Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on "Jean-Claude Deville's contributions to survey theory and official statistics"

by Marc Christine

Release date: January 3, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2023
Vol. 49, No. 2, pp. 305-312
Statistics Canada, Catalogue No. 12-001-X

Comments on "Jean-Claude Deville's contributions to survey theory and official statistics"

Marc Christine[1]

Abstract

This article discusses and provides comments on the Ardilly, Haziza, Lavallée and Tillé's summary presentation of Jean-Claude Deville's work on survey theory. It sheds light on the context, applications and uses of his findings, and shows how these have become engrained in the role of statisticians, in which Jean-Claude was a trailblazer. It also discusses other aspects of his career and his creative inventions.

Key Words: Sampling; Calibration; Balancing; Weight sharing; Quota; Statistical methodology; Statistician.

The announcement of Jean-Claude Deville's death in October 2021 was met with great emotion at the Institut national de la statistique et des études économiques (INSEE) and the Service statistique public (SSP) (France's official statistical system), as well as in international circles (e.g., Canada, European partners). He was a great personality – both personally and intellectually – and his colleagues were deeply saddened by his death, but his body of work speaks to his outstanding qualities. Yves Tillé compiled a bibliography of Jean-Claude Deville's work, available on INSEE's Journées de méthodologie statistique website (French only), which also contains the proceedings of all the Journées from 1991 to 2022, including the papers mentioned in this article.

Ardilly, Haziza, Lavallée and Tillé's article is a remarkable account of some of Jean-Claude Deville's major innovations in survey theory and statistical survey practice, particularly at the SSP, but also internationally – balanced sampling, calibration, weight sharing, calculation of complex estimator variance, quota sampling.

For those unfamiliar with Jean-Claude, the article seems to suggest that his career was essentially devoted to developing advanced fields of theoretical investigation, complex mathematical formulations and highly specialized research. However, as mentioned in the introduction, it is important to remember that Jean-Claude Deville worked at INSEE for several decades, part of which he spent working only with the Statistical Methods Unit, where survey questions had a key role. After this, he began working at the survey statistics laboratory of the École nationale de la statistique et de l'analyse de l'information (France's National School of Statistics and Information Analysis) in Rennes, France.

All this to say that he was driven by the need to find practical solutions to the real problems he or other statisticians faced in their day-to-day work, without necessarily knowing how to solve them. He also had a remarkable ability to describe and broaden the scope of the problems brought to his attention as well as to establish links between different disciplines.

1. Marc Christine, former Inspector General of INSEE. E-mail: mchristine7577@gmail.com.

Jean-Claude Deville also worked on subjects other than those strictly related to survey theory, including demography and data analysis – further proof of his diverse research. He was also the founder of and resident theoretical expert on a major overhaul of the French census system, transforming it into a continuous sample census that was implemented in 2004. Although this article does not discuss his work, it is worth remembering.

He also taught at the École nationale de la statistique et de l'analyse économique (France's National School of Statistics and Economic Analysis) (e.g., second-order random processes, in-depth surveys).

Ardilly, Haziza, Lavallée and Tillé produced an eloquent, yet highly detailed summary of the technical aspects of Jean-Claude Deville's work. *This discussion cannot begin without first highlighting the quality and comprehensive nature of their article.*

In addition, what follows is not intended as a commentary on the quality of this general article or on the intrinsic benefit of Jean-Claude Deville's work, but rather as a way to shed light on some of his work to recontextualize it according to the daily questions of the statisticians who gave rise to it or who applied it.

A man and his career

Discussing Jean-Claude Deville's work using the overview presented in the article also means discussing his day-to-day work as a statistician, which – as mentioned above – fuelled his theoretical reflections.

Here are a few examples that can help to remember the variety of his work and his pragmatism. He was an expert on many subjects. The following is a random list drawn from my personal experience:

Using the Vacances des Français survey (survey on vacations taken by the French), he helped answer a question from a group of Breton communes, which wanted to show that tourism in their areas had dropped following the sinking of the Amoco-Cadiz in 1978 and the resulting oil spill, with the aim of obtaining compensation in criminal court (1984).

In 1985, for the CAMME project (transforming the household economic survey into a monthly survey), he devised a method for drawing the telephone sample from the paper directory – the only one available at the time.

From 1993 to 1998, he worked extensively to transform the French employment survey – which was initially annual – into a quarterly, then ultimately monthly, survey. He was disappointed that not all his recommendations were followed, but he did support and defend his colleagues, which was appreciated. For example, when a service provider with no subject-matter knowledge wanted to get involved ("I appreciate your vigilant tone", he told me when I protested against this outside initiative).

