Challenges in the Production of Official Statistics with Different Methods of Data Collection

Danny Pfeffermann

Annual Workshop on Survey Methodologies,

Brazilian Network Information Center (NIC.br)

Sao Paulo, Brazil, May 2019

What is official statistics? Why is it important?

Publication by a **national statistical office** (**NSO**)**,** based on a survey, census, administrative data, (**big data?**)

- **OS** is what people hear of almost daily. Unemployment rates, price indexes, education attainments, poverty measures, population counts, health and environmental statistics**,…**
- For most people, **OS** is what statistics is all about**!!**
- **OS** is what policy makers use (**should use**) for planning and decision making.
- **Growing** demands for detailed **timely** data, huge technological developments, declining response rates, tightened budgets**,…**

$\Rightarrow$ **Big new challenges**

Main methods of data collection for official statistics

1- **Surveys** based on probability samples**;** still the most common, and in many ways the most reliable method, if applied properly.

2- **Administrative records;** often requires linking several big files, which can be problematic and increase privacy concerns.

3- **Big data (?)** despite of all the noise, **not really implemented yet for OS;** increased pressure on **NSO's** all over the world to **digitise** (**"modernise"**) their production systems.

4- **Combinations** of the methods above.

Major problems with the use of traditional sample surveys

Yields objective (**unbiased**) estimators under the randomization (sampling) distribution, without the need for statistical models. Accommodates calculation of measures of errors. **However,**

- Often requires **large samples** for needed level of accuracy, particularly for small domain estimation ⇒ can be very **costly**.

- People and businesses are less and less willing to participate in surveys ⇒ **declining rates of response**, often **NMAR** ⇒ risk of **biased inference** if not handled properly.

- Use of models may increase efficiency at the risk of model failure, and hence possibly **biased inference**.

Proxi surveys (one reports for many)

One of possible ways to deal with small sample sizes and nonresponse**;** very common in household (**HH**) surveys (**e.g., LFS or** even **census**).

One person of **HH** (whoever can be reached), provides information for **all other members** of the household.

Possible ethical problem: Do other **HH** members agree that their **personal** data **(e.g., medical information)** is provided to interviewer**?**

- Major problem in **non-mandatory** surveys.
- **High propensity** for **nonresponse: "Don't know"**.
- **High propensity** for **correlated measurement errors**.
- **Not efficient** statistically **(single-stage cluster sampling)**

Example of estimates from Labor Force survey in Israel

Estimates based on Total, Self- and Proxy respondents- LFS

1- Smoking by Age and gender, counts per 1,000 residents

	Age group				
	20–24	25–44	45–64	65+	Total 20+
	Males				
All Sample	284	320	299	134	284
Self Respondents	343	353	311	135	301
Proxy Respondents	276	298	290	132	273
	Females				
All Sample	109	129	163	63	126
Self Respondents	146	151	188	77	151
Proxy Respondents	101	105	124	44	101

Estimates based on Total, Self- and Proxy respondents- LFS

2- Participation in Labor force & employment by gender, percentages

	Participation		Employed		Unemployed	
	Male	Female	Male	Female	Male	Female
All Sample	70.9	60.3	68.0	57.7	4.1	4.3
Self Resp.	76.3	65.3	73.3	62.8	3.9	3.9
Proxy Resp.	67.4	56.9	64.5	54.3	4.3	4.7

- Proxy sample size/self-response sample size **~ 60:40**.
- Who responds **not found to be related** to employment status.
- **So, what should we do???**

Major problems with the use of traditional sample surveys (cont.)

❖ **Timeliness-** Traditional surveys often take many months- users nowadays require that data be collected and released **“in real time”**. **NSOs** need to **stay relevant** in a dynamic changing world.

❖ **But** sometimes, survey data are much quicker than, **e.g.,** administrative files. **Example: income information**.

❖ **Mode effects- mixed mode surveys:** different modes of response**;** **telephone**, **personal interview**, **email, internet,…** different modes often offered sequentially to non-respondents with a previous mode.

Mode effects (cont.)

Mode-effects encompass two confounded effects:

Selection effect; different characteristics of respondents with different modes ⇒ **possible differences in values of study variables,**

Measurement effect; effect of **responding differently by same person**, depending on mode of response.

