Introduction to the special issue: “Big data meets survey science” / Adam Eck, Ana Lucia Lucía Córdova Cazar, Mario Callegaro, & Paul Biemer

Abstract: Surveys have long been the primary means of data collection about peoples’ attitudes, beliefs, and opinions and are useful for measuring specific characteristics of individuals, as well as understanding public opinions and creating accurate and precise official statistics. Recently, artifacts of our increasingly digital lives have offered additional, broader information about our behaviors (e.g., purchase histories, personal interests captured through internet browsing) in the form of “big data”. Big data and surveys have great potential to complement one another (Baker, 2017; Callegaro & Yang, 2018) to allow scientists to better understand people and the world in which we live — for example, by combining the low cost per data point of big data (offsetting the rising costs of survey-based data collection) with the ability to collect very specific information addressing research questions using survey data.

Social media as an alternative to surveys of opinions about the economy / Fred Conrad, Johann Gagnon-Bartsch, Robyn Ferg, Michael Schober, Josh Pasek, & Elizabeth Hou

Social media as an alternative to surveys of opinions about the economy

Abstract: There is interest in using social media content to supplement or even substitute for survey data. O’Connor et al. (2010) report reasonably high correlations between the sentiment of tweets containing the word “jobs” and survey-based measures of consumer confidence in 2008-2009. Other researchers report a similar relationship through 2011 but after that time it is no longer observed, suggesting such tweets may not be as promising an alternative to survey responses as originally hoped. But, it’s possible that with the right analytic techniques, the sentiment of “jobs” tweets might still be an acceptable alternative. To explore this we first classify “jobs” tweets into categories whose content is either related to employment or not, to see if sentiment of the former correlates more highly with a survey-based measure of consumer sentiment. We then compare the relationship when sentiment is determined with traditional dictionary-based methods versus newer machine-learning based tools developed for Twitter-like texts. We calculated daily sentiment three different ways, and used a measure of association less sensitive to outliers than correlation. None of these approaches improved the size of the relationship in the original or more
recent data. We found that the many micro-decisions these analyses require, such as the size of the smoothing interval and the length of the lag between the two series, can significantly affect the outcomes. In the end, despite the earlier promise of tweets as an alternative to survey responses, we find no evidence that the original relationship in these data was more than a chance occurrence.

Where you at? Using GPS locations in an electronic time use diary study to derive functional locations / Anne Elevelt, Wim Bernasco, Peter Lugtig, Stijn Ruiter & Vera Toepoel
Abstract: Smartphones enable passive collection of sensor data alongside survey participation. Location data add context to people’s reports about their time use. In addition, linking GPS data to self-reported time use surveys can be valuable for understanding how people spend their time. This article investigates whether and how passive collection of geographical locations (coordinates) proves useful for deriving respondents’ functional locations. Participants of the ongoing Children of Immigrants Longitudinal Survey in the Netherlands (CILS4EU) were invited to participate in a time use survey administered with a smartphone app that also unobtrusively tracked respondents’ locations. Respondents reported their activities per 10-minute interval in a smartphone diary-app (N = 1339) and shared their geographical location data (N=1,264). The correspondence between the functional locations derived from the time use data and those derived from the geographical location data was assessed by calculating the percentage of intervals in which both measures are similar. Overall, results show that home locations can be automatically assigned reliably, but that respondent information is required to reliably assign work or school locations. In addition, location tracking data contain many measurement errors, making it difficult to record valid locations. Multilevel models show that the variability in correct classifications is intrapersonal and largely predicted by phone type, which determines location measurement frequency.

Capture-recapture techniques for transport survey estimate adjustment using permanently installed highway-sensors / Jonas Klingwort, Bart Buelens, & Rainer Schnell
Abstract: In this article survey, sensor, and administrative data are combined to correct for survey point estimate bias due to underreporting. The response to the Dutch Road Freight Transport Survey is linked to records from a road sensor network consisting of automated weighing stations installed on highways in the Netherlands. Capture-recapture methods are used to estimate underreporting in the survey. Heterogeneity of the vehicles with respect to capture and recapture probabilities is modelled through logistic regression and log-linear models. Six different estimators are discussed and compared. Results show a downward bias in the survey estimate due to underreporting, whereas the capture-recapture estimators yield larger estimates. This
research is a new example of multi-source statistics, a promising approach to improve the benefits of sensor data into the field of official statistics.

Income inequality and health: Expanding our understanding of state level effects by using a geospatial big data approach / Timothy Haithcoat, Eileen E. Avery, Kelly A. Bowers, Richard D. Hammer, & Chi-Ren Shyu

Abstract: Merging geospatial analytics with big data approaches provides a mechanism for leveraging and maximizing uses of traditional survey data to further extant work in meaningful ways. This study examines the income inequality hypothesis, which proposes that ecological (summary level) income inequality is harmful for population health. However, findings from extant work are inconsistent across health outcomes and levels of geography. We contribute to this debate by applying a big data geospatial approach to create three innovative measures that capture uniformity in income inequality across counties within U.S. states. Using data from the Behavioral Risk Factor Surveillance System and American Community Survey, we evaluate multilevel models of individuals within states to examine the ways that income inequality, operationalized as the Gini coefficient, and three spatial uniformity measures that capture the way income inequality is dispersed across space within states, are associated with several health outcomes. Specifically, the uniformity measures capture the extent to which 1) inequality is uniformly distributed spatially in states regardless of whether the level is high or low, 2) the extent to which states are more uniformly high in inequality across space, and 3) the extent to which they are more uniformly low in inequality. We conclude that state income inequality did not predict worse health across these outcomes (and indeed was associated with lower odds of depression and obesity). However, residents of states that have more uniformly high inequality across space are more likely to report below-average health, cardiovascular disease, difficulty concentrating, and that they have not sought care because it was too expensive. We conclude with a discussion of how a big data geospatial approach can further contribute to research on this, and other public health topics where scholars primarily rely on traditional survey data.

