ONLINE APPENDIX A. EXCERPT FROM THE U.S. NATIONAL SCIENCE FOUNDATION SOLICITATION 10-621 TO ESTABLISH THE NSF-CENSUS BUREAU RESEARCH NETWORK

[The full program solicitation can be found archived at https://web.archive.org/web/20170710231924/https://www.nsf.gov/pubs/2010/nsf10621/nsf10621.htm]

Some questions currently of interest related to data collection, analysis, and dissemination processes include the following (these topics are not exhaustive):

Traditional concepts of family and households, as well as traditional concepts of economic units, are rapidly evolving.

- What methods can improve universe frame coverage of persons with intermittent ties with households, for entrepreneurial activities leading to new economic units in economic unit frames?
- What data auxiliary to households and covered persons might be used to estimate the propensity to be covered, as a targeting tool for alternative ways of assembling universe frames?
- Can theories be developed to guide research decisions for sampling unit definitions (derived from frames) and measurement units (e.g., enterprises vs. establishments, households vs. persons) to improve overall designs?
- How can estimates of immigration (both documented and undocumented) be improved?
- Is the concept of an "establishment" still relevant given changing business models and increasingly heterogeneous economic activity?

Participation rates in sample surveys of households and economic units are declining.

- What theories can inform the linkage between nonresponse rates and nonresponse errors?
- What data might be collected or linked to traditional survey data to improve the postsurvey adjustment for nonresponse to reduce nonresponse errors?
- What mechanisms underlie the finding that offering choices of alternative modes of data collection depress overall participation? What antidotes might be created to reduce that effect?
- How can administrative records on persons, households, and economic units be used in conjunction with traditional sample surveys to reduce the nonresponse error of traditional surveys?

The complexity of economic units is increasing, with multiple establishments, loose alliances, multiple lines of business, virtual spatial attributes, and highly dynamic structures.

- How can administrative records be used to improve the tailoring of measurement techniques to diverse types of economic units?
- How can changes in key attributes of economic units be tracked over time to improve the collection of data from the units?
- In longitudinal measurement, how can deaths, mergers, and acquisitions of economic units be forecasted to permit real-time measurement of those phenomena?
- How can multiple modes of data collection facilitate measurement of complex economic units?
- How can we more accurately classify heterogeneous economic activity within business enterprises, individual locations, or aggregates of locations?

Editing and imputation techniques commonly used in sample surveys currently have few evaluative frameworks that guide decisions on what approaches maximally reduce bias in final

estimates.

- What logical or statistical approaches might offer guidance to the tradeoff decision of how much editing is optimal for diverse purposes?
- What editing algorithms might be developed to reduce the post-estimation review processes common in statistical estimation?
- What computer-assistance in editing might be developed to reduce the use of subject matter expertise in the review of data from longitudinal and other surveys?
- How can empirical diagnostic tools for evaluating auto-coding algorithms and large scale imputation approaches be improved?

Administrative records, when combined with survey data, may offer radically increased efficiencies in household and business surveys.

- What mathematical and statistical frameworks might be used to improve inference from probabilistically linked datasets?
- How can the social science community effectively monitor public attitudes toward administrative record usage?
- What conceptual frameworks might be developed to measure the error properties of linked survey and administrative record data?
- What imputation techniques can be created to deal with item missing data in linked files with variables common to multiple datasets?

While public use datasets have greatly benefited quantitative research in the social sciences, the data are increasing threatened by risk of inadvertent reidentification of sample members.

- What disclosure avoidance techniques can be developed to preserve pledges of confidentiality and maximize access to data?

- Can disclosure risk measurements be invented to guide practical decisions of data collectors regarding the release of data?
- How can synthetic data be produce that mimic the statistical properties of actual data but protect the identity of respondents?
- What effective analytic software approaches might be used to permit analysis of data without direct access to the data and protect pledges of confidentiality?

Small domain estimation using survey data offers the promise of greatly expanded useful estimates from sample surveys.

- How can model diagnostics be improved on small domain estimators?
- What small domain estimation approaches can exploit the longitudinal nature of surveys?
- What alternative approaches offer improved simultaneous estimation of small domains and higher-level aggregates?
- What practical estimators of total error of small domain estimates might be developed for public dissemination?

Cognitive and social psychological insights into respondent self-reports in social science research have reduced measurement errors.

- What questionnaire development tools are superior for detecting different mechanisms of response error?
- What diagnostic tools in instrument development can be enhanced through computer assistance?
- How do we identify optimal measurement approaches for a single construct using individual modes of data collection?
- What diagnostics can be developed to isolate translation errors as a distinct component of

measurement error in multi-language measurement?

The use of statistical models for large-scale descriptive statistics has advanced in important ways.

- How can diagnostic tools be advanced to measure potential model-specification errors within a total error framework for the estimates?
- What diagnostic tools might be developed using model-based approaches to identify errors in tabular data?
- What models might be useful to estimate sampling error covariances and auto covariances in longitudinal estimates?
- What statistical models might be useful to forecast final estimates based on preliminary measurements of a sample?

New approaches to disseminating census data to users are emerging, and new requirements for confidentiality protection will be required.

- What metadata approaches will be most useful in documenting census data, and how can existing metadata systems be improved?
- How can census data dissemination, including both tabular and microdata, be improved?
- What are the most significant risks in disseminating census data to user communities, and how can those risks be diminished?
- What approaches can be developed that will allow the user community to safely and securely access census and other administrative data that have been merged across multiple agencies or sources?