Big differences often observed in answers with different modes.

Reasons for mixed mode surveys: increased response rates, some modes **cheaper than others** (**internet!!**).

Example of mode effects- Agriculture Census, Israel, 2018

- **210** farmers responded both by internet and by telephone**!!**
- Ideal for assessing existence of **measurement effects.**

Study variables	**# Farmers T=I**	**# Farmers T>I**	**# Farmers T<I**
# of workers	**131**	**39**	**40**
Cultivated area	**139**	**38**	**33**

Study variables	**Mean Internet (I)**	**Mean Telephone(T)**	**Mean for T>I**	**Mean for T<I**
# of workers	**5.9**	**5.8**	**T=15.5** **I= 7.0**	**T= 7.5** **I=17.0**
Cultivated area	**108.5**	**105.9**	**T= 318.4** **I= 192.0**	**T= 88.3** **I= 144.5**

Mode effects (cont.)

A common approach to deal with mode effects**:** assume that one of the modes has **no measurement effect** ⇒ by restricting to this mode, the estimate of the population parameter is unbiased.

Uses **observational study** theory**;** requires knowledge of covariates satisfying strong ignorability conditions. (see **Pfeffermann, 2015** for details.)

❖ No such mode guaranteed - not clear how to test its existence.

Bayes-based Non-Bayesian Inference on Finite Populations from Non-representative Samples. A Unified Approach

(Pfeffermann, 2017, *Calcutta Statistical Association Bulletin*)

Bayes Theorem

For ***Y***, ***C*** random variables,

$$f_Y(y|C=c)=\frac{\Pr(C=c|Y=y)f_Y(y)}{\Pr(C=c)}=\frac{\Pr(C=c|Y=y)f_Y(y)}{\int \Pr(C=c|\tilde{y})f_Y(\tilde{y})d\tilde{y}}.$$

- **C is a conditioning variable, characterizing the sample or sample membership.**

Conditioning variables in special cases

$C_i = 1$ if population unit i is **sampled**, $C_i = 0$ otherwise √

$C_i = 1$ if sample unit i **responds**, $C_i = 0$ otherwise √

$C_i = 1$ if population unit i is an **internet user**, $C_i = 0$ otherwise **?**

$C_{it} = 1$ if sample unit i belongs to **treatment group** t √

$C_{im} = 1$ if sample unit i **responds** with **mode** m **?**

$C_i = 1$ if population unit i is included in **big data** **?**

❖ The target distribution in all the situations is the **unconditional population distribution**, $\Pr(Y = y_j \mid x_j)$, $j = 1, \ldots, J$ or $f_p(y \mid x_j)$.

Accounting for mode effects by use of Bayes Theorem

Suppose $\boldsymbol{M \geq 2}$ modes and denote by $\boldsymbol{G_i}$, the mode used by unit $i \in S$.

Denote by $\mathbf{x}_i$ covariates explaining $\boldsymbol{y}$.

Assump. 1. For every $j \in U$ exists a **true** value y_j with ***pdf*** $\boldsymbol{f_p(y_j \mid \mathbf{x}_j)}$,

Assump. 2. Every unit responds by one of the modes (but **see below**).

❖ **Not assumed** that y is measured accurately under any mode.

Accounting for mode effects (cont.)

By **Bayes theorem**,

$$f_M(y_i \mid \mathrm{x}_i, G_i = g) = \frac{\Pr(G_i = g \mid \mathrm{y}_i, \mathrm{x}_i)\, f_p(y_i \mid \mathrm{x}_i)}{\Pr(G_i = g \mid \mathrm{x}_i)}$$

$f_M(y_i \mid \mathrm{x}_i, G_i = g) \rightarrow$ accounts for **Selection/measurement** effects from using mode $\boldsymbol{g}$.

- Requires modelling $\Pr(G_i = g \mid \mathrm{y}_i, \mathrm{x}_i)$ (**e.g., multivariate logistic**).
- The covariates explaining the chosen mode not necessarily the same as the covariates explaining the outcome. (For model identification, the two sets of covariates need to differ in at least one variable.)

Further Remarks

1- The proposed approach does not require the existence of covariates that satisfy strong ignorability conditions.

