How to use big data for Official Statistics?

Li-Chun Zhang$^{1,2,3}$

$^1$University of Southampton (L.Zhang@soton.ac.uk)  
$^2$Statistisk sentralbyraa, Norway  
$^3$Universitetet i Oslo

Lyon, 25 October 2018
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (Kr)</td>
<td>%</td>
</tr>
<tr>
<td>Consumption in all</td>
<td>280078</td>
<td>100</td>
</tr>
<tr>
<td>01 Food, non-alcoholic drinks</td>
<td>33499</td>
<td>12,0</td>
</tr>
<tr>
<td>02 Alcohol, tobacco</td>
<td>8114</td>
<td>2,9</td>
</tr>
<tr>
<td>03 Clothing, shoes</td>
<td>16278</td>
<td>5,8</td>
</tr>
<tr>
<td>04 Housing, household energy</td>
<td>71278</td>
<td>25,4</td>
</tr>
<tr>
<td>05 Furniture, household art.</td>
<td>17321</td>
<td>6,2</td>
</tr>
<tr>
<td>06 Health</td>
<td>7717</td>
<td>2,8</td>
</tr>
<tr>
<td>07 Transport</td>
<td>56832</td>
<td>20,3</td>
</tr>
<tr>
<td>08 Post, telecommunication</td>
<td>5610</td>
<td>2,0</td>
</tr>
<tr>
<td>09 Culture, recreation</td>
<td>33634</td>
<td>12,0</td>
</tr>
<tr>
<td>10 Education</td>
<td>869</td>
<td>0,3</td>
</tr>
<tr>
<td>11 Restaurant, hotel, etc.</td>
<td>11379</td>
<td>4,1</td>
</tr>
<tr>
<td>12 Other goods or services</td>
<td>17547</td>
<td>6,3</td>
</tr>
</tbody>
</table>
Passing Paradigms

1. **Representative Method** (of statistical surveys)
   - A.N. Kiær (1895), J. Neyman (1934)
   
   *Census not necessary for descriptive statistics*

2. **Archive Statistics** [“arkivstatistiske systemer”]
   - S. Nordbotten (1966) *et al.*

   *Separation of data capture and statistics production*

| On the one hand, capture and curation as the data is generated; |
| on the other hand, processing and output as the need arises |

⇒ *secondary uses & combination of sources*
Errors with $n \times p$ statistical data

**Representation:** traditionally, “survey sampling”

- relationships among relevant populations & units
- e.g. frame coverage, sample selection, missing units

**Problem for big data:** non-probability sample

**Measurement:** traditionally, “survey sampling”

- subject/concept of interest vs. actual observations
- e.g. relevance, mode effects, mis-classification

**Problem for big data:** (machine) learning
B-sample simple expansion [B: big data]

Let \( \delta_i = 1 \) if \( i \in B \cap U \) [pop.] or 0 if \( i \in U \setminus B \). Observe \( y_i \) if \( \delta_i = 1 \).

Validity conditions (Smith, 1983) super-pop. (SP) approach

- \( \mu_i = E(y_i|\delta_i) = \mu \) \[“non-informative B-selection”\]
- \( E(N\bar{y}_B|B) = E(Y) \) where \( \bar{y}_B = \sum_{i \in B} y_i/n_B \)

Or, under quasi-randomisation (QR) approach

- \( p_i = Pr(\delta_i = 1; y_i) = p > 0 \) \[“non-info. B-selection”\]
- Then, \( E(\tilde{Y}) = Y \) where \( \tilde{Y} = \sum_{i \in B} \frac{y_i}{p} = \sum_{i \in U} \frac{\delta_i}{p} y_i \)
- Pluggin in \( \hat{p} = n_B/N \) yields the same \( \hat{Y} = N\bar{y}_B \)

q1: what if there exist \( i \in U \) and \( Pr(\delta_i = 1) = 0 \)?
q2: what if heterogeneous mean – model misspec.?