ONLINE APPENDIX B. OTHER OUTCOMES: STUDENTS, COURSES, AND SOFTWARE

Knowledge dissemination to a broader audience, and fostering of collaborations within the network, were an important component of the overall effort. Beyond the traditional academic research papers, each of the nodes also regularly presented new results in a "virtual" seminar, with researchers and students from all nodes, but also non-affiliated research institutes, actively participating through multi-site videoconferencing. Nodes added "official statistics" components to both undergraduate and graduate courses, often as "special topics." A multi-site course on "Understanding Social and Economic Data," led by researchers from the Cornell node, was taught as a hybrid distance-learning/remote-learning course, with typical attendance involving a dozen sites and over one hundred students and faculty, spread across the United States (course materials and video lectures are available at https://www.vrdc.cornell.edu/info7470/). Several other nodes created new course materials, workshops, and short courses (Michigan, Nebraska, Duke, Missouri) (see online Appendix B).

The University of Michigan offered a seminar for honors economics students, "Naturally-Occurring Data and the Macroeconomy" in 2016, wherein undergraduates did research using "big data" techniques advanced by the Michigan node. This course will be offered in future years. Aaron Flaaen used non-design data to create a new measure of the multi-national status of firms, linked it to the Census Business Register, and made it available to Census Bureau researchers and researchers in the FSRDC network (Flaaen 2015); his analysis using these measures received the World Trade Organization Award for Young Economists. Isaac Sorkin developed and implemented a method for measuring employer quality based on the firm's relative ability to hire and retain employees. This work used eigenvalue techniques that allow analysis of flows across all connected establishments in the United States (Sorkin 2015, 2018).

The Nebraska node created two new courses. The Interviewer-Respondent Interaction course explored different interviewing methods, methods to observe and analyze verbal behaviors during interviews, and methods to analyze these data (Belli 2012). The Survey Informatics course explored the role of technology throughout data collection, data management, and data analysis within survey research, as well as the increasing need for interdisciplinary teams within research to draw from the strengths of different disciplines (e.g., survey research and methodology, computer science and engineering, cognitive psychology, sociology, statistics, etc.); see Eck (2015a, 2015b) and Eck et al. (2015a, 2015b).

The nodes have also developed short courses, workshops, and modules for use in college courses. These include:

- Short course on spatio-temporal statistics taught at the Census Bureau but open to staff at other FSS agencies (Missouri).
- Short course, "Introduction to Privacy" (Carnegie Mellon).[1]
- Short course on record linkage (data matching) (Carnegie Mellon).[2]
- Short course on missing data for the Odum Institute (Duke).
- Short course on synthetic data for the Joint Program on Survey Methodology and the 2017 Joint Statistical Meetings (Duke).
- Topic modules on causes and statistical models for interviewer effects in survey data (Nebraska).
- Workshop on spatial demography and small-area estimation, "Measuring People in Place," at the University of Colorado (Colorado-Tennessee).

[1] http://www.stat.CMU/NCRN/PUBLIC/education.html#Priv
[2] http://www.stat.CMU/NCRN/PUBLIC/education.html#RLF13

- Workshops on using the SIPP and the synthetic SIPP (with matched earnings records from the Social Security Administration), conducted at Michigan, Duke, Census, and Population Association of America annual meetings, taught by Michigan and Census Bureau researchers (Michigan).[3]

A 2-day workshop on Spatio-Temporal Design and Analysis for Official Statistics, organized and hosted by the Missouri node in May 2016. More than 40 researchers invited from both inside and outside the NCRN were involved in a series of break-out discussions. A summary of those discussions was distributed to workshop participants and is archived at the Cornell University library (Holan et al. 2016).

One hope was that node-trained students would choose to work at a FSS agency upon graduation. Of course, successfully trained students also have other options, and it is difficult to assess empirically how many students gave the FSS consideration as an employment opportunity. As of this writing, we are aware of four NCRN-trained graduates at the U.S. Census Bureau, from the Duke and Missouri nodes, though several students have accepted positions at other agencies and companies that interact closely with the FSS. Based on the authors' experience in guiding students through the placement process, and based on interviews with colleagues and former students, a few observations emerge. First, students do consider the agencies comprising the FSS as potential and attractive employers. However, due to the widespread popularity of "data science," the salary structure of the federal government is not competitive enough to attract such individuals. Furthermore, while graduate students are drawn from many countries, and NSF funding is available to international students, those same students

[3] http://ebp-projects.isr.umich.edu/NCRN/training.html

cannot always be hired by federal agencies, due to legal restrictions that require an employee to be a U.S. citizen. Nonetheless, the exposure of such students to federal datasets and the challenges facing the federal statistical agencies likely still has benefits. As these individuals either continue their education or go on to academic jobs, they take with them an appreciation for federal statistical problems and may continue to focus on federal statistics as research topics.

These educational activities have been particularly important in increasing usage of new, innovative Census data products that are related to the NCRN research. For example, synthetic data (the SIPP Synthetic Beta and Synthetic Longitudinal Business Database datasets), have been available for several years, but the novelty of the data has limited its adoption by social scientists. The courses and the workshop organized by the Michigan node and supported by the Cornell node, described in online Appendix B, introduced graduate students and junior scholars interested in studying the causes and consequences of poverty using the synthetic SIPP data, and it culminated in a researcher-initiated panel at the 2016 American Social Science Associations-Labor and Employment Relations Association meeting.