2- The approach does not assume that the responses obtained by one of the modes are **correct**.

3- Nonresponse can be accounted for by viewing it as another mode.

4- The approach **requires** modelling $\Pr(G_i = g \mid \mathrm{y}_i, \mathrm{x}_i)$ and $f_{\boldsymbol{p}}(y_i \mid \mathrm{x}_i)$, but the model $f_{\boldsymbol{M}}(y_i \mid \mathrm{x}_i, G_i = \mathrm{g})$ **can be tested** using standard test procedures that compare observed and predicted values.

5- Model $f_{\boldsymbol{M}}(y_i \mid \mathrm{x}_i, G_i = g)$ can be fitted by **empirical likelihood**.

Use of Web-panels for Official Statistics?

Web Panel: big group of **volunteers**, **agreeing** to participate regularly in surveys via the internet, often in return to **money incentives**.

❖ Used extensively by private survey companies, **e.g.,** for opinion polls and election predictions. **Hundreds of thousands of persons**.

❖ **Web surveys** have huge advantages over traditional surveys.

Major problem: volunteers with access to the **internet** ⇒ at best represent the population of **web users**.

❖ **WP possibly** recruited by **probability sampling**, and samples from **WP often** selected by **probability sampling**.

Challenge: Estimate parameters of general population *P* from WP sample. Can we do it?

Inference based on Bayes theorem

Let $A_i = 1$ if unit $i \in U$ is in the web panel (**WP**), $A_i = 0$ otherwise.

Assumption: $\Pr(A_i = 1 | \mathrm{x}_i, y_i) > 0 \ \forall i \in U$ **;**

y-outcome (study) variable, $\mathbf{x}$-covariates.

$$f_{WP}(y_i | \mathrm{x}_i) = f(y_i | \mathrm{x}_i, A_i = 1) = \frac{\Pr(A_i = 1 | \mathrm{x}_i, y_i) f_p(y_i | \mathrm{x}_i)}{\Pr(A_i = 1 | \mathrm{x}_i)},$$

$f_p(y_i | \mathrm{x}_i)$ = distribution in target population ***U***,

$f_{WP}(y_i | \mathrm{x}_i)$ = distribution in **WP**.

Inference based on Bayes theorem (cont.)

In practice, not every **WP** member responds to every survey taken from its members. Let $R_i = 1$ if **WP member** i **responds**, $R_i = 0$ otherwise. The marginal distribution for **responding WP** i is then,

$$f_{WR}(y_i \mid \mathrm{x}_i) = f(y_i \mid \mathrm{x}_i, A_i = 1, R_i = 1)$$
$$= \frac{\Pr(R_i = 1 \big| y_i, \mathrm{x}_i, A_i = 1)\Pr(A_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(R_i = 1 \mid \mathrm{x}_i, A_i = 1)\Pr(A_i = 1 \mid \mathrm{x}_i)}.$$

- **Two conditioning variables.** Requires that covariates are partly different in the three models.
- Similar decomposition when considering informative sampling and nonresponse (**Feder & Pfeffermann, 2018**).

Use of administrative records

Supposed to provide timely, accurate data with good coverage, but this is not always the case.

- Israel's population register covers all the population residing in Israel but ~**15%** of the addresses are wrong.
- Tax records of businesses are obtained with a delay of ~**2 years**.
- No administrative data on opinions, attitudes, etc.
- Means or totals of administrative data often used to strengthen survey estimates by use of **statistical models** or **calibration**.
- If data are timely, accurate and contain all required information, avoids the use of a survey.
- Unfortunately, even Government agencies are often reluctant to transfer the data to **NSO's** because of data protection issues.

Integration (matching) of several administrative records

Unlike in planned surveys, desired information possibly contained in more than one record.

- **Matching problematic** if personal identifiers unknown, requires probabilistic algorithms based on information in all the records.
- **Coverage of records** might be different and may not apply to same time periods.
- **Definitions & accuracy** of information may differ between records.
- Possibly **Conflicting information** in different records, **e.g.,** different addresses in different records. (**Major problem** with the use of censuses based on administrative records.)
- Possibly magnified problems of **data protection** after integration.

