# Summary of "Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics"

by

NCRN Missouri Node

(PI: S.H.Holan; Co-PIs: N.Cressie, C.K.Wikle; Postdocs: J.R.Bradley, M.Simpson)

September 30, 2016

## 1 Introduction

On Friday, May 20 and Saturday, May 21, 2016, the Spatio-Temporal Statistics-NSF Census Research Network (STSN) at the University of Missouri hosted a workshop on spatial and spatio-temporal design and analysis for official statistics, sponsored by the NSF-Census Research Network (NCRN). The conference was designed to facilitate discussion about the use of spatial and spatio-temporal statistical methods for official statistics in a variety of capacities, including constructing estimates of population quantities and constructing sampling designs. Participants came from academia, government, and industry. The workshop was organized around three discussion topics. For each topic, participants were randomly selected into three separate groups. After each group met individually to discuss each topic, they came together and a group leader from each group summarized their discussion for all participants. Finally, the floor was opened to all participants to comment and expand on the summaries.

What follows is a summary of the discussions of each of the topics from the workshop. It is not intended to represent a consensus or a complete view (e.g., there is no bibliography), but rather it is a summary of the dialogue between participants in what we consider to be an ongoing discussion.

## 2 Topic 1: Doing More With Less

The first topic of the workshop is a perennial challenge in any organization: how to be more productive while using fewer resources. The discussion groups were asked to think about the role of several tools statistical agencies often use or could be using: surveys, big data, web-scraping, social media, and statistical dependence.

A common thread across all three discussion groups was that we need to define what we mean by "more" and "less." One group suggested trying to measure value by measuring the quality of estimates, disclosure risk and, of

course, cost. At a high level, everyone agreed that the goal is to provide higher-quality estimates and/or more estimands of interest at a lower cost. Measuring cost is the easier side of this equation, but measuring the benefits is more difficult, since data products have both use-value and option-value. Use-value occurs any time someone uses the data product, for example a business trying to understand consumers in a region where they are introducing a marketing campaign. Option-value is not so readily apparent unless data-product consumers decide to exercise the option. For example, economic indicators may have a small user-base in normal times, but many more firms are interested in them during a recession. The option to use these indicators when needed is valuable even if it never becomes necessary to use them.

Surveys are the most common way to obtain the information for these estimates, but the groups discussed several alternatives, including remote sensing and supplementary data sources such as social media, other web-based data, and crowd-sourced data. There is much of this type of data around, but it is not clear how useful it is. In particular, it might erode our trust in having long-term estimates if, for example, private companies like Facebook or Twitter change their partnership policies or go out of business. There are problems with using survey data as well, including resolution and extensibility, so what we need is a cost/benefit analysis of using different types of data. For both data sources we need protocols for assessing any models used to create estimates based on these data sources. This is expensive, but it helps ensure trust in the data products produced by the agencies.

A key piece of the puzzle for combining alternative data sources with survey data is leveraging dependence. One example of using dependence arises from borrowing strength across spatial units and over time when producing location-based estimates.

Leveraging spatial dependence allows us to develop change-of-support and regionalization methodologies, although more work needs to be done here. In this context, ongoing model assessments are important in order to catch when, for example, the dependence structure between different locations at different time points changes. But dependence is also important for combining data sources, for example for specifying the dependence structure over space for social-media data and between social-media data and survey data at nearby locations. When specified well, these dependence structures will allow for better estimates of quantities of interest, but the usual caveats about ensuring model fit apply. Using a complicated dependence structure necessitates using model-based estimates, and it can often be difficult to convince statistical agencies to incorporate model-based estimates in their products. Thus, we need to transition gradually, providing examples to prove that the model-based approach works. To accomplish this, and to convince users of the value of this approach, we need to do a better job of publishing methodologies and associated software.

Big Data has several different possible meanings, but the common thread during the discussions was that using such datasets for statistical purposes presents a computational challenge. New approaches to dimension reduction are needed to handle these datasets; either the dimension of the dataset itself

is reduced or the dimension of the model used to produce estimates based on the dataset is reduced. The latter is particularly important for fitting spatio-temporal models discussed in the previous paragraph.