The nodes have also taken on the task of creating software for others to use in both improving and analyzing federal datasets. The Colorado-Tennessee node developed open-source software for producing new statistical areas (out of existing census areas such as census blocks). This software reduces the variance in ACS estimates through intelligent aggregation.

The Cornell node produced software to edit Data Documentation Initiative (DDI)-formatted metadata, called the Comprehensive Extensible Data Documentation and Access Repository. No existing DDI editor could show the additional features that Cornell had incorporated into the existing (DDI-C) standard, thus requiring the creation of the editor to be able to edit and display the additional data. The 2018 version is CED²AR V2.9.0.

The Duke node has developed several R software packages implementing missing data techniques, including the stochastic edit-imputation for continuous data of Kim et al. (2015), the model for mixed categorical and continuous data of Murray and Reiter (2016), the non-ignorable imputation method of Paiva and Reiter (2017), and the model for categorical data with structural zeros of Manrique and Reiter (2014). It also developed software for generating synthetic values of the decennial census short form variables, using the methodology in Hu et al. (2018); the software ensures that structural zeros are respected (e.g., a daughter cannot be older than her biological father), and it captures within-household relationships.

The Michigan node developed software in STATA and SAS, and a related STATA command, to improve the standardization of employer names and thereby improve record-linkage software for businesses (Wasi and Flaaen 2015). It also improved software to impute tax liability to household surveys that are not linked to administrative data in order to compute the Census Bureau's alternative poverty measure.

The Missouri node is working on R software to implement customized geography and/or time periods (e.g., for the ACS). This software will automate the methodology of Bradley et al (2015). It is also collaborating with a private software company, Esri, on R software to quantify aggregation error from combining smaller geographies, allowing more efficient inferences (Bradley et al. 2017).

The Missouri node has developed R code for visualizing the uncertainty in (spatial) areal data. This software appears in the online supplement to Lucchesi and Wikle (2017) and in the VizU R package available on Github (https://github.com/pkuhnert/VizU).

The Nebraska node has developed a program to automate scrubbing of computer-assisted survey audit trails to ensure confidentiality of all text fields, implemented at the Census Bureau.

This program enabled release of thousands of audit trails by replacing costly and time-consuming human intervention with automated processes.

Links to the software listed, and other software products, can be found at https://www.ncrn.info/software.

REFERENCES – APPENDIX B

Belli, R. (2012), *Advanced Seminar – Interviewer-Respondent Interaction: Survey Research & Methodology Special Topics 898, Spring 2012*, Department of Sociology, University of Nebraska-Lincoln, https://digitalcommons.unl.edu/sociologyfacpub/490.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2015), "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates," *Stat*, 4, 255–270. https://doi.org/10.1002/sta4.94.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2017), "Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error.," *Journal of the Royal Statistical Society - Series B*, 79, 815–832. https://doi.org/10.1111/rssb.12179.

Eck, A. (2015a), *SRAM898 — Special Topics: Survey Informatics, UNL — Fall 2015 Course Syllabus*, Department of Sociology, University of Nebraska-Lincoln, https://digitalcommons.unl.edu/sociologyfacpub/489.

Eck, A. (2015b), "Teaching Survey Informatics for the Future of Survey Research," in *Annual meeting of the Midwest Association for Public Opinion Research*, Chicago IL.

Eck, A., Soh, L.-K., McCutcheon, A. L., and Belli, R. F. (2015a), "Predicting Breakoff Using Sequential Machine Learning Methods," in *Annual meeting of the American Association for Public Opinion Research*, Hollywood FL.

Eck, A., Soh, L.-K., Olson, K., McCutcheon, A. L., Smyth, J., and Belli, R. F. (2015b), "Understanding the Human Condition through Survey Informatics," *IEEE Computer*, 48, 110–114. https://doi.org/10.1109/MC.2015.327.

Flaaen, A. B. (2015), "Essays on Multinational Production and the Propagation of Shocks.," Ph.D., University of Michigan, http://hdl.handle.net/2027.42/111331.

Holan, S. H., Wikle, C. K., Bradley, J. R., Cressie, N., and Simpson, M. (2016), *Summary of "Workshop on Spatial and Spatio-Temporal Design and Analysis for Official Statistics.,"* NSF-Census Bureau Research Network, University of Missouri Node, , http://hdl.handle.net/1813/56543.

Hu, J., Reiter, J. P., and Wang, Q. (2018), "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data," *Bayesian Analysis*, 13, 183–200. https://doi.org/10.1214/16-BA1047.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015), "Simultaneous Editing and Imputation for Continuous Data.," *Journal of the American Statistical Association*, 110, 987–999.

Lucchesi, L. R., and Wikle, C. K. (2017), "Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation, and Glyph Rotation," *Stat*, 6, 292–302. https://doi.org/10.1002/sta4.150.

Manrique-Vallier, D., and Reiter, J. P. (2014), "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros," *Survey Methodology*, 125–134.

Murray, J. S., and Reiter, J. P. (2016), "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence.," *Journal of the American Statistical Association*, 111, 1466–1479.

