Proceedings of Statistics Canada Symposium 2016
Growth in statistical information: challenges and benefits

Data Science for Dynamic Data Systems: Implications for Official Statistics

Mary E. Thompson[1]

Abstract

Many of the challenges and opportunities of modern data science have to do with dynamic aspects: evolving populations, the growing volume of administrative and commercial data on individuals and establishments, continuous flows of data and the capacity to analyze and summarize them in real time, and the deterioration of data absent the resources to maintain them. With its emphasis on data quality and supportable results, the domain of Official Statistics is ideal for highlighting statistical and data science issues in a variety of contexts. The messages of the talk include the importance of population frames and their maintenance; the potential for use of multi-frame methods and linkages; how the use of large scale non-survey data as auxiliary information shapes the objects of inference; the complexity of models for large data sets; the importance of recursive methods and regularization; and the benefits of sophisticated data visualization tools in capturing change.

Key Words: Large-scale data; Combining data sources; Recursive methods; Dimension reduction; Visualization.

1. The dynamic nature of Official Statistics

I have chosen the subject of "dynamic data systems" because the title of the symposium itself is "Growth in statistical information: challenges and benefits", and growth is a dynamic concept. Even if we define Official Statistics very simply as "describing how things are" in society and the economy, Official Statistics has always been dynamic, because "how things are" is constantly changing, predictably and unpredictably.

I would like to consider a fairly broad definition of Official Statistics, as the production of summaries of data collected by or on behalf of government agencies, for the purposes of government. The data are collected under official auspices; they are collected with care, according to rigorous data collection designs and protocols; and they are expected to be available (at least as summaries) to users, along with measures of quality. Under this broad definition, Official Statistics could be thought of as providing a moving portrait of the state and the world: encompassing population, economy, health, environment, and society itself. In fact, recently, important societal and social justice initiatives have begun to include data as an important component of their work. Examples include the "data revolution" for the Sustainable Development Goals (United Nations, 2014), and the work of the Human Rights Data Analysis Group, e.g. Price and Ball (2015). In Canada, in the report of the Truth and Reconciliation Commission of 2015, Call to Action #55 is all about the provision of data.

There are many ways in which Official Statistics by this broad definition has been dynamic. A large component of its work has always been tracking and monitoring change in the social and economic circumstances of the state. Another component has been the construction and continual improvement of survey frames, of establishments, areas and households. Sampling designs have been regularly redesigned and refreshed, and estimates and forecasts updated. For decades, Official Statistics has had to deal responsibly with the accumulation of data, and with retention and release policies. Yet owing to large technological and cultural shifts, Official Statistics now faces more rapid change.

2. The changing scene

The 2016 Symposium featured many aspects of changing operations and their guiding principles. For example, there was discussion of censuses making increasing use of administrative data. Census design, once carried out in waves

[1]Mary E. Thompson, University of Waterloo, Waterloo, Ontario, Canada, N2L 3L9

with periodicity five or ten years, is now more and more a continuous operation. The recent article by Blumerman (2015), about preparations for the 2020 census in the US, addresses some of the same issues.

The tempo of statistics production has speeded up, and has the potential to increase even further. We have long been accustomed to seeing final economic or financial figures coming out several months after the reference period they describe. With greater automation in data collection, we may see some of these being produced, as several speakers noted, almost in real time -- facilitating economic "early warning systems", and increasing the possibility of using the information to change or control "how things are".

Of course, there are dangers in automatic processing, instantaneous "course correction" on the basis of random fluctuations, and the engendering of feedback, as for example reactions of labour markets to changing exchange rates.

Frames are becoming more dynamic in the sense of more and more frequent updates; well-known examples are the Statistics Canada Business Register, which is updated continuously from tax data and other sources; and the US Postal Service Delivery Sequence File, the basis for many Address-Based Sampling designs, which is updated weekly to monthly. In Statistics Canada's Household Survey Frames program, as described in a Symposium paper, the frame components are now refreshed at least quarterly from several sources.

Frames maintained by statistical agencies are not only keeping pace with population change but also growing in richness, with more auxiliary information, including information on relationships among the units. They are growing in linkability, with a concomitant vast expansion in their utility. At least for administrative purposes there are now databases linkable by student IDs, health card numbers, social insurance numbers, images, and even by DNA. The Symposium featured several interesting explorations of the challenges and potential of record linkage. Despite the promise and increasing use of unique identifiers, there are still some frontiers: applications where it is necessary to make the best possible use of traditional partial identifiers such as names, addresses, and birth dates. An interesting article by Fu et al (2014) develops a methodology for automatic linkage of persons through linkage of households in historic census data.

