

# The evolving role of non-survey data in social statistics

Jack G. Gambino<sup>1,2</sup>

## ABSTRACT

For decades, administrative data such as birth records and tax data have played a major role in official statistics. In recent years, efforts to exploit other administrative and non-survey data have increased significantly. Some statistical agencies have adopted an admin-first policy. To meet an information need, the agency first looks at existing data sources, turning to the survey option as a last resort. Perhaps this change would have occurred naturally over time, in part due to cost pressures, but it became essential due to the gradual decline in survey response rates. For some surveys, the decline has been substantial, leading to questions about the validity of their estimates. In this paper, we look at the role of survey and non-survey data and their integration in a statistical system. The focus is on a suitable infrastructure and how that infrastructure can be used. Finally we consider some risks involved in combining survey and non-survey data for official statistics.

## RÉSUMÉ

Pendant des décennies, les données administratives telles les registres des naissances et données fiscales ont joué un rôle prépondérant dans les statistiques officielles. Ces dernières années pourtant, on a redoublé d'efforts pour exploiter d'autres données administratives et des données non issues d'enquête. Certains organismes de statistique ont adopté une politique de « priorité à l'administratif » : pour répondre à un besoin d'information, ils exploitent d'abord les sources de données existantes avant de se tourner vers l'option enquête en dernier recours. Peut-être ce changement se serait-il fait naturellement avec le temps, en raison notamment de pressions économiques. Mais il est aujourd'hui devenu essentiel, en raison de la diminution progressive des taux de réponse. Dans le cas de certaines enquêtes, ce déclin a été considérable, donnant lieu à des questions quant à la validité des estimations. Dans cet article, nous examinons le rôle des données obtenues par sondage ou d'autres moyens et leur intégration dans un système statistique. L'accent est mis sur l'infrastructure nécessaire et comment exploiter cette dernière. Enfin, nous examinons les risques que présente l'inclusion de telles données dans les statistiques officielles.

## 1. Introduction

The second half of the twentieth century has been viewed by some as the golden age of survey sampling (Singer (2016), Kalton (2019), Rao (2019), Gambino (2016)). In that period, in developed countries, social surveys on a broad range of topics proliferated. In Canada, the Labour Force Survey (LFS) was introduced at the end of World War II. Gradually, other surveys were added, sometimes in the form of supplements to the LFS. The 1990s saw the advent of longitudinal surveys in Canada, similar to developments in other countries. Over the years, collection methods evolved, from face-to-face interviewing to telephone interviewing to internet-based response, and in another dimension, from paper questionnaires to computer-assisted interviewing to web-based questionnaires. For a more detailed overview of the evolution of social surveys, see Gambino and Silva (2009).

The growth of sample surveys is related to their utility. They have become important inputs to policy and decision making and, in some cases, are used in the allocation of funds in multi-billion dollar programs. However, for a variety of reasons, not all well-understood, household surveys have been experiencing a gradual decline in response rates in the past decade or so (see Figure 1 for Canadian examples). For some surveys, the decline has been substantial, leading to concern about the validity of their estimates. In this paper, we will look at some options for dealing with the nonresponse problem and other problems with “official” social statistics. We highly recommend to anyone reading the current paper to also read Brick (2011), Couper (2013) and Citro (2014) who do an excellent job of assessing the current situation and looking forward. The paper by Groves (2011) is also relevant, particularly its last section. We will try to avoid duplication with these papers, but some overlap is inevitable.

---

<sup>1</sup> Jack G. Gambino, Statistics Canada (retired), Ottawa, Canada K1A 0T6, jack.gambino@gmail.com

<sup>2</sup> This paper is an updated version of Gambino (2016), which was presented at the International Chinese Statistical Association (ICSA) conference in December of that year and at Statistics Canada's Advisory Committee on Statistical Methods in October 2017.

The fall in response rates is just one part of a broader problem. Undercoverage of certain parts of the population, and in some cases, difficulty in reaching certain groups, is also a serious concern. In the early part of the “golden age”, the use of area frames for household surveys was common. In principle, area frames have complete coverage of the population. In practice, there are coverage problems due to hidden dwellings and people missed when a dwelling is enumerated, but these are small in comparison to the coverage issues we now face. Gradually, area frames were used less frequently for new surveys due to their very high cost compared to the use of address lists, telephone number lists and random digit dialing.

