Data quality

Considerations when using linked data

Amelia B. Finaret, PhD | 8 October 2021

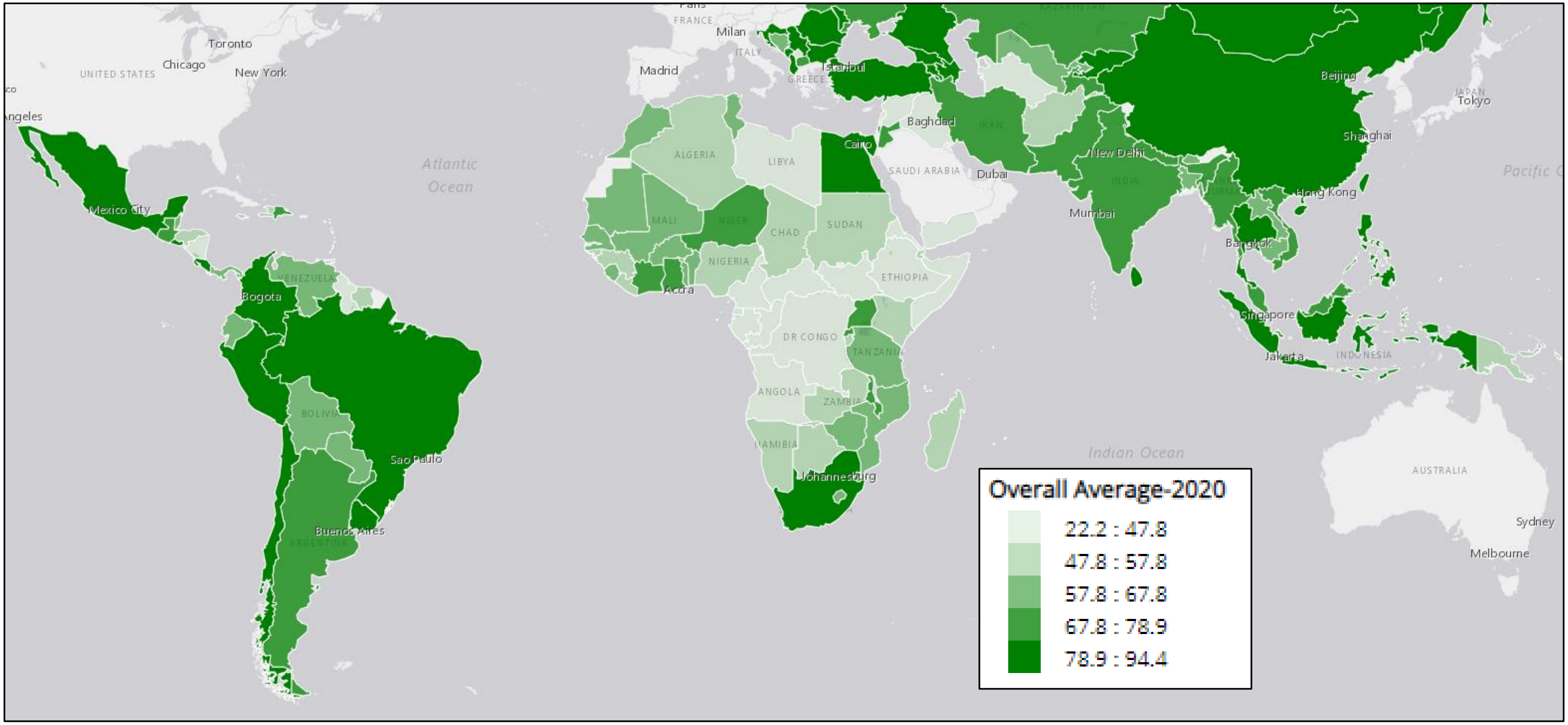
Evidence-based policymaking for applied economists | AAEA Workshop

ALLEGHENY COLLEGE

What is data quality and why should we care?