Survey design processes are becoming intentionally dynamic, with adaptive sampling designs and responsive data collection. These now can use instantly available paradata, for example for determining follow-up strategies and data collection modes. The new design processes present new analytic challenges, for example incorporating the timing of interviews and mode-related measurement effects into analyses.

New sources of auxiliary data have become abundant, and much of the latest research, including some described at the Symposium, is focussing on how to make use of these new sources in an economical way. Price data are a prime example, with important progress being made in the strategic use of scanner data and other automatically generated streams.

With new sources for price data, the concept of price index may also evolve. For many years, a typical Consumer Price Index (CPI) has been constructed from two sources of survey data: surveys of retail establishments (outlets) and surveys of family expenditures. But we seem to be increasingly close to being able to estimate the movement in a quantity like this:

$$\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{m_{jk}} p_{jki}\right)/J$$

where conceptually $j = 1, \ldots, J$ indexes households, $k = 1, \ldots, K$ indexes goods and services items in a "basket", and p_{jki} is the price paid by household j for its i-th purchase of item k in a certain year. This quantity, the overall expense per household in the year for goods and services in the basket, is not a price index in the traditional sense, but is closely related. The "basket" could be the CPI basket of goods and services, or a product class, and the quantity is the average price paid per household in the year for goods and services in the "basket". The numerator can be regarded as a sum over a population for which the elementary unit is a purchase. At least for some kinds of item, databases of purchase transactions will be available.

Methods for using auxiliary data are being revisited and adapted, for example in the use of "organic" data streams for "nowcasting" of economic time series (Castle et al, 2015), or for the use of credit databases and the like in the correction of non-probability samples (Rivers, 2007).

3. Data growth and inference

Given the changing scene as just described, what are the implications of dynamic data and data growth for inference and for Official Statistics practice? These depend both on the characteristics of the data and on the nature of inference.

The characteristics of the new data streams are variable. Some are low dimensional, while others are characterized by richness and high dimensionality. Some are very comprehensive, and collected densely in time or space, with high frequency of capture. In many cases they arrive in "real time". Some data are of very high accuracy, as in the case of data from a sensor, as long as the sensor is functioning properly. Other data may be full of errors, gaps or holes.

What is inference in Official Statistics? Much of it is description, relying mainly on design based inference, but using models to improve efficiency or account for non-sampling errors. Included in the area of description are surveillance and monitoring for anomalies. Other aims involve prediction and forecasting, which depend on modeling to connect future to past and present. Finally, modeling is important for understanding connections and relationships. Thus, despite the emphasis on description, inference in Official Statistics is closely related to inference in other areas of statistics, and models figure prominently.

In characterizing the statistical challenges, I would identify three, corresponding approximately to the three types of inference:

- Although descriptive aims are simple, well-fitting models for big data sets are complex.
- Prediction and forecasting from accumulating data require rolling capture and rolling inference techniques.
- High-dimensional data and complex models need dimension-reduction techniques for understanding and processing.

3.1 Model complexity

In an interesting recent paper, Cox (2015) argues that "So called big data are likely to have complex structure, in particular implying that estimates of precision obtained by applying standard statistical procedures are likely to be misleading." He puts forward a model illustrating the impact of accumulating sources of variability (e.g. from local to national to global) on the estimation of means and regression coefficients; under the model the variance of a mean or regression coefficient decays more slowly than the reciprocal of the "size" of the growing data set.

For traditional descriptive inference this complexity is not an issue. The Canadian population is large, with complex structure. It can be described with census data, and its labour force characteristics can be estimated using design-based inference from the Labour Force Survey with a probability sample of about 56,000 households. But for analytical purposes, the complexity of the population is highly relevant. To take an extreme example, one would not apply a simple logistic regression model to census data and expect the near-zero standard errors and results of tests from standard methods to reflect one's understanding of the dependences in the data.

3.2 Accumulating data

Dynamic sampling designs have a long history: the rolling samples of Kish (1961; 1979; 1998); designs for rotating samples such as the one by Fellegi (1963); the McLeod and Bellhouse (1983) technique for drawing a simple random sample of size n from a "single pass" through a population of size N. The use of permanent random numbers (Ohlsson, 1995) allows coordination of successive samples from a population so as to spread the response burden evenly. These concepts and techniques essentially allow the management of data growth by letting some of it go in a controlled manner, so as to maintain a sample of approximately constant size.

