Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data

Zachary H. Seeskin  
*NORC at the University of Chicago*

A. Rupa Datta  
*NORC at the University of Chicago*

Gabriel Ugarte  
*University of Chicago*

Follow this and additional works at: [https://repository.upenn.edu/admindata_conferences_presentations_2018](https://repository.upenn.edu/admindata_conferences_presentations_2018)

Seeskin, Zachary H.; Datta, A. Rupa; and Ugarte, Gabriel, "Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data" (2018). 2018 ADRF Network Research Conference Presentations. 32.  
[https://repository.upenn.edu/admindata_conferences_presentations_2018/32](https://repository.upenn.edu/admindata_conferences_presentations_2018/32)

DOI [https://doi.org/10.23889/ijpds.v3i5.1053](https://doi.org/10.23889/ijpds.v3i5.1053)

This paper is posted at ScholarlyCommons. [https://repository.upenn.edu/admindata_conferences_presentations_2018](https://repository.upenn.edu/admindata_conferences_presentations_2018)  
For more information, please contact repository@pobox.upenn.edu.
Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data

Abstract
State and local agencies administering programs have in their administrative data a powerful resource for policy analysis to inform evaluation and guide improvement of their programs. Understanding different aspects of their administrative data quality is critical for agencies to conduct such analyses and to improve their data for future use. However, state and local agencies often lack the resources and training for staff to conduct rigorous evaluations of data quality. We describe our efforts developing tools that can be used to assess data quality as well as the challenges encountered in constructing these tools. The toolkit focuses on critical dimensions of quality for analyzing an administrative dataset, including checks on data accuracy, the completeness of the records, and the comparability of the data over time and among subgroups of interest. State and local administrative databases often include a longitudinal component which our toolkit also aims to exploit to help evaluate data quality. While we seek to develop general tools for common data quality analyses, most administrative datasets have particularities that can benefit from a customized analysis building on our toolkit. In addition, we incorporate data visualization to draw attention to sets of records or variables that contain outliers or for which quality may be a concern.

Comments
DOI https://doi.org/10.23889/ijpds.v3i5.1053

This presentation is available at ScholarlyCommons: https://repository.upenn.edu/admindata_conferences_presentations_2018/32
Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data

Zachary H. Seeskin
A. Rupa Datta
Gabriel Ugarte

NORC at the University of Chicago

ADRF Network Research Conference
Washington, DC
November 13, 2018

Disclaimer: This research was supported by the Family Self-Sufficiency Research Consortium, Grant Number #90PD0272, funded by the Office of Planning, Research, and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services to the University of Chicago, with NORC at the University of Chicago as a sub-grantee. The views expressed are solely those of the authors and do not necessarily represent the views of the Office of Planning, Research, and Evaluation.
Motivation

- Understanding data quality critical for expanding informed use of state and local administrative data sources for research
- However, few resources exist to support evaluation of quality of state and local data
  - Literature largely based on federal statistical agencies
  - State/local data face particular challenges (Allard et al. 2017)
- We construct a data quality toolkit to help fulfill this need
  - Provide best practices
  - Incorporate descriptive statistics and multivariate visualization
Overview

1. Background
   a. State and local data
   b. Data quality

2. Toolkit and elements—based on quality dimensions

3. Challenges for constructing a toolkit

4. Conclusion
Issues Particular to State and Local Data

- **Challenges at state/local agencies:**
  - May have outdated IT systems for supporting traditional datasets
  - Common quality concerns: data entry errors, missing data, duplicate records
  - Sometimes have lack of clear metadata and documentation

- **Aspects particular to state/local data:**
  - May have varying quality for different variables based on their importance for program administration
  - Represent special populations without ready official statistics available
  - Subject to changes in eligibility rules over time with groups differentially affected by policy changes
  - Often track participants longitudinally
## Data Quality Dimensions from Literature