Integration of administrative records with surveys

Not all the required data possibly contained in administrative records.

Example: at **ICBS** we are attempting to match **retrospectively**, and in the future **concurrently** **LFS** data with administrative **social security (SC)** data on earnings, allowances etc.

- Attaching the **LFS** weights to the matched **SC** records, should allow evaluating the coverage of the **SC** data and identify possible **failures**.

- **Imputing** the **LFS** variables from the **SC** variables by a model estimated based on the matched records will allow (**if successful**) to have a **'complete'** administrative **SC** file.

- If all goes well, we shall end up with an enriched **longitudinal** administrative file.

Use of big data for production of official statistics (OS)

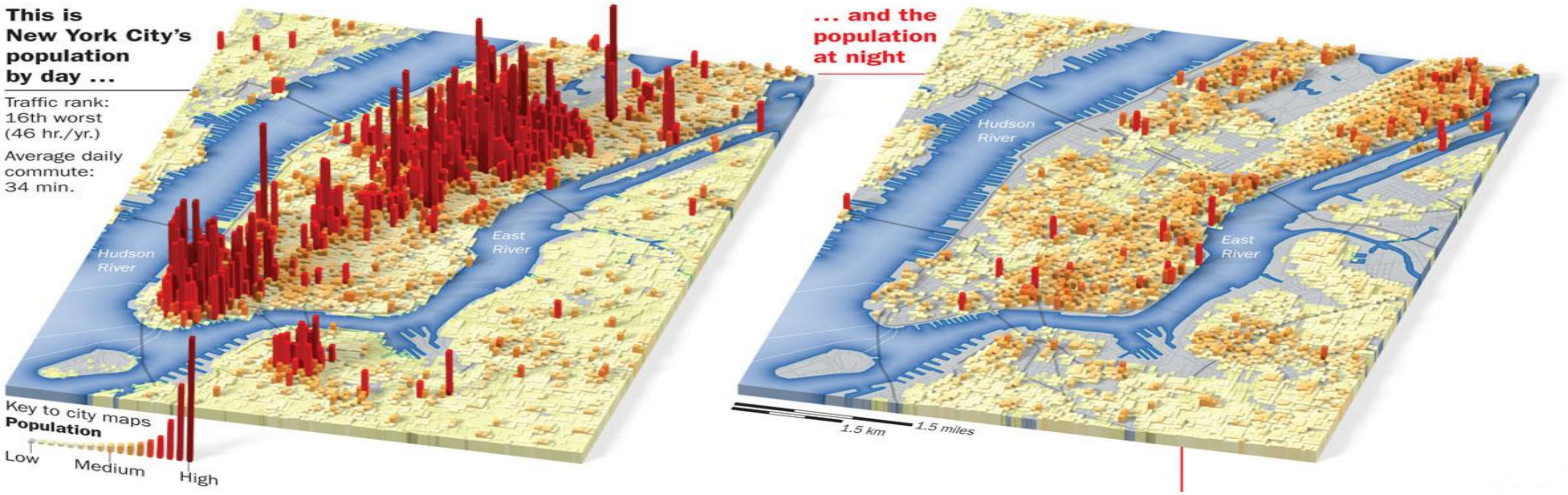
Differences between administrative files and big data (??):

Both are big!!

- Big data often unstructured, diverse, and appears irregularly (**e.g.,** data obtained from social networks, e-commerce**…**)
- Big data updated dynamically**/**timingly.
- Big data not prepared or maintained for administrative or statistical purposes. **It is a by-product, not produced for OS purposes!!**
- Big data can cease to appear at any time.
- Big data at risk of data manipulation.

Use of big data for production of official statistics (OS)

Location data from mobile phones


This is New York City's population by day ...
Traffic rank: 16th worst (46 hr./yr.)
Average daily commute: 34 min.
Hudson River
East River
Key to city maps
Population
Low
Medium
High
... and the population at night
Hudson River
East River
1.5 km
1.5 miles

Very hard to get this information from surveys for every location. Will telephone companies provide this information? Maybe in the future after proper legislation. But how will we learn, for example, about the purpose of the trip or type of transportation?

Use of big data for production of official statistics (cont.)