- SP approach: heterogeneous mean if

$$\mu_i = E(y_i | \delta_i) = E(y_i; i \in U) \neq \mu \text{ despite } \mu = \sum_{i \in U} \frac{\mu_i}{N}$$

Model $E(y_i | \delta_i) = \mu$ is still statistically ‘correct’, and

$$\sum_{i \in U} [E(y_i | \delta_i) - \mu] = \sum_{i \in U} \mu_i - N \mu = 0$$

- QR approach: heterogeneous mean if

$$p_i = E(\delta_i | y_i) = E(\delta_i; i \in U) \neq p \text{ despite } p = \sum_{i \in U} \frac{p_i}{N}$$

$$E\left(\sum_{i \in U} \frac{\delta_i y_i}{p}\right) - \sum_{i \in U} y_i = \frac{1}{p} \sum_{i \in U} (p_i - p) y_i \neq 0 \quad (!)$$
A non-parametric asymptotic (NPA) formulation

W.r.t. \( F_N = \{ \frac{1}{N}, \ldots, \frac{1}{N} \} \), we have \( \bar{y}_B = \bar{Y} \) provided

\[
\begin{align*}
\text{Cov}_N(\delta_i, y_i) &= \frac{1}{N} \sum_{i \in U} \delta_i y_i - \left( \frac{1}{N} \sum_{i \in U} \delta_i \right) \left( \frac{1}{N} \sum_{i \in U} y_i \right) = 0 \\
E_N(\delta_i) &= \frac{1}{N} \sum_{i \in U} \delta_i > 0
\end{align*}
\]

e.g. Rao (1966), Bethlehem (1988), Meng (2018). Assume

\[
\begin{align*}
\lim_{N \to \infty} \text{Cov}_N(\delta_i, y_i) &= 0 \quad [\text{non-informative B-selection}] \\
\lim_{N \to \infty} E_N(\delta_i) &= p > 0 \quad [\text{non-negligible B-selection}]
\end{align*}
\]

For SP: \( E(\text{Cov}_N(\delta_i, y_i) | \delta_U) = \frac{1}{N} \sum_{i \in U} \delta_i \mu_i - \left( \frac{1}{N} \sum_{i \in U} \delta_i \right) \left( \frac{1}{N} \sum_{i \in U} \mu_i \right) \)

For QR: \( E(\text{Cov}_N(\delta_i, y_i); y_U) = \frac{1}{N} \sum_{i \in U} p_i y_i - \left( \frac{1}{N} \sum_{i \in U} p_i \right) \left( \frac{1}{N} \sum_{i \in U} y_i \right) \)
General difficulty with validating validity conditions

Assume \( p_i = p > 0 \). Suppose known \( z_i \), for all \( i \in U \).

Two goodness-of-fit checks based on ‘held-out’ \( z_i \)’s:

\[
\begin{align*}
z_B &\equiv n_B \bar{z}_B = \hat{p}N \bar{Z} \\
Z &\equiv n_B \bar{z}_B / \hat{p}
\end{align*}
\]

\[
\begin{align*}
z_i &\equiv 1 \\
n_B &\equiv n_B = \hat{p}N \\
N &= n_B / \hat{p}
\end{align*}
\]

Setting \( \hat{p} = n_B / N \): we are simply checking if \( \bar{Z} = \bar{z}_B \)?

If \( z_i \) correlated with \( y_i \), non-info. selection corroborated; however, would be natural then to use \( z_i \) in estimation...

A dilemma: building the best model for estimation would at the same time reduce the ability to verify it?
General difficulty with validating validity conditions

...

Of course, the situation changes completely, provided we have an additional probability sample $S \subset U \setminus B$.

The bigger the $B$-sample, the greater gain it is then.

The NPA condition turns up many places otherwise...
Example: Register-Survey DSE [NB. model-based]

Dual System Estimator of unknown population size $N$:

$$\hat{N} = \frac{xn}{m} \quad \left[ x = \sum_{i \in U} \delta_iA \quad n = \sum_{i \in U} \delta_iB \quad m = \sum_{i \in U} \delta_iA\delta_iB \right]$$

Treat $A$-register $\delta_iA$’s as fixed, allow for heterogeneous survey $B$-capture $p_i = \Pr(\delta_iB = 1) \neq p$ and $p_i \in [0, 1]$:

$$\lim_{N \to \infty} E(\hat{N} - N)/N = \lim_{N \to \infty} \left( \sum_{i \in U} \delta_iA \left( \frac{\sum_{i \in U} p_i}{\sum_{i \in U} p_i \delta_iA} \right) - N \right)/N$$

$$= \lim_{N \to \infty} \frac{\left( \sum_{i \in U} \delta_iA \right)\left( \sum_{i \in U} p_i \right) - N \left( \sum_{i \in U} p_i \delta_iA \right)}{N \sum_{i \in U} p_i \delta_iA}$$

$$= -\lim_{N \to \infty} \text{Cov}_N(\delta_iA, p_i)/(\frac{1}{N} \sum_{i \in U} p_i \delta_iA)$$

Consistent $\hat{N}$ if NPA $B$-capture

NB. Constant $B$-capture not required; [over-count; match]
Example: Big-data proxy expenditure weights

