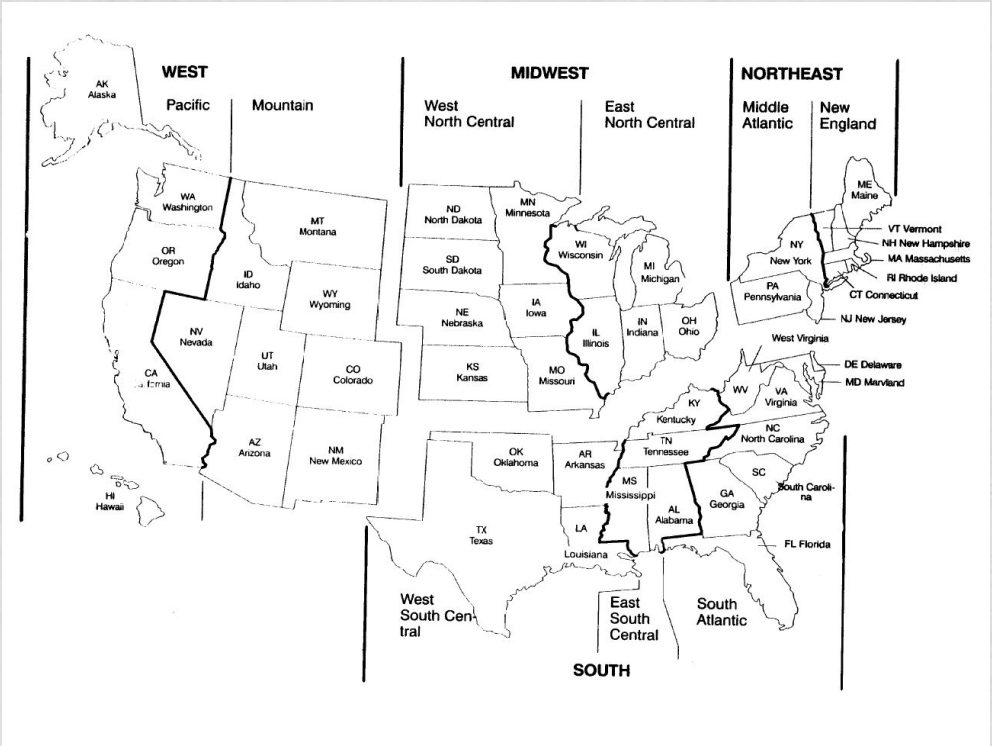


Figure: Nine U.S. Census Geographic Divisions

Figure: Nine U.S. Census Geographic Divisions

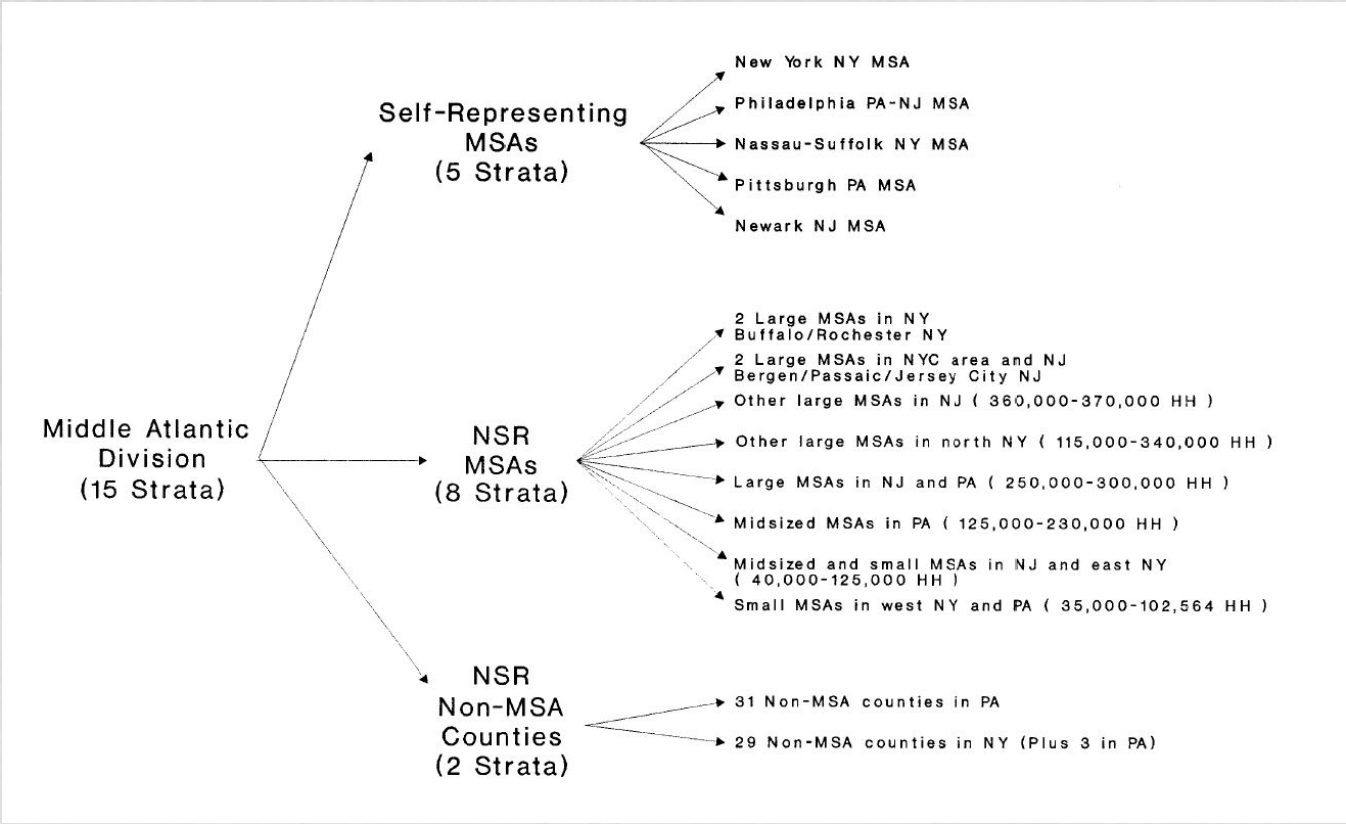


Figure: Fifteen SRC Strata in Middle Atlantic Division

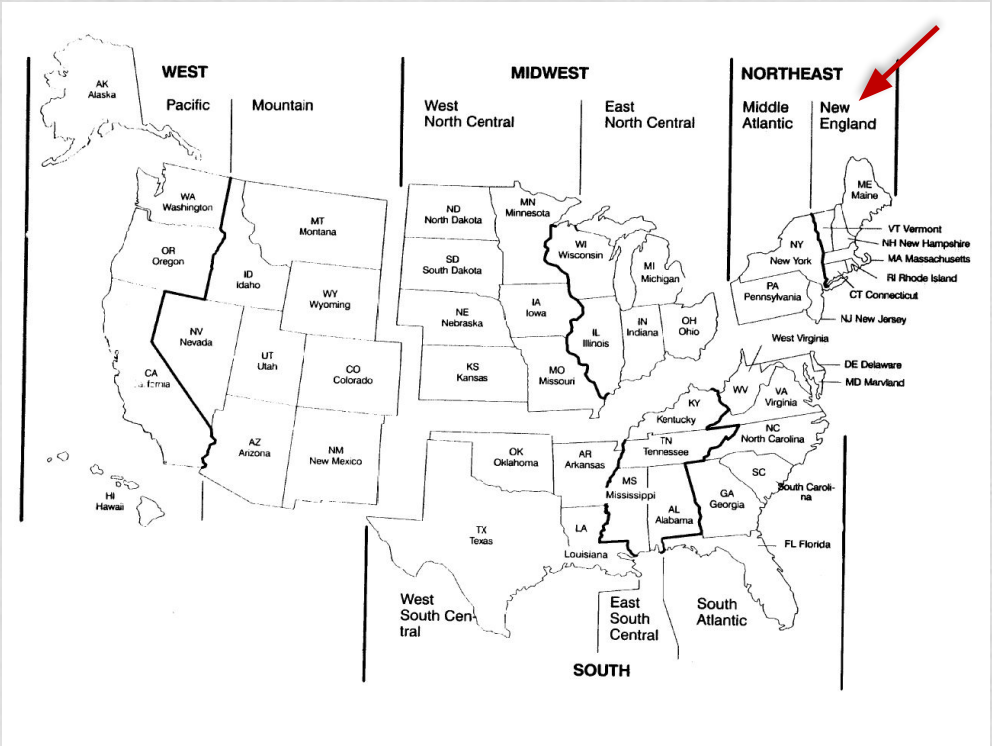


Figure: Nine U.S. Census Geographic Divisions

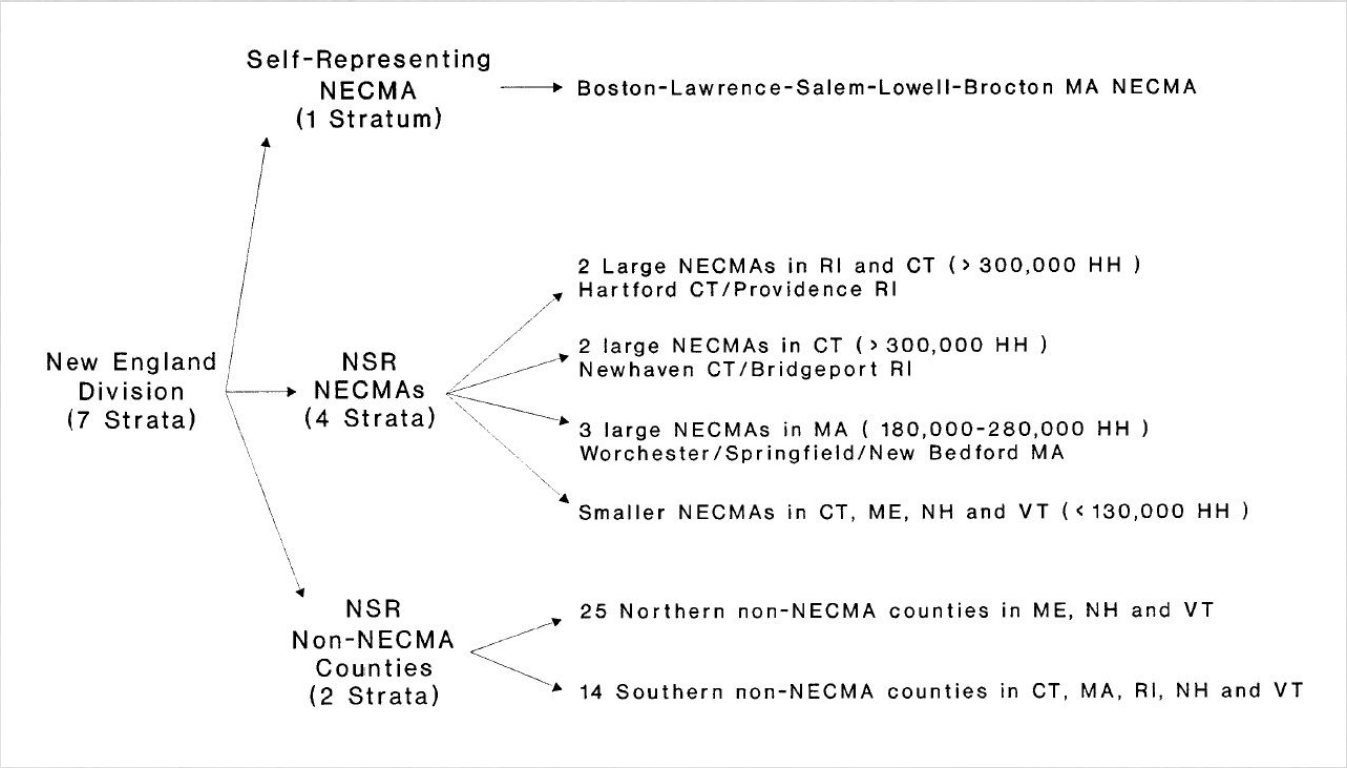


Figure: Seven SRC Strata in New England Division

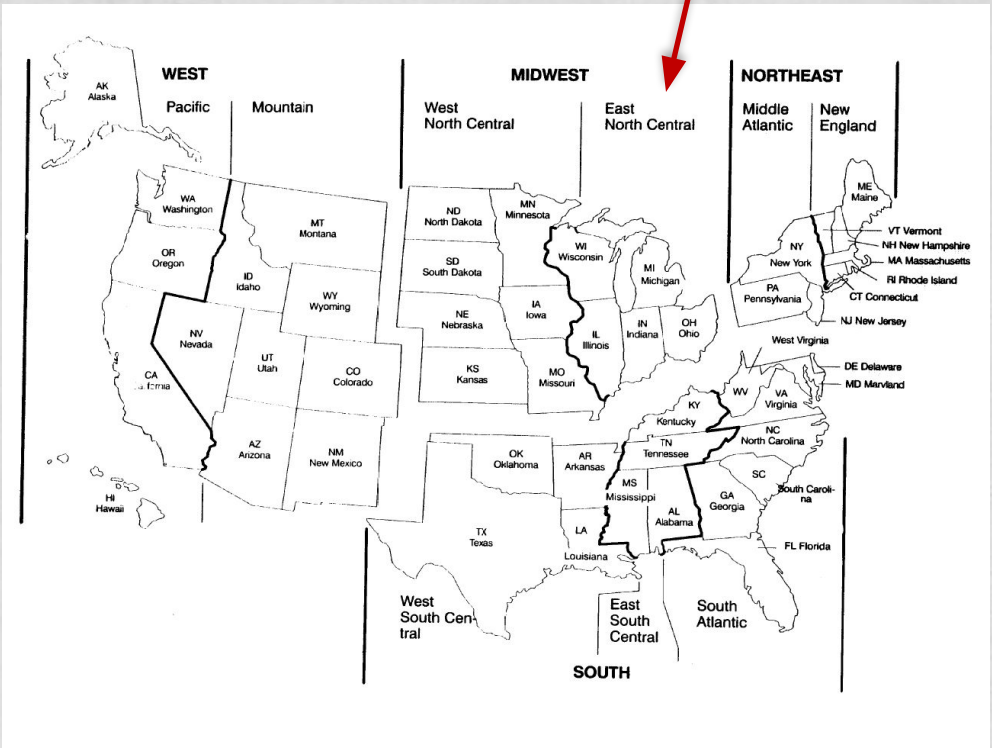


Figure: Nine U.S. Census Geographic Divisions