Paiva, T., and Reiter, J. P. (2017), “Stop or Continue Data Collection: A Nonignorable Missing Data Approach to Continuous Data.,” *Journal of Official Statistics*, 33, 579–599.
Simmer, C., Perry, B., Barker, B. E., Vilhuber, L., and Brumsted, K. (2018), *CEDAR*, NCRN-Cornell Node. https://doi.org/10.5281/zenodo.597000.
Sorkin, I. (2015), “Ranking Firms Using Revealed Preference and Other Essays About Labor Markets.,” Ph.D., University of Michigan, http://hdl.handle.net/2027.42/116747.
Sorkin, I. (2018), “Ranking Firms Using Revealed Preference,” *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjy001.

ONLINE APPENDIX E: SPATIO-TEMPORAL HIERARCHICAL STATISTICAL MODELS

In this appendix, additional technical details are provided to illustrate one aspect of spatio-temporal modeling and analysis that the Missouri node has undertaken. Data sources in official statistics are often multivariate (contain a large number of variables), are spatially referenced, recorded over discrete time and contain multiple spatio-temporal scales. Adding to this complexity, the datasets are often extremely large (the so-called "big data" problem with millions of observations) and non-Gaussian. Taking advantage of the inherent dependence structure is essential for increasing the precision of desired estimates, especially in under-sampled or unsampled geographies.

The broad approach proposed by the Missouri node for modeling the complex data arising in official statistics settings can be cast in its most general form as a spatio-temporal mixed effects model. The spatio-temporal mixed effects model includes a fixed effects term that accounts for spatial or spatio-temporal covariates, and a random effects term that is typically formulated in terms of the sum of spatial or spatio-temporal basis functions and associated random coefficients. While it is conceptually straightforward, in practice specific modeling choices must be made with the intent of capturing dependence, while delivering computational feasibility. Model development proceeds through the hierarchical-statistical-model paradigm (e.g. Cressie and Wikle 2011; Holan and Wikle 2016), wherein the basic hierarchical model can be written as a "data model" and a "process model." If the parameters are estimated, the hierarchical model is called an empirical hierarchical model; if instead a "parameter model" (i.e., a prior) is posited, the hierarchical model is called a Bayesian hierarchical model. Borrowing notation from the hierarchical-modeling literature, consider random variables U and V where $[U|V]$ denotes the conditional distribution of U given V, and let Z be an n_Z-dimensional data

vector, Y be an n_Y-dimensional latent random vector, θ_D the data parameters, and θ_P the process parameters. Then, a basic hierarchical model can be specified by $[Z|Y, \theta_D]$ and $[Y|\theta_P]$, with the Bayesian hierarchical model also including $[\theta_P]$. From the discussion above, it is these models that are called the data model, the process model, and the parameter model, respectively. Most of the hierarchical-modeling research in the Missouri node has been of the Bayesian type although, in a precursor to NCRN research, Sengupta and Cressie (2013b) developed empirical hierarchical models for high-dimensional spatial count data using a Poisson data model. This work was followed by NCRN-supported research in Sengupta and Cressie (2013a), where the Poisson data model was generalized to an exponential-family data model. In the remainder of this appendix, the Bayesian hierarchical model will be featured.

For illustration, we proceed with a description of the multivariate spatio-temporal mixed effects model (Bradley et al. 2015a). This model was originally used to model public-use Quarterly Workforce Indicators (QWI) data from the Longitudinal Employer-Household Dynamics Program of the U.S. Census Bureau. The Quarterly data are at the county level for both genders and different North American Industry Classification Sectors (NAICS). 10/3/2018 9:47:00 AMFor $\ell = 1, \ldots, L$, $t = T_L^{(\ell)}, \ldots, T_U^{(\ell)}$, and $A \in D_{P,t}^{(\ell)}$, the data model is defined by

$$Z_t^{(\ell)}(A) = Y_t^{(\ell)}(A) + \epsilon_t^{(\ell)}(A),$$

where $\{Z_t^{(\ell)}: \ell = 1, \ldots, L\}$ represents multivariate spatio-temporal data; $Y_t^{(\ell)}$ represents the ℓ-th latent variable of interest at time *t*; *t* indexes discrete time; and $\epsilon_t^{(\ell)}(\cdot)$ is an iid Gaussian process with mean zero and known variance $v_t^{(\ell)}(\cdot)$. The set *A* represents a generic areal unit on the predictive domain, $D_{P,t}^{(\ell)}$, at time *t* for variable ℓ.

The process model is defined by

$$Y_t^{(\ell)}(A) = \mu_t^{(\ell)}(A) + \boldsymbol{S}_t^{(\ell)}(A)'\boldsymbol{\eta}_t + \boldsymbol{\xi}_t^{(\ell)}(A).$$