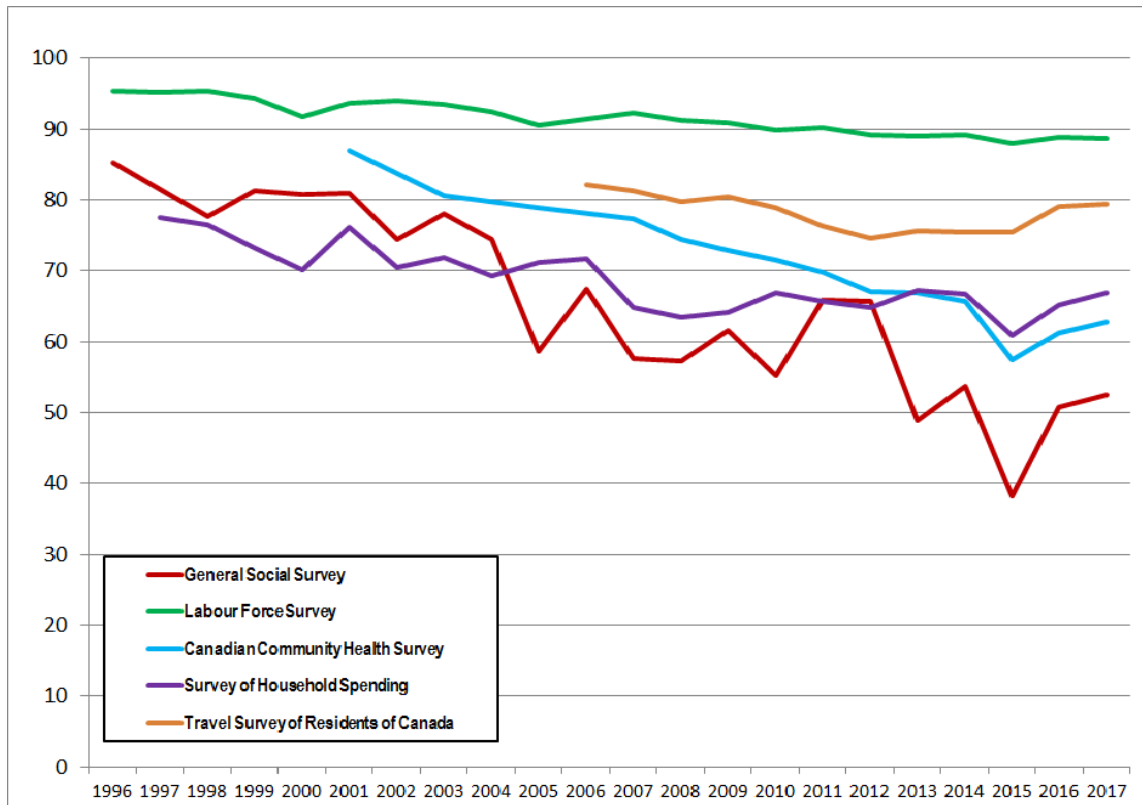


Figure 1: Response rates to some Canadian household surveys

Another part of the broader problem is budgetary constraints. For an ongoing survey, even if its overall budget remains stable, the amount of effort needed to obtain each response continues to increase. Therefore, to maintain a desired response rate, either efficiencies have to be found or some activities, such as certain quality assurance programs, have to be reduced or eliminated. If recent trends in cost per response continue, the increase in cost and effort needed to maintain current response rates will be unsustainable in the long term.

To deal with the situation, a number of initiatives are underway in Canada and around the world. Improvements in the management of data collection (responsive/adaptive design, better communication with potential respondents, etc.) are continuously sought. Options that were used sparingly by some national statistical organizations or offices (NSOs), such as the use of monetary incentives and making more surveys mandatory, are being reconsidered. However, we believe none of these will necessarily “solve” the problem. For example, there is evidence that making surveys mandatory and/or offering incentives will increase response rates, but there is no reason to expect these measures to reverse the general downward trend.

In this paper, we will look at possible directions for social statistics. We will begin by looking at the increasingly important role of non-survey data. Then we will look at where sample surveys may fit in the social statistics program of the future, and consider some possible methodologies. Our views are guided by a fundamental assumption: NSOs that have not done so already will develop a system of integrated databases (referred to as statistical registers by Wallgren and Wallgren (2014)) that will be maintained and updated frequently using administrative data. We will touch on some of the elements of this vision in sections 2, 3 and especially 4, but our focus will be on its implications rather than on how to attain it. Combining data from different sources is a key element of the developments we discuss. In section 5, we discuss some of the risks and

challenges this poses for official statistics. In section 6, we briefly discuss the role of statistical models in official statistics. We close with a brief summary of our key points.

