How to use big data for Official Statistics?

Li-Chun Zhang
University of Southampton & Statistisk sentralbyrå

Résumé L'échantillonnage probabiliste présente un certain nombre de vertus particulièrement intressantes dans le cadre de la Statistique Officielle comme la transparence, l'universalité et la flexibilité. Au cours des dernières décennies, l'utilisation de données provenant des sources administratives s'est accélérée au niveau international. La révolution numérique récente a grandement amélioré la disponibilité des données autres que celles des sources administratives, lesquelles peuvent également être mobilisées dans le cadre de la Statistique Officielle. Dans cette présentation, je discuterai des différentes façons d'utiliser des données massives en Statistique Officielle. L'accent sera mis sur le besoin de formuler des conditions de validité précises pour l'inférence descriptive à partir de données volumineuses non probabilistes et sur les difficultés associées à la vérification de ces conditions. En effet, dans la plupart des situations, le biais semble inévitable. Une approche basée sur l'évaluation de l'incertitude et sur la communication sera proposée et illustrée.

Limitations of survey sampling: An example To compute the Consumer Price Index (CPI), the price indices from different consumption groups need to be weighted together according their expenditure shares. Table 1 shows the weights at the most aggregated level, which were estimated based on the Norwegian Consumer Expenditure Survey (NCES) in 1998-2000 and 2012. The NCES sample size was about 7000 households, and the diary data were collected over two weeks. The expenditure shares calculated for the whole year had thus appreciable uncertainty due to sampling and non-sampling errors.

The monthly CPI requires actually weights at a much more detailed level. For instance, for "Food, non-alcoholic drinks", one would like to break down the two figures in Table 1 to 100+ subgroups, in order to reach the level of "homogeneous products" according to the definition in National Account. Also break-downs by geographic and/or demographic features of the consumers are of considerable interest to the users. As it can be seen in Table 1, at the most aggregated level, the food expenditure decreased for about 0.2% over this 10-year period. The actual need of statistical information is obviously beyond what the NCES can possibly satisfy, due to the limited sample size and unavoidable sampling and non-sampling errors.

Changing paradigms Historically speaking, one can readily discern at least 3 major paradigms for the production of Official Statistics.

I. In the beginning, there was only census, or administrative data collected in a census.

Table 1: Household expenditure (Source: 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 articles	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

II. The 20th century witnessed the emergence and maturing of survey sampling.

III. From the late 1960's, "achieve statistical systems" came into creation in the Nordic countries, based on data that are continuously updated from administrative sources.

Producing census-like statistics based on data originated from administrative sources has yet to be given a systematic and articulated scientific foundation. Nevertheless, "the comprehensive uptake of administrative data in the Nordic countries was inspired by the idea of *possible separation of data collection and production*: on the one hand, capture and curation as data is generated; on the other hand, more or less separately, processing and output as the need arises" (Zhang, 2018). The outlook has two essential characteristics, which at least in theory are unnecessary for the approaches based on census and survey sampling:

(a) reuse of data initially generated and recorded for a different purpose,

(b) combining multiple sources to amend the deficiencies of coverage and measurement.

Uses of other types of big data for Official Statistics share these two essential characteristics with the approach based on administrative data. Time will tell if the infusion of various forms of data can be accommodated within the existing paradigm — only differing from the past in certain technical and methodological aspects, important though they may be — or if it leads to fundamental reconceptualisation of Official Statistics.

Sources of relevant big data At the present stage, the most relevant sources of big data for Official Statistics seem to include the following:

- Administrative data from government agencies (i.e. admin data $\subset$ big data)
- Commercial and financial transactions
- Telecommunication networks
- Satellites, traffic monitors, smart meters and similar sensor systems
- Health registers (of disease, prescription, etc.)
- Internet (NB. impossible to exclude, but often as a last resort presently)

Types of usage Relevant big data can have *three* broad types of usage:

- Indirectly as part of the statistical infrastructure;
- Directly as statistical data, especially for replacing some of the existing ones;
- Facilitating new targets of Official Statistics.

Statistical infrastructure Big data can obviously provide additional *auxiliary* data to support survey sampling. In principle this does not differ from the long-standing tradition of using auxiliary data from administrative sources or past censuses for sampling, processing and estimation. The required methodology can still vary regarding how the big data can be assimilated into the statistical system. This pertains to unit identification and linkage, as well as derivation of relevant measure when the big data initially exists in other non-numeric forms such as text or image.

For example, big data provide various *sign-of-life* data, when they are generated by events that actually take place. There is typically a non-negligible degree of delay and incorrect registration in the administrative sources, such as when someone fails to notify the municipality authority of an address change in a timely manner, or files an address which is not where the person actually 'lives'. Sign-of-life data generated by concurrent events can help to bridge the gap between the 'formal' status in the administrative systems and the actual situation, and improve the frame for sampling and/or estimation.

Statistical data Big data is increasingly being reused as statistical data, possibly in combination with primary statistical data collected in censuses and sample surveys.

As an example, a number of countries currently aim to replace the population census by a statistical population dataset (SPD) compiled from multiple population-size administrative datasets, in combination with a suitable coverage survey. A prominent issue in this context is that the SPD suffers from appreciable over-counting, because there is often little incentive

for people to de-register or otherwise update quickly. Both the design of coverage survey and the associated estimation method will need to adapted as a result, whether or nor other sign-of-life big data are utilised in the construction of SPD.

