Competing Effects of Scale, Scope and Complexity in the Production, Dissemination and Use of Official Statistics

John L. Eltinge, United States Census Bureau John.L.Eltinge@census.gov

Key Words: administrative record data; constraints; data quality; efficiency; granularity; network density and complexity; record linkage; sample surveys; trade-offs

Abstract: National statistical organizations (NSOs) have encountered increasing needs to improve efficiency and data quality, while they also expand suites of statistical information products and services. Efforts to meet these goals often center on integration of data from multiple data sources, e.g., surveys, administrative records, sensors, and web scraping; and on expansion of platforms through which statistical information is disseminated and used.

The practical impact of these efforts can involve numerous trade-offs among competing effects of scale, scope, and complexity. These effects may in turn require re-examination of stakeholder priorities within the space of statistical information; features of prospective data sources; the architecture and adaptability of systems for ingest and management of data, and for production and dissemination of estimates; changing requirements for personnel in high-priority technical areas; the measurability and stability of related cost structures; and multi-dimensional criteria for data quality. This paper explores these issues, with emphasis on three points.

(1) Operational definitions of "scale," "scope" and "complexity" within the broad context of current and prospective systems for statistical information production, dissemination, and use.

(2) Production: Application of those definitions to current practice and literature for sample surveys, administrative record systems, sensor networks, web scraping and record linkage. These applications lead to exploration of trade-offs among scale, scope and complexity effects among multiple dimensions of efficiency and data quality in production of estimates for large-scale population aggregates and finer-scale small domains.

(3) Dissemination and use: Evaluation of scale, scope and complexity effects related to anticipated groups of data users; their patterns of data use; and related features of data dissemination systems. User groups include segments of stakeholders who access and use statistical tables, graphs and maps provided through public-domain websites; and more specialized researchers who may integrate and analyze microdata within restricted-access environments. These groups vary in their patterns of use of published results and related quality measures, and in the substantive context within which they interpret numerical results.

1. Introduction: "Doing More with Less" in Production, Dissemination and Use of Public-Stewardship Statistical Information?

National Statistical Organizations (NSOs) often are asked to "do more with less" in response to constraints on budgets, accompanied by requests for expansion of their portfolios of statistical information products and services, e.g., greater granularity of publication; fundamentally new estimands (e.g., based on new substantive measures Y); expanded dissemination tools; and customization to address an expanded data user base that may have heterogeneous needs for support. This has led NSOs to explore in depth a wide range of options for refinement of methodology and technology for data capture, integration and estimation (e.g., Beaumont, 2020; Lohr, 2021), and expanded capabilities for data dissemination and user support (e.g., Clark, 2020, and references cited therein).

Practical decisions about the development and implementation of these options (e.g., Holt, 2007; Pfeffermann, 2015) often involve the ways in which both stakeholder value and efficiency of operations can be associated with the effects of scale (anchored in the *number* of similar cases), scope (involving the *variety* of cases handled through a system) and complexity (from *dependencies, other constraints, and network effects* within the system). The effects of scale, scope and complexity have the potential to produce gains in efficiency based on specified metrics, but those prospective benefits are not guaranteed, and depend on case-specific empirical results.

For NSOs, four concepts can be useful in exploration of these issues. First, addressing the above-mentioned needs can involve an expansive approach to *design of statistical information systems*, including NSO decisions on priority stakeholder information needs; resulting priorities for target estimands; exploration and selection of survey and non-survey data sources; platforms for dissemination of aggregate estimates, modeled ("synthetic") data, and microdata; and all supporting actions related to methodology, technology and management. Second, NSO decisions also require careful attention to a large *space of environmental factors* that have important effects on stakeholder information priorities; availability and properties of current and prospective data sources; and stakeholder navigation and use of the statistical information produced by the NSO. Third, direct measurement of value delivered to stakeholders by NSOs can be very challenging, so we often use *multiple measures of quality* as partial indicators of stakeholder value. Biemer et al. (2017), National Academies (2017) and Brackstone (1999) give examples of quality measures. For this article, measures of accuracy, relevance, comparability, cross-sectional and temporal granularity, punctuality, interpretability and accessibility will be of principal interest. Fourth, evaluation of efficiency generally involves *numerous fixed and variable components of cost*. Groves (1989), Karr and Last (2006) and Wagner et al. (2020) give examples of cost models for NSOs. Cost models *for data users* also are important in the production and dissemination of public-stewardship statistical information.