## 2. Administrative and other non-survey data

Administrative data have always been used by statistical agencies, both as a direct source of information (e.g., vital statistics, crime data) and as complementary information for surveys (e.g., use of income data from tax files as an alternative to asking respondents questions on income). But sample surveys have been viewed as the preferred method for obtaining information on a wide variety of topics where administrative data either does not exist or is viewed as not quite suitable. That has begun to change. Increasingly, to meet an information need, the response to “conduct a survey” is viewed as a last resort to be considered only after exhausting other options such as finding relevant administrative data to (more or less) meet the need. This change is sometimes referred to as the “admin-first” approach, and is viewed as encompassing not just traditional administrative data but non-survey data in general.

To make this “admin-first” approach work, the NSO needs an appropriate infrastructure. Setting up this infrastructure involves not only changes within the statistical agency but also a concerted effort to collaborate with holders of non-survey data. It is also necessary to establish rules and protocols for the acquisition of external data, including methods to assess the quality and usefulness of such data for statistical purposes.

The use of non-survey data to replace a whole survey, or perhaps part of a survey, is obviously not its only possible use. New information (not currently collected) can be produced from such data as well. A systematic approach to assessing information needs and the data to meet those needs would involve listing, for each subject-matter area, the information needs and then determining the best way to satisfy them: survey data (existing, modified or new), non-survey data, or a combination of the two.

In an admin-first world, there is a need to be flexible: in some cases, we will have to live with data based on concepts that are not exactly what we want (e.g., the income categories from tax files may not correspond to the ones we would use in a survey questionnaire). This flexibility is needed not just on the part of the statistical agency but, more importantly, on the part of the external users of the information. We also need to recognize that certain variables can only be obtained from surveys (e.g., questions on physical activity, exercise habits, time use, feelings of satisfaction or depression). Finally, how the information will be used may be a deciding factor in choosing whether to go with a survey or to use non-survey data. For example, the Labour Force Survey unemployment rate is used to administer Canada’s employment insurance (EI) program. The administrative data that comes to mind as a possible alternative to the LFS is the file of EI beneficiaries, which is continuously updated. This data is interesting in its own right, but if the variable of interest is the unemployment rate, the EI data is not useful on its own: the correlation between being an EI beneficiary and being unemployed is not strong<sup>3</sup>. The only way to know who is truly unemployed (i.e., not employed and looking for a job) is to ask individuals. Thus there will always be a place for surveys.

## 3. Sample surveys

If we take the continued existence of sample surveys as a given, the obvious question is what they will look like in the future. In this section we consider some of the possibilities, keeping in mind the assumption that we are heading in the direction of an integrated system of databases (or statistical registers). A corollary to this assumption is that the integrated system can provide an infrastructure for sample surveys.

*Sampling households, sampling persons:* Statistics Canada is investigating the possibility of eventually replacing the “short-form” census, which is essentially a head count plus very basic information (age, sex, address), with a “virtual” or statistical person register created and maintained solely from administrative sources. An early test version of this database looked very promising, and work on developing a true virtual register, called the Statistical Population Register (SPR), is underway. Its successful implementation would put Canada in a position similar to that in countries with true population registers, and then the infrastructure discussed in Wallgren and Wallgren (2014) would become a realistic option. By linking the SPR to other databases (such as telephone numbers), new sampling options appear. In particular, it becomes more feasible to sample persons directly, rather than go via the dwelling (ultimately the household) as is currently the norm in Canada. This approach can be especially effective if the SPR is linked to socio-economic information that can then be used to stratify the population.

---

<sup>3</sup> An unemployed person may not be eligible for EI and an EI beneficiary may be employed.

For some surveys, there will continue to be advantages to sampling dwellings rather than people directly. A major one is that dwellings are fixed and, from a representativity perspective, it doesn't matter if the occupants of the dwelling have moved and been replaced by new occupants. Selecting people directly may require tracing movers to ensure that the sample properly represents the population. Tracing can be expensive, especially if field operations are involved. A special case where sampling individuals is likely to be preferable is for special populations such as children, seniors, diabetics and so on. Unless we have a good idea about which dwellings contain members of the special population, a large oversample, with many "wasted" dwellings, has to be selected to get enough respondents who belong to the special population. In each case, an analysis of costs (contact, collection, tracing) and benefits (effective sample size) would be needed to choose the most cost-effective option.