MOTIVATION AND DEFINITIONS

National statistical capacity varies around the world – SDG 17 focuses on improving national statistical capacity



Overall average of the World Bank's Statistical Capacity Indicator, National-level data for 2020
Source: World Bank Bulletin Board on Statistical Capacity

As we've been hearing, datasets were constructed with different goals in mind

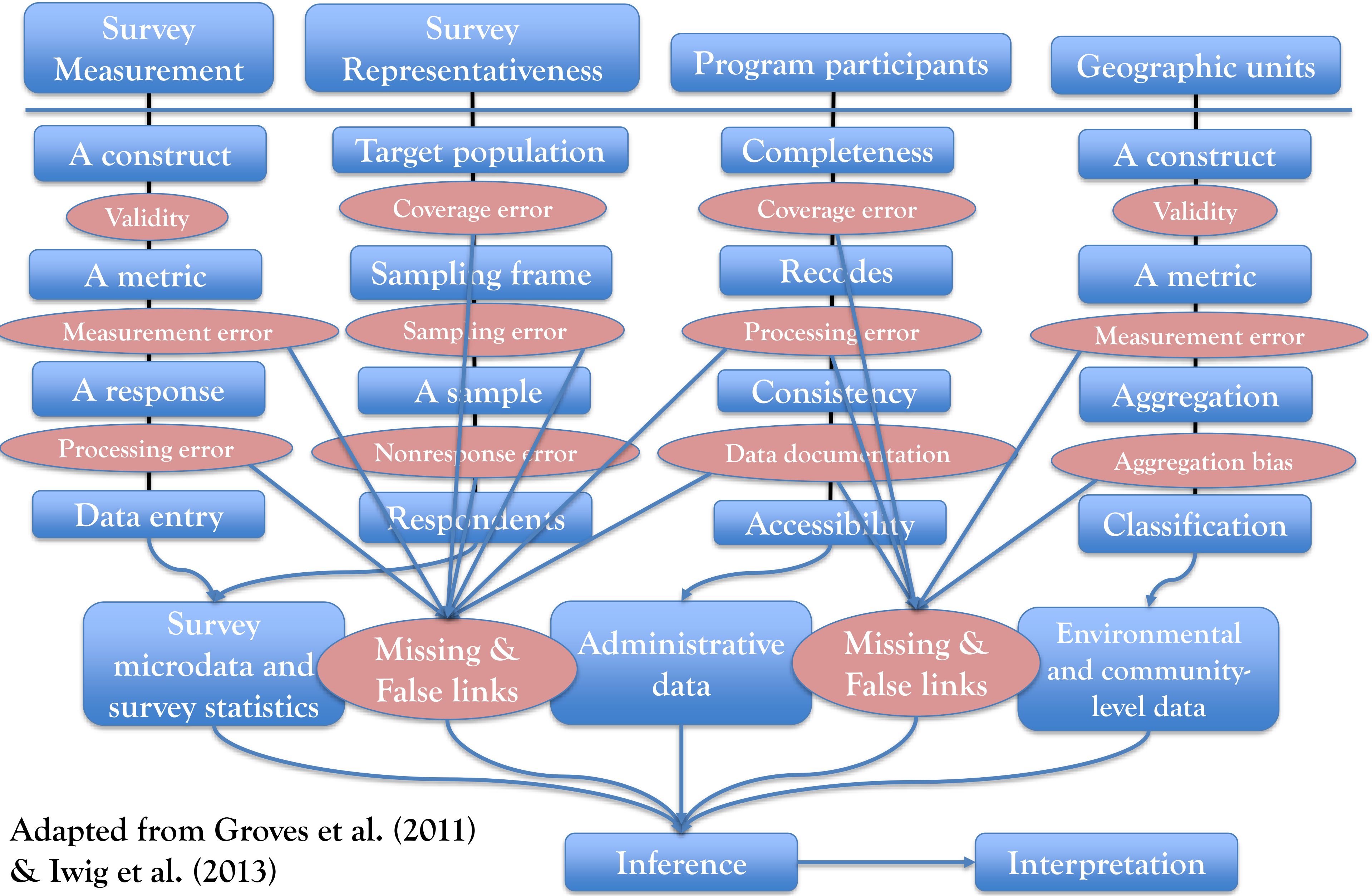
- Linking datasets can improve data quality directly (Citro 2014)
  - E.g., using admin records to correct non-response
  - E.g., to inform imputation or validation
  - E.g., to compare aggregate/prevalence estimates
- Linking data can also introduce new problems with data quality
- Potential to make incorrect inferences
- Newly linked datasets may have unique qualities
- Federal agencies have developed systems for examining data quality for CPS, ACS, etc.
  - But it is still *up to the researcher* to do this analysis for one's own project

How could we define ‘data quality’?