In this case, we set $\mu_t^{(\ell)}(\cdot) = x_t^{(\ell)}(\cdot)'\beta_t$,where $x_t^{(\ell)}$ is a known p-dimensional vector of covariates with associated unknown parameter vector β_t. In the process model above, $S_t^{(\ell)} \equiv \left(S_{t,1}^{(\ell)}, \ldots, S_{t,r}^{(\ell)}\right)'$, for $\ell = 1, \ldots, \mathrm{L}$, denote r-dimensional vectors of spatio-temporal basis functions, and $\{\xi_t^{(\ell)}\}$ represents fine-scale variability assumed to be i.i.d. with unknown variance, $\{\sigma_{\xi,t}^2\}$. In Bradley et al (2015a), these basis functions are specified to be the Moran's I (MI) basis functions. A rich class of areal basis functions was later introduced in Bradley et al. (2017b). For each t, it is assumed that the r-dimensional vector η_t follows a vector autoregressive process of order one; that is

$$\eta_t = M_t\eta_{t-1} + u_t,$$

where η_t is Gaussian with mean zero and unknown $r \times r$ covariance matrix K_t, M_t is an $r \times r$ propagator matrix, and u_t is Gaussian with mean zero and $r \times r$ covariance matrix W_t. After vectorizing $Y_t^{(l)}$ for $t = 1, \ldots, T$, by stacking, the process model can be rewritten to avoid spatial confounding. In fact, this representation leads to a modeling innovation referred to as the MI propagator matrix, which is defined analogously to the MI basis functions.

Due to issues with confounding, and because of the reduced-rank structure of the MI basis function and MI propagator matrix, various sources of variability may be inadvertently ignored. To address this concern, $\{K_t\}$ and $\{W_t\}$ are specified as positive-definite matrices that imply a spatio-temporal covariance matrix that is "close" to a target precision matrix that includes the various sources of variability. For comprehensive details, see Bradley et al. (2015a) and the references therein.

The methodology outlined above applies to Gaussian data. However, as previously

alluded to, many of the applications found in official statistics arise from non-Gaussian data. A typical approach to modeling such data is to specify a generalized linear mixed model using a latent Gaussian process (Diggle et al. 1998; Rue et al. 2009). That is, in the data-model specification, the Gaussian assumption would be replaced with a distribution from the exponential family. In high-dimensional settings, like those encountered in official statistics, estimation in the non-Gaussian setting is especially challenging. Sengupta and Cressie (2013a) give methodology in the spatial univariate empirical hierarchical model context. In the spatio-temporal multivariate Bayesian-hierarchical-model context, Bradley et al. (2017a, 2018) meet the challenge with new distribution theory that produces a latent conjugate multivariate distribution for the natural exponential family and then implements a multivariate spatio-temporal mixed effects model.

For example, in the case of a Poisson data model, a multivariate log-gamma distribution is proposed (Bradley et al. 2018). In particular, let the m-dimensional vector $w = (w_1, \ldots, w_m)'$ consist of m mutually independent log gamma random variables such that $w_i \sim LG(\alpha_i, \kappa_i)$ for $i = 1, \ldots, m$. Then, define

$$q = c + Vw,$$

where the $m \times m$ matrix $V \in \mathbb{R}^m \times \mathbb{R}^m$ and $c \in \mathbb{R}^m$. Then q is called a multivariate log gamma (MLG) random vector. For the sake of brevity, we do not include the expression of the pdf for the MLG random vector here; instead, for $\alpha \equiv (\alpha_1, \ldots, \alpha_m)'$ and $\kappa \equiv (\kappa_1, \ldots, \kappa_m)'$, we denote it as MLG$(c, V, \alpha, \kappa)$. Then, in the Gaussian process model, η and β are assumed to follow a MLG distribution and ξ_i $(i = 1, \ldots, m)$ is assumed to follow a log-gamma distribution. See Bradley et al. (2017a, 2018) for comprehensive details related to a Poisson data model and the natural exponential family data model cases, respectively.

The models described above are fully parametric. In principle, the classic Fay-Herriot nested error regression model for small area estimation can be thought of as a special case of the mixed effects models described above. In a spatial setting where it is of interest to relax the distributional assumption on the data model, one can take a semiparametric approach. Specifically, the data model can be specified using an empirical likelihood, and the process model can be specified as a latent Gaussian process. Detailed discussion of the semiparametric empirical likelihood approach can be found in Porter et al. (2015a; b).

Federal survey data are usually presented and analyzed over geographic regions. However, often inference is desired on a different spatial and/or temporal support than the support of the survey data. The problem of conducting statistical inference on spatial and/or temporal supports that differ from the support of the data is known as spatio-temporal change of support (ST-COS). The support of the data is typically referred to as the "source support" (e.g., census tracts), whereas the support of interest is designated as the "target support" (e.g., congressional districts). The majority of methodological contributions for spatial COS are based on assuming that the underlying data are Gaussian and consider spatial-only or count data without explicitly accounting for sampling uncertainty; see Bradley et al. (2016) and the references therein. Motivated by the problem of estimating discontinued 3-year period estimates for the ACS, Bradley et al. (2015b) present methodology that performs ST-COS for survey data with Gaussian sampling errors. In contrast, Bradley et al. (2016) propose methodology for count-valued data in which the change-of-support is accomplished by aggregation of a latent spatial point process that accounts for sampling uncertainty. Importantly, when changing spatial support, it is necessary to be concerned with the modifiable areal unit problem or MAUP (and the ecological fallacy). In other words, inferences made at one level of geography should be

consistent at other levels of geography. Bradley et al. (2017b) develop methods to determine when COS is appropriate, that is, when aggregation error is problematic. The proposed statistic is called *Criterion for Aggregation Error (CAGE).*

REFERENCES – APPENDIX C

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015a), "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics.," *Annals of Applied Statistics*, 9, 1761–1791.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2017a), *Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family.*, arXiv, , https://arxiv.org/abs/1701.07506.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2018), "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)," *Bayesian Analysis*, 13, 253--310. https://doi.org/10.1214/17-BA1069.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2015b), "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates," *Stat*, 4, 255–270. https://doi.org/10.1002/sta4.94.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016), "Bayesian Spatial Change of Support for Count-Valued Survey Data with Application to the American Community Survey.," *Journal of the American Statistical Association*, 111, 472–487.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2017b), "Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error.," *Journal of the Royal Statistical Society - Series B*, 79, 815–832. https://doi.org/10.1111/rssb.12179.

