Beyond Normal: Preparing Undergraduates for the Work Force in a Statistical Consulting Capstone

Byran J. Smucker[1] *and A. John Bailer*[1]

[1] Byran Smucker is corresponding author and Assistant Professor in the Department of Statistics at Miami University, Oxford, OH 45056 (e-mail: smuckerb@miamioh.edu). A. John Bailer is University Distinguished Professor and Chair of the Department of Statistics at Miami University, Oxford, OH 45056. The students in the data practicum classes in the spring semesters of 2010 and 2013 conducted the analyses summarized in the case studies reported in this paper and the figures included in this manuscript were extracted from their client reports. The Oxford Parking & Transportation Advisory Board and William Renwick were the clients who supplied the projects described herein.

Abstract

In this article we chronicle the evolution of the undergraduate statistical consulting course at Miami University, from canned to client-based projects, and argue that if the course is well designed and the students suitably mentored, they can perform remarkably sophisticated analyses for real-world problems that require solutions beyond the methods encountered in previous classes. We illustrate this with two case studies of such projects and relate this class model to the skills demanded in the modern statistical workforce.

KEY WORDS: Statistical education; case study; workforce preparation

1. Introduction

In recent years, it has become common to integrate a statistical consulting experience into the undergraduate statistics curriculum. Early courses (e.g. Spurrier, 2001; Miami University until approximately 2010) focused on canned projects, which allowed the instructor to control the structure of the class as well as the range of statistical methods encountered. Use of canned labs promotes the development of writing, presentation, and teamwork skills, but implies a known solution that the instructor possesses. Moreover, this model avoids unleashing undergraduates on real data—with all of the associated complications and subtleties in both the analysis and its communication—and denies them the opportunity to experience statistical consulting in its native state. More recent literature pushes back against this idea by emphasizing the importance of exposing students in such a class to the realistic-but-difficult reality associated with formulating a statistical problem, cleaning the data, and dealing with the inevitable difficulties that don't tend to arise in projects that have already been analyzed (Jersky, 2002; Mackisack and Petocz, 2002; Taplin, 2003, Boomer et al., 2007; Hooks and Malone, 2012; Kim et al., 2014).

In this article, we highlight and extend the latter idea by suggesting that under the appropriate faculty guidance, undergraduate students can sometimes succeed in surprising and sophisticated ways when challenged to undertake projects that require a significant investment in learning new computing or statistical techniques to adequately handle the data at hand. We first give an overview of the background of undergraduate consulting at Miami, and enlarge on the current approach to statistical consulting. We then discuss two specific case studies that have arisen in this context, in which undergraduates navigated challenging projects exceptionally. Finally, we consider the difficulties that accompany the model we advocate for, and conclude with some remarks relating it to workforce preparation.

2. History and Background: Undergraduate Consulting in Statistics at Miami

The graduate version of the Data Practicum class at Miami began in 1973, in response to recognition that consulting was often a part of a working statistician's job description, and an exposure to "real world" problems would provide valuable experience. The quotation marks indicate that despite this motivation, the early versions of the course were labs derived from consulting problems that had been solved previously by staff in Miami's statistical consulting center.

The undergraduate version of the course was developed around 1994, was inspired by and modeled after its graduate cousin, and serves as a senior capstone for statistics majors and statistical methods minors from a variety of majors (e.g. psychology). An important difference between the courses is the statistical preparation expected of the students. In the graduate course, all students will have at least regression, experimental design, and statistical programming while students in the undergraduate version may have no more than a statistical modeling course to supplement the introductory course they have taken. This creates a challenging-but-mutually beneficial environment that includes students with a wide range of

statistical and programming backgrounds, but students also with a variety of perspectives on data, since the minors often have more exposure to data as encountered in their disciplines.