There are examples where the dwelling approach may continue to make more sense. Some surveys, such as the Canadian LFS, have (i) a relatively short questionnaire and (ii) allow proxy response. In this situation, we get information on all household members very inexpensively (keeping in mind the implicit assumption that any proxy bias is small). For some surveys, even if we are interested in just one member of the household, it is desirable (and possibly essential) to collect basic information about all members of the household because it provides useful covariates. Finally, certain surveys involve concepts that are at the household level by their very nature. These include household expenditure surveys and wealth (assets and debts) surveys.

*Questionnaires:* It is generally acknowledged that questionnaires used in surveys conducted by NSOs are too long. When the transition from mostly in-person interviewing to mostly telephone interviewing occurred, in many cases, questionnaires were not shortened (or not shortened enough) to reflect the change in mode. Some surveys conducted over the phone by Statistics Canada can be up to one hour long. The increased use of cell phones exacerbates the problem. Moving forward, with the increased use of the internet as a response medium, on the one hand, questionnaire length may be less of an issue if the respondent is using a desktop or laptop computer, and on the other hand, is likely to be a serious problem if the respondent is using a smartphone.

There is, therefore, a clear need for shorter questionnaires. One approach that has been studied in the literature is the matrix survey sampling or split questionnaire approach (see Merkouris (2015) and Chipperfield and Steel (2009)). In this approach, all respondents get a common core of questions (module 1), and then different modules are administered to different subsets of respondents. Some respondents get modules 1 and 2, say, some get modules, 1, 3 and 5, and so on. A complete data set (or its equivalent) can then be derived using models (explicit or implicit) that exploit relationships between core variables and variables from the modules.

This approach can be extended by adding non-survey data to the mix. Ideally, if the information in some of the modules is available from administrative sources, the result is not only a shorter questionnaire, but possibly better-quality data (assuming the administrative data are based on suitable concepts and definitions). However, combining data from different sources has risks and challenges, as discussed in section 5.

*Collection modes:* Related to the questionnaire is the mode of collection. Many NSOs are in the midst of a push towards internet-based collection for social surveys. Two major challenges in this push are, first, that many people (particularly young people) access the internet primarily using their smartphone and, second, a non-ignorable percentage of people either own just a smartphone (and no computer or tablet) or own none of the three devices<sup>4</sup>. For the internet option to have maximum penetration, a way must be found to make surveys work well on a variety of devices. The alternative is to lump people with only smartphones with those who have no internet access, and collect their information using other modes.

#### **4. Methodologies for the future**

Before looking at some possible methodologies, we elaborate on infrastructure. The "ideal" infrastructure, or something close to it, already exists in some European countries (see Wallgren and Wallgen (2014)) but not in Canada. In an ideal situation, we would have the following: For each person, there is a permanently-associated unique identifier (UID) and every UID is linked to an address. When a person moves, the UID-address link is updated promptly. With this setup, we

---

<sup>4</sup> Up to date statistics for the U.S. are published by Pew Research ([www.pewinternet.org](http://www.pewinternet.org)). In early 2019, 81% of Americans owned a smartphone, but ownership is correlated with education and income. Almost half of people aged 65+ do not have a smartphone. For a detailed analysis, see Anderson (2019). Also, 10% of Americans do not use the internet (see Anderson et al. (2019)).

can sample dwellings (residential addresses) directly or persons directly. The need for something like a UID is well-recognized and, in fact, its development has been studied at Statistics Canada.

*Methodology I:* The first methodology for the future we will discuss is actually one from the past, namely, address-based sampling (ABS). We assume that all dwellings have a civic address and a mailing address; if the two are different, we assume they are linked. Note that this is currently a strong assumption for Canada because it does not hold for a small but non-ignorable part of the dwelling population. However, there are efforts to decrease the size of this problem, and we assume that it will become ignorable in the future. Then address-based sampling (see Iannacchione (2011), for example) becomes feasible for all surveys. Though ABS is not new, it has become very popular in recent years, partly due to the decline in the efficacy of random digit dialing (itself due to rapidly declining landline usage and increased call screening). We mention two implementations of ABS: mail-first and telephone-first.

