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

Example: Expenditure weights for CPI (ssb.no)

	1998 - 2000		2012	
	Total (Kr)	%	Total (Kr)	%
Consumption in all	*280078*	*100*	*435507*	*100*
01 Food, non-alcoholic drinks	33499	**12,0**	51429	**11,8**
02 Alcohol, tobacco	8114	2,9	11717	2,7
03 Clothing, shoes	16278	5,8	23618	5,4
04 Housing, household energy	71278	**25,4**	135982	**31,2**
05 Furniture, household art.	17321	6,2	24495	5,6
06 Health	7717	2,8	11421	2,6
07 Transport	56832	20,3	81574	18,7
08 Post, telecommunication	5610	2,0	8253	1,9
09 Culture, recreation	33634	12,0	43347	10,0
10 Education	869	0,3	985	0,2
11 Restaurant, hotel, etc.	11379	4,1	15557	3,6
12 Other goods or services	17547	6,3	27129	6,2

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 $\widehat{Y} = N\bar{y}_B$

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

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

- QR approach: heterogeneous mean if

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

$$E\Big(\sum_{i\in U} \frac{\delta_i y_i}{p}\Big) - \sum_{i\in U} y_i = \frac{1}{p}\sum_{i\in U} (p_i - p) y_i \neq 0 \quad (!)$$

A non-parametric asymptotic (NPA) formulation

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

$$\begin{cases} Cov_N(\delta_i, y_i) = \frac{1}{N}\sum_{i\in U} \delta_i y_i - \big(\frac{1}{N}\sum_{i\in U}\delta_i\big)\big(\frac{1}{N}\sum_{i\in U} y_i\big) = 0 \\ E_N(\delta_i) = \frac{1}{N}\sum_{i\in U}\delta_i > 0 \end{cases}$$

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

$$\begin{cases} \lim_{N\to\infty} Cov_N(\delta_i, y_i) = 0 & \text{[non-informative B-selection]} \\ \lim_{N\to\infty} E_N(\delta_i) = p > 0 & \text{[non-negligible B-selection]} \end{cases}$$

For SP: $E\big(Cov_N(\delta_i, y_i)|\delta_U\big) = \frac{1}{N}\sum_{i\in U}\delta_i\mu_i - \big(\frac{1}{N}\sum_{i\in U}\delta_i\big)\big(\frac{1}{N}\sum_{i\in U}\mu_i\big)$

For QR: $E\big(Cov_N(\delta_i, y_i); y_U\big) = \frac{1}{N}\sum_{i\in U} p_i y_i - \big(\frac{1}{N}\sum_{i\in U} p_i\big)\big(\frac{1}{N}\sum_{i\in U} y_i\big)$

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:

$$\begin{cases} z_B \equiv n_B \bar{z}_B = \hat{p} N \bar{Z} \\ Z = n_B \bar{z}_B / \hat{p} \end{cases} \overset{z_i \equiv 1}{\Longrightarrow} \begin{cases} n_B \equiv n_B = \hat{p} N \\ N = n_B / \hat{p} \end{cases}$$

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?

...

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:

$$\widehat{N} = \frac{xn}{m} \qquad \Big[x = \sum_{i \in U} \delta_{iA} \quad n = \sum_{i \in U} \delta_{iB} \quad m = \sum_{i \in U} \delta_{iA}\delta_{iB}\Big]$$

Treat A-register δ_{iA}'s as fixed, allow for heterogeneous survey B-capture $p_i = \Pr(\delta_{iB} = 1) \neq p$ and $p_i \in [0, 1]$:

$$\begin{aligned}
\lim_{N \to \infty} E(\widehat{N} - N)/N &= \lim_{N \to \infty} \Big(\sum_{i \in U} \delta_{iA} \frac{\sum_{i \in U} p_i}{\sum_{i \in U} p_i \delta_{iA}} - N\Big)/N \\
&= \lim_{N \to \infty} \frac{(\sum_{i \in U} \delta_{iA})(\sum_{i \in U} p_i) - N(\sum_{i \in U} p_i \delta_{iA})}{N \sum_{i \in U} p_i \delta_{iA}} \\
&= - \lim_{N \to \infty} Cov_N(\delta_{iA}, p_i)/\Big(\frac{1}{N} \sum_{i \in U} p_i \delta_{iA}\Big)
\end{aligned}$$

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

NB. Constant B-capture not required; [over-count; match]

Example: Big-data proxy expenditure weights

CPI given elementary aggregate $i = 1, ..., m$:

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

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

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

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

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

$$P^* = \sum_{i=1}^{m} w_i^* \widehat{P}_i \ \textbf{?}$$

Answer to Q1: NPA condition

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

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

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

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

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

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

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

$$a_i = w_i/w_i^* - 1$$

of big-data w^*-weights, and unbiased price index $\widehat{P}_i$

Combining transaction data relevant for CPI

Ideal data

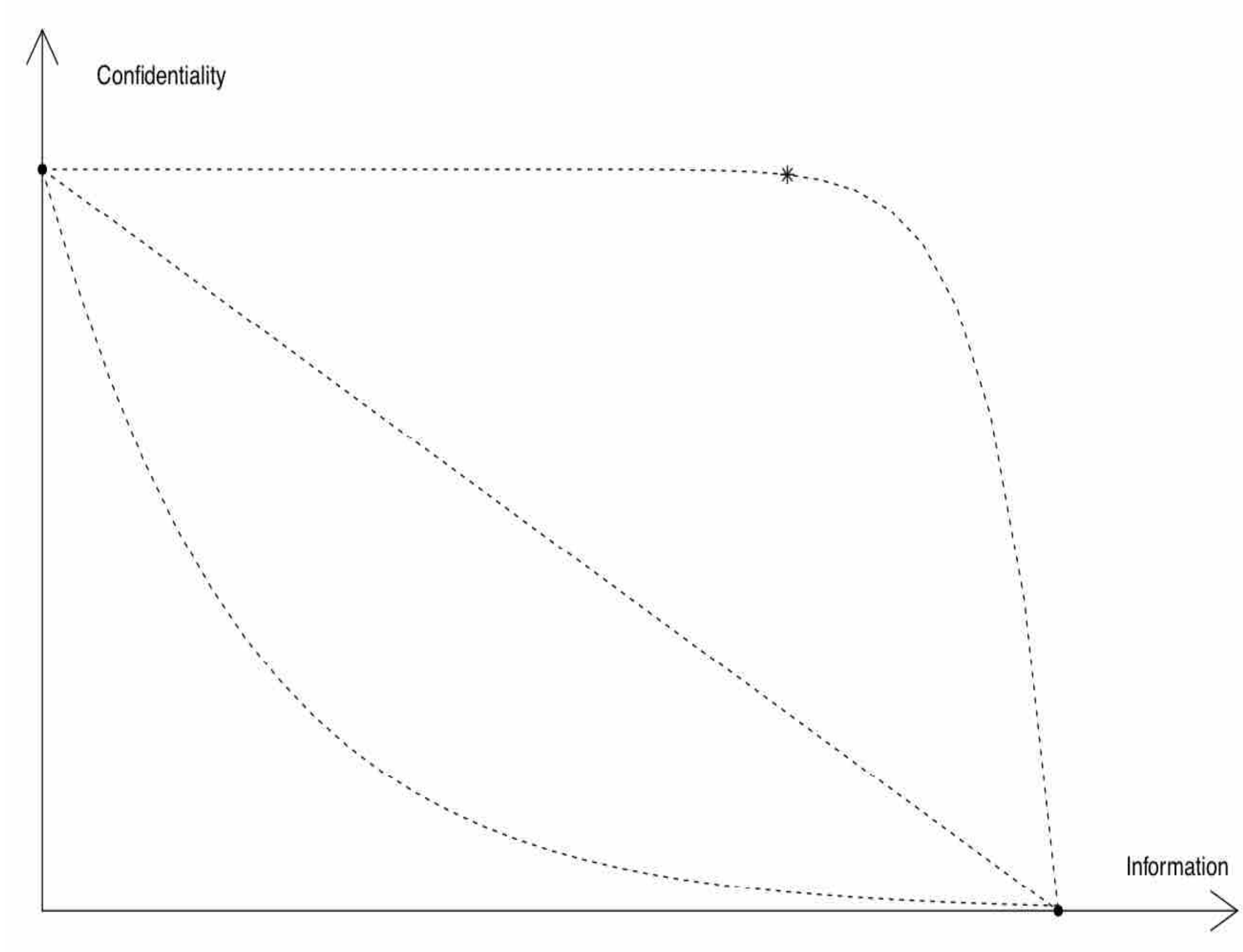
Who	**What** (COICOP-V/VI)			
	01.1.1.x	01.1.2.x	$\cdots$	12.7.1.x
(hush/pers) $\vdots$	***How much?*** $(Value = Price \cdot Quantity)$			