In Spring of 2010, the second author transitioned the undergraduate practicum class from canned labs to real projects after having done the same with the graduate class years previously. This change allows critical engagement of the students with clients and the client-defined tasks, and also requires that the students work to understand the client's subject matter, which is often key to understanding the underlying statistical questions. The new model also mimics the sort of environment in which most working statisticians find themselves—collaborative and iterative—which is crucial experience that can be used to great advantage in an increasingly competitive analytics job market. Indeed, students coming through these courses often highlight their practicum projects in their résumés.

Clients are solicited for the course from a list of contacts at Miami University. Most clients are internal, though occasionally we reach beyond the university, and hope to strike a balance between internal and external clients in the future. The course begins with some introductory material about statistical consulting, and an initial, simple, canned project is assigned for which they are to write a draft report. Feedback is given on this first draft, to set their expectations and convey important ideas for future reports, including the importance of structuring the report, revision, critical reading and graphical displays. The students find that the course instructor is likely to be the worst client they will encounter during the semester. A second assignment sometimes included is to have the class propose a mechanism to form teams with the only constraints that all students should have an opportunity to lead a team at least once each semester and that team membership should rotate between client projects. The students team assignment plan is reviewed and considered in light of special skills that are desired for each team such as making sure that at least one member of every team has had the statistical programming class.

A project life cycle goes something like this: (1) The client presents his/her project to the class, and is questioned by the faculty member and students; (2) Teams are formed to work on the project (this may be the whole class if the project is of sufficient size, or part of the class if that is all the project warrants); (3) Informal feedback via meetings with the instructor and reports to the class; (4) Draft report produced and given extensive feedback by instructor; (5) Final report and presentation to client.

In the following two sections, we present two challenging projects that undergraduates in this class have undertaken. They illustrate the extent to which these students can perform in a demanding consulting environment.

3. Case Study: Parking in a Small College Town

3.1 Motivation

Oxford, Ohio is the home of the main campus of Miami University. In addition to 17,700 students, the town has 21,000 other residents. In recent years, new mixed use spaces (retail on first floor and residential rental spaces on upper floors) were being built throughout town. It should come as no surprise that parking is at a premium along High Street, Oxford's main thoroughfare (see Figure 1), and that pressures for spaces were anticipated to increase.

The data practicum class was asked to conduct an occupancy study of the various parking areas in town so that the Oxford Parking & Transportation Advisory Board (OPTAB) could make data-based decisions regarding meter rates, hours, fines, etc. The OPTAB represents nearly every constituency in town: City Council, Miami University, Chamber of Commerce, Oxford Landlords, etc., and include as ex-officio members the Chief of Police and the City Manager.

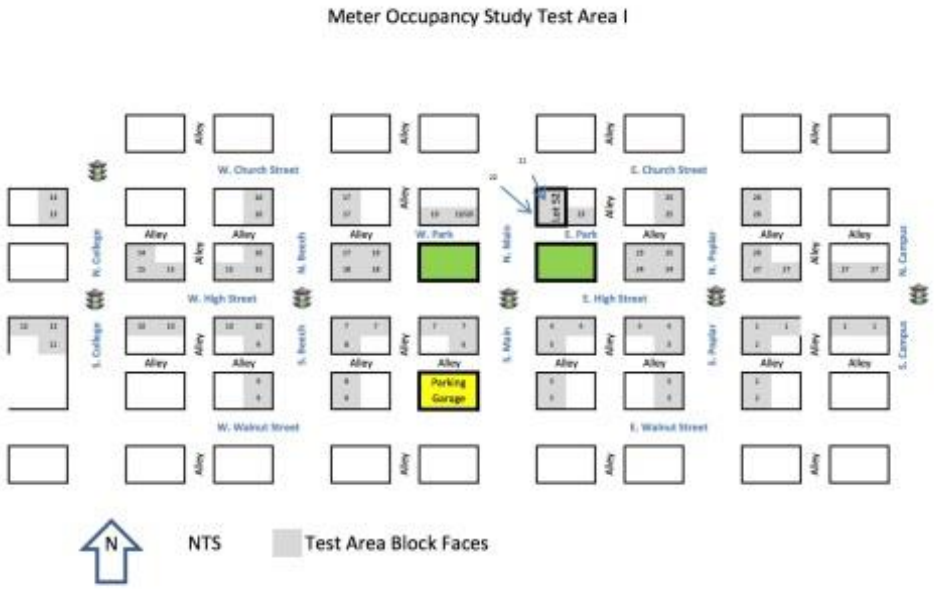

Figure 1. Parking in uptown Oxford. Figure extracted from STA 475 class presentation to client in Spring 2010.