- High dimensionality and extremely large size.
- Possible **coverage/selection bias** (we are talking of **OS**).
- Data accessibility, new legislation**?** Permission by the public**?**
- Increased risks of data disclosure.
- New sampling algorithms (to reduce size and control disclosure**;** sampling from **big, versatile dynamic data different** from sampling **finite populations**).
- Heavy computation, new algorithms and analytic tools.
- Integration of files from multiple sources in different formats appearing at different times.
- Risks of data manipulation or sudden unavailability.

Shall we really get what we need for our official statistics?

Two types of big data

Type 1. Data obtained from sensors, cameras, cell phones**…,** generally structured, accurate, relates to a particular population.

Type 2. Data obtained from social networks, e-commerce,**…,** generally diverse, unstructured and appears irregularly.

- Data from different sources may have different formats, arrive at different times with different degree of reliability, and defined differently.

- **No such problems with traditional surveys**.

- **NSOs need to be prepared that data may cease to exist**.

- **Big data is a by-product, not produced for OS purposes!!**

Other important issues

Non-representativeness- major concern in use of big data for **OS**. House sales advertised on the internet do not represent properly all house sales, web scraping for job vacancies does not represent all job vacancies, data from social media not representative of data held by **general public (e.g.,** "public sentiment").

❖ **Big not always better!!** Collecting enormous amounts of data **does not guarantee** getting right answer. A **smaller balanced sample** may provide better insights than a **large skewed sample.**

No problem when using big data as **predictors** of other variables.

e.g., use **BPP** to predict the **CPI**, **job adverts** to predict **employment** or **job vacancies**. Use **Satellite images** to predict **crops**.

Requires proper statistical analysis to identify and test (routinely) the prediction models.

Web (internet) scraping- sharing economy

❖**Potentially, one of the main uses of big data for official statistics**.

The rapid rise of electronic platforms link individuals with each other, offering the opportunity to share goods or assets without transfer of ownership, or to exchange services. This **"new economy"** has gained such importance that it is now a real source of economic activity.

❖ Statistical offices around the world are interested in assessing the volume of activity of the **sharing economy** (value added, incomes, prices, employment, etc.)

❖ Difficult to capture this new phenomenon with traditional data collection methods.

Feasibility study of web scraping: Measuring sharing economy in short-term rental

In Israel (and in many other countries), one of the most significant components of the sharing economy is **short-term rental**, and the leading company is **Airbnb**. We conducted an experiment to estimate the extent of Airbnb's activity through **web scraping**.

Collecting data in the **Airbnb website** provides information about the extent of the phenomenon in Israel.

Assessing the short-term rental activity from Airbnb website

We like to answer the following questions:

How many properties are offered for short-term rental**?**

Where are these properties**?**

What are the properties' characteristics (number of bedrooms, number of beds, maximum guest occupancy**,…**)**?**

Prices per night for different sizes of groups in different seasons**?**

How many people offer rental properties through this platform**?**

Country of origin of tourists staying in holiday apartments via Airbnb**?**

Ratio of Israelis to tourists**?**

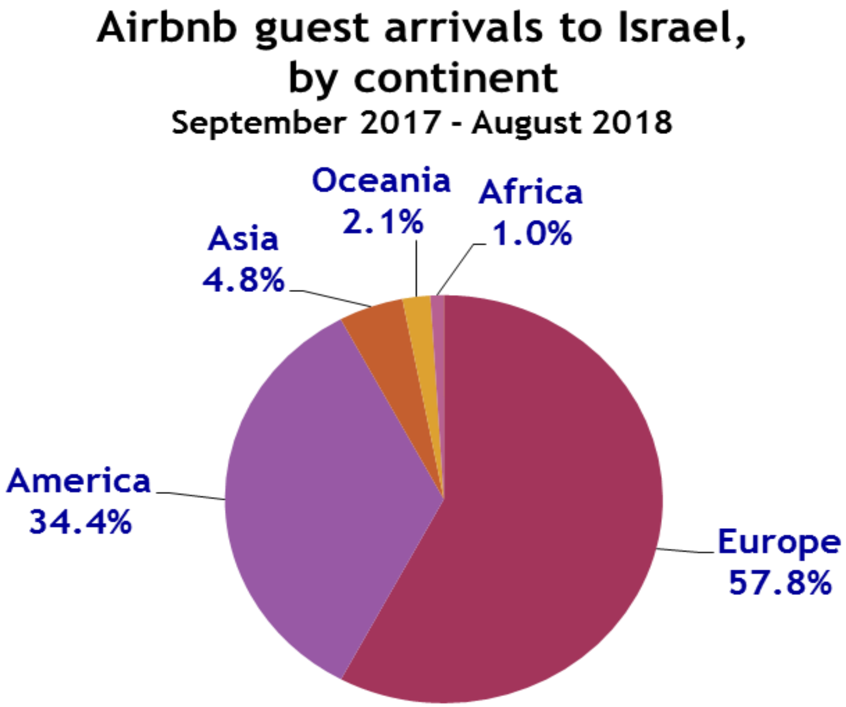
Assessing short-term rentals (cont.)