3.3 High-dimensional data and model selection.

The surveys of Official Statistics have always been multi-purpose and have typically collected multiple measures on businesses and households. Estimates are required for subpopulations defined by geography, sector, or demographic groups, and their production is straightforward. However, when modeling is used as a device for improving efficiency or as a basis for dealing with missing data, the need for model selection comes into play. This is all the more the case with models for forecasting, where the dimension of the explanatory or prediction variable space is constantly increasing.

The next two sections discuss in more detail some dynamic data methods which Official Statistics has in common with other areas. With moving or rolling samples, recursive techniques go hand in hand; and dimension reduction techniques are an important tool for model selection.

4. Recursive inferences

Recursive inference techniques, and more generally what we might call moving or rolling methods, where the estimation or forecasting algorithms are invariant under time shifts, are well established in Official Statistics. ARIMA time series methods and their extensions are a prominent example. Increasingly important are the production of models and forecasts for high dimensional or spatially distributed time series, ideally satisfying certain consistency requirements such as aggregation constraints (Quenneville and Fortier, 2012). Hyndman et al (2016) propose sparse matrix algorithms for reconciling forecasts for large numbers of grouped time series.

Other methods are related to state space models, to Kalman filtering and its generalizations, applied to mean estimates over time by authors such as Tam (1987) and Pfeffermann (1991); see also Tam (2015). A simple example of a state space model for a time series y_t (a special case of the Kalman filter model)

$$y_t = x_t\beta_t + u_t$$

where *t* denotes time, the state β_t is an AR(1) process, and u_t are independent and $N(0, \sigma^2)$. More generally, the state β_t is a Markov process, and the distribution of y_t is determined by the state, making it possible to estimate the state at each time *t* on the basis of observation of the time series up to that point. The state of the time series, which could for example be a moving population mean, is estimated to remove noise from the time series. The aim is to understand variation in the state or "signal" over time, so as to be able to summarize it cogently and to forecast it well.

State space models are models, but there is a design-based analogue in an application of composite estimation, as used in the Canadian Labour Force Survey, to enhance efficiency and smooth the component time series (Singh et al, 2001; Fuller and Rao, 2001). This application of regression composite estimation to a survey with a rotating panel design uses the previous month's data for continuing respondents to improve estimation in the current month for level and change. Both the composite estimation and Kalman filter can be derived from efficient combination of functions having expectation 0, and in that sense they are akin.

The state space approach is particularly helpful when a spatial or temporal trend has a complex model, based on physical or biological theory. An example is found in a paper by Shaman et al (2014), where the ebola outbreak in West Africa has been modeled with a SEIRX (susceptible-exposed-infectious-recovered-deceased) model which provides the form of the evolution of the state, which is the daily reproductive number. A generalized Kalman filter uses the weekly observations to update the estimate of the state, so as to be able to forecast the condition of the epidemic six weeks into the future. The purpose is to try to generate realistic and timely forecasts of this highly non-stationary phenomenon.

Other examples of the use of state space models include the seasonal forecasting of crop yield (Newlands et al, 2014) and the estimation of fish stock maturities (Xu et al, 2015), both for purposes important to decision-makers in government.

5. Dimension reduction and regularization

With the growth of available data, we can expect modeling to become more and more important for producing statistical summaries. At the same time, the data to be modelled will become more complex. They may be higher dimensional either in the sense that the measurements are high-dimensional or in the sense that there are more sources of variation. Thus dimension reduction will be increasingly important for Official Statistics.

A case in point is that of functional data, where a data point is a function over time, and thus in a conceptual sense possibly infinite dimensional. An interesting example is the estimation of mean electricity consumption curve over all customers in a region, where there are millions of consumers each with an electricity meter giving readings at very fine time scales (Lardin-Puech et al, 2014). Classical ways of effectively reducing the dimension include Fourier analysis, or other ways of approximating the function by a linear combination of basis functions. Another form of dimension reduction would be linear interpolation of the function between values at sampled time points.

Dimension reduction in the sense of sparse model selection can be increasingly important as we try to exploit richer data to repair missingness. Phipps and Toth (2012) have applied regression trees to finding a parsimonious model for response propensity in an establishment survey with a relatively rich frame. An interesting problem arises concerning how to use the same kind of auxiliary data for imputation.