3.2 Problem Statement

The liaison to OPTAB was a police officer, and he helped frame the problem. When a block face (i.e., the collection of parking spaces on a street for a city block) exceeds 85% occupancy it appears full. If the remaining 15% of block face spaces are vacant, this leads to a loss of earnings (if these are metered spaces). Additionally, the city parking garage has low occupancy.

The problem for the data practicum class was to conduct an analysis of a sample of the 790 metered spots in Oxford as well as the parking garage. The goal was to investigate different rates based on location to spread occupancy from "hot spots" to outer locations as well as into the unoccupied city garage. In addition, "meter feeding"—occupants of parking spaces extending their time for periods longer than the maximum time by adding coins to the meter before the time expires—could be removing spaces from

circulation for extended periods of time. Detecting the occurrence of such behavior was also of interest. A second issue of interest was extending meter hours past the current 6 p.m. deadline which could force many vehicles to park away from the High and Main intersection as well as produce additional income for the City of Oxford.

3.3 Solution

The class restricted attention to 27 block faces and the parking garage for a total of 377 parking spaces. The students developed a data collection plan that was implemented in March of 2010. Four days (Monday, Thursday, Friday and Saturday) during a one week period during the semester were sampled hourly (11 a.m.-3 p.m.; 4 p.m.-8 p.m.). Other days were omitted because parking occupancy patterns on Tuesdays and Wednesdays were believed to be similar to Monday and meter fees did not need to be paid on Sundays. Each of the 14 students in this class collected data on at least two occasions. Logistics of the data collection (order that spaces were checked—see Figure 2; data collection sheets; decisions about what information to record, such as license numbers) and appropriate documentation and emergency numbers (phone number of police contact if challenged when reviewing meter status) were key components of preparation before entering the field to collect the data.

For each metered spot, occupancy was recorded along with whether the meter was in violation. Summaries of the collected data include:

- Proportion of spaces occupied/available by time, day and block, on street parking or in garage
- Proportion of spaces with violations by time, day and block, on street parking or in garage
- Proportion of spaces with evidence of meter feeding

Figure 2: Paths for checking spaces for hourly data collection. The red (blue) path designates the route taken to check occupancy on the southern (northern) block faces. Figure extracted from STA 475 class presentation to client in Spring 2010.

Results reported to the OPTAB clients included a heat map of occupancy that was animated for display to the client, a presentation originated fully from the students; this proved to be a powerful depiction of how the occupancy varied over time (see Figure 3 for one map that was included in the animated set).

Other tables and graphs include occupancy/violation rate per day in block, occupancy/violation rate per day in garage, and length of stay. Plots faceted by day and block face provided insight into underutilized spaces (e.g. Figure 4).

