Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Admin-First as a Statistical Paradigm for Canadian Official Statistics: Meaning, Challenges and Opportunities[1]

Eric Rancourt[2]

Abstract

For decades, National Statistical Offices have claimed that they intend to use more and more administrative data; and they have actually used them to various degrees from one program to another. But with the advent of the data revolution, it is no longer a wish, a side issue, a marginal method, or an increasing trend; it has become the central focus of attention for the future of programs. Whether the objective is to enhance relevance, reduce response burden, increase efficiency, or produce faster with more details, the use of administrative data (in the broadest sense) is proliferating within and without statistical systems at a sky-rocketing pace. Statistics Canada is facing the new data world by modernizing itself and embracing an admin-first paradigm. This paper attempts to explain what this means from the statistical perspective, to highlight some of the theoretical challenges and to point out possible related opportunities. There are also legislative issues as well as important considerations of the elements of the social license such as privacy and respondent burden, but they are not the focus of this paper.

Key Words: Deductive Approach, Inductive Approach, Statistical Inference, Theoretical Framework.

1. Context

The World is always changing and National Statistical Offices (NSOs) being in the business of measuring society must adapt. In recent years, not only the socio-economic landscape has been evolving, but also - and actually more so - the data landscape. Today data are everywhere from the workplace, the marketplace, the air (sensors and satellites images), at home (computers, security systems, appliances, etc.), and even on and in our bodies (smart watches, pacemakers, DNA, etc.). In a nutshell, to be relevant, statistical systems and their associated methods must constantly be re-thought and adapted to ensure that society has at its disposal the best information to make sound decisions. Today's statistical paradigm cannot be the same as it was a few decades ago.

1.1 Historical context

Official statisticians have used both primary collection (a collection mechanism that the NSO controls, chiefly surveys) and secondary collection (obtaining data collected by another entity, e.g. administrative data) methods since the beginning of Official Statistics production. Before the Second World War, attempts at statistical inference capacity in the context official statistics essentially meant attempting with as much discipline as possible to get data from all units by means of a census or by obtaining transaction files (e.g. vital statistics) from related entities. Sometimes pseudo-censuses were carried out when the largest units of a population covered a percentage that was deemed large enough (e.g. business revenues of manufactures).

[1] The content of this paper presents theoretical views that could evolve before being implemented in a statistical program.

[2] Eric Rancourt, Director General of Methodology, Statistics Canada. 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. Eric.rancourt@canada.ca.

In 1934, a seminal paper by Jersey Neyman (Neyman, 1934) provided the foundation of a valid statistical inference approach known today as survey sampling theory. This approach rapidly proliferated throughout the World after the Second World War. In Canada, the Labour Force Survey started using sampling in 1945 and eventually almost all statistical programs became survey-based. This is the situation that has mostly prevailed until now. Increasing the use of administrative data in statistical programs has been a strong priority in the past and recently (Statistics Canada, 1984; 2012; 2018).

1.2 Current context

Nowadays, data are everywhere. With the Data Revolution (see for example United Nations, 2014), the context for official statistics and statistical programs has completely changed. The appetite for data has immensely increased and the variety of data being produced (whether they are accessible or not by a third party) has snowballed. So as decision-makers, businesses and people want to have information more timely and free, there are expectations that citizens would have to be even less involved/burdened in the provision of data and that the data are well protected. Hence the need to re-think the current approach for producing official statistics. So is the survey sampling theory still going to be appropriate and useful? Very much. But is the survey sampling theory sufficient to answer new needs? Not at all. Then, how are new needs going to be supported by theory? In all likelihoods by expanding existing parts of the theory and/or by developing new theories.

Recently (Arora, 2018), Statistics Canada has embarked on a modernization journey that is indeed aimed at better positioning the National Statistical Office to answer new statistical demands in the context of the modern data world. The strategy is built along five pillars: User-centric delivery service; Leading-edge methods and data integration; Statistical capacity building and leadership; Sharing and collaborating; Modern workforce and flexible workplace.

In terms of the overarching approach used to support statistical programs, an admin-first approach to using and/or collecting information has been identified to be the working statistical paradigm.

2. An admin-first paradigm: What does this mean?

This section defines the concept of the admin-first paradigm and presents concepts and terms that are needed to provide meaning to the approach.