Cressie, N., and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: John Wiley and Sons.

Diggle, P. J., Moyeed, R. A., and Tawn, J. A. (1998), "Model□Based Geostatistics," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47, 299–350.

Holan, S. H., and Wikle, C. K. (2016), "Hierarchical Dynamic Generalized Linear Mixed Models for Discrete-Valued Spatio-Temporal Data," in *Handbook of Discrete–Valued Time Series*, eds. R. A. Davis, S. H. Holan, R. Lund, and N. Ravishanker, Boca Raton, FL: CRC Press, pp. 327--348.

Porter, A. T., Holan, S. H., and Wikle, C. K. (2015a), "Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models.," *Journal of Statistical Planning and Inference*, 165, 78–90.

Porter, A. T., Holan, S. H., and Wikle, C. K. (2015b), "Multivariate Spatial Hierarchical Bayesian Empirical Likelihood Methods for Small Area Estimation.," *STAT*, 4, 108–116.

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Sengupta, A., and Cressie, N. (2013a), "Hierarchical Statistical Modeling of Big Spatial Datasets Using the Exponential Family of Distributions.," *Spatial Statistics*, 4, 14--44.

https://doi.org/10.1016/j.spasta.2013.02.002.
Sengupta, A., and Cressie, N. (2013b), “Empirical Hierarchical Modelling for Count Data using the Spatial Random Effects Model,” *Spatial Economic Analysis*, 8, 389–418. https://doi.org/10.1080/17421772.2012.760135.

ONLINE APPENDIX D: SPATIAL VISUALIZATION

In this appendix we provide additional details related to the methodology provided in Lucchesi and Wikle (2017)10/3/2018 9:47:00 AM; note that it is not intended as an overview of spatial visualization. The simultaneous presentation of spatial data (or predictions) along with their uncertainties is important for conveying the quality of a spatial map. However, there has long been a concern that adding an uncertainty measure to a map will simply clutter the visualization and make the map more difficult to interpret (e.g., MacEachren et al. 2005). Uncertainty visualization for spatial and spatio-temporal data has been gaining increased attention from statisticians and is providing an opportunity to make use of new tools in statistical software (e.g. Genton et al. 2015). The Missouri node considered several tools to visualize the uncertainty of spatial data, including new formulations of (1) bivariate choropleth maps, (2) map pixelation, and (3) rotated glyphs, as described in Lucchesi and Wikle (2017). This appendix only discusses bivariate choropleth maps in detail, though illustrations of the other two techniques are shown.

The Census Bureau produced some of the first known bivariate choropleth maps in the late 1970s (Fienberg 1979; Olson 1981) 10/3/2018 9:47:00 AM. These maps were designed to visualize two variables, such as death rate and population density. However, they were somewhat controversial in that they were widely considered to be difficult to interpret (e.g. Wainer and Francolini 1980). Suggestions to improve these maps included limiting the color bins, selecting more interpretable colors, and adding more description to the map caption.

Bivariate choropleth maps have been typically used to visualize two variables; in contrast our interest is in visualizing a variable and its associated uncertainty. There have been previous attempts to perform such a visualization, for example using a diverging color scheme to

represent uncertainty and the relative contrast to represent the variable (e.g., Howard and MacEachren 1996). In addition, Retchless and Brewer (2016) used a 4 x 5 grid to represent the variable with color and its uncertainty with the saturation value of those colors. These are not choropleth maps.