The data collected has the potential of improving the official statistics in the following areas:

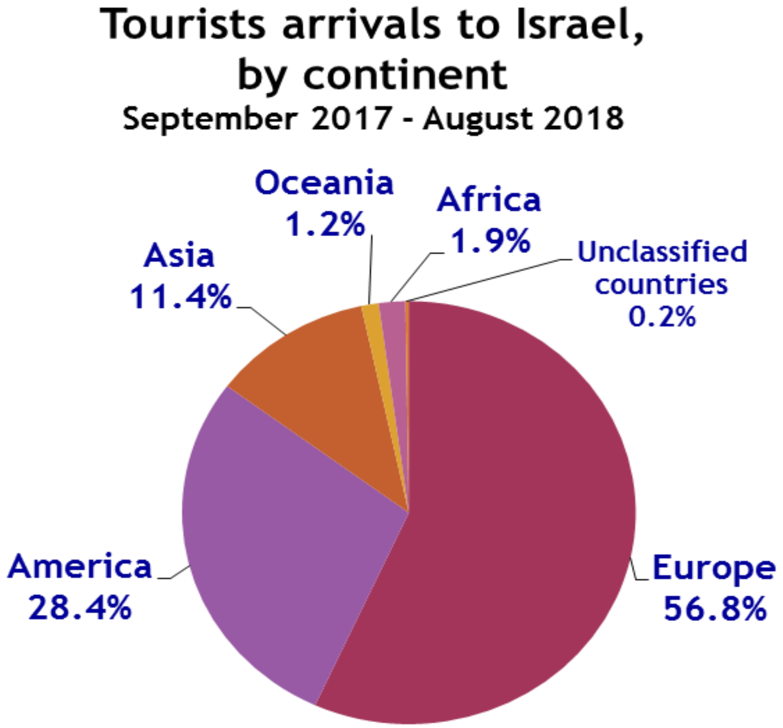
- Evaluate the monthly price changes so that it can be added to the **basket of goods** of the Consumer Price Index (**CPI**).
- Estimate the extent of revenue so that it can be added to the **national accounts** and **GDP**.
- Expand official statistics of tourism.

Comparison between distributions of tourists to Israel and tourists staying in Israeli Airbnb vacation rentals

Based on Airbnb



Based on border police reports



Example: estimating the revenue of Airbnb activity in Israel (2018)

Data on tourist's arrivals (from border police)

Data from annual tourism survey: **%** tourists to Israel staying in rented apartments.

Data from Airbnb website: number of guests that stayed in Israel through **Airbnb** (**90%** of total short term rentals).

From scraping guest reviews: relative share of Israeli guests and tourists (**20%** Israeli tourists out of **425,000** tourists).

Data from tourism survey: average length of stay of a tourist in a rented apartment **&** distribution of group size.