# 3 Topic 2: Official Statistics From a Spatio-Temporal Perspective

The second discussion topic was official statistics from a spatio-temporal perspective, with the "prompts," design, analysis, change-of-support, and synthetic data. At present, spatial and spatio-temporal models are not widely used in official statistics, so it seems that there are opportunities. However, we should make sure that there is actually something to be gained by doing spatio-temporal modeling. For example, it may be that using the right covariates removes the benefit of the spatio-temporal component of the model. It is key to remember what statistical agencies do — they construct a variety of statistical estimates for a variety of reasons but, crucially, high degrees of quality and precision are expected in estimates produced by them. In whatever capacity we use spatio-temporal models, we should make sure that we attain these quality and precision goals. If we can, then we will have achieved more with with less, given the relative inexpense of modeling and computing compared to trying to do it all with large, complex, resource-consuming surveys.

There are some useful diagnostics we can use in order to assess the quality of our models, such as posterior-predictive checks, nonparametric sources of error, and graphical approaches. Also, statisticians should make sure they have compelling evidence that the spatio-temporal methodology they are using works (e.g., through cross-valuation and simulation experiments). Furthermore, many of the methodologies applicable in spatial and spatio-temporal domains may be applicable when the relevant space is not physical space.

One of the areas where spatio-temporal modeling looks promising is for doing change-of-support or regionalization. Change-of-support is using estimates at one set of areal units to make inference on another set at a coarser or finer level of geography. There are some existing methods for doing this, but often they are computationally expensive. Using wavelets or Gaussian processes may help, but each has its own unique challenges. Regionalization is choosing which geography to use initially. Datasets may have different spatial and/or temporal supports, which makes combining them difficult. This is another area where more research is needed. Ideally, we would produce general-purpose methodology and software that allows users to create estimates at whatever spatial or temporal supports they want. This may entail a general change-of-support methodology, or customized tabulations, or more likely some combination of the two.

All of the groups agreed that the design-versus-model based dichotomy is a distraction. The crucial question is understanding when a statistical methodology works (at the very least, estimates of bias and variance should be ob-

tainable) and what its vulnerabilities are. To this end, we should be trying both approaches on the same problems, in order to understand their strengths and weaknesses. On the analysis side, this means comparing estimates from model-based and design-based approaches. On the design side, this means trying both model-based and design-based approaches to construct sample designs. Adaptive designs are an attractive approach in the literature, and model-based methods are well suited to implementing them. Another nice feature of model-based analysis is that it can be applied to analyze data that was collected using a combination of different survey-sample designs.

Synthetic data often offer a nice solution to data-combination problems from multiple sources, for example from multiple levels of spatial or temporal support. The statistical agency can then release the synthetic data to the public, and users can often more easily analyze the synthetic data than multiple sources of somewhat similar data. However there are problems. Ideally, the agency would release synthetic microdata consistent with aggregate estimates, but it needs to ensure that privacy constraints are not violated. Good models using synthetic data may, to some extent, still be able to identify individuals. Second, releasing synthetic data requires infrastructure that currently may not exist in many statistical agencies and may be expensive to acquire. Finally, the synthetic data need to accurately reflect relevant relationships in the population. To preserve all possible interesting relationships is an impossibly high-dimensional problem, but we should do our best to capture the interactions in the latent processes, especially those important to data-users.

# 4 Topic 3: Thoughts for the Future

The final discussion topic was an invitation to think about how we might implement spatio-temporal methodologies in official statistics and where future opportunities lie. The groups were prompted to think about the role of maps and visualization, the interface between various institutions such as government agencies and academic institutions, and the role of official statistics in future societal challenges.

When it comes to maps and visualization, everyone seemed to agree that existing methods are not used very well and that there is wide variety of opinions on how to use them properly. For example, no one knows how to visualize margins of error well, or other sources of variability such as variability within an areal unit. Mapping origin-destination dynamics would also be useful, but it is an extremely difficult problem. In general, we should be doing more work testing visualizations to see which ones help people understand the data better, which may involve a psychological component. We also have a responsibility to make data products accessible to the visually impaired, but it is not clear how to do this usefully. For example, how do you make a map using only sound? Indeed, sound could be a useful way to communicate data to users more generally, if we can find a way to leverage it. Similarly, tactile maps via 3-D printing are an under-utilized way to communicate features of data to users.