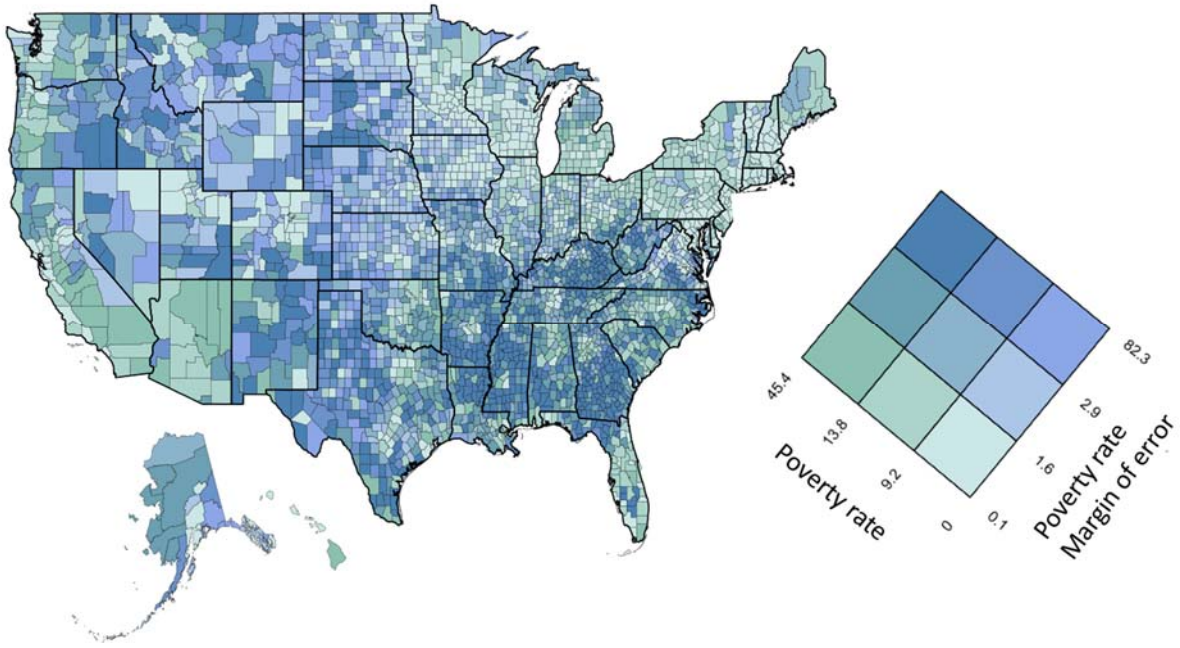
The bivariate-choropleth map approach that Lucchesi and Wikle (2017) developed is novel in that it visualizes uncertainty and improves visualization of traditional bivariate choropleth maps. In particular, they use a low-dimensional and interpretable 3 x 3 color scheme that is a natural additive blend of two single-hue red-green-blue color palettes. In addition, the associated key is rotated 45 degrees so that the highest values for both the variable and the uncertainty are at the top of the grid, which is easier to interpret.

This approach is demonstrated here using U.S. county-level poverty rates from the 2011-2015 ACS (see Figure E.1). In this case, each county is assigned one of nine colors depending on the poverty rate and the associated 90% margin of error (MOE). In this case, the counties with the lowest poverty rates and the smallest MOEs are represented by the lightest blue/green color at the bottom of the grid, which is an average of the lightest blue and lightest green color. In contrast, the darkest color is an average of the darkest blue and darkest green color, and it represents counties with the highest poverty rate and the largest MOE. Spatially contiguous clusters and trends in poverty rate and the associated MOEs are apparent in this map.

The VizU R package (https://github.com/pkuhnert/VizU) developed by P. Kuhnert and L. Lucchesi allows users to easily investigate different color palettes to aid in the interpretability of a particular map and its uncertainty. The package also allows for other spatial-uncertainty visualization approaches, including map pixelation (see Figure E.2), and glyph rotation (see Figure E.3). Note that the package also allows for the animation of the map pixelation to

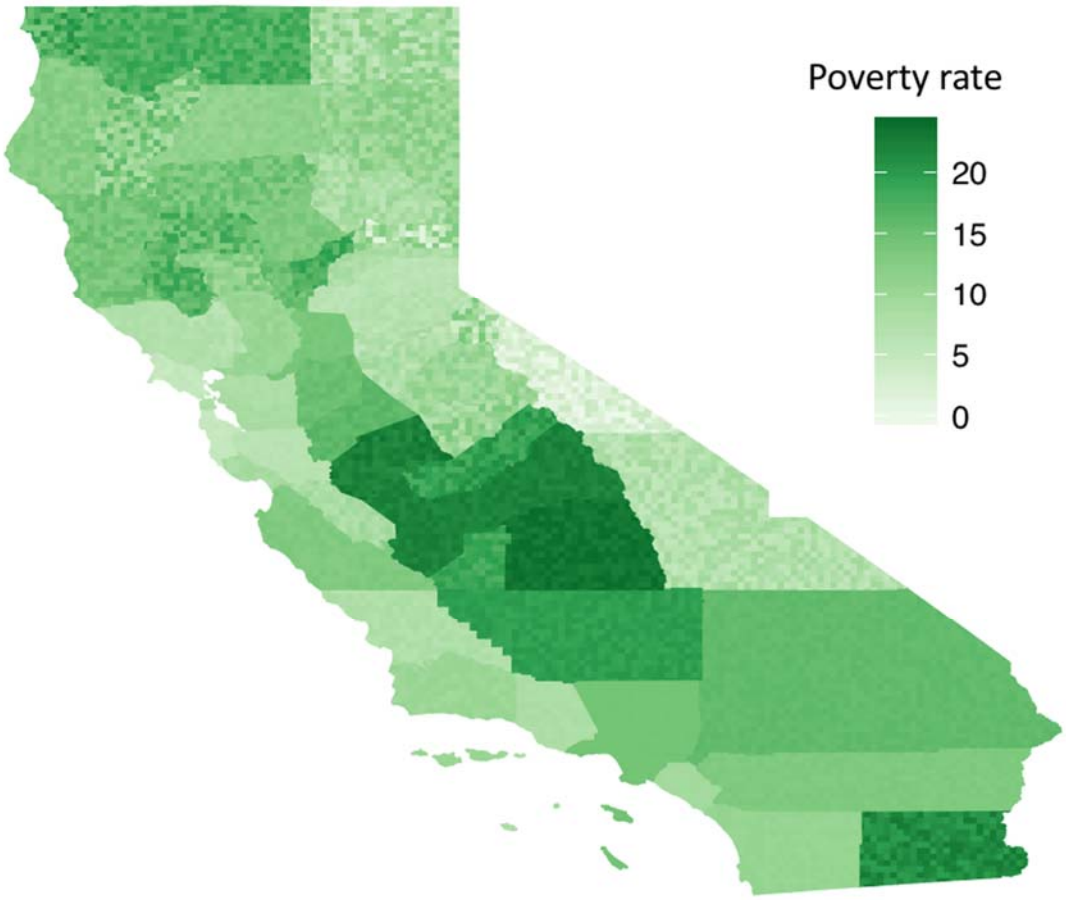
accentuate the uncertainty.