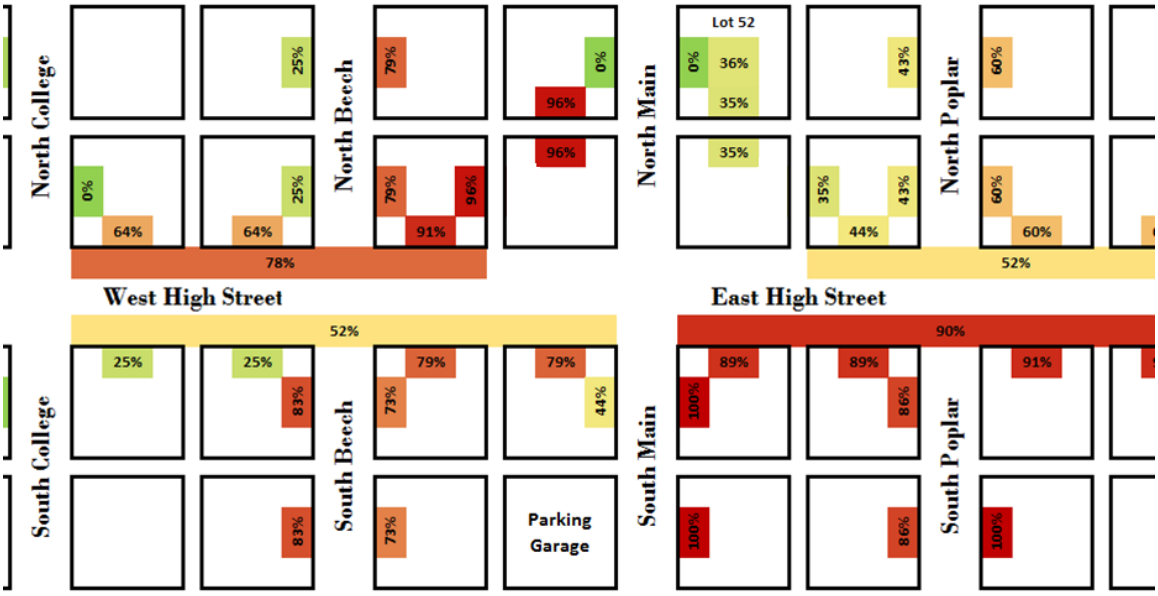

Figure 3: Example of Heat Map of occupancy for one sampling time (Monday at 1 p.m.) where red signifies high occupancy, yellow moderate occupancy and green low occupancy for each block face. Figure extracted from STA 475 class presentation to client in Spring 2010.

3.4 Impact and Follow-up

Students presented the results of this data collection and analysis effort to the OPTAB. The board was very impressed with the depth, quality and insight provided by the students, and mentioned that these data would be relevant for setting parking meter rates that might include differential rates for underutilized spaces.

The OPTAB members also commented that this level of work may have exceeded the value of a previous report that cost over $20,000 to conduct. At this point, the students were left speechless.