The following notation can help in use of these four concepts to explore the effects of scale, scope and complexity. Let the vector Z represent the environmental conditions within which the statistical organization is operating; and let the vector $X = (X_1, X_2)$ represent the full set of design decisions by the statistical organization, where X_1 is the sub-vector of design factors for which we explore scale, scope and complexity effects; and X_2 is the remaining sub-vector of X. In addition, let the vectors $X_B = (X_{B1}, X_{B2})$ and Z_B represent the "baseline" design settings and environmental conditions, respectively, in which we will anchor our analyses; and let C_B and $V_{B\theta}$ equal the expected cost incurred, and stakeholder value delivered, respectively by the statistical organization in producing and disseminating estimates of the parameter vector θ under the baseline conditions. We consider a schematic model:

$$C(X, Z;\ \gamma) = C_B + g(X_1, X_2, Z;\ \gamma) + e_C \qquad \text{(M-C)}$$

where $g(\cdot)$ is a function with known form such that $g(X_{B1}, X_{B2}, Z_B\ ;\ \gamma) = 0$; the error term e_C has a mean zero and distribution function that may depend on X and Z; and γ is a parameter vector.

In addition, we use a related vector V_θ of the multiple dimensions of stakeholder value conveyed by production and distribution of estimates of θ using a system with design X under environmental conditions Z

$$V_\theta(X, Z;\ \alpha) = V_{B\theta} + h_{V\theta}(X_1, X_2, Z;\ \alpha) + e_{V\theta} \qquad \text{(M-V)}$$

where $h_{V\theta}(\cdot)$ is a function with known form such that $h_{V\theta}(X_{B1}, X_{B2}, Z_B\ ;\ \alpha) = 0$; the error term e_V has mean zero and distribution function that may depend on X and Z; and α is a parameter vector.

Within the context defined by models (M-C) and (M-V), one could obtain a preferred scale, scope or complexity effect if a specified change in X_1 leads to a very substantial improvement in the stakeholder value profile $V_\theta(X, Z;\ \alpha)$ while incurring little or no increase in the cost profile $C(X, Z;\ \gamma)$. Conversely, a scale, scope or complexity effect may reduce aggregate efficiency if the change in X_1 leads to little or no improvement in the stakeholder value while incurring substantial increases in cost. For example, such problematic cases may occur for sample surveys in which major increases in the nominal sample sizes may reduce sampling error variance, and incur correspondingly large increases in fieldwork costs, while not necessarily reducing other sources of error arising from problems with, e.g., the sampling frame or the survey instrument. In addition, the framing provided by (M-C) and (M-V) makes clear that in some cases, a prospective scale, scope or complexity effect attributed to a given change in X_1 may be contingent on the other design sub-vector X_2 and the environmental factors Z, e.g., through interaction effects captured in the functions $g(X_1, X_2, Z;\ \gamma)$ and $h_{V\theta}(X_1, X_2, Z;\ \alpha)$; and also may be subject to uncertainties reflected in the error terms e_C and $e_{V\theta}$. Similarly, if there are strong interactions with Z, and if Z is highly volatile, then it may be difficult or impossible to make productive use of observed scale, scope and complexity effects, even if those effects nominally are quite large for some Z. Consequently, reports of empirical results for scale, scope and complexity effects in a given pilot study or other setting warrant careful evaluation within the context of the above-mentioned models.