Figure D.1. U.S. county-level poverty estimates and their uncertainty, 2011-2015, using bivariate chloropleth map approach



Further details: The bivariate choropleth map shows U.S. county-level 2011-2015 American Community Survey poverty estimates (percentage of families whose income was below the poverty level) and associated uncertainties (90% margin of error, or MOE). The estimates and MOEs are divided into 3 categories by terciles. Each square in the 3 x 3 color key is an average of green, representing poverty rate, and blue, representing MOE.

Figure D.2. State of California county-level poverty estimates and their uncertainty, 2011-2015, using pixelated map approach



Further details: The pixelated map shows county-level 2011-2015 American Community Survey poverty estimates for California and their associated MOEs. Each pixel in a county is assigned a color within the county estimate's MOE. Areas of high uncertainty appear pixelated because the MOE covers a wide range of colors within the palette. Areas of low uncertainty appear smoother because the differences in color between pixels is much smaller.

Figure D.3. State of Colorado county-level poverty estimates and their uncertainty, 2011-2015, using glyph approach

Further details: The glyph map shows county-level 2011-2015 American Community Survey poverty estimates for Colorado and their associated MOEs. The color of each glyph represents the estimated poverty rate among families, and its rotation represents the estimate's MOE.

REFERENCES - APPENDIX D

Fienberg, S. E. (1979), "Graphical Methods in Statistics," *The American Statistician*, 33, 165–178.

Genton, M. G., Castruccio, S., Cripps, P., Dutta, S., Huser, R., Sun, Y., and Vettori, S. (2015), "Visuanimation in Statistics," *Stat*, 4, 81–96.

Howard, D., and MacEachren, A. M. (1996), "Interface Design for Geographic Visualization: Tools for Representing Reliability," *Cartography and Geographic Information Systems*, 23, 59–77.

Lucchesi, L. R. an. C. K. W. (2017), *Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation, and Glyph Rotation.* https://doi.org/10.1002/sta4.150.

Lucchesi, L. R., and Wikle, C. K. (2017), "Visualizing Uncertainty in Areal Data with Bivariate Choropleth Maps, Map Pixelation, and Glyph Rotation," *Stat*, 6, 292–302. https://doi.org/10.1002/sta4.150.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. (2005), "Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know," *Cartography and Geographic Information Science*, 32, 139–160.

Olson, J. M. (1981), "Spectrally Encoded Two-Variable Maps," *Annals of the Association of American Geographers*, 71, 259–276.

Retchless, D. P., and Brewer, C. A. (2016), "Guidance for Representing Uncertainty on Global

Temperature Change Maps," *International Journal of Climatology*, 36, 1143–1159.
Wainer, H., and Francolini, C. M. (1980), "An Empirical Inquiry Concerning Human Understanding of Two-Variable Color Maps," *The American Statistician*, 34, 81–93.

ONLINE APPENDIX E. ACTIVE AND IMPLEMENTED NCRN-FSS COLLABORATIONS BASED ON NCRN RESEARCH PUBLICATIONS

Below is a list of the research publications that have had a substantial impact on methods and activities at the U.S. Census Bureau. "Active collaboration" means that there is a current research project at the Census Bureau or another statistical agency based on this work, and one of the NCRN researchers is a current collaborator. "Implemented" means that techniques originally developed or elaborated in the cited research are being or have been engineered into at least one production system. Citations refer to the main article's reference list.

Active Collaborations (as of April 2018)

Belli et al. (2016)

Bradley et al. (2015a, b; 2016a, b; 2017a, c; forthcoming)

Flaaen et al. (2017)

Green et al. (2017)

Kirchner and Olson (2017)

Manrique-Vallier and Reiter (2018)

Olson and Smyth (2015)

Olson et al. (2016)

Olson et al. (forthcoming)

Porter et al. (2014, 2015c)

Quick et al. (2015a)

Seeskin and Spencer (2015, 2018)

Simpson et al. (2018)

Smyth and Olson (forthcoming)

Spielman and Folch (2015)

Sorkin (2016)

Steorts et al. (2016)

Wasi and Flaaen (2015)

White et al. (2018)

Wood et al. (2015)

Implemented Collaborations (as of April 2018)

Abowd et al. (2012)

Abowd and Schmutte (2016, 2017)

Chen et al. (2017)

Kim et al. (2015)

Kinney et al. (2011, 2014)

Lagoze et al. (2013a, b; 2014)

McKinney et al. (2017)

Miranda and Vilhuber (2016)

Murray and Reiter (2016)

Sadinle and Reiter (2017, 2018)

Vilhuber and Schmutte (2017a, b)

Vilhuber et al. (2016)