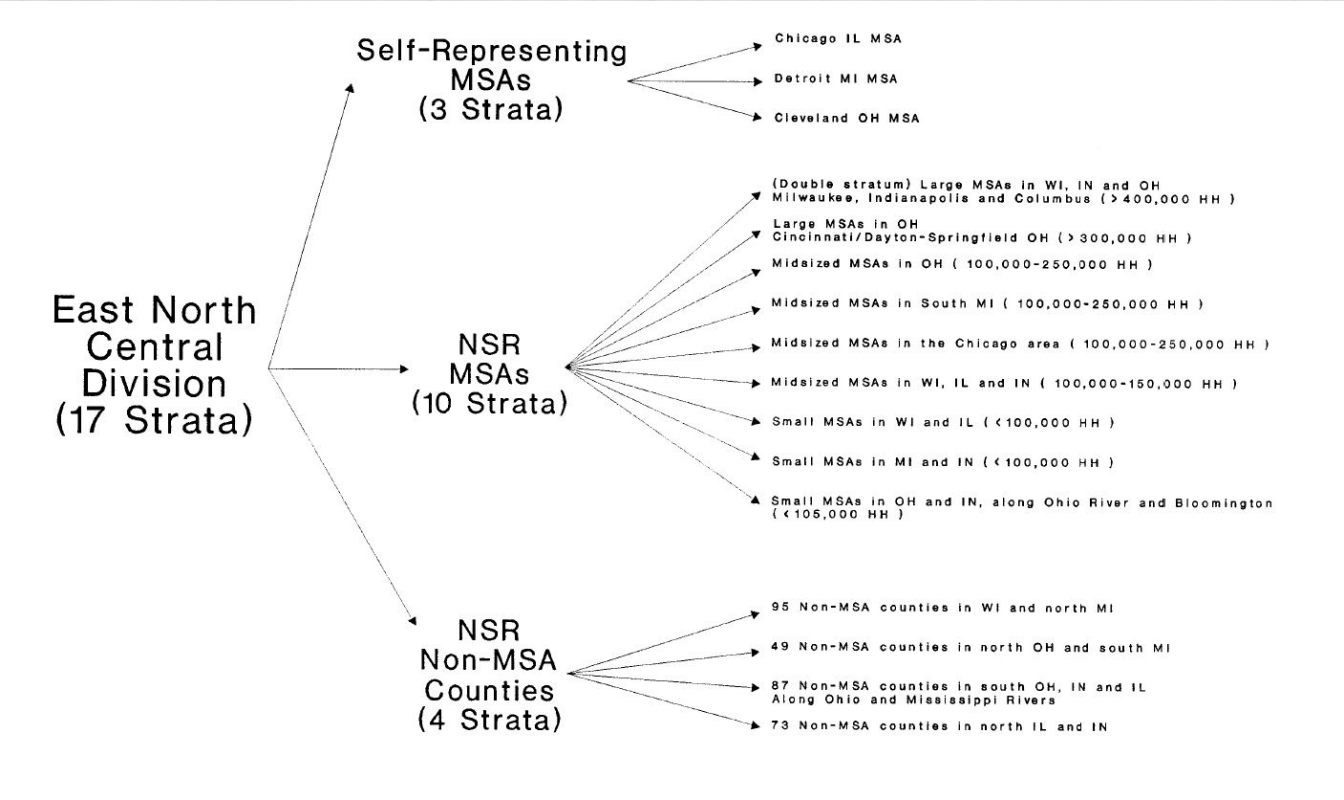


Figure: Seventeen SRC Strata in East North Central Middle Atlantic Division

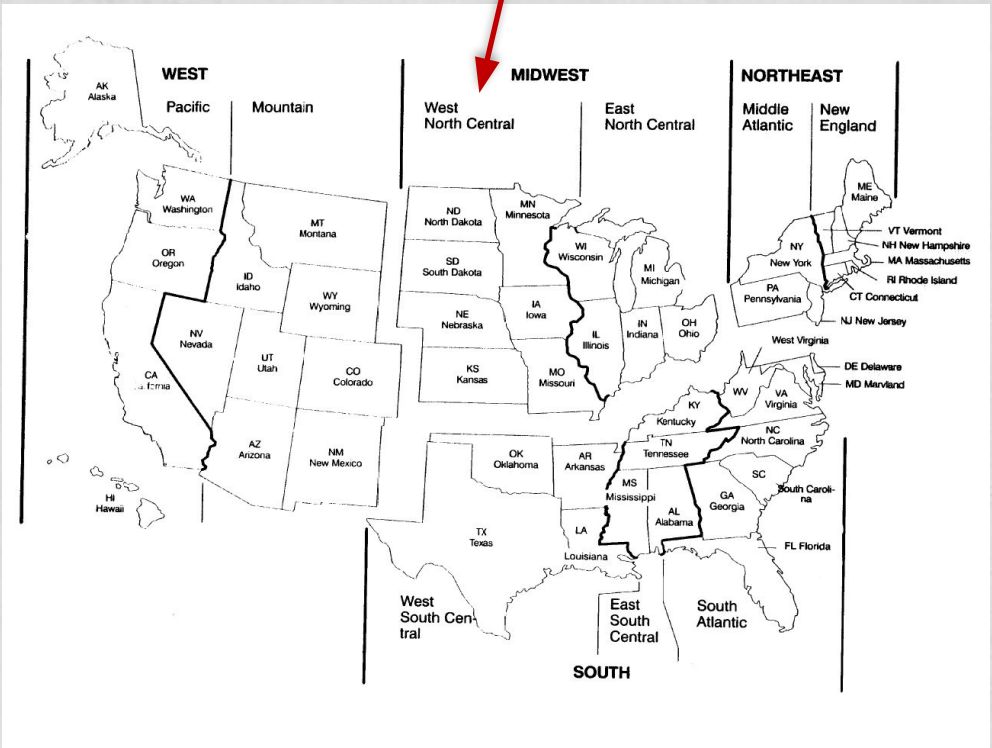


Figure: Nine U.S. Census Geographic Divisions

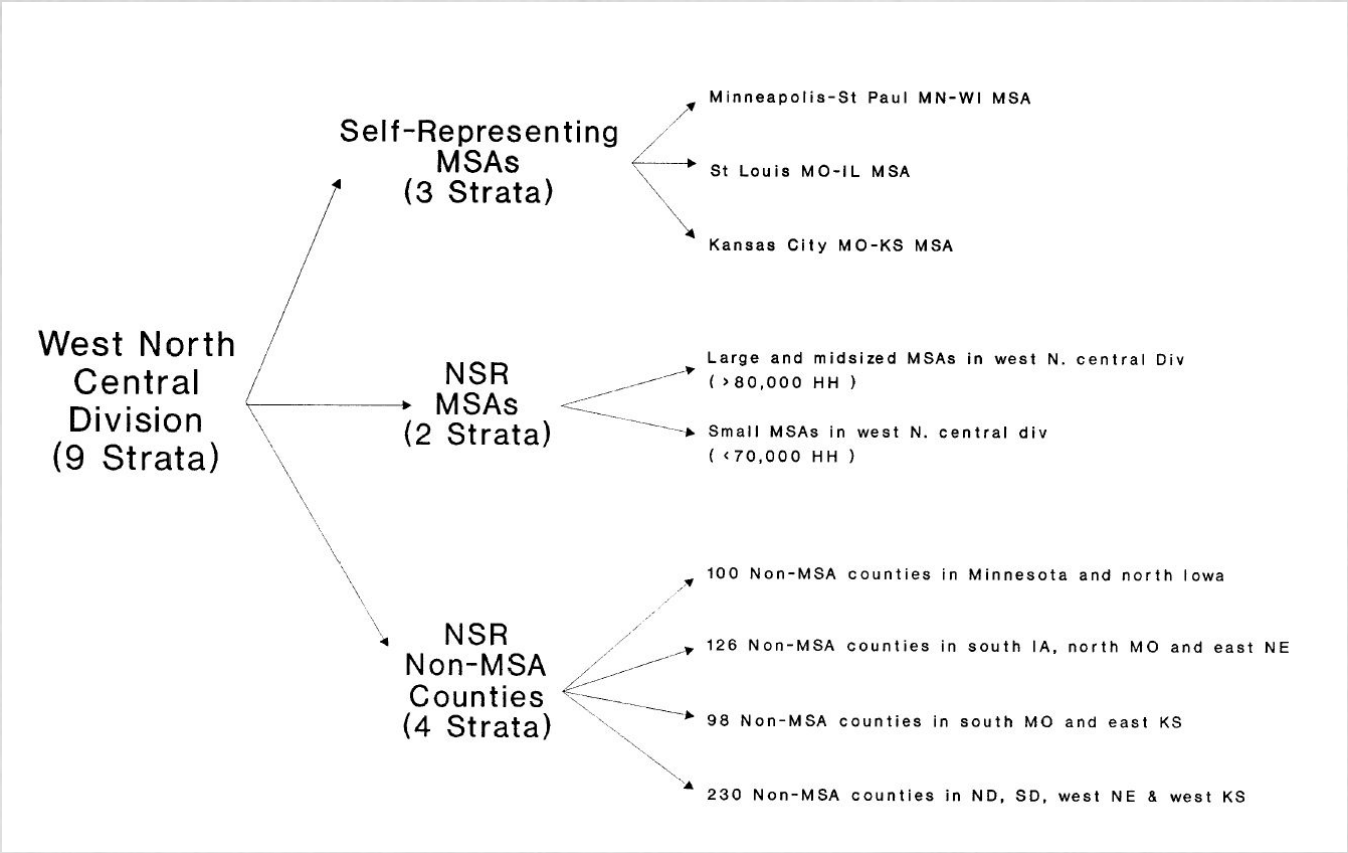


Figure: Nine SRC Strata in West North Central Division

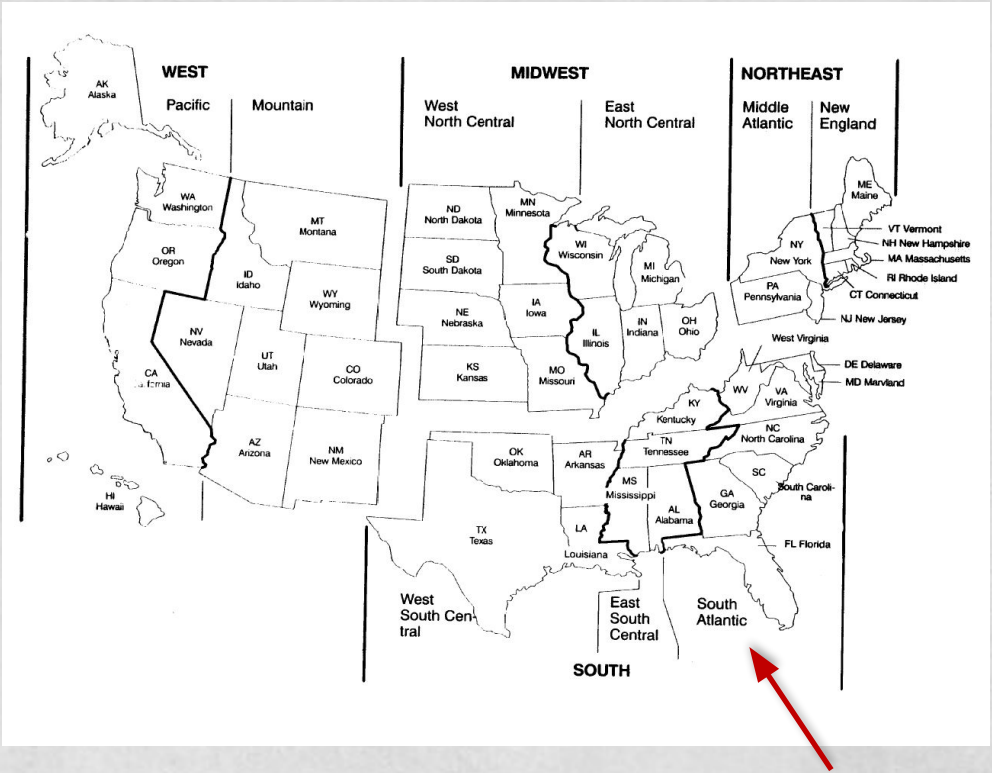


Figure: Nine U.S. Census Geographic Divisions

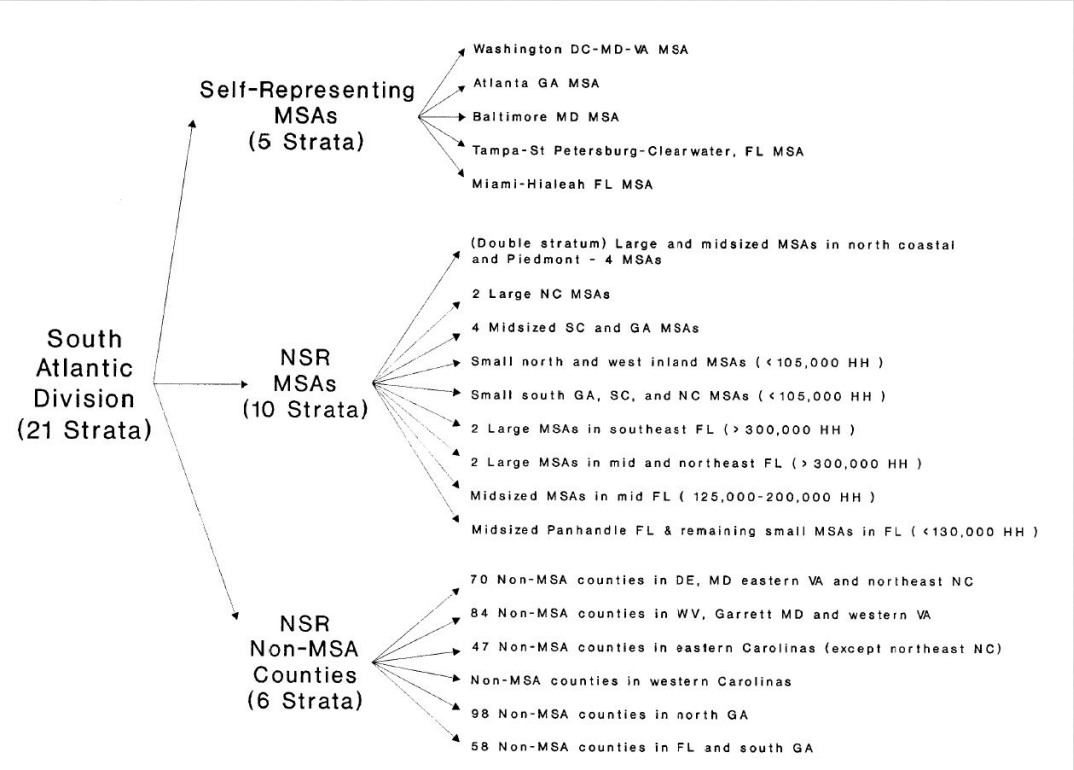


Figure: Nine SRC Strata in South Atlantic Division

Figure: Nine U.S. Census Geographic Divisions