#### Mail-first survey

- All sampled dwellings are contacted by mail
- The respondent is asked to complete a paper questionnaire (and may be given the option to complete it online instead)
- For nonresponse follow-up, dwellings with an associated phone number are contacted by phone
  - For telephone nonrespondents, a subset are visited in person
- For nonresponse follow-up, dwellings without a phone number (or a subset thereof) are visited in person

#### Telephone-first survey

- An introductory letter may be sent to sampled units (continuing a current practice at Statistics Canada)
- Dwellings with an associated phone number are contacted by phone for an interview (and may be given the option to respond online instead)
- Dwellings without a phone number (or a subset thereof) are visited in person
- For telephone nonrespondents, a subset are visited in person

In both the mail and telephone approaches, we want to emphasize *the importance of following up a subset of nonrespondents* in person. This is the best (and, in some cases, the only) way to deal with nonresponse bias. The hope is that the cost of this follow-up can be offset by the savings that accrue from the use of inexpensive collection modes (internet, telephone) for a large portion of the sample.

Under the ideal infrastructure mentioned above, assuming there are good links to rich data sources (such as tax, health and education data), this information can help us deal with nonresponse, at least for certain variables. For example, if there are linked variables that are highly correlated with some of the survey variables of interest, this can be used to improve the imputation method or nonresponse adjustment. However, it is also possible that there are no such external variables, or worse, that the survey variables of interest are related to the propensity to respond. Particularly for the case where suitable external variables don't exist *and* the survey variables of interest are related to the propensity to respond, we emphasize once again the need to follow up at least some nonrespondents, even if this is expensive.

*Methodology II:* We consider a panel-based, parametrized modular framework for household surveys<sup>5</sup>.

To illustrate this framework, assume that a very large probability sample is given a short but mandatory survey questionnaire that collects basic demographic information (and perhaps a handful of key variables such as overall income and highest level of education). Given the mandatory nature of this short survey, we may decide to not offer monetary incentives to the sampled units (though arguments can be made for the opposite as well). This establishes a panel.

Once the panel is established, subsets of this big sample get different modules (questionnaires) over time. Depending on factors such as content and response burden, some modules may have an associated monetary incentive. Alternatively, borrowing an idea from online panels used in the private sector, it may be preferable to always give a monetary incentive

---

<sup>5</sup> The purpose of having modules here is different from the split questionnaire approach we mentioned in section 3. Here, we do not necessarily want to create a complete data set, covering all possible modules, for all respondents.

for these modules, with the amount varying according to content and burden. In any case, completion of most (perhaps all) of these modules would be voluntary<sup>6</sup>.

We assume that people (or households) will stay in the panel for a long time (measured in years). To maintain representativity, the panel would have to be replenished periodically (at least annually). Therefore it is natural to implement a sample rotation scheme. For example, to keep units in the sample for two years, we can rotate one-eighth of the sample each quarter. For a longer time frame, we can use an annual rotation, such as replacing one-quarter of the sample every year (in which case each unit would be in the panel for four years).

In the past, the above approach would have been prohibitively expensive. However, if we assume that the vast majority of data collection will be done over the internet, it may now (or soon) be viable. The obvious “big ticket” parts of this approach are (1) the cost of the monetary incentives, (2) collecting data from units that don’t have internet access and (3) nonresponse follow-up. We view a proper *nonresponse follow-up program* including telephone and in-person components as essential to making this approach retain its validity. Currently, (2) is also essential because it involves a non-trivial proportion of the population. The greatest leeway may be in (1) since we can choose to not (or rarely) offer incentives.

Finally we note that a large part of the current household survey program can be made to fit into this framework. This includes major surveys such as the LFS and the Canadian Community Health Survey (CCHS). In this respect, there are similarities with the master sample idea that was studied by Statistics Canada a decade ago and ultimately rejected (see Tambay et al. (2009)). What has changed since then is the feasibility of using the internet for the bulk of data collection. This change, combined with the implementation of the infrastructure vision mentioned above, makes this model much more viable than it was in the past.

*Methodology III:* Given the increasing amount and availability of non-survey data, including big data, it becomes important to explore its use for statistical purposes. One possibility is to use such data as a source of imputation variables to fill gaps (unintended or intended) in survey data. The sample matching approach due to Rivers (2007) has been studied as one way to achieve this. Though Rivers focused on the use of web panel data as the external data source (possibly from a self-selected, nonrepresentative set of respondents), the idea has wider applicability. Therefore, in this section, rather than referring to “web-based” or “administrative” data, we will use the term “external” data. More precisely, external data here refers to any data not collected by the survey itself and that is not on the sampling frame.

