Training Undergraduates, Graduate Students, Postdocs, and Federal Agencies: Methodology, Data, and Science for Federal Statistics

Noel Cressie*, **Scott H. Holan, and Christopher K. Wikle**

Department of Statistics, University of Missouri

`http://stsn.missouri.edu/`

NCRN Spring Meeting
National Academy of Sciences, 8 May 2015

* Also Distinguished Professor at University of Wollongong

Federal Statistics

- This could be interpreted as everything from Intelligence, Defense, Astronomy, Geoscience, Ecology, Agriculture, Economics, and Socio-Demographics.

- Our emphasis in this talk is training in the latter area, where the state of the population of the United States of America is the core (echoing the original meaning of the word "statistics").

- We suggest a vision for training in Federal Statistics that can surely be improved with your input and that from other relevant stake-holders

Training in Federal Statistics: Vision From the Top Down

- Federal Agencies: Directors, Associate Directors,. . ., group managers, line statisticians. Training modules need to be tailored to those who “decide" and those who “do."
- Postdocs: Federal agencies need fresh postdocs: US citizens with a PhD and at least one year’s independent research experience. Non-citizens have opportunities through universities and industry.
- Graduate Students: Masters and PhD candidates with training that includes design-based and model-based inference.
- Undergraduates: Need to develop strong computing and mathematical skills, as well as basic statistical knowledge.

Training Pyramid From the Bottom Up

- Undergraduates: Strong educational foundation, with a major in some version of "Data Science."
- Graduate Students: Variety of courses including probability, mathematical statistics, Bayesian methodology, design, survey methodology, statistical methods (applied statistics), dependent data, and computing – motivated by exposure to economic and socio-demographic studies.
- Postdocs: Economic and socio-demographic research questions motivate strong statistical methodology.
- Federal Agencies: Hire completed undergraduates, graduate students, and postdocs; incentivize research as a part of their regular duties.

University of Missouri Node: Training Structure

- Vertically Integrated Structure
- PI and Co-PIs (Senior Personnel): Responsible for overall node productivity and for mentoring postdocs, GRAs, and URAs.
- Postdocs have the opportunity to assist with mentorship of graduate and undergraduate students.
- Graduate students have the opportunity to assist with the mentorship of undergraduate students. (For example, a postdoc and graduate student mentored an undergraduate who presented a node-related paper at the National Conference for Undergraduate Research (NCUR, April 2015). Presentation selected out of over 3700 abstracts!)
- Weekly meetings of node participants, including a Spatio-Temporal Reading Group.

Horizontal Integration into Federal Statistics

- At any given stage of the training process, how can NCRN students (and university students in general) experience what a career in Federal Statistics would be like?
- The training pyramid needs to be horizontally integrated with a career pyramid in the federal agencies.
- Traditional internships are one answer. Non-traditional internships: Introduce flexibility for shorter-term visits and telecommuting in combination with an RDC.
- Challenges:
  - Often recruiting delegated to HR Departments
  - "Not my job"
  - Focused mentoring essential (in both pyramids)
  - US Citizenship

Methodology, Data, Science

- Here the science is economics and socio-demographics, but that can clearly interact with the physical sciences. Climate change will almost certainly be a driver in changes in the US population.
- Data inform science but science allows intelligent data collection.
- Methodology is needed to transition from Data to Information to Knowledge to Decisions.
- Uncertainty is a part of all of this!

Uncertainty Quantification

- Probabilities (more precisely, conditional probabilities) are a powerful way to quantify uncertainty. They are unitless, lie between 0 and 1, and satisfy certain laws, such as Bayes' Rule.

- Design-based probability models should be contrasted with model-based probability models. For a modern perspective on this, see Little (2011, *Statistical Science*) and Chambers (2014, *Proc. Stat. Canada Symp.*).

- The path to answering important questions about US subpopulations and hence the "state of the population of the United States of America" is through model-based inference.

Hierarchical Modeling

- We believe that any graduate program directed towards training current and future federal statisticians should teach hierarchical statistical modeling, based on conditional-probability models.
- Conditional thinking is at the core of modern statistics and is fundamental to coherent uncertainty quantification.
- The type of hierarchical models that we believe should be taught can be described through the following conditional-probability models: [data| process, parameters] and [process| parameters].
- In this context, it is a choice whether the parameters are estimated or a prior, [parameters], is placed on them.
- Policy decisions and science depend on the latent process. The data are an imperfect view of that process. Hierarchical modeling recognizes this.

Data Science and Big Data

- These two terms are often confused.
- Data Science is being used now in a way that appears to subsume Statistical Science, (i.e., "the science of uncertainty"). Any program in Data Science that does not have "uncertainty" at its core is incomplete!
- Big Data refers to an opportunity our society is faced with: Turn those large, many, and often cheap datasets into knowledge and smart decisions.
- This opportunity can only be fully realized in the context of uncertainty quantification.

Big Data Goals

- We cannot (and should not) obtain characteristics (including space-time coordinates) of every person in the US.
- Governments need timely, aggregated information to make decisions for its population and sub-populations. Businesses need the same information in order to be competitive and react to the marketplace. The amount of aggregation depends, *inter alia*, on the sub-populations of interest.
- "Representative" data are collected for the sub-populations, mandates are met, services are provided, and planning decisions are made.
- The goal is still to estimate economic and socio-demographic characteristics in the presence of uncertainty. "Big Data" may or may not reduce that uncertainty.

The Signal in the Noisy Big Data

- The fog of "Big Data." Size matters but so does noise, missingness of individuals, and missingness of variables.
- Design principles: "Stratify, Cluster, Randomize" to get at the signal. Add "Aggregate" to this list, to help clear the fog.
- Computational considerations: Data archives are distributed; analytics are done at data nodes; Moore's law needs parallelization; memory size constrains analyses; software.
- "Big" can also mean "Many" datasets. Confidentiality applies to the whole and should not be treated piecemeal.

What are we striving for?

"What we measure affects what we do. If we have the wrong metrics, we will strive for the wrong things." (Joseph Stiglitz in "Towards a better measure of well-being," Financial Times, September 13, 2009)

- Our metrics have been high accuracy (i.e., small bias) and high precision (i.e., small variance), within a cost constraint.
- Have our metrics changed? Do we now want low cost within a quality constraint (e.g., bias/variance)?
- While the cost of "Big Data" is going down, are they complete and to be trusted? Clearly, no!

Metrics

- Sample-based quality metrics such as bias and variance seem to be waning in importance for Federal Statistics. Their model-based versions, particularly those based on hierarchical statistical models, are becoming more prominent.

- Just because we have "Big Data," it does not imply that we have removed the uncertainties surrounding the question being answered. If n is large, it does not imply it is large for the sub-population of interest; and even if it is, it does not imply that the mean-squared-error metric,

$$E(\theta - \widehat{\theta})^2 = (\text{bias})^2 + \text{var}(\widehat{\theta})$$

  is necessarily small.

Big Data Center at the U.S. Census Bureau

- The Center has been recently established but no Chief has been named yet.

- Formulate goals, then formulate priority areas to work on, then formulate problems within areas. Take a look at NSF's solicitation "Critical Techniques and Technologies for Advancing Foundations and Applications of Big Data Science & Engineering":

  `http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767`

- The Center's success will depend on whether it is, or is not, another "unfunded mandate" for those working in it. It should be resourced with serious FTEs.

- Opportunity: Make training/internships one of the missions of the center

Conclusions

- Training should have uncertainty quantification at its core.
- Computing skills are essential, but the metrics are still statistical.
- We have to be smart about realizing the potential of Big Data.