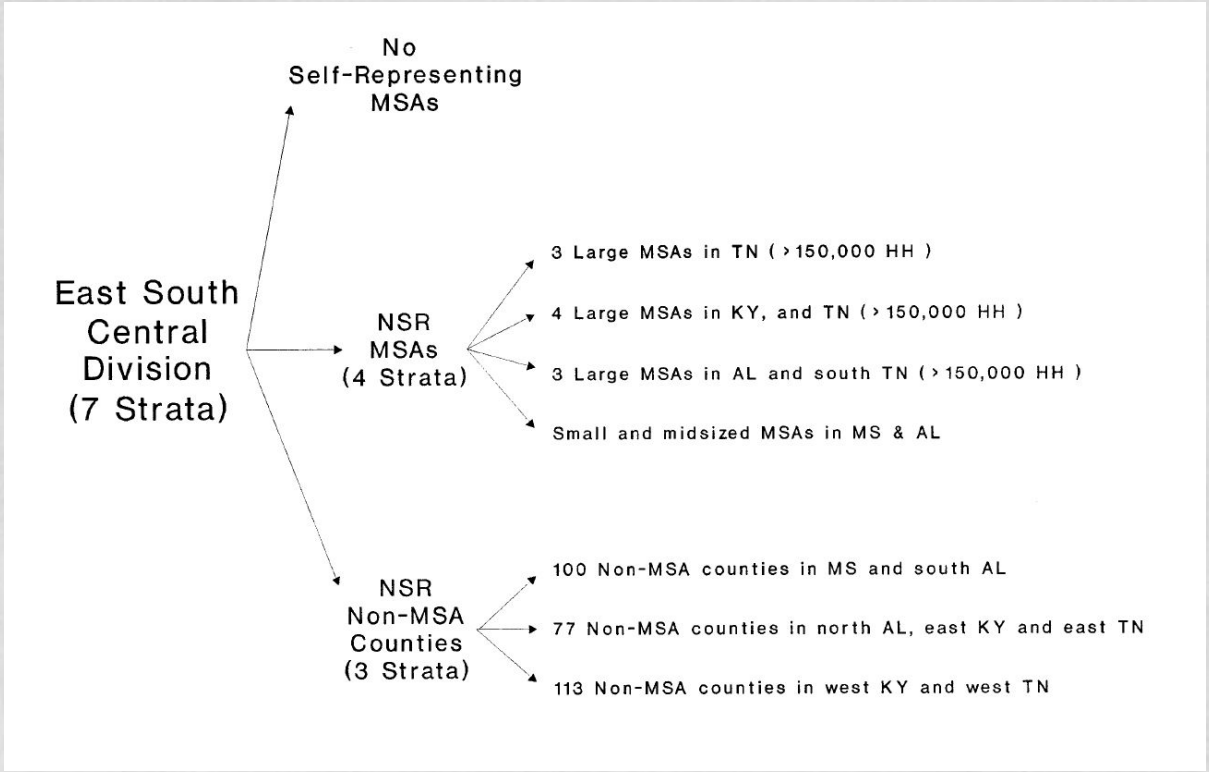


Figure: Seven SRC Strata in East South Central Division

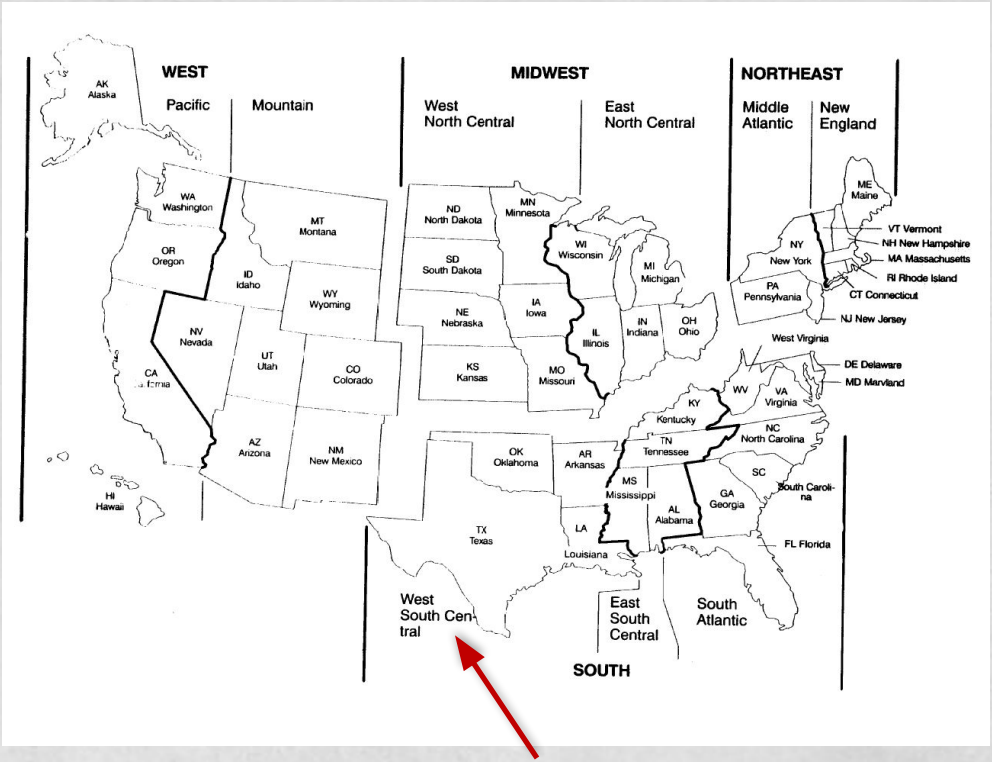


Figure: Nine U.S. Census Geographic Divisions

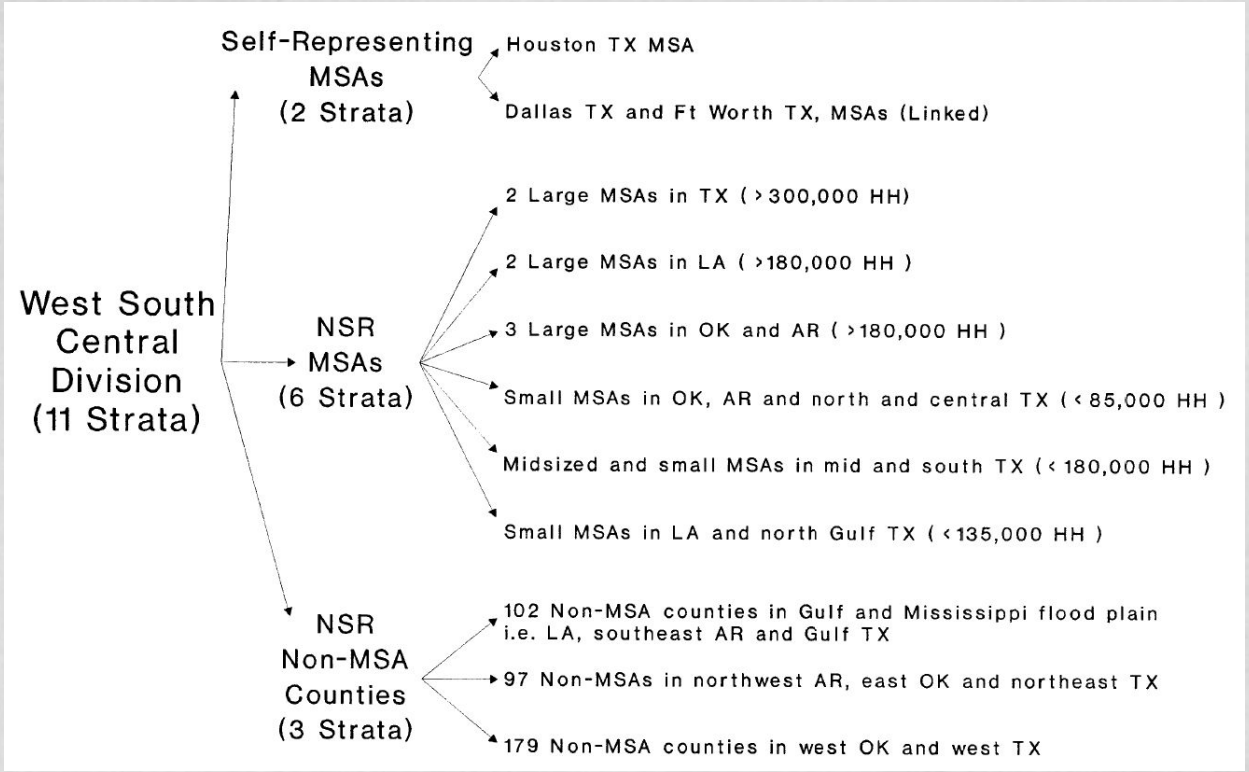


Figure: Seven SRC Strata in West South Central Division

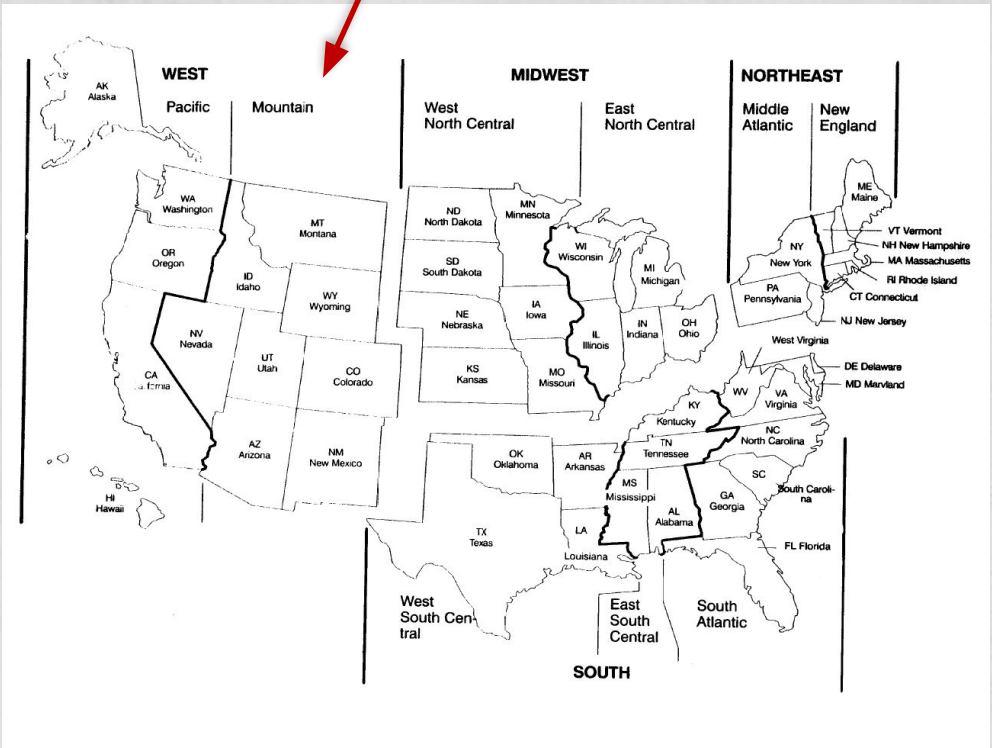


Figure: Nine U.S. Census Geographic Divisions

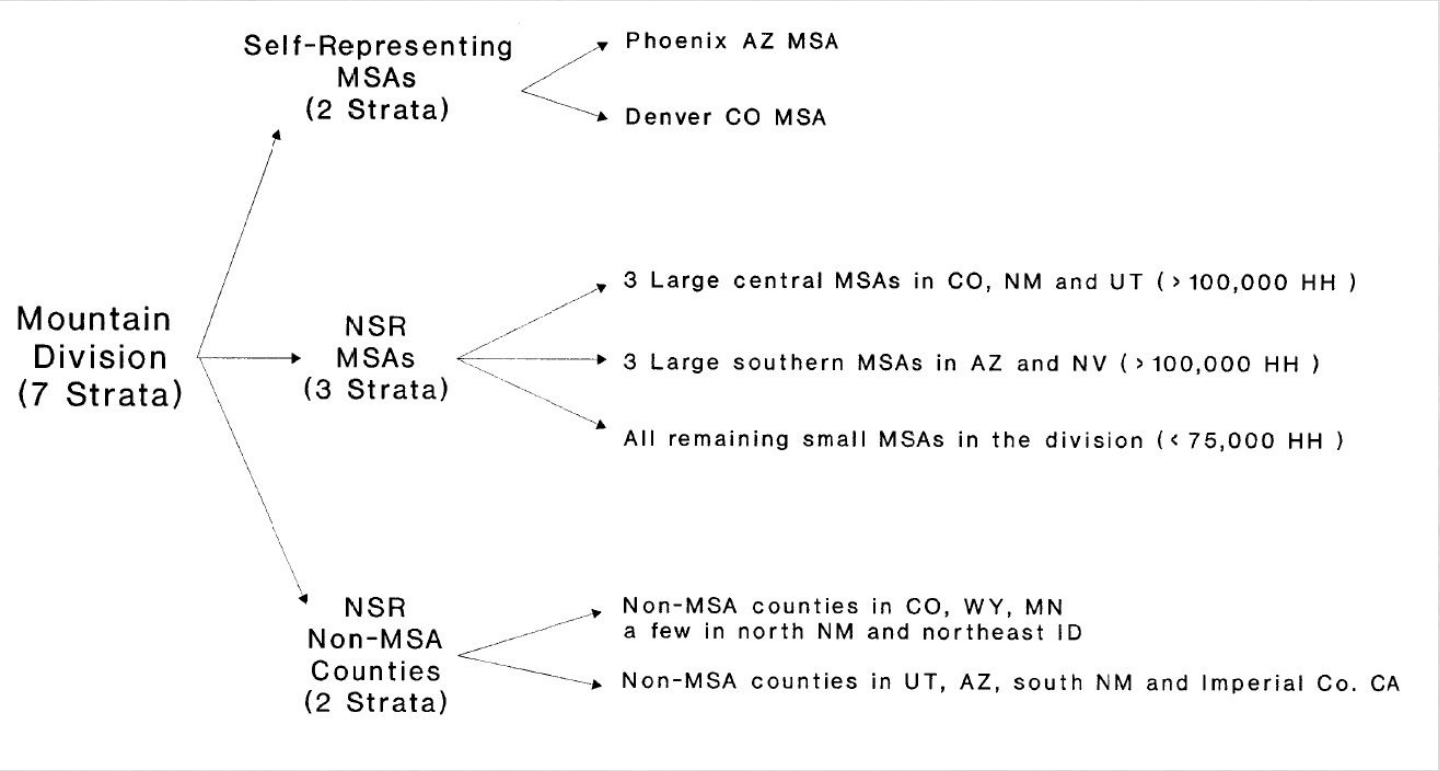


Figure: Seven SRC Strata in Mountain Division

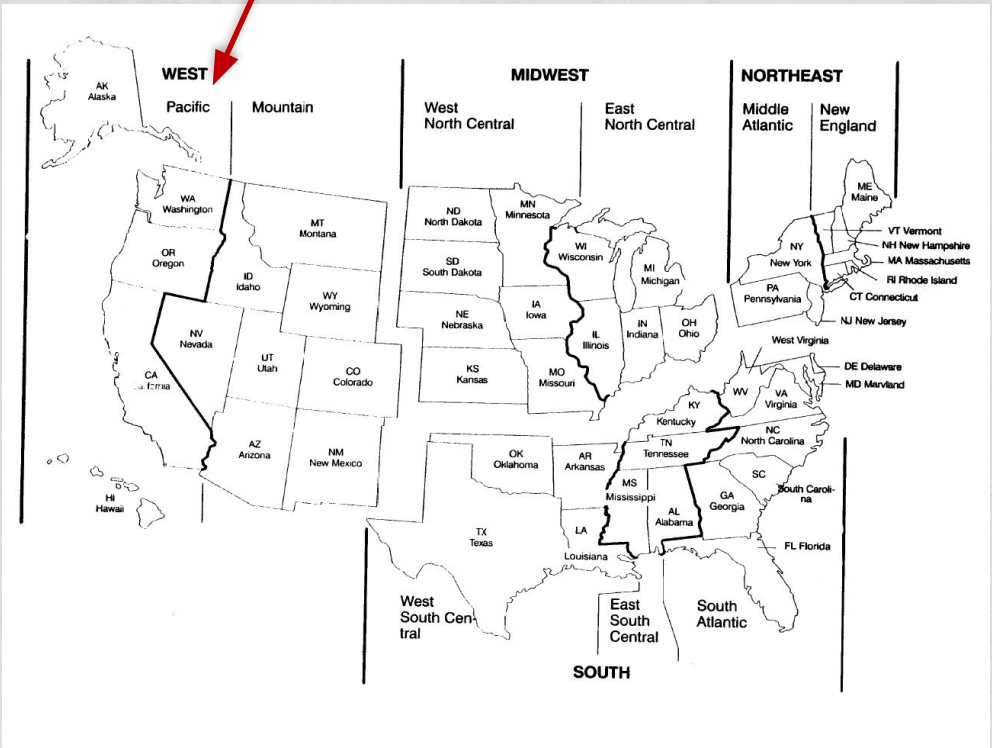