In our context, we would start with a proper probability sample and then find “donors” in an external data file to complete the data set. The basic idea is to find units in the external source that are close in some sense to the units in the sample. Note that the goal is *not* to find the same unit (person or household) in both files, but to find a similar unit. The machinery to do this, such as nearest-neighbour imputation, already exists and is relatively easy to adapt for this purpose. The challenge is to find variables on both the original sample file and the external sources that can be used to do the matching in a useful way. One option is to use the panel approach (Methodology II) to begin, since its first step involves administering a basic questionnaire that can be designed to include variables that are useful for sample matching.

Ultimately the usefulness of the sample matching method will depend not only on the availability of variables for matching, but more importantly on their relationship with the variables that the survey sponsor is most interested in. The approach has been tested at Statistics Canada with mixed results (Chatrchi et al. (2018)). They applied the method to health data obtained from web panel respondents (and matched to a probability sample) and compared the resulting estimates to estimates for the same variables from the Canadian Community Health Survey. The quality of the “imputed” estimates was not of sufficiently good quality to be useful, possibly because the matching variables were not sufficiently informative of the outcomes. Nevertheless, because this work is in its early stages, it is worth exploring further. While the method may not be useful for some variables, further research might show that it is worthwhile for others.

*Combining the methodologies:* The three methodologies can be combined into a global model for social surveys. The address/mail infrastructure in Methodology I can be used to reach sampled units, whether they be dwellings (selected using Address Based Sampling) or persons (selected via the household or directly). These units then form a panel (or perhaps panels), as in Methodology II. Then the sample matching methodology in Methodology III can be used to complement survey data with data from non-survey sources, at least for the variables where it has been demonstrated that the method

---

<sup>6</sup> Therefore the parameters for a module are incentives (yes/no) and voluntary/mandatory.

works well. In practice, we are likely to implement elements of each methodology to a lesser or greater extent, depending on factors such as cost, quality of the results and statistical efficiency.

## 5. Combining data: Beware of black boxes

Combining data from different sources is at the heart of some of the methods being considered to supplement, replace or improve sample surveys. Data from two sources, say A and B, can be combined in several ways: record linkage (where the goal is to find the *same* unit in A and B), statistical matching and imputation (where the goal is to find *similar* units) and modelling (where a statistical model is used to “predict” values; e.g., data set B is used to fit a model  $y = f(x)$  which is then applied to  $x$  values in A to predict  $y$  values). In this section, we discuss some of the challenges and risks *for official statistics* when combined data are used to produce information. By official statistics we mean information that is produced by an NSO for policy makers, decision makers, researchers and the public (see United Nations (2014)). We argue below that the criteria for *acceptability* or *fitness for use* for official statistics are not the same as those for other endeavours such as research on relationships among variables.

When record linkage between data sets is successful, the benefit for official statistics is that we obtain a data set with a richer set of variables for the matched units. We do not discuss record linkage further in this paper. Rather, we focus on the risk of making invalid inferences when the other methods mentioned in the previous paragraph are used. The risk is illustrated by considering the simplest possible example, simple linear regression. Consider a population for which there are two variables  $x$  and  $y$  for each person. We assume  $x$  is available for everyone. The  $y$  values are not available but we have a fitted model  $\hat{y} = a + bx$  from some other source. Every statistics student learns that naively using the  $\hat{y}$  values from a fitted regression line to “impute”  $y$  values will result in inappropriate conclusions, even if the (true) model  $y = \alpha + \beta x + \varepsilon$  is valid and  $(a, b)$  are excellent estimates of  $(\alpha, \beta)$ . This is true even if the correlation between  $x$  and  $y$  is good, e.g., in the range 0.5-0.8. While the estimate of the mean of  $y$  will be good, other quantities of interest, such as the first and fifth quintile of  $y$ , will be wrong (they will be closer to the mean than they should be). This is because the variance of the predicted values is too small by a factor  $\rho^2$ , where  $\rho$  is the correlation coefficient. In this particular example, and under ideal conditions (e.g., knowledge of the population variance of  $y$ ), we can fix the problem by adding just the right amount of noise to the predicted values to make their distribution the same as the population distribution.

