Computational Statistics (2019) 34:1443–1447
https://doi.org/10.1007/s00180-019-00923-w

EDITORIAL

The 2013 Data Expo of the American Statistical Association

Heike Hofmann[1] · Hadley Wickham[2] · Dianne Cook[3]

Received: 6 September 2017 / Accepted: 19 September 2019 / Published online: 18 October 2019

1 Introduction

Since 1983, the Sections on Statistical Computing and Statistical Graphics of the American Statistical Association (ASA) have held a Data Exposition competition (usually called "Data Expo") as part of the Joint Statistical Meetings (JSM). These competitions presented participants with a data set and the challenge to produce a comprehensive analysis of the data. Entries were in poster form, with an emphasis on graphical presentation of the results. Early Data Expos were held roughly every two years, but there was a hiatus after 1997. At the 2006 JSM conference in Seattle, the Data Expo competition was revived. Since then, there have been competitions every several years again, with meteorological data from the National Aeronautics and Space Administration (NASA) in 2006 (Murrell 2010), airline ontime data from the Bureau of Transportation Statistics in 2009 (Wickham 2011), and data from several sources in relation to the Deepwater Horizon oil spill in 2011 (Cook 2014). A full list of competitions, data and winners can be found at http://stat-computing.org/dataexpo/. In 2013, the focus was on data from the Knight Foundation (http://www.knightfoundation.org/), using data collected by Gallup that examined the emotional attachment of people to their community. Twelve entries were submitted to the 2013 Data Expo competition, with the top entries invited to submit articles describing their analysis of the data.

✉ Dianne Cook
dicook@monash.edu

Heike Hofmann
hofmann@iastate.edu

Hadley Wickham
h.wickham@gmail.com

[1] Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA

[2] RStudio, Inc., Boston, USA

[3] Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia

2 Overview of the data

Why do some communities thrive while others do not? Are there specific community attributes that attract people to certain communities, tempt them to set down roots and commit to the community for the long term? If so, then this is valuable information for community leaders who wish to grow their communities by attracting both employers and employees and improving the local economic climate.

As part of the project "Soul of the Community" (SOTC), the Knight Foundation in cooperation with Gallup collected data from 43,000 people over three years (2008–2010) in 26 communities across the United States. The 26 communities did not constitute a random sample of communities across the United States. Participating communities were those where the Knight Foundation was already active: Aberdeen, SD; Akron, OH; Biloxi, MS; Boulder, CO; Bradenton, FL; Charlotte, NC; Columbia, SC;. Columbus, GA; Detroit, MI: Duluth, MN; Fort Wayne, IN; Gary, IN; Grand Forks, ND; Lexington, KY; Long Beach, CA; Macon, GA; Miami, FL; Milledgeville, GA; Myrtle Beach, SC; Palm Beach, FL; Philadelphia, PA; San Jose, CA; State College, PA; St. Paul, MN; San Jose, CA; Tallahassee, FL; and Wichita, KS.

The web site http://www.knightfoundation.org/sotc/ provides access to background information, results, and reports created by the Knight Foundation that are measuring the emotional attachment of people towards the community they live in. According to http://www.knightfoundation.org/sotc/about-knight-soul-community/, participants were surveyed in eleven different domains related to their personal feelings about their community:

- Aesthetics: physical beauty and green spaces
- Basic services: community infrastructure
- Civic involvement: residents' commitment to their community through voting or volunteerism
- Education systems
- Emotional wellness: the mixture of mental and physical well-being
- Leadership and elected officials
- Local economy
- Openness/welcomeness: how welcoming the community is to different people
- Safety
- Social offerings: opportunities for social interaction and citizen caring
- Social capital: social networks between residents

3 The Data Expo challenge

The aim of the 2013 Data Expo was to provide a graphical summary of important features of the SOTC data set. This was intentionally vague in order to allow different entries to focus on different aspects of the data, but there were a few ideas to get everyone started:

- What attaches people their community?

– What are key drivers behind emotional attachment? Are the key drivers all similarly important? What effect does their composition have on attachment?
– How different are the communities?

The 2013 Data Expo web site at http://stat-computing.org/dataexpo/2013/index.html provides access to the winning posters, the data files and data descriptions used for the competition, and some additional details related to the competition.

4 The winning entries

First place was awarded to Andee Kaplan and Eric Hare (Iowa State University) for "Putting down roots: A graphical exploration of community attachment" (Kaplan and Hare 2019). Their entry was unique in providing an interactive interface for viewers to explore the data in different ways, which you can see at http://andeek.shinyapps.io/CommuniD3. Three second places were awarded to Angela Minster (Temple University) for "Seeing the soul of the community"; Karsten Maurer, David Osthus (Iowa State University), and Adam Loy (Lawrence University) for "A tale of four cities: Exploring the soul of Biloxi, Detroit, Milledgeville, and State College" (Maurer et al. 2019); and Xiaofei (Susan) Wang, Cynthia Rush, and William Brinda (Yale University) for "Soul of the community".

Five more entrants also provided submissions for this special issue: Samuel Ackerman "Consistency of survey opinions and external data" (Ackerman 2019), Amelia McNamara "Community engagement and subgroup meta-knowledge: Some factors in the soul of a community" (McNamara 2019), Natalia da Silva and Ignacio Alvarez "Clicks and cliques. Exploring the soul of the community" (da Silva and Alvarez 2014), Anna Quach, Jürgen Symanzik, and Nicole Forsgren "Soul of the community: An attempt to assess attachment to a community" (Quach et al. 2019), and Jessica Orth "Drivers of community attachment: An interactive analysis" (Orth 2019). Early versions of two of the articles published in this special issue can be found on the 2013 JSM Proceedings CD (Orth 2013; Quach et al. 2013).