Figure: Nine U.S. Census Geographic Divisions

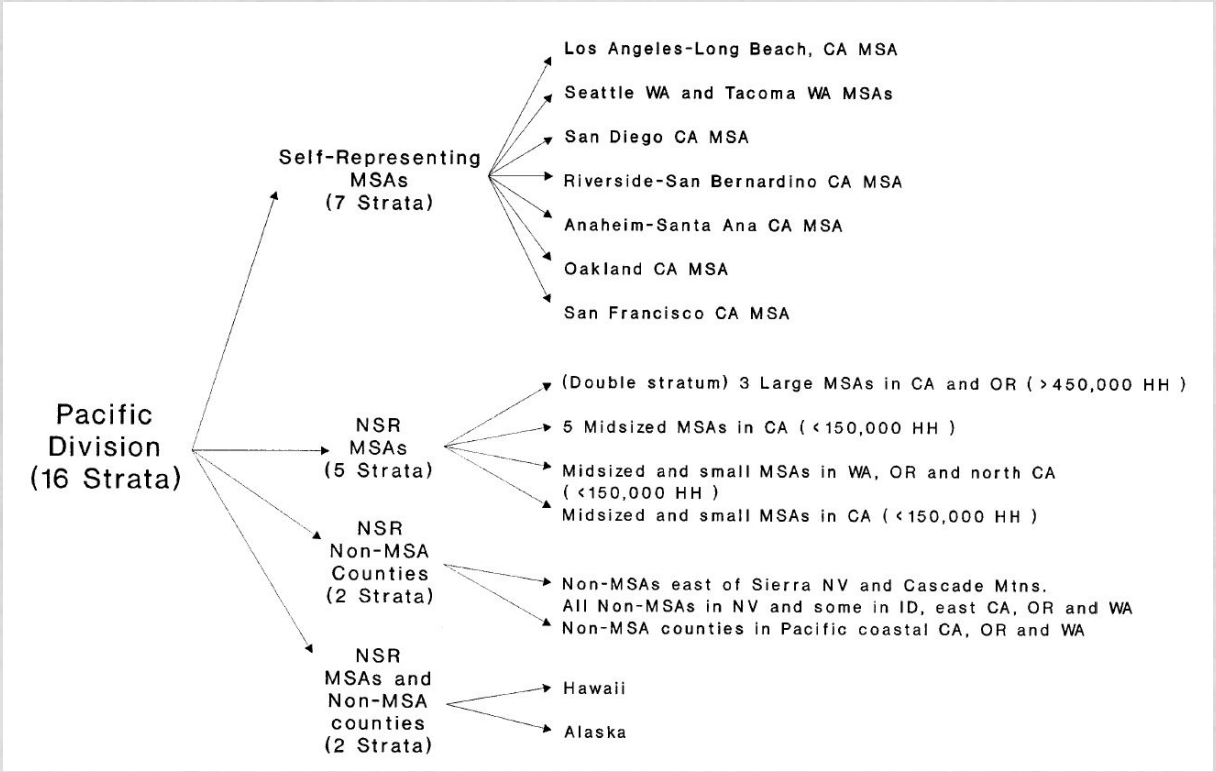


Figure: Sixteen SRC Strata in Pacific Division

Monitoring the Future: Updated Sampling

- Advantageous to have geographic regions – and not schools – as a level of analysis.
- Would like to preserve this, if possible

Monitoring the Future: Updated Sampling

- Choosing major population centers with 100% certainty is advantageous
  - Example of Los Angeles Unified School District

Monitoring the Future: Updated Sampling