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Degree to which statistics meet needs of user, including whether data provide what is needed for use or research topic.</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Whether data values reflect true values and are processed correctly.</td>
</tr>
<tr>
<td>Completeness</td>
<td>Whether data cover population of interest, include correct records, and do not contain duplicate or out-of-scope records. Additionally, whether cases have information filled in for all appropriate fields without missing data.</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Whether the data are available in time to inform policy matters of interest.</td>
</tr>
<tr>
<td>Accessibility</td>
<td>The conditions in which users can obtain and work with the data, including physical conditions and legal requirements for access.</td>
</tr>
<tr>
<td>Clarity/Interpretability</td>
<td>Whether data are accompanied by sufficient and appropriate metadata to understand the data and their quality.</td>
</tr>
<tr>
<td>Coherence/Consistency</td>
<td>Data from different sources are based on the same approaches, classifications, and methodologies, with enough metadata available to support combining information from different sources.</td>
</tr>
<tr>
<td>Comparability</td>
<td>Extent to which differences between statistics reflect real phenomena rather than methodological differences. Types of comparability: over time, across geographies, among domains.</td>
</tr>
</tbody>
</table>
Analyses reflect recommended practices from the literature (Daas et al. 2011, Laitila et al. 2011, Iwig et al. 2013, Office for National Statistics UK 2013, Statistics Canada 2018)

Analyze data as standalone data source

Implemented using R Markdown

Supports variety of analyses: Cross-sectional, Time series, Longitudinal

Provide recommendations on how to adapt findings from toolkit for research question
## Toolkit Components

### Accuracy

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity of units</td>
<td>Assesses validity of identification keys for units in the dataset.</td>
</tr>
<tr>
<td>Validity of variable values</td>
<td>Assesses sensibility of values of single variables and among variables using the metadata.</td>
</tr>
<tr>
<td>Trustworthy variable values</td>
<td>Determines values in data that, while valid, are suspicious from judgment or experience.</td>
</tr>
</tbody>
</table>

### Completeness

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage of units</td>
<td>Assesses whether there are units that are missing or not available for the analysis.</td>
</tr>
<tr>
<td>Duplicates</td>
<td>Looks at the occurrence of multiple registrations of identical units in the dataset.</td>
</tr>
<tr>
<td>Missing values</td>
<td>Looks at the absence of values for the variables and analyzes whether characteristics of the units with missing data are different from those of units with complete data.</td>
</tr>
</tbody>
</table>

### Comparability

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution of variables</td>
<td>Assesses distribution of relevant variables to look for incongruences with expected distributions.</td>
</tr>
<tr>
<td>Relationships between variables</td>
<td>Looks for unexpected patterns in relationships among variables.</td>
</tr>
<tr>
<td>Consistency over time</td>
<td>Looks for unexpected patterns in variables over time.</td>
</tr>
<tr>
<td>Spell characteristics</td>
<td>Studies the characteristics of the spells in longitudinal analysis, such as duration and churn.</td>
</tr>
</tbody>
</table>
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplot sorted by time (last column)

Note: From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Example on Simulated Data: Tableplots

Example: Tableplots (Tennekes et al. 2011)

**Note:** From simulated data source with about 1.5 million observations representing five year range with 100,000 cases
Examples on Simulated Data: Letter-Value Plots

Example: Letter Value Plots (Hofmann et al. 2015)

Left: Variable number of recipients from simulated data source with about 1.5 million observations representing five year range with 100,000 cases

Right: From simulated data with 10,000 observations each for two groups
Examples on Simulated Data: Letter-Value Plots

Example: Letter Value Plots (Hofmann et al. 2015)

**Left:** Variable number of recipients from simulated data source with about 1.5 million observations representing five year range with 100,000 cases

**Right:** From simulated data with 10,000 observations each for two groups
Examples on Simulated Data: Letter-Value Plots

Example: Letter Value Plots (Hofmann et al. 2015)

**Left:** Variable number of recipients from simulated data source with about 1.5 million observations representing five year range with 100,000 cases

**Right:** From simulated data with 10,000 observations each for two groups
Examples on Simulated Data: Letter-Value Plots

Example: Letter Value Plots (Hofmann et al. 2015)

Left: Variable number of recipients from simulated data source with about 1.5 million observations representing five year range with 100,000 cases

Right: From simulated data with 10,000 observations each for two groups
Challenges

1. Indirectness of measures of data quality

2. Clarity of metadata and documentation

3. Importance of understanding legal and programmatic changes

4. Limits to generalizability
Conclusion

- Provide much-needed resource to enhance usability of state and local data for research

- Value of R Markdown
  - Potential to provide guidance to an array of kinds of users

- Future question: Toolkit for comparison to external data (via linkage or otherwise)


Thank You!

Zachary Seeskin
Seeskin-Zachary@norc.org