Old and new data*

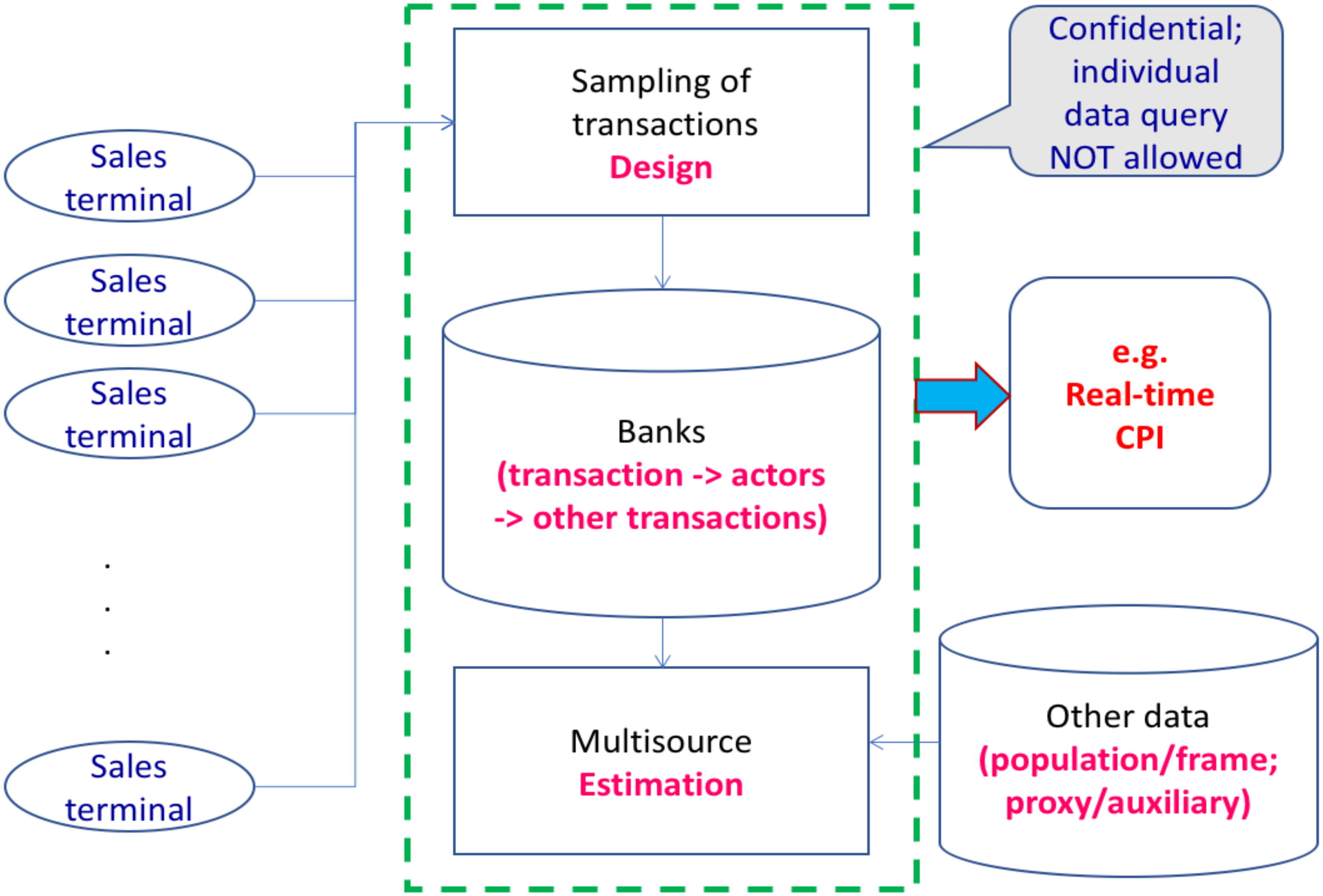
Source	**Who**	**What**	Remark
Survey	+	+/-	Non-sampl. err./Variance
Scanner	NA	+	Unstable GTIN
Receipt	+/-	+	Uncertain person ident.
Bankcard	+	NA	Changing platforms

$\star$ 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: $G = (U, A)$ = (nodes, edges) [digraph by default]

Network: $\mathcal{N} = (G, X)$ = (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 = locality

Def.: For each person k, $a_{ij}^k \in A$ iff $i \to j$ 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 = home, j = work
Same multigraph from different sources:

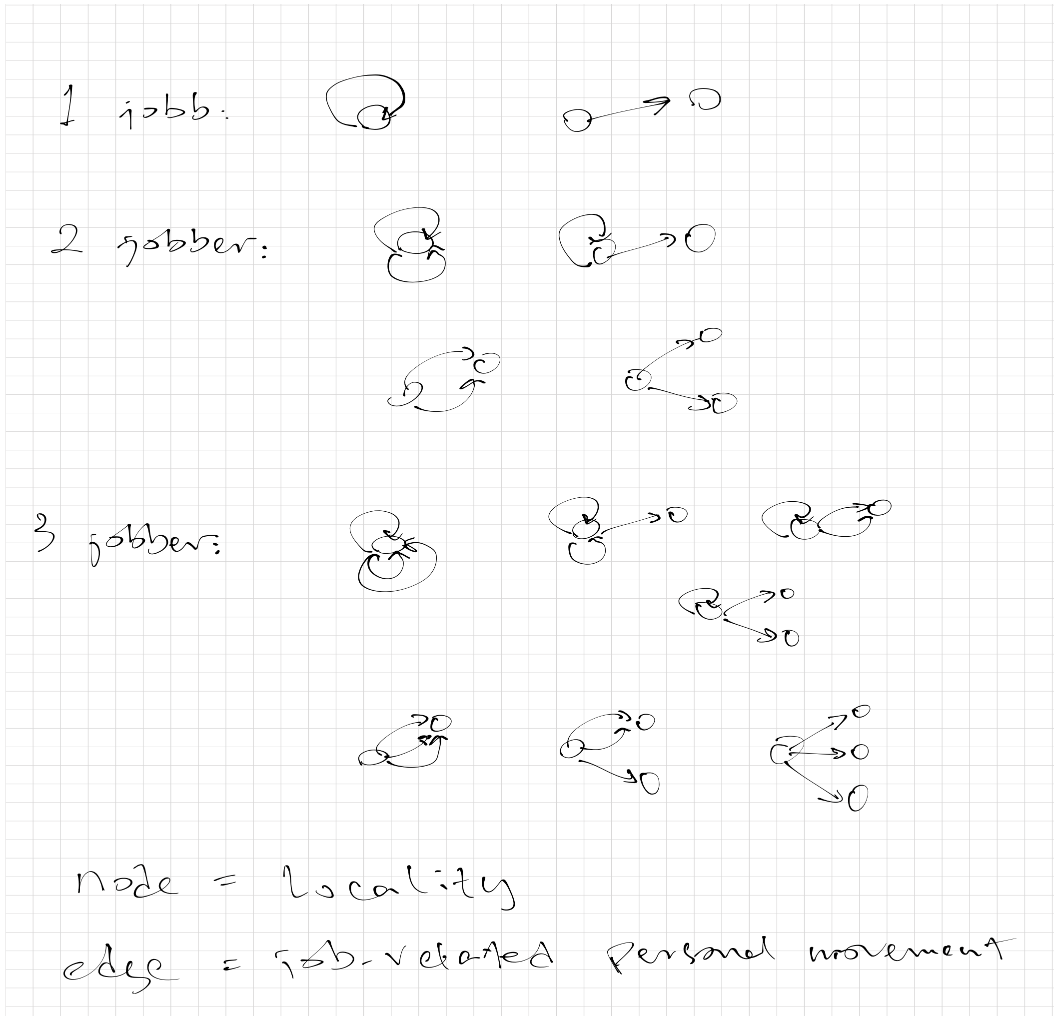
- admin: normative $\neq$ real work location j
- telecom: machine learnt $\neq$ real work location j

Some statistics: Norwegian admin data

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

NB. stronger growth of people with 2+ jobs

Motifs in job-related personal movement multigraph



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 <u>digraph</u> $G = (U, A)$, i.e. $|A_{ij}| \equiv 1$