↓

Scrape prices for different periods and different group sizes for estimating average income per night

↓

Yearly Revenue

Possible issues when using web scraping to measure the sharing economy in the short-term rental sector

A- Country of origin:

Coverage bias - Do tourists reporting their countries of origin in the Airbnb guest reports represent properly non-reporting tourists**?**

B- Prices

B1. Consistency - are the prices collected for tourists from different countries the same, or do they differ between currencies and users**?**

B2. Coverage bias – prices were collected for **available dates.** What about prices which are not available**?** How seasonal are the prices**?**

C- Employment

Hard to assess whether a person advertising his asset in a digital platform should be considered as **“**employed**”**. Does he/she spend working hours renting out the property**?** If so, is this his/her main source of earnings**?**

We found from **LFS** that **3,800** persons are engaged in short term rental as their main occupation. In the Airbnb website we find **6,046** persons for which there was at least one guest review. (Under estimation**!!**)

D- Legal Issues

Web scraping may be in **contrast** to the terms of use of some websites. On the one hand, there is a concern that web scraping could breach database rights of website owners. On the other hand, national statistics laws **empower NSOs** to collect necessary data for statistical purposes. Is web scraping consistent with this objective**?**

Other important issues in the use of big data (cont.)

Sampling: **random sampling** will continue to play a major role in the era of big data.

❖ Reduces **storage space**, helps protecting **privacy**, produces **manageable data sets** on which algorithms can run to produce **estimates**, and **models** can be fitted.

❖ Sampling from **big, versatile dynamic** data **different** from sampling **finite populations**, requiring **new** sampling algorithms; **e.g.,** sampling from social networks **(?)**

❖ If no sampling ⇒ **no sampling errors**. Which measures of error should be computed**? Measure of bias? How? Compare to traditional estimates? Measurement errors? Only sampling errors when sampling from the big data?**

Other important issues with the use of big data (cont.)

Big Data for sub-populations: NSOs publish estimates for

sub-populations; age, **gender**, **ethnicity**, **geography,…**

Big data may not contain this information. Requires **massive linkage if** missing information available in other big files.

Data on **sales from supermarkets** contains **no information** on buyers ⇒ cannot compare consumption patterns (or types of commodities) between different types (**e.g.,** age) of buyers.

Possible solution: Link sales to buyers by use of **credit card numbers**. **Will credit card companies provide them?**

- ❖ **Will traditional sample surveys always be needed?**

Computer engineering for OS from big data

No longer **Gigabytes** (~10^9 bytes). **Terabytes** (~10^{12} bytes) and **petabytes** (~10^{15} bytes) **new standards**.

❖ Available computing facilities at **NSOs** cannot store and handle such huge volumes of data.

Possible solution: Use **cloud** storage, management and processing facilities (**Amazon, Microsoft,…**)

Big problem with **Data protection**. **Many users**, data distributed over a **large number of processors**.

Another solution: Data centre. Incorporate **all local computers; central management** of storage space **&** processing power of separate servers. **Major challenge**.

Use of Big data for OS- summary remarks

New expensive computing facilities**,** **new** data processing techniques**,** **new** linkage methods**,** **new** visualization methods**,** **new** sampling methods**,** **new** analytic methods**,** **new** measures of error**,** **new** disclosure control procedures**,** **new** legislation**,** **new** employees (**data scientists**)**,…**

Big **potential advantages**: timeliness, much broader coverage (possible **coverage bias**), **no** need for sampling frames, **no** questionnaires, **no** interviewers**,** answer new questions**,…**

❖ Constant **decline in response** rates in traditional surveys and tightened budgets ⇒**use of big data inevitable**.

“Good news”: Big data will just grow bigger and bigger.

Expect no miracles for OS in the near future.

Accounting for non-representativeness of big data

Major concern in use of big data for **OS**.

Kim (2017) proposes **3** different procedures to account for **non-representativeness** of the data**:**

Reservoir sampling, Inverse sampling, Survey integration.

Survey integration: combine big data with survey data.

Basic assumption: membership of **sample elements** in big data **(B) known**. (Match**?** Ask sample members**?**)

Let $\delta_i = 1(0)$ if $i \in B (i \notin B)$. **Sample data:** $\{(\mathrm{x}_i, z_i, \delta_i)$**;** $i = 1,...,n\}$**;**

$\mathbf{x}_i$ = model covariates**,** z_i =variables explaining **B-**membership.

Procedure: Model $\pi_i = \Pr(\delta_i = 1 | \mathrm{x}_i, z_i)$ from sample data$\Rightarrow$ $\hat{\pi}_i$.

Use $w_i = (1/\hat{\pi}_i)$ as weights for analysing the **big data**.

