Statistical uses of Administrative data with applications

Jongho Im

Joint work with Zhegyuan Zhu, Emily Berg and Colin C. Lewis-Beck

November 11, 2016
Outline

1. Definition of administrative data
2. Benefits/Uses of administrative data
3. Applications
4. Discussion
Administrative data
FAO project

“Improving the methodology for using administrative data in agricultural statistics system"

1. Reviewing the relevant literature and studies on the quality and use of administrative sources for agricultural data.
2. Administrative data and the statistical programmes of developed countries.
3. Analysis of agriculture administrative data gaps and ways of improving the quality and use of administrative data sources for agricultural data.
4. Critical analysis of agricultural administrative sources being currently used by developing countries.
5. Strategy and methodology for improving the use of administrative data.
6. Improving the methodology for collecting and using administrative data in an agricultural statistics system.
7. (Not published yet) Strategy and methodology with applications.
In this research, we define *administrative* data as

“Information collected primarily for administrative (not statistical) purposes by governments and other organizations (including private sectors) usually during the delivery of a service or for the purposes of registration, record keeping or documentation of a transaction".
Administrative data

Example

- Soil information
- Land registration and cadastral records
- Crop insurance and subsidy programmes
- Private sources
  - Ontario Wheat producers marketing board
- Taxation data
- Reporting system (with expert opinions)
Administrative data

Benefits 1

1. Cost savings.
   - Less than the cost of a survey and a census.
   - Register-based surveys or censuses.

2. Reduced burden on respondents.
   - In the CEAP, administrative data substitute survey with a view to reducing the number of questions.

3. Improvement in micro-data quality.
   - Admin data are intrinsically linked to the identity of the individual unit in the target population.
   - tax data.
4. Improvements in the efficiency of macro level estimators
   ▶ Used for incorporating additional data information.
   ▶ Ex) calibration.
5. Small area statistics
   ▶ Improve small area estimates.
6. Timeliness
   ▶ Many developing countries do not have regular surveys or census on the agricultural section.
Administrative data

Uses 1

1. Direct tabulation
   - Quality of administrative data is important.
   - NASS publishes administrative information on the hog slaughter obtained from inspections by federal and state officials.

2. Frame construction and improvement
   - Statistics Sweden uses taxation information to analyze the coverage of their business register.

3. Survey design
   - Used as auxiliary information in sample designs. (Ex. pps)

4. Model-assisted calibration estimators
   - NRI(18 land categories) incorporates estimates obtained from federal and state authorities.
5. Nonresponse adjustments and imputation
   - Used as auxiliary information in imputation model or response model

6. Model based small area estimation

7. Cut-off surveys
   - Statistic Canada used a cut-off survey design to reduce the burden on small business respondents in the late 1990s.

8. Data collection.
   - Administrative data provide list of names and addresses.
Three major crops (Maize, Millet, Sorghum) planted area.

**Figure:** Left panel: Millet; Right panel: Sorghum
Multiple data sources

- Agricultural census/survey data conducted by Namibia Statistics Agency (NSA).

- Administrative data collected by Ministry of Agriculture, Water and Forestry (MAWF).
Crops are mainly planted at the north area of Namibia.

Zambezi (Caprivi), Kavango West and East, Omusati, Ohangwena, Oshana, Oshikoto.
Application: Measurement error model

Namibia agricultural data 4

- NSA data

- MAWF data
  - A non-random sample (farms) data collected from early 90s.
  - Collected data for the same 6 regions (Kavango East and West are combined).
  - Reporting data-the crop assessment checklist
    - Subjective changes compared to last year are reported.
    - Aggregated to regional level by extension officer.
Features

▶ In the AAS data, crop specific estimates are available in (six) regional level.
▶ In the MAWF data, only aggregated estimates are available in (six) regional level.
▶ 40 (=7 overlap years \times 6\ regions-2\ not\ eligible)\ data\ lines\ with\ four\ dimensions: \ AAS\ Maize,\ AAS\ Millet,\ AAS\ Sorghum\ and\ MAWF\ Total.
Figure: Scatter plots between the AAS crops planted area survey estimates and MAWF crops total planted area estimates.
Application: Measurement error model