Regularization is a generalization of parameter dimension reduction wherein an object of inference with an unstable representation such as a density is approximated by one with a more stable representation (Bickel and Li, 2006) . It is a technique for avoiding over-fitting. This has applications in time series where the object of inference would be the autocorrelation structure or a spectral density. It enables Bickel and Gel (2011) to approximate a possibly nonlinear time series by a "long autoregressive series" and Burr et al (2015) at Health Canada to model associations between time series of environment and mortality statistics. Bornn and Zidek (2012) use Bayesian regularization in spatial modeling of crop yields over time in the Canadian prairies, for the purposes prediction by climate-related and soil-related variables.

6. Data visualization

Although the visual portrayal of statistical information can often mislead, data visualization could have increasing importance for Official Statistics, for data exploration and for the enhancement of communication.

Data visualization in other fields has important applications in monitoring and control. For example, engineers working in manufacturing and the chemical industries have developed sophisticated techniques using not only visualization by the "naked eye" but automatic image analysis (Duchesne et al, 2012). Traffic congestion real or synthetic data can be overlaid on Google earth maps (Kwoczek et al, 2014). It is possible to imagine many such applications where visualization can help with decisions on how, where and when to intervene.

At a time of heightened anticipation of an influenza epidemic, the SIMID project produced a prototype dashboard for monitoring the progress of an influenza epidemic in Peel Region in Ontario (Ramírez Ramírez et al, 2012). We constructed an assumed contact network (families, schools and workplaces) and ran microsimulations of a stochastic epidemic model on the network, showing the progress of each on a map of the region for various settings of the epidemic parameters like the latency period and the control parameters like the vaccination rate.

Mapping can also help in portraying survey and population data, including the sampling design strata and units, for the purpose of data exploration. For example, when the outcomes of interest are related to geography, it may be helpful to overlay the strata and primary sampling units for a survey on a map of population density from the census.

Small area estimation is now using "big data sources" (Marchetti et al, 2015), and visualization of small area means on maps is increasingly possible, subject to the usual cautions. The US Census Bureau has a web site on Small Area Health Insurance Estimates 2005-2013, animating the results of small area county-level estimation using spline interpolation within states from the geometric centres of counties. Sangalli et al (2013) also use a spline interpolation technique to portray census data on population density in Montreal.

7. Conclusion

This paper has presented a review of methods related to data growth, and to some of the work presented at the Symposium. It is clear that we are making rapid strides in harnessing the power of new data sources. Data collection and data management methods have changed radically, leading to new opportunities and analytical challenges. Evolution of analysis methods is leading to more points of contact between emerging practices and other branches of statistics and computer science.

References

Bickel, P. J. and Gel, Y. R. (2011), "Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series", *Journal of the Royal Statistical Society Series B*, 73, pp. 711-728.

Bickel, P. G. and Li, B. (2006), "Regularization in statistics", *Test*, 15, pp. 271-344.

Blumerman, L. (2015), "Planning for the 2020 census: A new design for the 21st century", *Amstat News*, Issue 462, pp. 12-13.

Bornn, L. and Zidek, J. V. (2012), "Efficient stabilization of crop yield prediction in the Canadian Prairies", *Agricultural and Forest Meteorology*, 152, pp. 223-232.

Burr, W., Takahara, G. and Shin, H. H. (2015), "Bias correction in estimation of public health risk attributable to short-term air pollution exposure", *Environmetrics*, 26, pp. 298-311.

Castle, J. L., Hendry, D. F. and Kitov, O. (2015), "Forecasting and nowcasting macroeconomic variables: a methodological overview", *Handbook on Rapid Estimates*, Eurostat.

Cox, D. R. (2015), "Big data and precision", *Biometrika*, 102, pp. 712-716.

Duchesne, C., Liu, J. J. and MacGregor, J. F. (2012), "Multivariate image analysis in the process industries: a review", *Chemometrics and Intelligent Laboratory Systems*, 117, pp. 116-128.

Fellegi, I. (1963), "Sampling with varying probabilities without replacement: rotating and non-rotating samples", *Journal of the American Statistical Association*, 58, pp. 183-201.

Fu, Z., Boot, H. M., Christen, P. and Zhou, J. (2014), "Automatic record linkage of individuals and households in historic census data", *International Journal of Humanities and Arts Computing*, 8, pp. 204-225.