More generally, there is a risk that the distribution of modelled or imputed values will be quite different from the distribution of the variables of interest in the population. The risk is greater, of course, when the modelling mechanism is a black box.

Our simple regression example can be used to illustrate the main point of this section, namely that the criteria for acceptability or fitness for use for official statistics are not the same as those in other domains. In the regression example, we noted that even in the case of good correlations between variables, estimates of population parameters other than the mean will be incorrect. For official statistics, a  $\rho$  of 0.5 is not very useful to predict  $y$  (e.g., income) from  $x$ , unless we are in the unlikely situation where we can confirm that the distribution of the predicted values is close to that of the population values. But *this same result* can be important in other contexts because it tells us that 25 percent of the variation in  $y$  is explained by  $x$ . For example, if  $y$  is a child’s intelligence and  $x$  is a diet-related variable for the mother during pregnancy, then this would be a major finding if it were to be replicated!

Finally, we cannot discuss the role of non-survey data without mentioning “big data” (transaction data, tweets, Google searches and so on). The analysis of big data has produced interesting results (and it is likely that the most useful results are not known to us because they are proprietary). A recent example that received media attention is the analysis of a massive amount of Spotify data (Park et al. (2019)). The authors selected “a stratified random sample of one million worldwide Spotify users, matching each country’s age and gender distribution” and found interesting patterns along different dimensions (time of day, demography, culture, etc.). We mention this example to emphasize our final point: the analysis of big data does indeed produce very interesting and useful information, but they are not official statistics. This does not imply that such data has no role in official statistics, only that its role still needs to be worked out. The recent post by Groves (2019) has some relevant insights for this discussion. He highlights some of the differences between planned data (such as most official statistics) and “data produced by others” and encourages data scientists to devote more attention to the creation process for the latter.

## 6. The use of statistical models

We have not addressed statistical inference or the use of models, except peripherally in the previous section. Over the years, the use of models in official statistics has increased significantly, as evidenced by their use in weighting, imputation and small area estimation. In this paper we have described a world where the role of non-survey data is much more prominent than it has been thus far. In such an environment, where the relationships among variables from different sources are so important, the need for model-based methods is certain to grow. A recent view, from the perspective of an advocate of a model-based approach, is given by Chambers (2014). The increasing importance of models is recognized by other statisticians such as Groves (2016), who discusses their use in the context of combining data from different sources.

Along with the greater role for model-based approaches, we expect the use of Bayesian methods to increase as well. In the past, even though repeated surveys such as the LFS had a wealth of information that could have been used to develop objective priors for estimation, the use of Bayesian methods in official statistics was relatively modest, with small area estimation being a notable exception. In the environment envisaged in the current paper, with data coming from numerous survey and non-survey sources, we expect that the implementation of Bayesian approaches, such as the calibrated Bayes approach described by Little (2012), will become feasible and perhaps necessary.

Increasing the use of models, whether in a Bayesian or non-Bayesian context, will require a great deal of work such as developing, or adapting, model building and validation tools for the complex environment we envisage. It should include research into covariates that will make the models useful. This work is imperative if we are to successfully exploit the vast array of data that will be at our disposal.

## 7. Summary

In this closing section, we summarize some key points made in this paper.

We emphasize the importance of a suitable infrastructure that will allow us to conduct surveys efficiently and also combine data from both surveys and other sources. Under this infrastructure, we will be able to move seamlessly between households and persons, addresses and telephone numbers, and among collection modes.

In an “admin-first” world, there is a need to be flexible and willing to compromise. In some cases, we will have to live with data based on concepts that are not exactly the same as what we would want ideally. Furthermore, there are certain variables that can only be obtained directly from individuals, i.e., there is no prospect for an “admin-only” world.

With an improved infrastructure, the feasibility of sampling persons directly will increase significantly. But there will still be cases where it is either more cost-efficient or necessary, due to subject matter, to sample households.

Questionnaires for some of our major surveys are too long and burdensome. Alternative solutions need to be investigated. This research will take place in the context of the increasing use of internet-based collection, which has its own challenges (coverage, variety of devices).

Some methodologies for social surveys under the assumed infrastructure are presented in this paper. Elements of these methodologies can be combined in different ways, but before deciding on which ways work well in practice, studies, including field tests, need to be conducted.