Remarks on proposed procedure

Neat idea but with important limitations:

Assumes existence of a sample with required data.

Assumes knowledge of membership in **B** of sample elements.

Assumes existence of variables **$\mathbf{x}$** and **$\mathbf{z}$** explaining **B- membership**

Assumes $\Pr(\delta_i = 1 | \mathrm{x}_i, z_i, \boldsymbol{y_i}) = \Pr(\delta_i = 1 | \mathrm{x}_i, z_i)$**;**

(“**noninformative sampling**”).

Accounting for coverage bias by use of Bayes theorem

Population model: $f_p(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for target population outcomes (**census model**),

Big data (B) model: $f_B(y_i \mid \mathrm{x}_i) \rightarrow$ model holding for **B** data.

Denote, as before, $\delta_i = 1(0)$ if $i \in B (i \notin B)$.

$$f_B(y_i \mid \mathrm{x}_i) \overset{def}{=} f(y_i \mid \mathrm{x}_i, \delta_i = 1) \overset{\text{Bayes}}{=} \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

$$\Downarrow$$

$$f_B(y \mid \mathrm{x}_i) = f_p(y \mid \mathrm{x}_i) \textbf{ iff } \Pr(\delta_i = 1 \mid y_i, \mathrm{x}_i) = \Pr(\delta_i = 1 \mid \mathrm{x}_i) \forall y_i. \quad \textbf{(**)}$$

❖ **If (**) satisfied, feel free to use B to analyse population data.**

Alternative procedure (cont.)

$$f_{B}(y_i \mid \mathrm{x}_i) = \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)\, f_{p}(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}.$$

Target *pdf* is $f_p(y \mid \mathrm{x})$**;** observations only available from $f_B(y \mid \mathrm{x})$.

The two distributions connected via **probability link function** $\Pr(\delta \mid y, \mathrm{x})$**;** enables estimating **target population pdf** from observations obtained for **Big data.**

- $f_B(y_i \mid \mathrm{x}_i)$ can be estimated from **B** (or **sample** thereof).
- $\Pr(\delta_i = 1)$ allowed to depend on target variable, **y**. May depend also on variables ***z***, but **only need** modelling $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ (or include $\mathbf{z}$ among the $\mathbf{x}$-variables).

Alternative procedure (cont.)

$$f_{\boldsymbol{B}}(y_i \mid \mathrm{x}_i) = \frac{\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)\, f_{\boldsymbol{p}}(y_i \mid \mathrm{x}_i)}{\Pr(\delta_i = 1 \mid \mathrm{x}_i)}$$

- Inference requires modelling $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ (and possibly $f_{\boldsymbol{p}}(y_i \mid \mathrm{x}_i)$), but **no survey data required**.

- Models assumed for $\Pr(\delta_i = 1 \mid \mathrm{x}_i, y_i)$ and $f_{\boldsymbol{p}}(y_i \mid \mathrm{x}_i)$ **testable** by testing the implied model for $f_{\boldsymbol{B}}(y_i \mid \mathrm{x}_i)$, using **conventional model testing** procedures, since the big data are **known**.

Concluding remarks on use of big data for OS

- Use of big data for **OS** is **not straightforward** and requires overcoming many legal, ethical and computational problems **+** development of new methodologies.
- Use of big data for **OS** **inevitable** in the long run. Promises huge advantages, which cannot be ignored.
- **Non-representativeness** of big data is a major concern in their use.
- The procedures outlined in this presentation to deal with the **non-representativeness** problem are only **first** steps.
- Much more theoretical and applied research required.

Final comments (quotations from other authors)

Holt (2007): Five formidable challenges for official statistics:
wider, deeper, better, quicker, cheaper.**

Citro (2014): Official statistical offices need to move from probability sample surveys paradigm to **mixed data source paradigm.**

Kalton (2018):

- Unlikely that social surveys will be replaced by administrative data, although these data can be valuable addition to surveys.
- Quality of estimates from internet surveys is a concern.
- Interface between design based and model dependent inference is needed for inference from non-probability samples.

**** NSOs** need to be much **quicker** in their production in order to stay **relevant**. **Impossible** to publish estimates in **2019** that relate to **2015**.