Fuller, W. A. and Rao, J. N. K. (2001), "A regression composite estimator with application to the Canadian Labour Force Survey", *Survey Methodology*, 27, pp. 45-51.

Hyndman, R. J., Lee, A. J. and Wang, E. (2016), "Fast computation of reconciled forecasts for hierarchical and grouped time series", *Computational Statistics and Data Analysis*, 97, pp. 16-32.

Kish, L., Lovejoy, W. and Rackow, P. (1961), "A multi-stage probability sample for traffic surveys", *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 227-230.

Kish, L. (1979), "Samples and censuses", *International Statistical Review*, 47, pp. 99-109.

Kish, L. (1998), "Space/time variations and rolling samples", *Journal of Official Statistics*, 14, pp. 31-46.

Lardin-Puech, P., Cardot, H. and Goga, C. (2014), "Analysing large sets of functional data from a survey sampling point of view," *Journal de la Société Française de la Statistique*, 155, pp. 70-94.

Kwoczek, S., Di Martino, S., Nejdl, W. (2014), "Predicting and visualizing traffic congestion in the presences of planned special events", *Journal of Visual Languages and Computing*, 25, pp. 973-980.

McLeod, A. I. and Bellhouse, D. R. (1983), "A convenient algorithm for drawing a random sample", *Applied Statistics*, 32, pp. 182-184.

Newlands, N. K., Zamar, D. S., Kouadio, L. A., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S. and Hill, H. S. J. (2014), "An integrated, probabilistic model for improved seasonal forecasting of crop yield under environmental uncertainty", *Frontiers in Environmental Science*, 2, pp. 1-21.

Ohlsson, E. (1992), *SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden – A Methodological Description*, Stockholm, Sweden: Statistics Sweden.

Pfeffermann, D. (1991), "Estimation and seasonal adjustment of population means using data from repeated surveys", *Journal of Business and Economic Statistics*, 9, pp. 163-175.

Phipps, P. and Toth, D. (2012), "Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data", *Annals of Applied Statistics*, 6, pp. 772-794.

Price, M. and Ball, P. (2015), "Selection bias and the statistical patterns of mortality in conflict", *Statistical Journal of the IAOS*, 31, 263-272.

Quenneville, B. and Fortier, S. (2012), "Restoring accounting constraints in time series: methods and software for a statistical agency", In: W. R. Bell et al. (eds) *Economic Time Series: Modeling and Seasonality*, Boca Raton, Florida: CRC Press, pp. 231-253.

Ramírez-Ramírez, L. L., Gel, Y. R., Thompson, M., de Villa, E. and McPherson, M. (2012), "SIMID: SIMulation of Infectious Diseases using Random Networks", *Computer Methods and Programs in Biomedicine*, 110(3), pp. 455-470.

Rivers, D. (2007), "Sampling for web surveys". *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Sangalli, L. M., Ramsay, J. and Ramsay, T. (2013), "Spatial spline regression models", *Journal of the Royal Statistical Society Series B*, 75, pp. 681-703.

Shaman, J., Yang, W. and Kandula, S. (2014), "Inference and forecast of the current West African ebola outbreak in Guinea, Sierra Leone and Liberia", *PLOS Currents Outbreaks*, doi: 10.1371/currents.outbreaks.3408774290b1a0f2dd7cae877c8b8ff6.

Singh, A. C., Kennedy, B. and Wu, S. (2001), "Regression composite estimation for the Canadian Labour Force Survey with a Rotating Panel Design", *Survey Methodology*, 27, pp. 33-44.

Tam, S. M. (1987), "Analysis of repeated surveys using a dynamic linear model", *International Statistical Review*, 55, pp. 67-73.

Tam, S. M. (2015), "A statistical framework for analysing big data", *The Survey Statistician*, July 2015, pp. 36-51.

United Nations (2014), *A World that Counts: Mobilising the Data Revolution for Sustainable Development.* Report of the UN Secretary-General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development.

Wang, Q. and Rao, J. N. K. (2002), "Empirical likelihood-based inference under imputation for missing response data", *The Annals of Statistics*, 30, pp. 896-924.

Xu, X., Canton, E., Mills Flemming, J. and Field, C. (2015), “Robust state space models for estimating fish stock maturities”, *Canadian Journal of Statistics*, 43, pp. 133-150.