Comput Stat (2014) 29:117–119
DOI 10.1007/s00180-013-0474-x

EDITORIAL

The 2011 data Expo of the American Statistical Association

Dianne Cook

Published online: 12 January 2014

1 Introduction

Since 1983, the Sections on Statistical Computing and Graphics of the American Statistical Association has held a Data Exposition competition 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 results. Early Data Expos were held roughly every 2 years, but there was a hiatus after 1997. At the 2006 Joint Statistical Meetings (JSM) conference in Seattle, the Data Expo competition was revived (Murrell 2010), with help from the Section on Statistics and the Environment, using a data set of atmospheric measurements from NASA. Since then there have been competitions every two years again. A full list of competitions, data and winners can be found at http://stat-computing.org/~dataexpo/. In 2011 the focus was on data made available in association with the BP oil spill of 2010. Twelve entries were submitted to the competition, with the four award winning invited to submit articles describing their analysis of the data. Three participants agreed and their articles follow.

2 Overview of the data

The BP oil spill arose from an explosion at the Deepwater Horizon rig, at the location 28.44°N, 88.23°W on April 20, 2010. There were three types of data available to participants, all taken from public data repositories: routine monitoring data of water temperature and salinity, water chemistry and affected wildlife counts.

The water temperature and salinity monitoring data was collected by the National Oceanic and Atmospheric Association (NOAA) using equipment on gliders, boats

D. Cook (✉)
Department of Statistics, Iowa State University, Ames, IA 50011-1210, USA
e-mail: dicook@iastate.edu

and floats (http://www.noaa.gov/sciencemissions/bpoilspill.html). In addition there was some very limited data on fisheries. These were the most extensive measurements taken, with equipment operating out into the Gulf waters on a reasonably regular basis.

The water chemistry data was collected by the Environmental Protection Agency (EPA) and contain measurements of petrochemical products. The sampling was limited to near coastal regions. Original data was drawn from this web site http://www.epa.gov/bpspill/download.html.

The wildlife data was collected by the US Fish and Wildlife Service (USFWS) and made available at http://gomex.erma.noaa.gov/erma.html. Reports of oiled birds, turtles and mammals were recorded, and whether they were alive or dead. The collection was not systematic.

The publicly available data was very limited, and reveals some serious inadequacies in the collection systems. It would have made far more sense if the NOAA equipment could also measure water chemistry where they were measuring temperature and salinity. Then it may have been possible to determine the extent of the oil spill, and perhaps even determine if salinity and temperature could adequately estimate oil contaminant levels. But the EPA was restricted to measuring only areas accessible on the coast, and NOAA was restricted to measuring out in the Gulf. We were told this was customary that different government organizations could not overlap in their missions. In addition, the wildlife observations were purely observational data—no systematic sampling was conducted, and no baselines for usual counts of wildlife were provided. We were also told that more measuring was conducted by both the Federal Government and BP but for legal reasons each party kept their data private. The sparseness made analyzing the data a huge challenge.

3 The data Expo challenge

The aim of the data Expo is to provide a graphical summary of important features of the data set. The guidelines are intentionally vague in order to encourage participants to focus on different aspects of the data. These were a few questions posed about the BP oil spill data at the time of the competition:

- Are the extents of the oil spill visible in measurements on temperature and salinity?
- Are the temperature and salinity measurements consistent between measuring devices?
- Is there a spatiotemporal pattern in the temperature and salinity measurements that might indicate presence of oil?
- Can you find the disappearing oil?
- Is there evidence of contamination in the fisheries?
- What locations along the coastline are most in danger of contamination from oil?
- What species of birds were the most affected and where were they when found?

4 The winning entries

The 2011 entrants were primarily graduate, and undergraduate students, which differed substantially from previous years where academic faculty and researchers from indus-

try submitted entries. Three prizes were awarded, along with one entry being highly commended. The first prize winner was Lendie Follett, an undergraduate student from Iowa State University advised by Heike Hofmann and Ulrike Genschel. Second prize went to Walter Hickey and Bimal Parakkal, from William and Mary College, and third prize to Aida Yazdanparast and Tony Tran, California State University, East Bay. An honorable mention was given to Tianxi Li, Stanford University.

Three of the four groups elected to submit a paper for publication on their data analyses. Lendie Follett's paper "A Graphical Exploration of the Deepwater Horizon Oil Spill" (Follett et al. 2014) provides a comprehensive look at all of the data, with well-designed graphics produced using the ggplot2 package in R. You will see an approach to using the temperature and salinity to estimate the oil extent, and an excellent discussion of the issues arising from observational only measurements on the impact on the wildlife. The paper by Tran, Yazdanparast and Suess "Effect of Oil Spill on Birds: A Graphical Assay of the Deepwater Horizon Oil Spill's Impact on Birds" (Tran et al. 2014) focus on the impact on the wildlife, and utilized Google Fusion Tables and Maps to explore the data. The paper by Li, Gao, Xu and Rajaratnam "Detecting the Impact Area of BP Deepwater Horizon Oil Discharge: An Analysis by Time Varying Coefficient Logistic Models and Boosted Trees" (Li et al. 2014) illustrates how modeling can be used to understand the spread of the oil.

References

Follett L, Genschel U, Hofmann H (2014) A graphical exploration of the deepwater horizon oil spill. Comput Stat 29. doi:10.1007/s00180-013-0432-7

Li T, Gao C, Xu M, Rajaratnam B (2014) Detecting the impact area of BP deepwater horizon oil discharge: an analysis by time varying coefficient logistic models and boosted trees. Comput Stat 29. doi:10.1007/s00180-013-0449-y

Murrell P (2010) The 2006 data Expo of the American Statistical Association. Comput Stat 25:551–554. doi:10.1007/s00180-010-0207-3

Tran T, Yazdanparast A, Suess EA (2014) Effect of oil spill on birds: a graphical assay of the deepwater horizon oil spill's impact on birds. Comput Stat 29. doi:10.1007/s00180-013-0472-z