Survey: A survey is a tool designed by the NSO to carry out primary collection activities aimed at obtaining information on clearly defined concepts. It is assumed here that the survey would be designed, conducted and used to enable valid statistical inference under the survey sampling theory. And as such, surveys are normally probabilistic in nature. Other tools to carry out primary collection could be a census, focus groups, mobile applications, electronic devices, etc.

Administrative data: Data that one uses to administer a program / organization. It is understood in the broadest sense in that it could be data from government departments, but also from any other organization whether public or private. As these data are not produced for inference (but rather to manage a program, offer services, or regulate), their design is almost always non-probabilistic. For the conceptual discussion of a data framework, we assume that administrative data are used in a secondary purpose, in opposition to data obtained through primary collection activities.

Survey-only paradigm: A framework whereby the inference to a population about concepts is made from and only from a survey.

Survey-first paradigm: A framework whereby the inference to a population about concepts is made from a survey, aided in various ways by administrative data. Ways in which administrative data can improve the survey are many (see for example Brackstone, 1987 or Beaumont, 2018). In terms of design, the assumption in this paradigm is that a survey is first considered and then administrative data are used to improve it. In terms of inference, it could be with respect to the design, assisted or not by a model, or to a model.

Admin-only paradigm: A framework whereby the inference to a population about concepts is made from administrative data only. It could be that multiple sets of administrative data be required.

Admin-first paradigm: A framework whereby the inference to a population about concepts is made from administrative data, aided in various ways by surveys (or more generally by primary collection activities). In other words, before going ahead and carrying out surveys, all reasonable attempts are made to try to make use to the fullest of already-existing information that can be made available to / accessible by the NSO. In terms of design, the assumption under this paradigm is that administrative data are first considered, and then a survey may or may not be conducted to complement the administrative data in terms of scope or to evaluate quality. In terms of inference, it is likely that it be with respect to a model, but the choice of this paradigm does not necessarily imply a sequential choice of administrative data and then survey data. As the design can be developed by jointly considering both, inference could also be designed-based in some cases. An example of this paradigm is in New Zealand's quarterly business statistics (Liken et al, 2018).

3. Some challenges and opportunities with an admin-first paradigm

In the context of production of official statistics, work normally starts with the expression of a data need by decision-makers and/or society to the NSO. Citro (2014) takes such a viewpoint about the role of NSO's statistical program. Under the survey-first paradigm, this is by definition the approach taken. The admin-first paradigm takes a more general approach and offers new opportunities in that it now becomes possible to explore existing data to find out what could be accessed/used. However, the use of data is supposed to stem from an identified purpose hence the challenge of expanding data possibilities while respecting the purpose.

Administrative data are defined with concepts that are appropriate and/or specific to the program for which they were produced. Such concepts could 1) be different from the targeted concepts in the statistical program, and 2) be different from concepts used through surveys. There is then the challenge of conciliating such discrepancies. As the variety of data sources increases, the likelihood of encountering different concepts increases as well. Perhaps the use of modelling or other techniques can be used to account for differences but at least discrepancies will have to be acknowledged and ideally measured.

The complete compendium of all existing data file in a country may not be something that exists. As a result, designing a statistical program under the admin-first paradigm presupposes the awareness of the existence of data potential data sources that could be related to a given statistical program. Further, even with full awareness of the data holdings, one then has to determine / develop protocols for accessing the data.

Once data can be accessed, their structure, nature and quality should be assessed. As administrative data are far from perfect, one is faced with the problem of understanding the messiness (see for instance Baker et al., 2013 about the limits of administrative data) and put in place mechanisms to cope with these. Meng (2018) provided a striking example of how non-probability data of apparently high quality could in fact lead to incorrect conclusions if one is not careful. Hence the importance of defining and adopting a total quality framework that will adequately inform managers of statistical programs and users of the information they produce. But more completely, there is the need to ensure that estimates be produced under an explicit theoretical framework that enable valid statistical inference.

1. Admin-first paradigm → Data → Tools → Conclusions

2. Survey-first paradigm → Concept → Tools → Inference

→ Need to reconcile concepts & inference

With a survey-first paradigm, the inference is deductive, in that first, a concept to measure is identified and then tools are built to measure it. Inference is drawn from the resulting data up to the population through a clear survey sampling theory. In the context of the admin-first paradigm, sometimes the same type of inference is desired and combining administrative and survey data will allow for the deductive approach. However, some people are interested in starting with administrative data, building tools and trying to generate conclusions through an inductive approach. This is not wrong in itself, but the difficulty lies in trying to combine this approach with traditional survey-based approaches which are more deductive. This poses two significant challenges. Firstly, if the resulting information / estimates are produced by both approaches, then the challenge is to ensure that appropriate signals and explanations are provided

to users for them to understand the inference context and be able to draw the right conclusions. Secondly, there is a challenge in the integration of the two approaches in order to maximize the joint power of the survey and administrative data. For this, a more complete and /or general theoretical framework is needed.