- Move away from Metropolitan Statistical Areas (MSAs)?
  - Official definition: An MSA consists of one or more counties that contain a city of 50,000 or more inhabitants, or contain a Census Bureau-defined urbanized area (UA) and have a total population of at least 100,000 (75,000 in New England).
    - MSA boundaries may be shaped more by political than demographic factors, especially in more recent decades.

Monitoring the Future: Updated Sampling

- Move away from Metropolitan Statistical Areas (MSAs)?
  - On January 19, 2021, OMB submitted a regulation for public comment that would increase the minimum population needed for an urban area population to be a metropolitan statistical area to be increased from 50,000 to 100,000.
    - Substantial change!

Monitoring the Future: Updated Sampling

- Use analyses of current data to inform decisions on future sampling?
  - Find factors strongly associated with drug use, and incorporate them into the sampling design
    - E.g., we could oversample by race/ethnicity

Monitoring the Future: Updated Sampling

- Use analyses of current data to inform decisions on future sampling?
  - I am a heretic on this issue
    - MTF samples over 70 drugs.
      - Easy for me to envision that the analyses would not provide much direction; e.g. some factors important for some drug but not others
    - Documentation of emerging drug use, such as vaping, crack cocaine, and ecstasy, is a major charge of MTF. How does analysis of current data help us survey drugs that do not yet exist?
    - Would rather just be able to say that data are nationally-representative, and let large sample size do the rest.

Monitoring the Future: Updated Sampling

- Current plans:
  - 9 Census divisions, by 3 levels of population size (e.g., city, suburban, rural), by 3 levels of school size
    - This amounts to 9*3*3=81 strata
  - Geographic unit could be counties, school districts, municipalities, etc.
    - Would like to take out largest population centers and form similar-sized strata with remainder
      - Use AI algorithm for remainder, or expert judgement?