The approaches to tackling the data were quite varied. Kaplan and Hare (2019) facilitated an understanding of why people feel attachment to their communities through the use of interactive and web-based visualization, using the R package `Shiny` and the JavaScript library `D3`. Maurer et al. (2019) focused on four communities that stood out after their initial exploration of the data set: State College, PA; Detroit, MI; Milledgeville, GA; and Biloxi, MS. They used survey-weighted binned scatterplots to graphically explore the association between an individual's community attachment and the perceived economic outlook. Ackerman (2019) focused on one community, Long Beach, CA, describing the city in terms of geographic distribution, income race, and available resources. He found that ratings for the public schools appeared to be inconsistent, and proposed a solution to fix this. McNamara (2019) focused on the idea of community meta-knowledge, which is essentially majority group empathy or understanding of how minorities experience their community. Three minority groups were explored: seniors, families with young children, and racial minorities. da Silva and Alvarez (2014) used graphical, supervised, and unsupervised learning tools to

answer the primary research questions. Instead of only using the original data provided for the Data Expo, they used combined data from the U.S. Census Bureau and the Knight Foundation. Quach et al. (2019) provided an assessment via various machine learning algorithms (such as random forests, support vector machines (SVM), multiple linear discriminant analysis (LDA), and recursive partitioning and regression trees (RPART)) which factors may have an effect on attachment to a community. They also used an archetypal analysis to characterize communities with similar attachment status. Orth (2019) investigated several different approaches and methods to displaying multivariate data (such as time-series data and the results of multidimensional scaling and principal component analysis (PCA)), using the R package `googleVis`.

5 Reproducible research

A big difference between this special issue of *Computational Statistics* related to the 2013 Data Expo and past special issues related to previous versions of the Data Expo is that the analyses presented in the articles in this issue are supplemented with computer code at https://github.com/COSTDataExpo2013. This was the first experiment in reproducible research (Stodden et al. 2014; Gandrud 2015; Xie 2015) conducted for *Computational Statistics*. We wanted readers to not only see the final analyses but also to be able to understand how they were created. To this end, the github repository was constructed, the code was thoroughly tested and github issues were used to provide feedback to authors on code quality issues. Different approaches to software were used by the authors: da Silva and Alvarez (2014), McNamara (2019), and Quach et al. (2019) used the R package `knitr` (Xie 2014, 2015, 2016) to embed code with their articles, Orth (2019), Maurer et al. (2019), Ackerman (2019) provided R scripts, and Kaplan and Hare (2019) provided a web app written in JavaScript. It was a lengthy process checking code, but an interesting exercise.

References

Ackerman S (2019) Consistency of survey opinions and external data. Comput Stat 34(4). https://doi.org/10.1007/s00180-019-00882-2

Cook D (2014) The 2011 Data Expo of the American Statistical Association. Comput Stat 29(1–2):117–119

da Silva N, Alvarez I (2019) Clicks and cliques: Exploring the soul of the community. Comput Stat 34(4). https://doi.org/10.1007/s00180-019-00881-3

Gandrud C (2015) Reproducible research with R and RStudio, 2nd edn. Chapman and Hall/CRC, Boca Raton

Kaplan A, Hare E (2019) Putting down roots: a graphical exploration of community attachment. Comput Stat 34(4). https://doi.org/10.1007/s00180-018-0850-7

Maurer K, Osthus D, Loy A (2019) A tale of four cities: Exploring the soul of State College, Detroit. Milledgeville and Biloxi. Comput Stat 34(4). https://doi.org/10.1007/s00180-018-00863-x

McNamara A (2019) Community engagement and subgroup meta-knowledge: some factors in the soul of a community. Comput Stat 34(4). https://doi.org/10.1007/s00180-019-00879-x

Murrell P (2010) The 2006 Data Expo of the American Statistical Association. Comput Stat 25(4):551–554

Orth JM (2013) Dynamic graphics: an interactive analysis of what attaches people to their communities. In: 2013 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 3013–3025

Orth JM (2019) Drivers of community attachment: an interactive analysis. Comput Stat 34(4). https://doi.org/10.1007/s00180-018-00862-y
Quach A, Symanzik J, Forsgren Velasquez N (2013) Soul of the community: a first attempt to assess attachment to a community. In: 2013 JSM Proceedings, American Statistical Association, Alexandria, VA, pp 4053–4067
Quach A, Symanzik J, Forsgren N (2019) Soul of the community: an attempt to assess attachment to a community. Comput Stat 34(4). https://doi.org/10.1007/s00180-019-00866-2
Stodden V, Leisch F, Peng RD (eds) (2014) Implementing reproducible research. Chapman and Hall/CRC, Boca Raton
Wickham H (2011) ASA 2009 data expo. J Comput Graph Stat 20(2):281–283
Xie Y (2014) knitr: a comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD (eds) Implementing reproducible research. Chapman and Hall/CRC, Boca Raton, pp 3–31
Xie Y (2015) Dynamic documents with R and knitr, 2nd edn. Chapman and Hall/CRC, Boca Raton
Xie Y (2016) knitr: a general-purpose package for dynamic report generation in R. http://yihui.name/knitr/, R package version 1.15