Modified ME model 1

Use factor model (See section 1.5 of Fuller, 1987).

\[ Y_{t,r} = \beta_0 + \beta_1 \sum_{k=1}^{3} x_{t,r,k} + e_{t,r} \quad (1) \]
\[ X_{t,r,k} = x_{t,r,k} + u_{t,r,k}, \quad k = 1, 2, 3 \quad (2) \]

- \( x_{t,r,k} \): True crop \( k \) planted area at region \( r \) and year \( t \).
- \( Y_{t,r} \): MAWF aggregated crops planted estimates in regional level.
- \( X_{t,r,k} \): AAS crop specific regional estimates.
Application: Measurement error model

Modified ME model 2

Figure: Scatter plot between the aggregated AAS and the MAWF aggregated estimates.

- First survey estimates 1996-1997 are denoted by asterisk.
- Almost the identical values across regions and years.
Application: Measurement error model

Modified ME model 3

- Generate a pseudo data that has a similar structure compared to the Namibia data.

**Figure:** Scatter plots for the illustrative example data
Identifiable conditions (factor model)

- $\sigma_{uu} = \sigma_{uu,kk}$, for $k = 1, 2, 3$. (Same survey).
- $\sigma_{uu,ij} = 0$ for $i \neq j$.
- Two sets of errors are independent each other, $\sigma = \text{cov}(u_{t,r,k}, e_{t,r}) = 0$.
- True values $x_{t,r,k}$ are also uncorrelated with two errors.

13 parameters

$$\theta = (\beta_0, \beta_1, \mu_{x,1}, \mu_{x,2}, \mu_{x,3}, \sigma_{uu}, \sigma_{ee}, \Sigma_{xx}),$$

where $\Sigma_{xx}$ is the covariance matrix of $x$. 
Application: Measurement error model
Modified ME model 5

14 sufficient statistics

\[ \hat{\eta} = (\bar{Y}, \bar{X}_1, \bar{X}_2, \bar{X}_3, \hat{\Sigma}_{XX}, \hat{\Sigma}_{XY}, \hat{\sigma}_{YY}) \]

- Identifiable and estimable with 14 sufficient statistics
- Suggest a least squares estimator obtained by minimizing

\[ \{\hat{\eta} - \eta(\theta)\}^T \{\hat{\eta} - \eta(\theta)\}. \]

- Does not require any parametric model.
Estimation
The least square estimator $\hat{\theta}$ can be also obtained by solving the following equation,

$$A(\hat{\theta} - \theta) = \hat{\eta} - \eta(\theta),$$

where $A$ is the first-order derivative of $\eta(\theta)$.

Parameter updates

$$\hat{\theta}(t+1) = \hat{\theta}(t) + R^{-1} Q^T \{\hat{\eta} - \eta(\hat{\theta}(t))\},$$

where $Q$ (Orthogonal matrix) and $R$ (Upper triangular matrix) are the matrices obtained from the QR decomposition of $A$. 
Application: Measurement error model
Modified ME model 7

Table: Average and (standard error) for the illustrative data

<table>
<thead>
<tr>
<th></th>
<th>AAS Maize</th>
<th>AAS Millet</th>
<th>AAS Sorghum</th>
<th>MAWF Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11,863 (5,687)</td>
<td>49,471 (14,561)</td>
<td>34,721 (14,150)</td>
<td>91,575 (29,470)</td>
</tr>
</tbody>
</table>

Results obtained from the modified ME model

\[
(\hat{\beta}_0, \hat{\beta}_1) = (2509, 0.93),
\]

\[
(\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3) = (11,863, 49,471, 34,721),
\]

\[
(\hat{\sigma}_{uu}, \hat{\sigma}_{ee}) = (4516^2, 16454^2),
\]

\[
(\hat{\sigma}_{xx11}, \hat{\sigma}_{xx22}, \hat{\sigma}_{xx33}) = (3662^2, 13443^2, 13758^2),
\]

\[
(\hat{\rho}_{xx12}, \hat{\rho}_{xx13}, \hat{\rho}_{xx23}) = (-0.08, 0.58, 0.70).
\]

- **Point estimates** are the same between the direct estimates and the ME model estimates.
- **Standard errors** are smaller compared to the direct estimates.
Application: Small Area Estimation

SAE: outline

- Administrative data often have little or no sampling errors at detailed levels of geographic detail.
- Survey data can have large estimated sampling variance at granular levels of geography due to small realized sample sizes.
- MAWF aggregated crop planted area estimates are reported in six regional level.
- Our goal is to produce the minimum mean squared error (MMSE) predictor of the major crops planted area on the six regions.
Received the 2013/2014 Namibia Census of Agriculture data.