4. Progress in expanding and defining the theoretical framework

Much progress has been made on how to more fully use administrative data for statistical programs in the last few decades. Since the early days of the survey sampling theory, the context has expanded. Further, methods to estimate survey quality and data quality in general have proliferated and the statistical community has made strides towards a total quality framework.

In terms of theory, just to name a few advances, one could think of Särndal, Swanson and Wretman (1992) who have provided a comprehensive approach to using administrative data through the model-assisted approach. Considering data integration Zhang (2012) advances the theory in the context of registers and Lohr and Raghunathan (2017) present how to combine surveys with other data sources. Looking at the problem of using administrative data (or auxiliary information) through models for small areas, Rao and Molina (2015) provide a comprehensive account of small area estimation approaches. Then considering how to combine probabilistic and non-probabilistic sample together, Beaumont (2018) provide a solid review of existing practices.

In terms of quality framework, much has been developed and Beamer (2016) describes the theory and practice of a total survey error paradigm.

Several other authors have contributed to the advancement of the use of administrative data for valid inference (for example Mercer et al. (2017)) and the estimation of multiple dimensions of quality (for example Bosa et al. (2018) for the non-response variance; Pankowska et al. (2018) for measurement error), but the objective here is simply to provide the general context that leads to attempting to specify the elements of what could form a theoretical framework for the admin-first paradigm.

5. Attempting to specify elements of a theoretical framework

This section succinctly presents elements that could be used jointly to constitute a skeleton of what could be called a "socio-statistical theory" for inference under an admin-first paradigm. This is a modest attempt to unify the survey-first and admin-first statistical paradigms with data stewardship activities and elements such as data access and statistical registers. Further, the framework could also provide a context to measure quality.

If the following nine assumptions were found to be valid in a specific context, then one would be positioned to make "perfect" estimates. In practice, there will inevitably be departures from these assumptions to a certain extent. Inference must then be made by taking into account the measurement of the departure from these assumptions. Implicitly, this provides a structure coherent with that of Biemer (2016) for assessing the quality of the data and Berka et al. (2011) provides the Austrian example of a quality framework under administrative data.

<u>Elements of a Socio-statistical theory:</u>

Assumption 1: Data for all units exist in digital form. Assuming that people/businesses will find a compelling reason for at least one of the programs that exist in both the private and public sectors, they will adhere to it perhaps along the lines of the utility theory of von Newmann and Morgenstern (1953). Otherwise, some sensor or satellite or other device, including social media would catch the elements of the target population.

Assumption 2: There exist complete frames / registers available to the NSO. Whether the population is made of businesses, people, farms, animals, buildings, etc., having a complete list enables one to ensure complete coverage.

Assumption 3: The NSO either has access to the all the data through a mechanism that it controls (e.g. probabilistic sampling), or it is able and allowed to gather the needed information for valid statistical inference. It can be in the form of a single file or multiple file and / or sources.

Assumption 4: All units can be matched without error. This enables the avoidance of duplicates and full knowledge of the coverage.

Assumption 5: Data concepts equal target concepts, meaning that no conceptual discrepancy exist either between the collected data and the target objective, or between data sources, or between administrative data and survey data.

Assumption 6: There is no non-response error.

Assumption 7: There is no measurement error.

Assumption 8: There exists strong links between variables. In other words, models can be built for estimation or analysis purposes.

Assumption 9: Standard errors are small and/or the quantity of information is large enough to allow for precise inference.

From these assumptions, one could draw a number of corollaries. Some of them are:

Corollary 1: Registers should be built and maintained in real-time (or sufficiently frequently to answer the needs of statistical programs as they surface).

Corollary 2: The NSO should aim at getting access to all data.

Corollary 3: Under a fully implemented admin-first paradigm, primary collection should be at the service of evaluating uncertainties / departures from all the above assumptions.

Corollary 4: Improving matching quality and measuring its errors are central to inference.