- **Many definitions**
  - The definition for you as a researcher depends on *your* specific project, research question, and data used.
  - A good definition would include issues related to:
    - Completeness
    - Systematic errors (bias)
    - Variance (noise)
- **The Eurostat Quality Assurance Framework** describes five broad dimensions:
  - Relevance
  - Accuracy and reliability
  - Timeliness and punctuality
  - Accessibility and clarity
  - Coherence and comparability

What does "data quality" mean?

- **A high-quality linked dataset has:**
  - a high degree of completeness
  - no systematic differences between linked and non-linked observations
  - an acceptable amount of random noise
  - no differences in precision or accuracy across any observable characteristics
  - the potential to answer your research question



Adapted from Groves et al. (2011)
& Iwig et al. (2013)

Types of data errors in linked data

- Missing true links
  - Type II error / false negative
- False links
  - Type I error / false positive
- Dataset(s) might have missing values
  - Data might be missing at random, or not
  - Reduces sample size and increases standard errors
- Dataset(s) might have incorrect values
  - Some incorrect values can be spotted, others not
  - For health data, biological plausibility is helpful here
- Syntactic errors can prevent true links or cause false links
  - Spelling, data entry, format of data
- Semantic errors can prevent true links or cause false links
  - Meaning of the variable is unclear

National household surveys are an amazing source of data, but the quality still needs to be assessed when doing estimates or inference

NATIONAL HOUSEHOLD SURVEYS

National household surveys

- Demographic and Health Surveys
  - USAID and national governments
- Living Standards Measurement Study
  - World Bank and national governments
- Multiple Indicator Cluster Survey
  - UNICEF and national governments

A systematic review of articles which explored data quality for household surveys in Africa*

- 47 articles included
- Majority focus on anthropometric data and immunization data.
- Main issues explored:
  - Age heaping – people like 0s & 5s for years; 6s & 12s for months
  - Cleaning criteria – what is biologically plausible?
  - Garbage codes
  - Intra and inter observer error
  - Field work – weather, roads, security
  - Skip patterns
  - Data completeness, accuracy, plausibility
- Many claims of "poor data quality"
  - This declaration/judgement should be based on the quality of research and inference that can be done using the data, not on the raw dataset itself without context.

*My student A. Smith, Allegheny College '22 conducted the systematic review

Types of survey data quality issues: Sampling and non-sampling error

- **Sampling errors**
  - Seemingly random, happen even if no mistakes were made
  - Can be reduced if sample size is increased
  - Attenuates coefficient estimates towards the null hypothesis
  - Increase likelihood of a Type II error; wider confidence intervals
  - Random measurement error can reduce the impact of a study for funders or policymakers
- **Non-sampling errors**
  - Can be systematic or random
  - Harder to estimate/know about than sampling error
  - Hard to understand the underlying cause of
  - Causes attenuation towards the null, or inflation away from the null
  - Most data quality research focuses on sampling errors

Modeling the uncertainty of estimates by linking geospatial data with survey data