He was the architect of the modernized (continuous) census of population (1999-2000) and led a working group where he enjoyed participating in "spirited" scientific debate.

He closely followed the OCTOPUSSE project (Organisation Coordonnée du Tirage Optimisé Pour Une mise en œuvre StatiStique des Échantillons, i.e., France's coordinated optimized selection for statistical sample processing project) to build the new master sample from the continuous census, while ironically illustrating his concern (in the form of a "parable of local leaders") that INSEE's regional directors would be asked for their opinion on the samples once they had been selected (2009).

Calibration

This corpus should be first because, on the one hand, these studies are now more than 30 years old and, on the other, it could be said that calibration procedures are the most popular theories in that they have become both commonplace and widespread in most statistical surveys. In addition, of the studies cited in Ardilly, Haziza, Lavallée and Tillé, it is undoubtedly the oldest theory that has been developed, implemented and normalized in tools, such as the SAS CALMAR macro.

The value of calibration is well established – not only does it reduce estimator variance, but it also ensures the coherence of estimated totals based on a statistical survey with external data deemed to be accurate and indisputable, on a certain number of sociodemographic, geographic and economic reference variables.

To use it requires only available auxiliary information. Such information exists in most statistical systems and is usually correlated with survey variables of interest, but the calibration variables and those observed in the survey need to be conceptually coherent.

However, one downside is the apparent ease with which the technique can be used and the temptation to calibrate to "everything" available. The price to pay is a more delicate adjustment of calibration weights and the risk of obtaining outlier weights, hence the need to limit the weights or, more specifically, the ratio between calibration weights and initial survey weights. It is also important to consider the definition of the most relevant calibration variables, without trying to multiply them, as there is a risk of introducing random collinearity.

For example, at INSEE, the "Groupe Marges" (calibration working group) is working to define a policy on survey calibration methods, both in terms of the variables to use – whether mandatory or optional – and the sources best suited to providing reference margins.

The method is so popular that it can be used "without weights": empirical data can be calibrated to totals to make them "more representative" of the reference universe, without knowing the exact process of origin, then it is implicitly agreed to act as if there were initial survey weights that are all equal.

Calibration can also be done on "nothing", if no auxiliary information is available, which is rare. Generally, there are two dual formulas to keep in mind:

$$\begin{cases} \sum_{i=1}^{N} \pi_i = E(\operatorname{card} s) \\ E\left(\sum_{i \in s} \frac{1}{\pi_i}\right) = N \end{cases},$$

for a sample s selected from a population of size N, with inclusion probabilities π_i. At the very least, we can therefore calibrate to the size of the population to have weights that add up exactly to this size.

Calibration procedures are used when there are several levels of nested statistical units, e.g., individuals–households, establishments–businesses, when calibration data on the different levels are available and when it is necessary to ensure the weightings between these levels are consistent.

Calibration has also often been used as an implicit way to correct non-response within a general procedure commonly referred to as "survey adjustment," which eliminates the need for an explanatory non-response model, which requires knowing explanatory variables at the individual level for both respondents and non-respondents. Françoise Dupont (Dupont, 1993) illustrated the equivalence of an integrated method and a two-stage method (non-response correction, then calibration) under certain assumptions.

More specifically, the two approaches were combined in the generalized calibration method for treatment of non-response. With generalized calibration, non-response in a survey can be adjusted even when the individual characteristics most correlated with non-response are known only in the sample of people who responded to the questionnaire. This is particularly true when the non-observance of auxiliary calibration variables among non-respondents is the result of a delay between the information in the frame and the survey itself.

The method consists more specifically in writing the calibration equations as follows: $\sum_{k \in r} d_k F\left(z_k' \lambda\right) x_k = X,$ where z_k denotes the vector of non-response explanatory variables, known in the sample r of respondents, and x_k denotes the vector of calibration variables, well correlated with the variables of interest, known in the sample and for which the totals (vector X) are known in the population. F is a calibration function.

As a result, the method can correct non-response even when the explanatory variables are observed only in the sample of respondents, particularly when these variables are variables of interest. This has been programmed in a macro named "Calmar 2" (Le Guennec and Sautory, 2022), which is less popular than the original Calmar macro.

Balanced sampling

In some respects, balanced sampling – first formalized in the work of Jean-Claude Deville and Yves Tillé in the early 2000s – may appear to be a mirror image of calibration. It goes back to the idea Jean-Claude Deville himself put forth as a statistician's classic dream of building a "representative" sample, a perfect reduced model of the population from which it was drawn. While it was already known that a stratified sampling plan with proportional allocation conformed to the population structure in the sample, in terms of a certain number of descriptive characteristics, the aim was also to work with representativeness with regard to continuous variables.