For another example, the calculation of price indices belong to the CPI used to be based on price data collected for a sample of representative goods (or services), where the number of items surveyed in this way is negligible compared to the universe of consumption items. Over the last decade, a number of countries have started to compute the food price index based on scanner data delivered from the main supermarket chains in the country. Typically, the data consists of the total transaction value and quantity of each food product that has been bought and sold over the period of, say, a week. Such a 'census' has paradoxically raised the issue of index formula, because the traditional 'small-data' formulae do not use quantity data at the product level, which were simply impossible to obtain. Moreover, the traditional formulae require exact matching of the items over time, which would have been extremely resource-demanding if it had been pursued for all the available items.

New targets A network representation can be more appropriate for combining multiple datasets than the traditional population matrix representation of a list of units and their associated measures. For instance, a network of person, household, family, work place and residence locality can be constructed based on relevant administrative/big data. Genuine network parameters can be envisaged as the target of estimation, which more meaningfully summarise the reality than separate simple statistics associated with each type of units.

Remark It seems unavoidable for one's formulation of target parameter to depend on the data that are at all obtainable. It is thus always an act of balance between how much one adapts the target (i.e. its definition and interpretation) to the available data and how much one adapts the data (i.e. its design, collection and precessing) to a pre-fixed target.

Valid descriptive inference from big data Descriptive inference is concerned with one or several summary statistics of a given finite population, such the the population size, the mean of a value associated with each population unit, the number of pair of nodes that are connected with edges in both directions in a given finite digraph.

- Insofar as the relevant big data does not arise from a probability sampling design, there may be problems of coverage and selection, e.g. a unit may have zero probability of being observed or the probability may be positive but unequal and unknown.

- Insofar as the available measures do not conform to that of a designed survey, the big data may be subjected to measurement errors, i.e. a discrepancy between the observed

measure and that which could have been collected in a designed survey.

Validity conditions are assumptions which ensure valid descriptive inference from big data, despite the presence of coverage, selection or measurement issues, e.g. that consistent estimation is possible of the target parameter and the association uncertainty.

Remark Compared to ideal survey sampling, universality is lost with non-probability samples, unless a sample has neither coverage nor selection problems. One may be able to specify the validity conditions in general terms, but these may be impossible to verify with the available data, or one may even be convinced that the assumptions cannot possibly hold in a given situation. In contrast, the validity of design-based descriptive inference from probability sampling can be establish generally, because the conditions are not 'assumptions' in the same sense, but the actual procedure of sampling.

An example Suppose the big dataset contains and identifies all the population units (U), i.e. no issues with coverage or selection. Let $B = \left(\sum_{i\in U} x_i x_i^T\right)^{-1}\left(\sum_{i\in U} x_i y_i\right)$ be the target parameter, which is the census OLS of β in the linear regression

$$y_i = x_i^T\beta + \epsilon_i \qquad \text{and} \qquad E(\epsilon_i) = 0.$$

However, suppose that due to measurement error, one observes y_i' instead of y_i. Valid inference of B is feasible provided

$$y_i' = y_i + e_i \qquad \text{and} \qquad E(e_i) = 0,$$

but not if

$$y_i' = \alpha y_i + e_i \qquad \text{and} \qquad E(e_i) = 0 \qquad \text{and} \qquad \alpha \neq 1.$$

Big-data proxy expenditure weights Table 2 shows a break-down of the CPI weights for "Food, non-alcoholic drinks" at a level immediately below that of Table 1. The CPI weights were derived from the National Account data on retail turnovers. Also shown are the CPI weights for September 2016, and the *proxy* weights based on the purchase data from the retail chain Coop. The proxy weights are further calculated based on the purchases recorded for the members and the non-members, respectively. We notice the following.

- The most important use of the NCES in the past has been to provide the CPI weights. However, it is extremely burdensome, has a very high nonresponse rate, and is known to suffer from misreporting errors for various types of consumption. As noted before, the actual information need is beyond what the NCES can provide.

- The proxy CPI weights have negligible variance due to sampling. But they are biased due to coverage errors and selection errors. The latter is because in practice one is unable to classify automatically all the products in the reported purchases.
- The proxy weights exhibit clear discrepancies to the CPI weights, whether overall or separately by membership. However, the discrepancies do not appear to be larger than those between the NCES and CPI weights in 2012.

Table 2: Food expenditure shares (in percent)

	Year 2016, Sept.				Year 2012	
	Member	Non-member	Coop (all)	CPI	NCES	CPI
Meat	17.3	12.3	15.0	14.9	17.3	17.9
Dairy	15.7	13.1	14.5	12.1	14.3	10.0
Corn	12.6	12.2	12.4	8.9	12.0	10.0
Tobacco	6.4	11.3	8.6	10.6	6.8	12.9
Sweats	7.6	8.5	8.0	10.4	8.3	10.1
Soft drink	6.6	8.8	7.6	7.2	6.8	8.9
Other food	7.3	6.8	7.1	8.4	5.3	3.3
Beer	5.1	8.3	6.6	6.8	4.5	4.4
Vegetable	6.4	5.8	6.1	6.4	9.0	6.1
Fruit	5.0	5.4	5.2	6.2	6.8	5.0
Fish	5.2	3.8	4.6	4.2	5.3	7.4
Tea	2.7	2.2	2.5	2.8	2.3	2.7
Fat, oil	2.2	1.5	1.9	1.2	1.5	1.4

To be detailed in the talk Two questions will be be given special attention:

1. What are the potential effects of using proxy instead of survey CPI weights? Put in the other way, what are the conditions under which the proxy weights do not cause bias?

2. How to quantify the error associated with the resulting proxy CPI, since one can be quite certain that the required validity conditions will *not* be met exactly?