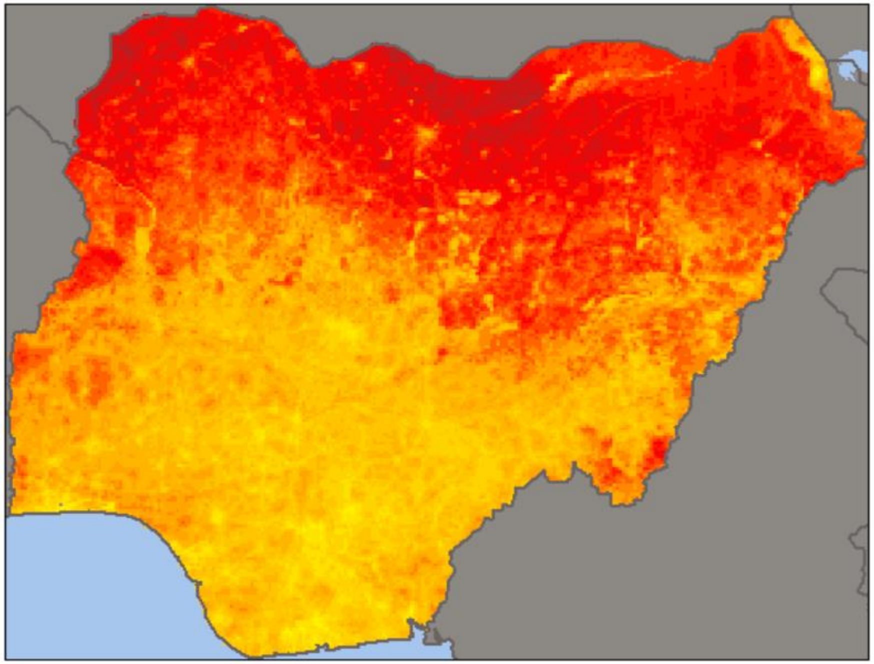
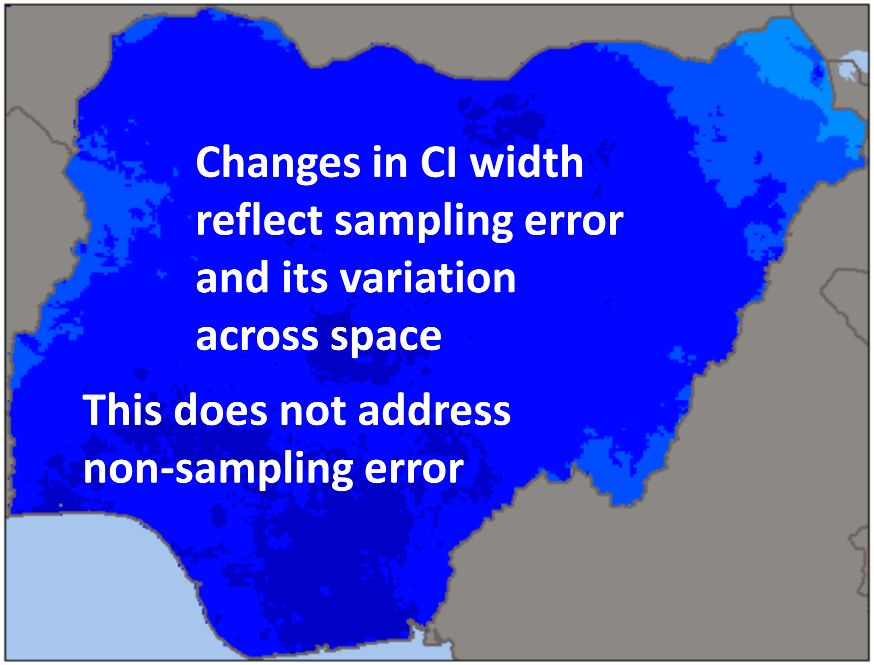


Figure 1. Interpolated surface for the indicator. The map plots the point estimate for each 5x5 km pixel based on geo−located cluster−level data from the survey.



Changes in CI width reflect sampling error and its variation across space

This does not address non-sampling error

Figure 2. Uncertainty surface for the indicator. The map plots the uncertainty for each pixel, measured using the width of the 95% credible intervals.

*Spatial Data Repository, The Demographic and Health Surveys Program. Modeled Surfaces. ICF International. Available from: https://spatialdata.dhsprogram.com/modeled-surfaces/

How can data quality be improved?

Before and during data collection
It starts here!

- Training enumerators
- Using hard measures vs. soft
- Validate survey instruments
- Interview mode
- Enumerator gender
- Question wording
- Question order
- Reduce survey fatigue
- Deal with heaping
- Understand incentives for respondents and enumerators

Before and during data analysis
Can fix some things, but not all!

- Cross-check by linking
- Estimate reliability of instruments
- Perform sensitivity analysis
- Account for sampling design
- Estimate design-based sampling errors
- Imputation, weighting, Bayesian approaches
- Assess heaping

Steps to analyzing the quality of your linked dataset

- Check for accuracy of links given the linkage process that you used
- Check for differences in observable characteristics between linked and non-linked observations
- Analyze each variable and assess missingness, heaping, SDs, and other errors
- Fix what you can fix, analyze and interpret accordingly

A particular concern for data linkage projects

MISSING DATA

On missing data as a data quality concern: What type of missing data do you have?