Additional complications arise because in work with statistical information systems, we often do not have detailed measures of all components that contribute to stakeholder value, nor even to a given dimension of data quality; and also lack detailed measures of some fixed and variable components of cost. Thus, discussion of scale, scope and complexity effects may be limited to analysis of certain proxy variables intended to reflect some notable dimensions of quality or cost, but acknowledged to present only a partial picture of cost-quality trade-offs.

The remainder of this article explores these ideas in additional depth for cases that may be of special current interest to NSOs. Sections 2, 3 and 4 focus on scale, scope and complexity effects, respectively; each section considers examples in data collection and information production; and in dissemination and use of statistical information dissemination and use. Section 5 suggests some areas in which additional study may be warranted.

2. Scale: Measured Units, Estimands and Data Users - Beyond $n^{1/2}$

Over the past decade, the statistical community has developed increasing interest in the application of data science and statistical methodology "at scale" and in assessment of the resulting benefits conveyed to stakeholders through improved efficiencies of scale. See, e.g., Jordan (2019), Meng (2018) and the extensive literature on "big data."

In one sense, NSOs have considered scale effects for many decades, largely within the context of sample surveys. One simple example arises in two-stage sampling of *n* out of *N* primary sampling units, and *m* out of *M* secondary sampling units from each selected primary unit. The resulting sample mean has a sampling error variance equal to $n^{-1}(1 - N^{-1}n)S_1^2 + n^{-1}m^{-1}(1 - M^{-1}m)S_2^2$, where S_1^2 and S_2^2 are measures of dispersion within and across primary sampling units, respectively (per expression (10.8) in Cochran (1977)). For the same design, a variant on Cochran (1977, Section 10.6) leads to the relatively simple linear cost model $C = C_0 + C_1 n + C_2 nm$, where C_0, C_1, and C_2 are constants. Thus, for population structures with constant M and m, and increasing *N* and *n*, data quality (as reflected in the proxy measure of the inverse of the sampling standard error) is improving relatively slowly with $n^{1/2}$, while costs are increasing linearly in *n*. Similar comments apply to extensions of such analyses to account for design effects attributed to use of stratification, additional stages of sampling, and unequal weighting; and also for "fraction of missing information" analyses arising from work with incomplete data.

NSOs also encounter many other scale-effect issues. Within the area of data capture and production of estimates, NSOs have directed special attention toward sources (e.g., administrative records, or other non-survey data sources like web scraping or sensors) that can produce data for a very large number of units at a very small incremental cost per unit, thus leading to large reductions in the nominal sampling error variance for estimators computed from such data. However, extending section 1comments on surveys with large nominal sample sizes, thoroughgoing analyses of scale effects for non-survey data sources generally will require balanced consideration of cost functions for additional design features that contribute to the "accuracy" dimension of data quality (e.g., population coverage, linkage errors, incomplete-data patterns, and variable-definition issues), and that also may contribute to other dimensions of data quality like relevance, comparability and interpretability. In addition, both scale and scope effects warrant careful attention during the development, operation, and maintenance of edit and imputation procedures for data from both sample surveys and non-survey data sources. Of special note are edit procedures that require substantial costly person-time for follow-up with a responding unit to verify the quality of a specified survey response or administrative-record entry. Also, both quality and cost measures can be influenced by edit and imputation procedures based on machine learning methods

In the dissemination and use of statistical information products, scale effects can depend heavily on the nature of the data to be disseminated, and features of the dissemination channel. For previously prepared data products (e.g., tables, graphs and maps), there can be high fixed costs for preparation and storage, while the incremental cost of an additional user's access can be relatively modest. Evaluation of that incremental cost becomes more complicated when a high-profile data release leads to a very high short-term surge in system access requests; buffering such a surge can prevent system overload, but in itself can incur substantial costs. Data dissemination and usage through a restricted-access facility can have a very different cost profile, especially if the NSO must expend substantial effort on, e.g., review of access applications; security monitoring of usage patterns; and disclosure-protection review of analysis results before clearance for public release. Such cases may display relatively limited efficiencies of scale, and it can be of interest to determine whether workflow management tools, or alternative approaches, can help in cost management.