To some extent, the rethinking of how social statistics are produced has been inspired by concerns about decreasing response rates for surveys. The infrastructure we have assumed will alleviate some of these concerns, both directly (by providing other sources of data) and indirectly (by improving sampling, collection and estimation). However, good response rates, and an understanding of nonresponse biases, continue to be important. We emphasize the need for a nonresponse follow-up program for surveys to ensure that the validity of survey results is not jeopardized.

We recognize the value of non-survey data, including big data, and of combining survey and non-survey data. We also recognize the usefulness of increasingly sophisticated computer-based statistical tools, such as machine learning algorithms. However, we also need to be aware of the risks involved, particularly when methods are not clearly defined or well-understood. Methods and models need to be understood and validated. We emphasize that the criteria for what is suitable

or acceptable depend on the intended use of the results. The criteria for official statistics are different from those of other types of research because its products are not put to the same use.

### Acknowledgements

I am grateful to François Brisebois and Andrew Brennan whose comments and suggestions improved the readability of the paper.

### Disclaimer

The content of this article represents the views of the author and does not necessarily represent those of Statistics Canada.

### References

Anderson, M. (2019). Mobile technology and home broadband 2019. Article from Pew Research: <https://www.pewinternet.org/2019/06/13/mobile-technology-and-home-broadband-2019/>

Anderson, M., A. Perrin, J. Jiang and M. Kumar (2019). 10% of Americans don't use the internet. Who are they? Article from Pew Research: <https://www.pewresearch.org/fact-tank/2019/04/22/some-americans-dont-use-the-internet-who-are-they/>

Brick, J.M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75, 872-888.

Chambers, R. (2014). Survey Sampling In Official Statistics - Some Thoughts on Directions. *Proceedings of Statistics Canada Symposium 2014*, Ottawa, Canada.

Chatrchi, G., J.-F. Beaumont, J. Gambino and D. Haziza (2018). An investigation into the use of sample matching for combining data from probability and non-probability samples. Presented at the Statistical Society of Canada annual meeting, Montreal.

Chipperfield, J.O., and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics*, 25, 227-244.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, 137-161.

Couper, M.P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, 7, 145-156.

Gambino, J. (2016). Thoughts on the future of household surveys at Statistics Canada. Internal document, Statistics Canada, April 13, 2016. Presented at the International Chinese Statistical Association conference in December 2016 and at Statistics Canada's Advisory Committee on Statistical Methods in October 2017.

Gambino, J.G. and Silva, P.L.d.N. (2009). Sampling and Estimation in Household Surveys, in C.R. Rao and D. Pfeffermann (Eds.), *Sample Surveys: Design, Methods and Applications*, Vol. 29A, Elsevier.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75, 861-871.

Groves, R.M. (2016). Towards a Quality Framework for Blends of Designed and Organic Data. Waksberg Award presentation at Statistics Canada's 2016 Methodology Symposium.

Groves, R.M. (2019). Derivative inquiry: Danger facing fields that use data without producing data. Posted February 13, 2019 at [blog.provost.georgetown.edu](http://blog.provost.georgetown.edu).

Iannacchione, V.G. (2011). The changing role of address-based sampling in survey research. *Public Opinion Quarterly*, 75, 556-575.

- Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87, S10-S30 (special issue).
- Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics*, 28, 309-372.
- Merkouris, T. (2015). An efficient estimation method for matrix survey sampling. *Survey Methodology*, 41, 237-262.
- Park M., J. Thom, S. Mennicken, H. Cramer and M. Macy (2019). Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour*, 3, 230-236.
- Rao, J.N.K. (2019). On making valid inferences when combining data from surveys and other sources. *Sankhya* (in press; special issue).
- Singer, E. (2016). Reflections on surveys' past and future. *Journal of Survey Statistics and Methodology*, 4, 463–475.
- Tambay, J.L., Laflamme, G. and Gambino, J. (2009). The Canadian Experience in Creating a Master Sample. *Proceedings of the 57th Session of the International Statistical Institute*, Durban, South Africa.
- Rivers, D. (2007). Sampling for web surveys. In the *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*, American Statistical Association, Alexandria, VA.
- United Nations (2014). *Fundamental Principles of Official Statistics*. Adopted by the UN General Assembly on March 3, 2014.
- Wallgren, A. and Wallgren, B. (2014). *Register-based statistics: statistical methods for administrative data*. Second edition. John Wiley and Sons.