CPI given elementary aggregate $i = 1, \ldots, m$:

$$ P = \sum_{i=1}^{m} w_i P_i \quad \text{where} \quad \sum_{i=1}^{m} w_i = 1 $$

Let $(w_i, \hat{w}_i, w_i^*) = (\text{true, survey, big-data proxy})$ weights

Let $(P_i, \hat{P}_i) = (\text{true, calculated})$ price indices

**Q1:** Source effect $\sum_{i=1}^{m} w_i^* \hat{P}_i - \sum_{i=1}^{m} \hat{w}_i \hat{P}_i$?

**Q2:** Suppose bias dominates variance: $w_i^* \approx E(w_i^*) \neq w_i$

and $V(w_i^*) \approx 0$, how to measure/describe the error of

$$ P^* = \sum_{i=1}^{m} w_i^* \hat{P}_i $$
Answer to Q1: NPA condition

Let $b_i = \frac{\hat{w}_i}{w_i^*} - 1$. No source effect provided

$$\sum_{i=1}^{m} w_i^* \hat{P}_i - \sum_{i=1}^{m} \hat{w}_i \hat{P}_i = - \sum_{i=1}^{m} w_i^* b_i \hat{P}_i$$

$$= - \text{Cov}(b_i, \hat{P}_i; w^*) = 0$$

w.r.t. $\mathcal{F}_{w^*} = (w_1^*, ..., w_m^*)$. [NPA non-info. discrepancy]

$$E(P^*) - P = \sum_{i=1}^{m} w_i^* E(\hat{P}_i) - \sum_{i=1}^{m} w_i P_i \quad [E(w_i^*) = w_i^*]$$

$$= E(\psi_i; w^*) - \text{Cov}(a_i, P_i; w^*) \quad [\psi_i = E(\hat{P}_i) - P_i]$$

i.e. unbiased $P^*$ provided NPA non-informative error

$$a_i = \frac{w_i}{w_i^*} - 1$$

of big-data $w^*$-weights, and unbiased price index $\hat{P}_i$
Combining transaction data relevant for CPI

### Ideal data

<table>
<thead>
<tr>
<th>Who</th>
<th>What (COICOP-V/VI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01.1.1.x</td>
<td>01.1.2.x</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>12.7.1.x</td>
<td></td>
</tr>
</tbody>
</table>

(hush/pers) : 

**How much?**

(Value = Price · Quantity)

### Old and new data*

<table>
<thead>
<tr>
<th>Source</th>
<th>Who</th>
<th>What</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td>+</td>
<td>+/-</td>
<td>Non-sampl. err./Variance</td>
</tr>
<tr>
<td>Scanner</td>
<td>NA</td>
<td>+</td>
<td>Unstable GTIN</td>
</tr>
<tr>
<td>Receipt</td>
<td>+/-</td>
<td>+</td>
<td>Uncertain person ident.</td>
</tr>
<tr>
<td>Bankcard</td>
<td>+</td>
<td>NA</td>
<td>Changing platforms</td>
</tr>
</tbody>
</table>

* overall coverage issues according to target of interest
A big problem of combining big data: Confidentiality
(I) Control what one can look, not what one can link!

NB. Sampling frame by time and location, not individual
(II) Reconceptualisation: building on non-disclosive data

Data structure: **Network = valued graph**

Graph: \( G = (U, A) = (\text{nodes, edges}) \) [digraph by default]

Network: \( \mathcal{N} = (G, X) = (\text{graph, values}) \)

Values: \( X = (X_U, X_A) \) associated with nodes, edges

Example: Cellphone data

- node = person, edge = calls in-between [confidentiality]
- node = locality, edge = connection in-between
  - locality: region, municipality, post code, etc.
  - connection-1, same person: movement
  - connection-2, between two persons: call/text
A multigraph of personal movements

Multigraph $G = (U, A) = (\text{nodes, edges}), \ U = \text{locality}$

Def.: For each person $k$, $a_{ij}^k \in A$ iff $i \rightarrow j$ by person $k$

NB. Distinguish between edges due to different persons

Motif $[M]$: $M \subset U$ of specific characteristics, whereby characteristics def. in terms of edges, of order $q$ if $|M| = q$

Example (cont’d): $a_{ij}^k \in A$ with $i = \text{home}, j = \text{work}$

Same multigraph from different sources:

• admin: normative $\neq$ real work location $j$

• telecom: machine learnt $\neq$ real work location $j$
Some statistics: Norwegian admin data