3. Scope: Variety of Data Sources, Processing Options, and Patterns of Stakeholder Information Usage

For some general background on scope effects, and distinctions between scale and scope effects, see, e.g., Panzar and Willig (1981), Goldhar and Jelinek (1983), Hernandez-Villafuerte et al. (2017) and references cited therein. In some cases, expansion of scope can improve per-unit efficiency through amortization of certain fixed costs over a wider range of units. Conversely, expansion of scope can degrade per-unit efficiency if the added units require a high level of expensive customization. Consequently, evaluation of scope effects in statistical information systems can require careful attention to the specific types of heterogeneity under consideration; and the impact of each type of heterogeneity on quality and cost profiles in data capture, estimation, dissemination, and information use.

As with scale effects, some areas of NSO work have considered variants on scope effects for many decades. It is arguable that classical randomization-based design and inference produce some extraordinary scope effects, in the sense that the resulting point estimators and interval estimators have satisfactory properties for a very wide range of populations, under mild mathematical conditions. In other cases, efficient NSO management of scope effects can depend heavily on timely incorporation of information about the heterogeneity of units. For example, customary stratified sampling can produce substantial efficiency gains, provided the sample design incorporates stratification information that includes strong predictors of conditional means at the unit level. Other examples include work with analysis of incomplete survey data, e.g., formation of weighting cells from groups of units with approximately equal estimated response propensities; and formation of imputation cells from groups of units with approximately equal estimated conditional means. In addition, development and estimation of generalized variance functions for sample surveys often requires classification of estimators into groups, within which variance function coefficients are approximately equal. We also encounter scope effects in several dimensions of survey data collection. For instance, in work with online survey instruments or other forms of self-response, one must address heterogeneity in the quality of responses; and in the amount of online "help desk" support required, when applicable. Also, collection of data across multiple languages can lead to changes in both cost functions and data quality profiles; special interest may center on trade-offs between increased costs and improved data quality that might result from efforts to collect survey responses in one additional language (i.e., moving from $L-1$ languages to L languages).

In capture and integration of data from non-survey data sources, scope effects can be especially notable in consideration of cost-quality trade-offs when an NSO considers movement from $S\text{-}1$ to S administrative record data sources. Cost issues may include effort required in negotiation with external partner organizations that might provide those data (which might improve overall accuracy through enhancement of population coverage); and also may include technical work related to ingestion of such data from heterogeneous administrative-record systems, and related to navigation of heterogeneous patterns of metadata (which might improve accuracy, comparability and interpretability of results). In addition, scope effects can arise in small domain estimation, for which use of expanded sets of data sources may improve accuracy and granularity of published estimates, but may require increased costs to carry out more refined modeling and related diagnostic work (e.g., Rao and Molina, 2015).

For data dissemination and use within a public-goods environment, scope effects can arise from heterogeneity in, respectively, the principal data users and uses; the dimensionality and structure of the data products; the nature of privacy protection required for the data; and the required extent of in-depth interaction between the NSO and individual data users. Some users focus on a small number of cells in specific published tables (e.g., monthly sales or price-index values for particular product categories), so their interest in quality may center on accuracy, relevance, and comparability across categories and time. Other data users may have strong interest in highly exploratory analyses of many tables, and of related maps and high-dimensional graphs; their quality focus may include the above-mentioned criteria, but also may emphasize accessibility and discoverability of many cell entries, related metadata features, and incorporation of high-quality graphical options. Still other data users may use a restricted-access facility to produce specialized analyses, potentially including linkage with other microdata sources. The latter case may involve highly variable user needs for guidance on available data and metadata; for computational capabilities; and for review of prospective product releases to ensure compliance with applicable disclosure-protection requirements. For all of these options, the associated scope effects related to costs may depend heavily on the prevalence of distinct cases requiring highly labor-intensive, customized support.