In fact, attempting to ensure the smallest departure from each of the nine assumptions and measuring how large the departures are could all constitute a corollary.

6. Conclusion

We have presented what is meant by an admin-first paradigm for statistical programs. This paradigm provides numerous possibilities for the production of information and is supported by a number of statistical methods with inferential capacity. We presented challenges and opportunities as well as elements of a theoretical framework, but there is still much scope to expand the theoretical framework to include a broader range of possibilities. For example, one could conceive of using administrative data to build priors and then design surveys to be optimal for likelihood estimation given this prior. Early work by Rao and Ghangurde (1972) can be related to this. Conversely, perhaps priors could be estimated from small quick surveys complemented by comprehensive administrative data. Issues of selection bias and impact of the size of variances would have to be carefully considered. Similarly, surveys could be designed to be at the service of filling data gaps present in administrative data and conceived optimally to measure one or many quality aspects of the administrative data. It could also be that a complete theory may not exist or that it may not be probabilistic. In this paper we have (almost) only considered quantitative approaches. One could / should also go farther and attempt to mix the use of qualitative data with quantitative data through mixed methods approaches (see for instance Poth, 2018).

Acknowledgements

The author would like to thank Martin Beaulieu, Jean-François Beaumont, Linda Howatson-Leo, Andrea-Leigh MacMillan, J.N.K. Rao and Siu-Ming Tam for their comments and insightful inputs to this paper.

References

Arora, A. (2018). *Modernizing the National Statistical System – Stakeholder Consultations*. Catalogue 89200003. Statistics Canada. ISBN 978-0-660-31580-5.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K., and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, pp. 90-143.

Beaumont, J.F. (2018). *Les enquêtes probabilistes sont-elles vouées à disparaitre pour la production de statistiques officielles*. Presented at the 10th Colloque francophone sur les sondages, Lyon, France.

Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H., Schwerer, E (2011). A quality framework for statistics based on administrative data sources using the example of the Austrian census 2011. *Austrian Journal of Statistics*, Volume 39 (2010), Number 4, pp. 299-308.

Biemer, P.P. (2016). Total survey error paradigm: Theory and practice. In *the SAGE Handbook of Survey Methodology*. Eds. Wolf, C., Joye, D., Smith, T. W., and Yang-chih, F. London. Sage.

Bosa, K., Godbout, S., Mills, F., Picard, F. (2018) How to decompose the non-response variance: A total survey error approach. *Survey Methodology*, 44, pp. 291-308

Brackstone, G. J. (1987). Issues in the use of administrative records for statistical purposes. *Survey Methodology*, 13, pp. 29-43.

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

Liken, C., Page, M. and Stuart, J. (2018). Realisation of 'administrative data first' in quarterly business statistics'. *Statistical Journal of the IAOS* 34, pp. 567-576. DOI 10.3233/SJI-180456.

Lohr, S., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, pp. 293-312.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*

Mercer, A.W., Kreuter, F., Keeter, S., and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, pp. 250-271.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558-625.

Pankowska, P., Bakker, B., Oberski, D.L., Pavlopoulos, D. (2018). Reconciliation of inconsistent data sources by correction for measurement error: the feasibility of parameter re-use. *Statistical Journal of the IAOS*, Volume 34, issue 3.

Poth, C. N. (2018). *Innovation in mixed methods research. A practical guide to integrative thinking with complexity*. London, Sage.

Rao, J.N.K., and Molina, I. (2015). *Small area estimation*. Second edition. Wiley. Hoboken, NJ.

Rao, J.N.K., and Ghangurde, P.D. (1972). Bayesian optimization in sampling finite populations. *Journal of the American Statistical Society*. Vol. 67. pp. 439-443.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-assisted survey sampling*. Springer-Verlag, New York.

Statistics Canada (1984). *Annual report 1983-1984*. 1-0000-502, Catalogue 11-201. Statistics Canada. Ottawa.

Statistics Canada (2012). *Statistics Canada - 2012-2013 Report on Plans and Priorities*.

Statistics Canada (2018). *Statistics Canada – 2018-2019 Departmental Plan*. Catalogue 11-635-X. ISSN 2371-7718.

United Nations, (2014). *A World that Counts – Mobilising the data revolution for sustainable development*. Report prepared at the request of the United Nations Secretary-General, by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development.

Von Neumann, J., Morgenstern, O. (1953). *Theory of games and economic behavior*. 3rd ed., Princeton University Press.