<table>
<thead>
<tr>
<th>Year Quarter</th>
<th>Normal &amp; freelance</th>
<th>Multiple jobs</th>
<th>2 jobs</th>
<th>3+ jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015 1. quarter</td>
<td>2,393,815</td>
<td>199,179</td>
<td>178,072</td>
<td>21,107</td>
</tr>
<tr>
<td>2015 2. quarter</td>
<td>2,427,443</td>
<td>209,038</td>
<td>186,247</td>
<td>22,791</td>
</tr>
<tr>
<td>2015 3. quarter</td>
<td>2,461,126</td>
<td>207,526</td>
<td>186,866</td>
<td>20,660</td>
</tr>
<tr>
<td>2015 4. quarter</td>
<td>2,434,718</td>
<td>219,689</td>
<td>194,778</td>
<td>24,911</td>
</tr>
<tr>
<td>2016 1. quarter</td>
<td>2,408,879</td>
<td>205,230</td>
<td>183,830</td>
<td>21,400</td>
</tr>
<tr>
<td>2016 2. quarter</td>
<td>2,434,789</td>
<td>214,961</td>
<td>191,965</td>
<td>22,996</td>
</tr>
<tr>
<td>2016 3. quarter</td>
<td>2,468,435</td>
<td>214,979</td>
<td>194,831</td>
<td>20,148</td>
</tr>
<tr>
<td>2016 4. quarter</td>
<td>2,455,903</td>
<td>226,880</td>
<td>203,354</td>
<td>23,526</td>
</tr>
<tr>
<td>2017 1. quarter</td>
<td>2,431,623</td>
<td>212,420</td>
<td>191,085</td>
<td>21,335</td>
</tr>
<tr>
<td>2017 2. quarter</td>
<td>2,458,160</td>
<td>222,301</td>
<td>199,789</td>
<td>22,512</td>
</tr>
<tr>
<td>2017 3. quarter</td>
<td>2,504,081</td>
<td>220,951</td>
<td>201,035</td>
<td>19,916</td>
</tr>
<tr>
<td>2017 4. quarter</td>
<td>2,491,555</td>
<td>234,901</td>
<td>209,979</td>
<td>24,922</td>
</tr>
</tbody>
</table>

NB. stronger growth of people with 2+ jobs
Motifs in job-related personal movement multigraph

Node = Locality
Edge = job-related personal movement

Non-disclosive data: motif counts instead of individuals
Combination of sources

1. Enriched Employment statistics
   - commuting, part-time work life, etc.
   - breakdown by motifs (and motif-variations) from telecom data

2. Flash estimates of Labour Market dynamics
   - reducing observation lag of Labour Market transition
     [employed $\rightarrow$ unemployed, active $\rightarrow$ leave-from work, etc.]
   - based on changes of personal motifs in telecom data

Q: how to adjust for coverage-relevance error in data?

Topic: Estimation of motif counts in the presence of ...
Simple digraph & network of labour flows

Aggregation of edges of different types in multigraph

\[\downarrow\]

Simple labour-flow digraph \( G = (U, A) \), i.e. \(|A_{ij}| \equiv 1\)

- \(U\) = locality
- \(A\) = existing between-flow of labour

Simple labour-flow network \( \mathcal{N} = (G, X) \)

- \(X_A\): weighted sum of multi-type edges in \(G\)
  
  [can be measured in no. persons, trips, time, etc.]

- \(X_U\): e.g. no. employed pers. anchored at each \(i \in U\)
Adjust for coverage-relevance error in separate sources?
Topics: Network calibration, motif mis-classification, etc.
Example analysis: Do cellphone calls redraw the maps?

NB. spatial connectivity based on phone-call relationships; maps remarkably well to administrative division
NB. Similar results in Belgium; England (by another team)
Example analysis: Do cellphone calls redraw the maps?

Network ‘cluster analysis’:

- Clusters in a population of units: list partition
- Clusters in a graph: connected components
- Clusters in a network, where the different clusters can still be connected in the underlying graph?

*Modularity maximization*, so-called Louvain Method:

“... increased density of links between the members of a given group [of nodes] with that obtained in a random group with the same overall characteristics”
## Characterisation of some relevant new sources

<table>
<thead>
<tr>
<th>Source</th>
<th>Measure</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartmeter</td>
<td>El-usage</td>
<td>Simple digraph</td>
</tr>
<tr>
<td>Sensor</td>
<td>Carrier position</td>
<td>Multi digraph</td>
</tr>
<tr>
<td></td>
<td>Check-point</td>
<td>Multi digraph</td>
</tr>
<tr>
<td>Cellphone</td>
<td>Call/text</td>
<td>Simple digraph</td>
</tr>
<tr>
<td></td>
<td>Position</td>
<td>Multi digraph</td>
</tr>
<tr>
<td>Transaction</td>
<td>Scanner (what)</td>
<td>List</td>
</tr>
<tr>
<td></td>
<td>Receipt (what-who)</td>
<td>List</td>
</tr>
<tr>
<td></td>
<td>Payment (who)</td>
<td>Simple digraph</td>
</tr>
</tbody>
</table>

- E.g. Sampling of payments for disaggregation of CPI; confidentiality if \( \text{Who} = (\text{geography, demography}) \neq \text{individual} \);
- E.g. Sampling of payments for disaggregation & timeliness of SNA
Back to Q2. What if bias dominates variance?