4. Complexity: Dependencies, Other Constraints, Network Effects and Related Uncertainties

Within the context of schematic models (M-V) and (M-C), evaluation of dependencies, other constraints, and network effects can lead to further analyses. These include examination of interactions between *X* and *Z*, and among components of *X;* constraints imposed on specified features of the design; constraints imposed on individual observations; threshold requirements for specified quality or cost measures; and uncertainties related to the measurement of stakeholder value, quality, cost and environmental factors.

NSOs have longstanding experience with some complexity effects arising from sample surveys. For example, data collection instruments may include additivity checks, e.g., for reported components of income or expenditures. In addition, personal-visit interview surveys often encounter notable constraints on time-specific availability of data collection personnel with training in the use of a specified survey instrument. For this scheduling constraint, one also encounters uncertainties due to unscheduled absences or unexpectedly long travel times.

In work with the capture and integration of survey and non-survey data sources, additional complexity effects can arise from the need for systems that make extensive use of modeling; and model lack of fit can have a substantial impact on the both the quality and cost of the resulting statistical information products. For example, indications of a poor model fit may make it inadvisable to publish estimates at the fine level of cross-sectional granularity that may be preferred; and efforts to improve model fit through use of additional regressors may increase production costs.

For data dissemination and use, important complexity effects can arise through networks of data users, especially when they actively share, and build upon, colleagues' predominant empirical results, as well as related well-documented code, edited data, and metadata. In addition, there can be special interest in complexity effects related to transparency, reproducibility and replicability (NASEM, 2019) of empirical results obtained through work within restricted-access microdata enclaves. Both cases can have effects on costs for both data users and NSOs.

Also, in keeping with ideas developed by Perrow (1999) and others, one can evaluate the extent to which the use of "complex and tightly coupled systems" in production of statistical information may increase the risks of specified system failures that may lead to substantial degradation of the stakeholder value in (M-V) and substantial increases in costs in (M-C). The cost issues may include direct costs of mitigation efforts after a failure has occurred; costs of monitoring for early detection of system degradation or changes in *Z*; and costs of more fundamental system changes to provide a more fault-tolerant design. Such phenomena can occur in systems for capture and integration of data (e.g., with complexity arising from uncertainties in features of ingested administrative record data; and limited capacity for real-time checking of the resulting input data quality); and for dissemination and use (e.g., with complexity arising from incomplete or insufficient curation of edited microdata, code and documentation).

5. Closing Remarks

One could explore several additional issues related to scale, scope and complexity effects in statistical information production systems. First, in some cases one could extend the schematic models (M-V)-(M-C) and the qualitative comments in this paper to specific measures of size (for scale effects), heterogeneity (for scope effects) and complexity, aligned with particular indicators of stakeholder value, estimates of fixed and variable cost components, as well as design factors X and environmental variables Z. Second, much of the current discussion of value focused on measures of data quality, essentially as proxies for *use value*, i.e., value conveyed to stakeholders through current use of the product. One also could consider assessment of scale, scope and complexity effects anchored in *option value*, i.e., value conveyed to stakeholders by prospective future use of a given suite of statistical information products, or related production processes. Third, in keeping with standard analyses of "disruptive innovations" and "sustaining innovations" (e.g., Christensen et al., 2015), one could expand the "scope effect" analysis of Section 3 to develop separate analyses of scale, scope and complexity effects for separate suites of statistical information products intended for fundamentally different groups of data users. Differences among these suites might include separate vectors of estimands θ based on distinct subpopulations or distinct measurements; separate stakeholder groups with very distinct expectations regarding *prospective* quality priorities; and separate production and dissemination process designs X that may lead to very different *realized* quality profiles and cost functions. Finally, one could connect many of these statistical information production system issues with some of the broad general literature on scale, scope and complexity effects in computer science and information technology management.