Compared with calibration, this technique also has the advantage of not modifying the design weights, as calibration does, but of imposing balancing conditions for a given set of inclusion probabilities, and therefore, design weights.

However, it should be noted that a survey design, as a law on the set of possible samples in a finite population, cannot be reduced to obtaining a fixed set of inclusion probabilities π_i. Selecting balanced sampling that respects the set of the $\pi_{i,}$, as opposed to one that does not, means modifying the initial design while respecting the values of the π_i and imposing so-called balancing constraints, which could give the impression of altering the randomness of the process.

Naturally, in extreme cases, too many constraints would reduce the number of admissible samples, or even eliminate them all. Conversely, obtaining a balanced sample may be only approximate, hence the need to put variables into order of their importance in balancing so as to eventually relax constraints.

For more than 20 years now, balancing techniques have spread through the statistical community and become commonplace. Some of the first applications in France involved selecting rotation groups of communes for the continuous census (to which Jean-Claude Deville contributed significantly), with various constraints on population counts. The technique was then used to draw primary units from the master sample from the census for household surveys, and it has also been used to ensure "geographical" balancing, but it would be impossible to list all its current applications. It's also thanks to the fact that the initial macro, called CUBE, saw many developments, extensions and kinetic improvements, first in SAS, then in R.

As with calibration, it can be said that balancing can be carried out "even on nothing", i.e., when no particular variable of interest is available or relevant to the survey topic. As a result, we use basic properties of balancing:

- Balancing on the *constant variable equal to 1* is equivalent to balancing the sample on population size N.
- Balancing on the *inclusion probability* variable is equivalent to ensuring a fixed-size design.
- A stratified design with simple random sampling within each stratum is balanced on the *stratum membership indicator variables*, and, with *constant* inclusion probabilities *per stratum*, balancing on these *variables* amounts to stratified sampling with a fixed sample size per stratum.
- With *constant* inclusion probabilities *in the population*, balancing on *stratum membership indicator variables* is equivalent to a stratified sampling with a fixed sample size per stratum and *proportional allocation*.

In addition, balanced sampling using the Cube macro often replaces the basic sampling procedures described above. However, if variables of interest are added, the balancing method obviously becomes more specific than a "basic" selection.

Weight sharing

The weight sharing method seems to be old (1980s) yet fairly intuitive. Philippe Brion (Clairin and Brion, 1997) cites examples of its application: the "weighted segment" method used in agricultural statistics involves selecting a sample of segments (determined from natural boundaries, e.g., rivers, roads), then surveying each farm with land in the segment. It then becomes clear that a farm can be reached several times from several segments and that this repetition must be taken into account to ensure the weights are calculated appropriately. This method was inspired by Julien and Maranda (1989).

Other examples are also given: to survey a population of nomadic herders, a list of the country's water points must be drawn up and a sample drawn, then we survey the individuals who visited these water points. The risk of double counting can be eliminated by retaining only the first visit during a given period, which requires knowledge of the details of these visits, including how many and when.

On a different note, "basic" weight sharing is used when there is a survey of individuals drawn from a sample of households, with characteristics relating to each of the two levels, and we want to produce consistent statistics from the two levels (e.g., a distribution of individuals according to a given qualitative characteristic must be consistent with a household-level statistic that counts individuals with this characteristic). Therefore, the household and individual weights need to be articulated via a weight sharing method. This field was then extensively investigated and theorized by Pierre Lavallée in collaboration with Jean-Claude Deville.

Jean-Claude Deville – a Breton both by adoption and at heart – had a keen interest in tourism in Brittany. One field of application for the weight sharing method was surveys of visitors to tourist sites. By surveying tourists at a given site for a given period, information about their stays (e.g., location, duration, accommodation), sociodemographic profiles and spending can be obtained. However, conducting these surveys at several sites could result in potentially interviewing the same tourist multiple times, thereby considerably biasing the resulting estimates.

By identifying links (in this case, the places where a tourist surveyed at one location had previously been elsewhere), using the weight sharing method corrects this risk of bias. However, it is very important to clearly and precisely identify all desirable links when using this method and anticipate what questions should be added to the questionnaire for this purpose, whether in these or other surveys.

Examples of similar uses of the method include surveys of homeless people (interviewed at a food establishment or accommodation) and attendance at cultural festivals (e.g., the opera festival in Aix-en-Provence or Beaune, the short film festival in Clermont-Ferrand), where audiences may attend several shows, and therefore be surveyed multiple times.

Calculating the variance of complex estimators

As a reminder, below are a few areas in which Jean-Claude Deville's techniques can be applied:

- fractiles;
- poverty indicators, estimated using SILC (Statistics on Income and Living Conditions) surveys, which have a major political impact;
- concentration indexes (e.g., GINI).