What’s lacking of mean squared error (MSE)?

• $\text{MSE}(\hat{\theta}_1) = \text{MSE}(\hat{\theta}_2) > 0$, which is better? Depends...

• True $\theta_0 \neq E(\hat{\theta}_1) = E(\hat{\theta}_2)$ and $V(\hat{\theta}_1) > V(\hat{\theta}_2) = 0$: but is $\hat{\theta}_2$ always better than $\hat{\theta}_1$? Depends...

• $(\hat{\theta} - \mu_{\hat{\theta}})/se(\hat{\theta}) \sim N(0,1)$: what to do with $\text{MSE}(\hat{\theta})$?

Well, a conservative CI, say, $(\mu_{\hat{\theta}} \pm 1.96\sqrt{\text{MSE}(\hat{\theta})})$.

Now, given big data, suppose $se(\hat{\theta}) = 0$, what then?

NB. One cannot estimate $\text{bias}(\hat{\theta})$ unbiasedly.

Rephrase Q2: *How to communicate uncertainty then?*
Back to Q2. What if bias dominates variance?

100\(\alpha\)% confidence interval of true parameter value \(\theta_0\):

\[
A_{\sigma,\alpha} = Z \pm \kappa_\alpha \sigma \quad \text{for} \quad Z \sim N(\theta_0, \sigma^2)
\]

\[
\Pr(\theta_0 \in A_{\sigma,\alpha}) = \kappa_\alpha \equiv (1 + \alpha)/2 \text{ quantile of } N(0, 1)
\]

**Coverage ratio (CR)** of \(\hat{\theta}^*\) with respect to \(A_{\sigma,\alpha}\) is

\[
\gamma_{\sigma,\alpha}(\hat{\theta}^*) = \frac{\Pr(\hat{\theta}^* \in A_{\sigma,\alpha})}{\Pr(\theta_0 \in A_{\sigma,\alpha})} = \frac{\alpha^*}{\alpha}
\]

where \(\alpha^*\) is the coverage of \(\hat{\theta}^*\) by \(A_{\sigma,\alpha}\) ['checking device']

NB. CR varies with \((\sigma, \alpha)\): how stringent checking is

NB. Works for big-data proxy estimate \(\hat{\theta}^* = \theta^* = E(\hat{\theta}^*)\)
Back to Q2. What if bias dominates variance?

Monte Carlo coverage ratio $\tilde{\gamma}_{n,\alpha}$ with $\alpha = 0.95$, where $K = 1000$, $B = 1000$: $\hat{\omega}_i^{(b)} \sim N(\omega_i, \sigma_{i,n}^2)$ with median($\sigma_{i,n}/\omega_i$) = 0.202 if $n = 250$, or 0.253 if $n = 160$; or $\hat{\omega}_i^{(b)} = (1 + r_i^*)\omega_i$ with $r_i^* \sim \mathcal{F}_r$ and median($|r_i|$) = 0.254.

<table>
<thead>
<tr>
<th>Proxy index $P^*$</th>
<th>Hypothetical survey index $\hat{P}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 250$  $n = 160$  $\mathcal{F}_r$</td>
</tr>
<tr>
<td>$\tilde{\gamma}$</td>
<td>0.873     0.913     0.930</td>
</tr>
<tr>
<td>s.e.($\tilde{\gamma}$)</td>
<td>0.005     0.003     0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proxy index $P^*$, scaling 1.4</th>
<th>Hypothetical survey $\hat{P}$, scaling 0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 250$  $n = 160$  $\mathcal{F}_r$</td>
</tr>
<tr>
<td>$\tilde{\gamma}$</td>
<td>0.754     0.822     0.862</td>
</tr>
<tr>
<td>s.e.($\tilde{\gamma}$)</td>
<td>0.008     0.006     0.005</td>
</tr>
</tbody>
</table>

NB. $A_{\sigma,\alpha}$: $\sigma$ varies over columns; scaling alters error magnitude

28