Acknowledgements and Disclaimer

The author thanks many colleagues in universities, private organizations, and government agencies throughout the world for many insights related to scale, scope and complexity in the design of statistical information systems; and thanks Wendy Martinez and Rob Sienkiewicz for very insightful comments on an earlier draft of this article. The views expressed here are those of the author and do not reflect the policies of the U.S. Census Bureau.

References

Beaumont, J-F. (2020). Are Probability Surveys Bound to Disappear for the Production of Official Statistics? *Survey Methodology* **46**, 1-28. Are probability surveys bound to disappear for the production of official statistics? (statcan.gc.ca)

Biemer, Paul P., Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, Brady T. West (Editors) (2017). *Total Survey Error in Practice.* New York: Wiley.

Christensen, C.M., M. Raynor and R. McDonald. (2015). What is Disruptive Innovation? *Harvard Business Review.* Innosight_HBR_What-is-Disruptive-Innovation.pdf

Clark, Cynthia Z.F. (2020). COPAFS-Hosted Tiered Access Workshops. Presentation to the Council Brackstone, Gordon (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology* **25**, 139-149. Managing data quality in a statistical agency (statcan.gc.ca)

of Professional Associations on Federal Statistics, March 6, 2020. https://copafs.org/wp-content/uploads/2020/03/CLARK-COPAFS-hosted-Tiered-Access-Workshops-rev.pdf

Cochran, W.G. (1977). *Sampling Techniques, Third Edition.* New York: Wiley.

Goldhar, J.D. and M. Jelinek (1983). Plan for Economies of Scope. *Harvard Business Review.* Plan for Economies of Scope (hbr.org)

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York: Wiley.

Hernandez-Villafuerte, K., Sussex, J., Robin, E. *et al.* (2017). Economies of Scale and Scope in Publicly Funded Biomedical and Health Research: Evidence from the Literature. *Health Research Policy and Systems* **15** (3) https://doi.org/10.1186/s12961-016-0167-3

Holt, D. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper (with discussion). *The American Statistician* **61**, 1-15. https://doi.org/10.1198/000313007X168173

Jordan, Michael I. (2019). Artificial Intelligence – The Revolution Hasn't Happened Yet. (with discussion and rejoinder). Harvard Data Science Review **1** (1). https://doi.org/10.1162/99608f92.f06c6e61

Karr, Alan F. and Michael Last (2006). Survey Costs: Workshop Report and White Paper. Available through: https://www.niss.org/research/technical-reports/survey-costs-workshop-report-and-white-paper-2006

Lohr, S.L. (2021). Multiple-Frame Surveys for a Multiple-Data-Source World. *Survey Methodology* **47**, 229-263. Multiple-frame surveys for a multiple-data-source world (statcan.gc.ca)

Meng, Xiao-Li (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox and the 2016 U.S. Presidential Election. *Annals of Applied Statistics* **12** (2) 685-726. https://doi.org/10.1214/18-AOAS1161SF

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: National Academies Press. https://doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

Panzar, J.C. and R.D. Willig (1981). Economies of Scope. *The American Economic Review*, **71**(2), 268-272. https://www.jstor.org/stable/1815729

Perrow, Charles (1999). *Normal Accidents: Living with High-Risk Technologies.* Princeton, New Jersey: Princeton University Press. Normal Accidents | Princeton University Press

Pfeffermann, D. (2015). Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture (with discussion). *Journal of Survey Statistics and Methodology*, **3** (4), 425-483. https://doi.org/10.1093/jssam/smv035

Rao, J.N.K. and I. Molina (2015). *Small Area Estimation, Second Edition.* New York: Wiley.

Wagner, James, West, Brady T., Elliott, Michael R. and Coffey, Stephanie (2020). "Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context" *Journal of Official Statistics*, **36** (4), 907-931. https://doi.org/10.2478/jos-2020-0043