He addressed the following other areas on this theme in an innovative way, but they were not cited in Ardilly, Haziza, Lavallée and Tillé:

- Eigenvalues in a principal component analysis of sample data;
- Estimators derived from selection points from an infinite set, where discrete probabilities are replaced by *measures* in the sense of measure theory.

Quotas

Jean-Claude Deville's work on this subject should be seen more as an exercise in style. He was very reluctant about empirical versus random surveys, as the former were suitable only for private institutions that lacked appropriate frames, whereas official statistics had the means to manage such frames, draw samples from them, and calculate estimators and their variances based on the definition of a survey design.

However, he wanted to build a theoretical approach to the quota method and define best practices.

There are two possible approaches:

- A model approach on the variable of interest, which is the best way of theoretically justifying the quota method. However, Jean-Claude Deville was, by nature, very reticent of models and of the risks official statistics would be taking by resorting to them. A model is based on assumptions, but are they verified, how do we know and can they be imposed on the user?
- We try to place quotas in the context of design-based sampling and pretend it was stratified simple random sampling.

In all cases, the key issue is the variance of the selected estimators. Can it be calculated theoretically and estimated and how can rules be derived for choosing between a random method and a quota method when either is possible but with different cost constraints and sample size scope?

Conclusion

In all the studies mentioned, and many others, Jean-Claude Deville was not only an innovator, but he also paved the way for many other areas of future development and research, which many researchers are now pursuing.

We cannot recall Jean-Claude Deville's memory without mentioning another major project he was involved in, one that goes hand in hand with his theoretical work: INSEE's Journées de méthodologie statistique. He started this event in March 1991 and organized them until 1998. They are still held but are organized by other teams. Michel Glaude, Director of Demographic and Social Statistics at INSEE, highlighted the following at the end of the seventh edition of the Journées on December 5, 2000:

"There was a passing of the baton between a team [*Jean-Claude's*] *that had organized the Journées de Méthodologie Statistique in the past and had devoted a great deal of energy and affection to it, and a new one.* [...] *I think it's a fine lesson in how to pass on and pursue a goal."*

Jean-Claude was obviously a major contributor to each Journées event, and some studies mentioned in Ardilly, Haziza, Lavallée and Tillé were presented there for the first time (jms-insee.fr). His last two papers are from 2015 and 2018 (Deville, 2015; Deville, 2018).

To understand his motivations and objectives in creating the Journées, we must go all the way back to the introduction he gave in 1991, which appeared in the Proceedings of the first edition:

"The first goal of the Journées was to give the statistical community a tangible way to express themselves by bringing it together in one place for two days.

The second was to make "practising statisticians" feel that statistical methods are useful for doing statistics.

The third was to put on an event of a high scientific level, even if it meant setting the bar a little high for some of the public.

Inspired by the example and the contacts he had established with Statistics Canada, Jean-Claude Deville's Journées were vector of knowledge sharing and transfer and an opportunity to invite leaders in international statistics to come and have lively, productive debates with them. The events have become a place for statisticians from the public and private sector, professional and academic statisticians, and French and foreign statisticians to strike up a conversation and innovate.

Had it not been for the COVID-19 pandemic, he might have been able to take part in the 30th anniversary in March 2021, but the most recent edition had to be postponed to March 2022.

Jean-Claude Deville had an inquisitive mind, always seeking to expand his knowledge and looking for improvements. He paved the way for others and proved that statistics is far from a static discipline.

References

Clairin, R., and Brion, P. (1997). Manuel de sondages. Applications aux pays en voie de développement, Insee-Ceped, 1996. PDF version at: https://www.pseau.org/outils/ouvrages/ceped_manuel_de_sondages_1993.pdf.

Deville, J.-C. (2015). Quelques éléments de géométrie et d'algèbre pour comprendre la nature d'un échantillonnage équilibré. *Actes des Journées de méthodologie statistique*, Insee, Paris.

Deville, J.-C. (2018). Calage à poids bornés : Que fait-on au juste ? *Actes des Journées de méthodologie statistique*, Insee, Paris.

Dupont, F. (1993). Calage et redressement de la non-réponse totale, validité de la pratique courante de redressement et comparaison des méthodes alternatives pour l'enquête sur la consommation alimentaire de 1989. *Actes des Journées de méthodologie statistique*, Insee, Paris.

Julien, C., and Maranda, F. (1989). Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988. Working paper No. BSMD-89-012F, Statistics Canada, Ottawa.

Le Guennec, J., and Sautory, O. (2002). CALMAR 2 : une nouvelle version de la macro Calmar de redressement d'échantillon par calage. *Actes des Journées de méthodologie statistique*, Insee, Paris.