- Missing completely at random (MCAR)
  - True random missingness
- Missing at random (MAR)
  - Random missingness can be fully explained with observable factors (but not the missing variable itself)
    - Can't prove this, but can do inference
- Missing not at random (MNAR)
  - Differences between missing and non-missing remains after taking observable differences into account
    - Can prove that data is MNAR if missingness is not fully explained by observable factors

How should I deal with missing data? (Allison 2001)

- In general, don't:
  - Delete all observations with missing values
    - There is information in those observations still!
  - Impute with sample mean
- Instead, try:
  - Finding the missing information from another source
    - Like a data linkage!
  - Multiple imputation
    - Create new versions of the dataset which impute missing data with predicted distributions based on observed data. Can average total estimated effects; they will be different depending on the dataset.
    - On the "mi" commands in Stata from UCLA - IDRE
  - Maximum likelihood
  - Random effects/mixed-effects models
  - Sensitivity analyses
  - Weighting the analysis

When using multiple imputation methods, be careful! (Sterne et al. 2009)

- Don't omit the outcome variable from modeling the variables w/ missingness
- Don't assume everything is normal
- Are your data *really* missing at random?
  - This cannot be proven, just hypothesized. What could be driving missingness in your newly linked dataset?
- Multiple imputation is computationally demanding
- Always report results with clear description of any imputation process and an analysis of missingness.

What if I have missing data not at random?

- Interpretation of your results – knowing matters, even if you can't fix the data quality issues
  - Root cause analysis
  - Consequence of missing data depends on the study
- Reducing respondent and enumerator burden
- Sensitivity analysis for direction of bias(es)
- More resources for national statistics offices
- More resources for household survey enumeration
- Incentivize accuracy instead of reaching a threshold

Stata commands & packages for understanding missing data

- Most Stata commands will automatically drop missing values
  - Default for linear regression is to ignore whole observation listwise
- **misstable**
- **mdsec** counts the number of missing values for all variables in a table, and calculates proportion of missing
- **rmiss2** (for string variables and numeric variables) is an extension to the "generate" command to calculate the number of missing values for each observation.
- **mvpatterns** creates tables that show patterns of missing values across observations.
- **misschk** makes the same calculations as the above packages but is for numerical data only (variable label can still be string).

See detailed descriptions here from UCLA IDRE

R resources, commands & packages for understanding missing data

- Julie Josse, Nicholas Tierney, and Nathalie Vialaneix put together a guide to R resources on missing data (*Comprehensive R Archive Network*)
- Split into topics:
  - Exploration of missing data
  - Likelihood based approaches
  - Single imputation
  - Multiple imputation
  - Weighting methods
  - Specific types of data
  - Specific fields
    - fastLink for admin records/surveys – from Enamorado, Fifield, and Imai (2019), "Using a probabilistic model to assist merging of large-scale administrative records."

Conclusion: Always analyze the quality of your data when linking datasets

- New data quality and coverage issues may arise when linking data
- There are tools to deal with some data quality problems, but other problems are harder to solve
- Understanding how the limitations of your dataset will affect inference is essential
- Aim for clear communication and description of data quality when you interpret results

REFERENCES AND RESOURCES

Useful references for understanding administrative and survey data quality

- Iwig, W., Berning, M., Marck, P. and **Prell, M.**, 2013. Data quality assessment tool for administrative data. *Prepared for a subcommittee of the Federal Committee on Statistical Methodology, Washington, DC (February).*
  - The authors develop 43 questions for researchers on the quality of the data they are using, organized into three phases of data work
- Davern, M., Roemer, M. and Thomas, W., 2014. Merging Survey Data with Administrative Data for Health Research Purposes. *Health Survey Methods*, pp.695-716.
  - Chapter which contains great background material in textbook style, with citations to relevant scholarly literature and examples. Additional chapters are helpful too.

Additional references

- Allison, P.D., 2001. *Missing data*. Sage publications. Series: Quantitative Applications in the Social Sciences
- Citro CF. From multiple modes for surveys to multiple data sources for estimates. Survey Methodology. 2014 Sep 3;40(2):137-61.
- Davern, M., Roemer, M. and Thomas, W., 2014. Merging Survey Data with Administrative Data for Health Research Purposes. *Health Survey Methods*, pp.695-716.
- Iwig, W., Berning, M., Marck, P. and Prell, M., 2013. Data quality assessment tool for administrative data. *Prepared for a subcommittee of the Federal Committee on Statistical Methodology, Washington, DC (February).*
- Spatial Data Repository, The Demographic and Health Surveys Program. Modeled Surfaces. ICF International. Available from: https://spatialdata.dhsprogram.com/modeled-surfaces/
- Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M. and Carpenter, J.R., 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj, 338*