Automatic classification of open-ended questions: Check-all-that-apply questions / Matthias Schonlau, Hyukjun Gweon, & Marika Wenemark

Abstract: Text data from open-ended questions in surveys are challenging to analyze and are often ignored. Open-ended questions are important though because they do not constrain respondents’ answers. Where open-ended questions are necessary, often human coders manually code answers. When data sets are large, it is impractical or too costly to manually code all answer texts. Instead, text answers can be converted into
numerical variables and a statistical / machine learning algorithm can be trained on a subset of manually coded data. This statistical model is then used to predict the codes of the remainder. We consider open-ended questions where the answers are coded into multiple labels (all-that-apply questions). For example, in the open-ended question in our Happy example respondents are explicitly told they may list multiple things that make them happy. Algorithms for multi-label data take into account the correlation among the answer codes and may therefore give better prediction results. For example, when giving examples of civil disobedience, respondents talking about “minor non-violent offenses” were also likely to talk about “crimes”. We compare the performance of two different multi-label algorithms (RAKEL, CC) to the default method of binary relevance (BR) which applies single-label algorithms to each code separately.

Performance is evaluated on data from three open-ended questions (Happy, Civil Disobedience, and Immigrant). We found weak bivariate label correlations in the Happy data (90th percentile: 7.6%), and stronger bivariate label correlations in the Civil Disobedience (90th percentile: 17.2%) and Immigrant (90th percentile: 19.2%) data. For the data with stronger correlations we found both multi-label methods performed substantially better than BR using 0/1 loss (“at least one label is incorrect”) and had little effect when using Hamming loss (average error). For data with weak label correlations, we found no difference in performance between multi-label methods and BR. We conclude that automatic classification of open-ended questions that allow multiple answers may benefit from using multi-label algorithms for 0/1 loss. The degree of correlations among the labels may be a useful prognostic tool.

Additional articles not part of the special issue

Slider bars in multi-device web surveys / Angelica M. Maineri, Ivano Bison, & Ruud Luijkx

Abstract: This study explores some features of slider bars in the context of a multi-device web survey. Using data collected among the students of the University of Trento in 2015 and 2016 by means of two web surveys (n = 6,343, 4,124) including two experiments, we investigated the effect of the initial position of the handle and the presence of numeric labels on answers provided using slider bars. It emerged that the initial position of the handle affected answers, and that the number of rounded scores increased with numeric feedback. Smartphone respondents appeared more sensitive to the initial position of the handle, but also less affected by the presence of numeric labels, resulting in a lower tendency to rounding. Yet, outcomes on anchoring were inconclusive. Overall, no relevant differences have been detected between tablet and PC respondents. Understanding to what extent interactive and engaging tools such as slider bars can be successfully employed in multi-device surveys without affecting data quality is a key challenge for those who want
to exploit the potential of web-based and multi-device data collection without undermining the quality of measurement.

Reports and communications

Analyzing citizen participation and engagement in European smart cities / María E. Cortés-Cediel, Iván Cantador, & Manuel Pedro Rodríguez Bolívar

Abstract: With the advent of smart cities, governance has been placed at the core of the debate on how to create public value and achieve a high quality of life in urban environments. In particular, given that public value is rooted in democratic theory and new technologies that promote networking spaces have emerged, citizen participation represents one of the principal instruments to make government open and close to the citizenry needs. Participation in urban governance has undergone a great development: from the first postmodernist ideals of countering expert dominance to today’s focus on learning and social innovation, where citizen participation is conceptualized as co-creation and co-production. Despite this development, there is a lack of research to know how this new governance context is taking place in the smart city arena. Addressing this situation, in this paper we present an exhaustive survey of the research literature and a deep study of the experience in participative initiatives followed by smart cities in Europe. Through an analysis of 149 smart city initiatives from 76 European cities, we provide interesting insights about how participatory models have been introduced in the different areas and dimensions of the cities, how citizen engagement is promoted in smart city initiatives, and whether the so-called creative smart cities are those with a higher number of projects governed in a participatory way.

The dynamics of political elections: A big data analysis of intermedia framing between social media and news media / LO Wai Han, LAM Shu Yan Benson, & CHEUNG Mei Fung Meily

Abstract: The paper examines the news framing of the 2017 Hong Kong Chief Executive Election using a big data analysis approach. Analyses of intermedia framing of over 370,000 articles and comments are conducted, including news published in over 30 Chinese press media, four prominent Chinese online press media, and posts published on three candidates’ Facebook pages within the election period. The study contributes to the literature by examining the rarely discussed role of intermedia news framing, especially the relationship between legacy print media, online alternative news media, and audience comments on candidates’ social network sites. The data analysis provides evidence that audiences’ comments on candidates’ Facebook pages influenced legacy news coverage and online alternative news coverage. However, this study suggests that legacy news media and comments on Facebook do not necessarily have a reciprocal relationship. The implication of the findings and limitations are discussed.