Table: The estimates of the number of household and the crop planted area between the NCA report and the received data.

<table>
<thead>
<tr>
<th>Major crop</th>
<th># HH(Report)</th>
<th>Area(Report)</th>
<th># HH(Data)</th>
<th>Area(Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maize</td>
<td>17,620</td>
<td>34,991</td>
<td>18,629</td>
<td>1,777,828</td>
</tr>
<tr>
<td>Sorghum</td>
<td>24,646</td>
<td>7,043</td>
<td>25,760</td>
<td>7,415</td>
</tr>
<tr>
<td>Millet</td>
<td>129,029</td>
<td>421,212</td>
<td>113,419</td>
<td>4,470,970</td>
</tr>
</tbody>
</table>

# HH denoted the number of household estimates.

- Exists extreme values in area estimates.
- May caused by processing error, simple typos or other reasons.
Table: The estimates of the number of household and the crop planted area estimates with the refined area estimates.

<table>
<thead>
<tr>
<th>Major crop</th>
<th>N.HH(Report)</th>
<th>Area(Report)</th>
<th>N.HH(Data)</th>
<th>Area(Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maize</td>
<td>17,620</td>
<td>34,991</td>
<td>18,629</td>
<td>31,401</td>
</tr>
<tr>
<td>Sorghum</td>
<td>24,646</td>
<td>7,043</td>
<td>25,760</td>
<td>7,415</td>
</tr>
<tr>
<td>Millet</td>
<td>129,029</td>
<td>421,212</td>
<td>113,419</td>
<td>332,722</td>
</tr>
</tbody>
</table>

- Delete extreme cases whose estimates are larger than 250ha.
- Do not conduct additional adjustments, because our goal is to illustrate how SAE can be applied by incorporating the admin data.
Figure: Scatter plot between the direct estimates and the MAWF estimates.
Application: Small Area Estimation

SAE: model 1

\[
\begin{align*}
\log(y_i) &= \beta_0 + \beta_1 \log(x_i) + u_i, \quad (3) \\
\log(\hat{y}_i) &= \beta_0 + \beta_1 \log(x_i) + u_i + e_i, \quad (4)
\end{align*}
\]

where \( u_i \sim N(0, \sigma_u^2) \) and \( e_i \sim N(0, \hat{\sigma}_{ei, \log}^2) \).

- \( y_i \): True (aggregated) crops planted area at region \( i = 1, \ldots, 6 \).
- \( x_i \): MAWF crops planted area at region \( i \).
- \( \hat{y}_i \): Direct estimator of \( y_i \) for region \( i \). May exposed to have a larger variance.

\[
\hat{y}_i = \sum_{j \in A_i} w_{ij} \hat{y}_{ij},
\]

where \( A_i \) is a index set of samples for the region \( i \), \( w_{ij} \) denotes the sampling weight and \( \hat{y}_{ij} \) is the averaged crop planted area of the \( j \)th unit at the region \( i \).
Application: Small Area Estimation

SAE: model 2

\[
\log(y_i) = \beta_0 + \beta_1 \log(x_i) + u_i, \\
\log(\hat{y}_i) = \beta_0 + \beta_1 \log(x_i) + u_i + e_i,
\]

where \( u_i \sim N(0, \sigma^2_u) \) and \( e_i \sim N(0, \hat{\sigma}^2_{ei,\log}) \).