A common thread across all of the discussions of how academia, government, and industry interface, was their respective time scales. Academics are on a different time scale than government employees, and in turn both are on different time scales than statisticians employed in the private sector. Each has different types of deadlines for different types of projects. In order to see benefits from these interactions, they need to be fostered more deliberately, since there are many reasons why cross-institutional collaborations will be beneficial to both parties. For example, academics may gain access to new sources of data or expertise about how data products are constructed, while a government agency may benefit from their expertise in order to construct better data products. Internships are a great way to accomplish this, but so is bringing senior academics into an agency in a consulting role. Student internships in companies work very well, and companies fund senior expertise in a consultancy relationship when they need it. But more mechanisms are needed to initiate and maintain these relationships. Resources are needed for course releases, to pay for internships and for other ways of combining efforts across institutions. The Census Bureau's NCRN has been a great source for this type of support.

Finally, there are several future challenges that provide opportunities for official statistics and official statisticians. Currently, official statistics are already used as a benchmark in the stock market, and perhaps with the use of alternative data sources this will expand. CDC uses population statistics as the denominator in the calculation of disease rates. Food security is a major concern for the USDA, and it is a global problem since the wheat harvest in the Midwest will impact food prices worldwide. NOAA and many others are interested in climate change, but they have major challenges with credibly constructing projections into the future. Another interesting issue is that many of the types of data that government agencies traditionally saw as confidential is now in the public domain because it was collected by private companies who are willing to publish them. In the past, privacy concerns drastically hampered the ability of government agencies to share data with each other and it is not clear if this will change. We must respect these constraints and, at the same time, we should emphasize the utility of an agency's data products to the public.

# 5 Conclusions

The discussions generated by the workshop were stimulating and useful, with many threads worth pursuing. We look forward to future workshops on related topics. With the hope of moving the conversation forward, we compiled this document so that interested parties may understand the state of the discussion. However, we reiterate that it does not represent a consensus view among participants or even the views of any given participant. Instead, it is a summary of the discussion with many details left out. Finally, on behalf of STSN, the Missouri Node of the NCRN, we wish to thank the participants for generating excellent discussion, and the NSF-Census Research Network for sponsoring the workshop.

# Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics
# PARTICIPANT LIST

**John Abowd**
Cornell University/
U.S. Census Bureau

**Linda Beale**
Esri

**Emily Berg**
Iowa State University

**Jonathan Bradley**
University of Missouri

**Kate Calder**
The Ohio State University

**Paul Conn**
NOAA

**Carol Crawford**
NASS, USDA

**Noel Cressie**
University of Wollongong/
University of Missouri

**Renee Ellis**
U.S. Census Bureau

**John Eltinge**
Bureau of Labor Statistics

**David Folch**
Florida State University

**Peter Glynn**
Stanford University

**Matthew Graham**
U.S. Census Bureau

**Daniel Griffith**
University of Texas at Dallas

**Chris Hassett**
University of Missouri

**Patrick Hayward**
U.S. Census Bureau

**Scott Holan**
University of Missouri

**Rebecca Hutchinson**
U.S. Census Bureau

**Ron Jarmin**
U.S. Census Bureau

**Jae-Kwang Kim**
Iowa State University

**Emily Leary**
University of Missouri

**Wendy Martinez**
Bureau of Labor Statistics

**Tucker McElroy**
U.S. Census Bureau

**Sakis Micheas**
University of Missouri

**Balgobin Nandram**
Worcester Polytechnic Institute

**Jean Opsomer**
Colorado State University

**Trevor Oswald**
University of Missouri

**Harrison Quick**
CDC

**Terrance Savitsky**
Bureau of Labor Statistics

**Matthew Simpson**
University of Missouri

**Eric Slud**
U.S. Census Bureau

**David Steel**
University of Wollongong

**Daniell Toth**
Bureau of Labor Statistics

**Jay Ver Hoef**
NOAA

**Lars Vilhuber**
Cornell University

**Suojin Wang**
Texas A&M University

**Dan Weinberg**
U.S. Census Bureau

**Christopher Wikle**
University of Missouri

**Linda Young**
NASS, USDA

**Alan Zaslavsky**
Harvard University

**Zhengyuan Zhu**
Iowa State University