- U = locality
- A = existing between-flow of labour

Simple labour-flow <u>network</u> $\mathcal{N} = (G, X)$

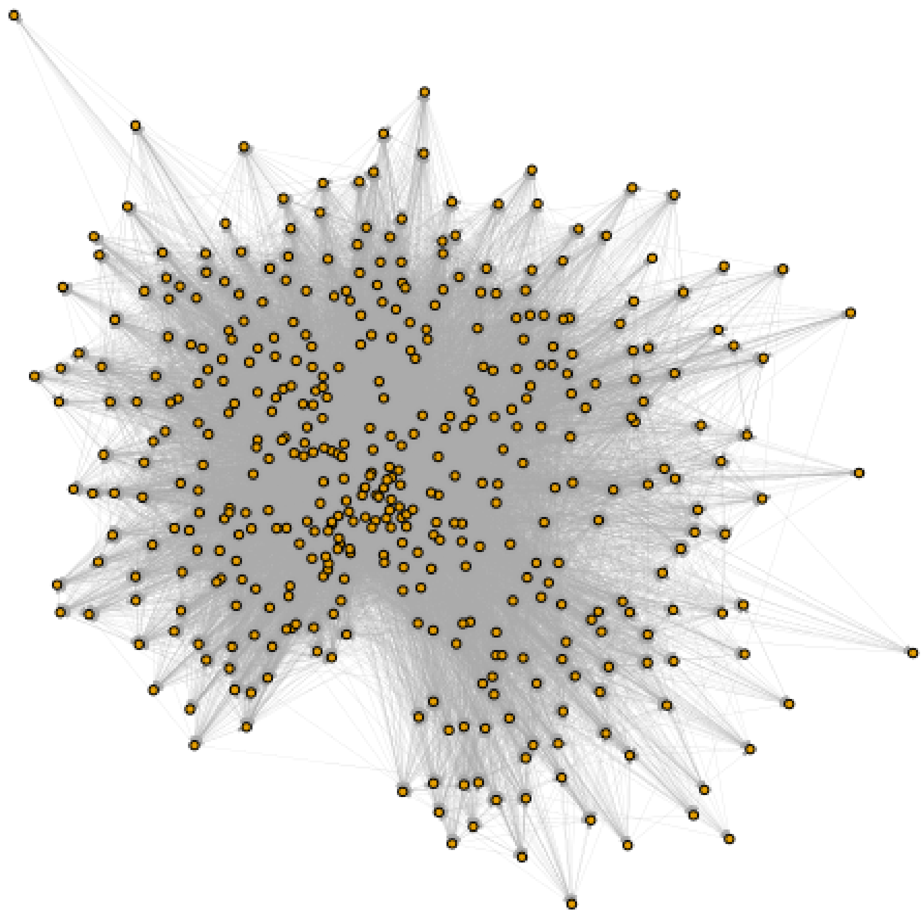
- X_A: weighted sum of multi-type edges in G

  [can be measured in no. persons, trips, time, etc.]
- X_U: e.g. no. employed pers. anchored at each $i \in U$

Simple digraph & network of labour flows

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

Source	Measure	Structure
Smartmeter	El-usage	Simple digraph
Sensor	Carrier position	Multi digraph
	Check-point	Multi digraph
	⋮	
Cellphone	Call/text	Simple digraph
	Position	Multi digraph
Transaction	Scanner (what)	List
	Receipt (what-who)	List
	Payment (who)	Simple digraph

E.g. Sampling of payments for disaggregation of CPI; confidentiality if Who = (geography, demography) $\neq$ 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, $\left(\mu_{\hat{\theta}} \pm 1.96\sqrt{\text{MSE}(\hat{\theta})}\right)$. 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 θ_0:

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

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

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 α^* is the coverage of $\hat{\theta}^*$ by $A_{\sigma,\alpha}$ [‘checking device’]

NB. CR varies with (σ, α): 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 $\bar{\gamma}_{n,\alpha}$ with $\underline{\alpha = 0.95}$, where $K = 1000$, $B = 1000$: $\hat{w}_i^{(b)} \sim N(w_i, \sigma_{i,n}^2)$ with $\text{median}(\sigma_{i,n}/w_i) = 0.202$ if $n = 250$, or 0.253 if $n = 160$; or $\hat{w}_i^{(b)} = (1 + r_i^*)w_i$ with $r_i^* \sim \mathcal{F}_r$ and $\text{median}(|r_i|) = 0.254$.

Proxy index P^*				Hypothetical survey index $\widehat{P}$			
	$n = 250$	$n = 160$	$\mathcal{F}_r$		$n = 250$	$n = 160$	$\mathcal{F}_r$
$\bar{\gamma}$	0.873	0.913	0.930	$\bar{\gamma}$	0.747	0.825	0.866
s.e.$(\bar{\gamma})$	0.005	0.003	0.003	s.e.$(\bar{\gamma})$	0.008	0.006	0.005
Proxy index P^*, scaling 1.4				**Hypothetical survey $\widehat{P}$, scaling 0.7**			
	$n = 250$	$n = 160$	$\mathcal{F}_r$		$n = 250$	$n = 160$	$\mathcal{F}_r$
$\bar{\gamma}$	0.754	0.822	0.862	$\bar{\gamma}$	0.863	0.914	0.929
s.e.$(\bar{\gamma})$	0.008	0.006	0.005	s.e.$(\bar{\gamma})$	0.005	0.004	0.003

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