- \( u_i \) is region specific random effect.
- \( e_i \) is sampling error.
- \( \hat{\sigma}^2_{ei,\log} \) is estimated variance of the direct estimator. Obtained using a replication method.
- \( \hat{\sigma}^2_{ei,\log} \) are the variance estimates of \( \log(\hat{y}_i) \).
- Delta method is used for the log transformation.
- Parameters are estimated using the REML method through "sae" R package.
The minimum mean squared error predictor of $y_i$ for the model in (3) and (4) is

$$\tilde{y}_i^{MMSE} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 \log(x_i) + \hat{u}_i + 0.5\hat{\gamma}_i\hat{\sigma}^2_{e, \log}\},$$

where

$$\hat{u}_i = \gamma_i\{\log(\hat{y}_i) - \hat{\beta}_0 - \log(x_i)\hat{\beta}_1\},$$

and

$$\gamma_i = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u + \hat{\sigma}^2_{e, \log}}.$$

$$\hat{V}(\tilde{y}_i^{MMSE}) = \exp\{2\hat{\beta}_0 + 2\hat{\beta}_1 \log(x_i)\}\exp(2\hat{u}_i + 2\hat{\gamma}_i\hat{\sigma}^2_{e, \log}) - (\tilde{y}_i^{MMSE})^2.$$
**Application: Small Area Estimation**

**SAE: result**

**Table:** Regional level crop planted area estimates with the MMSE predictor.

<table>
<thead>
<tr>
<th>Region</th>
<th>NCA (C.V.)</th>
<th>MAWF</th>
<th>MMSE (C.V.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zambezi</td>
<td>15,904 (0.128)</td>
<td>19,384</td>
<td>16,823 (0.122)</td>
</tr>
<tr>
<td>Kavango</td>
<td>51,302 (0.090)</td>
<td>21,588</td>
<td>49,999 (0.088)</td>
</tr>
<tr>
<td>Omusati</td>
<td>109,673(0.051)</td>
<td>78,030</td>
<td>109,492(0.051)</td>
</tr>
<tr>
<td>Ohangwena</td>
<td>81,337 (0.051)</td>
<td>79,828</td>
<td>81,649 (0.051)</td>
</tr>
<tr>
<td>Oshana</td>
<td>40,021 (0.198)</td>
<td>35,100</td>
<td>41,600 (0.177)</td>
</tr>
<tr>
<td>Oshikoto</td>
<td>68,481 (0.043)</td>
<td>58,080</td>
<td>68,568 (0.042)</td>
</tr>
</tbody>
</table>

C.V. denotes coefficient of variation of the estimator.
Budget is always an important issue in the developing countries.

Two phase sampling design can be applied to reduce cost.

- First phase sample: Short questionnaire $Q_s$.
- Second phase sample: Full questionnaire $Q_f \supset Q_s$.

How can we select a subset of $Q_f$ for the $Q_s$?

- Hold important variables or variables of interested in $Q_f \setminus Q_s$.
- Hold variables which requires a relatively higher measurement cost in $Q_f \setminus Q_s$. 

Application: Two phase sampling design

overview 2

▶ Problems
  ▶ Too many variables of interest.
  ▶ Not easy to distinguish variables in terms of measurement costs.
▶ Administrative data or previous survey data can be used to construct the two phase sampling design.
▶ Possible solution
  ▶ Logical structure: figure out a hierarchical structure of variables. Ex) A+B+C=D.
  ▶ Imputation model/Calibration model
  ▶ Space filling design:
    ▶ Maxmin design: Find a survey design that maximize a minimum distance between variables.
    ▶ Understood by setting up the tables in a restaurant such that one wants to minimize the chances to eavesdrop on another table talk. (Johnson et al., 1990)
Application: Two phase sampling design


$$\phi(S) = \left( \sum_{j=1}^{M} p_j d_j^{-s} \right)^{1/s}$$

where

- $S \in S^*$, where $S^*$ is a class of sampling designs.
- $M$ is the total number distinct number of distances between two data points.
- $d_j$ is the $j$th distance value, $j = 1, \ldots, M$. The distances are determined by Euclidean distance.
- $p_j$ is a probability measure for the distance $d_j$, where $\sum_{j=1}^{M} p_j = 1$.
- $s$: a nuisance parameter can be understood as the penalty term on the distance.
Discussion

1. We review the FAO project by focusing on definition, benefits and selected statistical uses.

2. We introduce some applications using the tailored measurement error model and the small area estimation with Namibia agricultural data, and also introduce an idea for the two phase sampling design.

3. For the full technical reports, see the FAO website: http://gsars.org/en/tag/administrativedata/