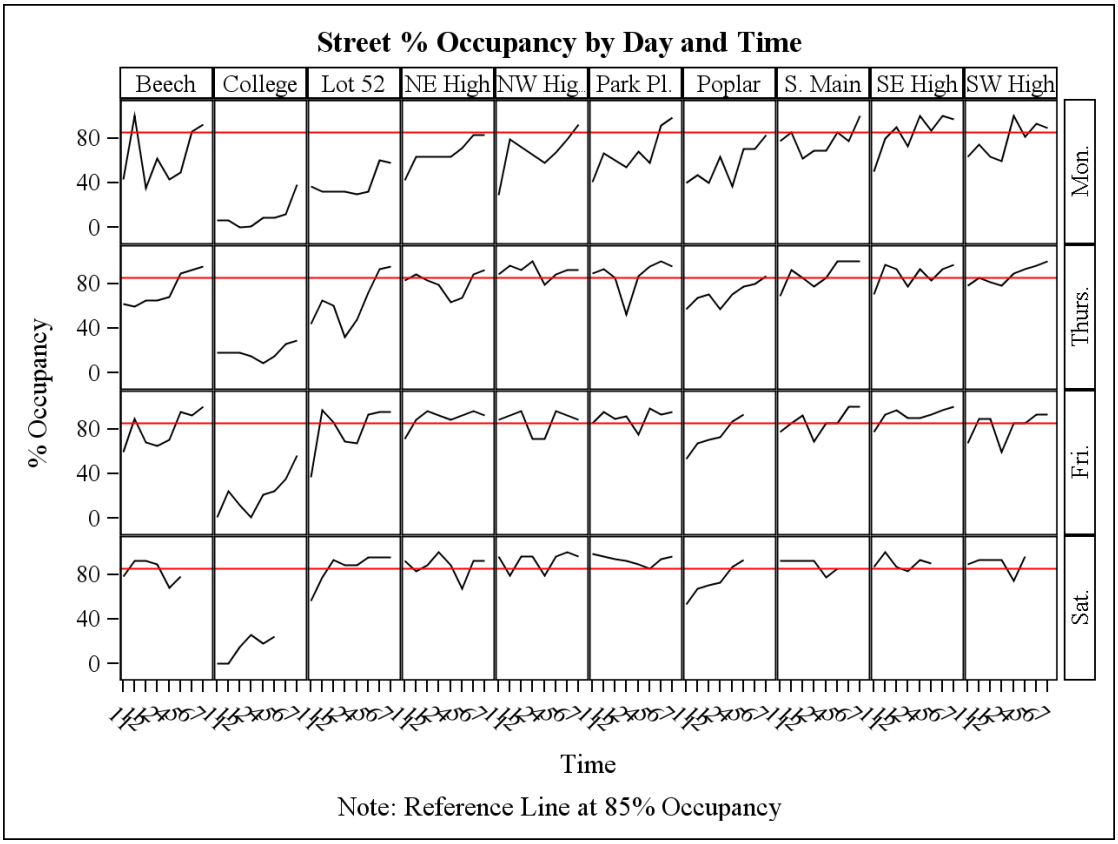

Figure 4: Occupancy of block face over time on different days for different days (rows) and for different block faces (columns). A horizontal red reference line at 85% occupancy is included in each graph. Figure extracted from STA 475 class presentation to client in Spring 2010.

In addition to the experience of working with a local government organization and contributing to information needed to support decision making, the students had to coordinate an extensive data collection effort, process the data into an analysis data set, construct displays and summaries that gave insight and develop a presentation and report that captured all of this work. The students did an outstanding job with this effort and were praised accordingly by the client. The comment about the cost of previous parking studies led to an interesting class exercise in which the students developed a cost estimate of how much billable work was reflected in the analysis, and what they would charge if they did this as a consulting company.

4. Case Study: Analysis of Reservoir Sedimentation Rates

4.1 Motivation

Reservoirs store water and are critical for human activities like drinking, irrigation, recreation, and flood control. Over time, these bodies of water tend to accrue sediment due to a variety of factors, and this reduces reservoir capacity. Of interest to geographers, then, is the rate at which the sediment accumulates in reservoirs around the U.S., in order to monitor erosion and evaluate water supplies.

4.2 Data Description

Dr. Bill Renwick, Professor of Geography at Miami University, was the client for this project in early 2013. The publicly-available dataset of interest included about 3,900 observations on roughly 1,900 different reservoirs, collected between 1755 and 1992. This analysis undertaken was something of a trial run, since a batch of newer data was expected to be made available soon. Within the dataset, five variables are of interest (Table 1).

A measurement in this dataset specifies a particular reservoir (RESSED ID) in a particular region (HUC2; Figure 5), and includes an estimated sedimentation rate (SedRate) that is calculated based upon estimated reservoir volumes measured at two different times (specified by Duration). A sedimentation rate is associated with the midpoint between the beginning and ending measurements (MidYear).

Table 1. List of variables, descriptions, and purpose of each variable in the model. Adapted from STA 475 written report to client in Spring 2013.

Variable Name	Description	Purpose in Model
MidYear (M)	Midpoint between beginning and ending measurements	Predictor
HUC2 (R)	Geographic region (1 to 18)	Predictor (Categorical)
SedRate (S)	Sedimentation yield (in cubic meters of sediment per square kilometers of drainage area per year)	Response
RESSED ID	Reservoir identification number	Correlation structure
Duration (D)	Time between beginning and ending measurements (in years)	Weighting

4.3 Problem Statement

There is a general assumption among reservoir managers that modern sedimentation rates are more-or-less unchanged when compared to the past. However, some hypothesize that due to improved environmental conservation, sedimentation rates have decreased. The goal of this project was to model the sedimentation rate as a function of year, to determine if there is a significant change in sedimentation rates across different regions within the United States.

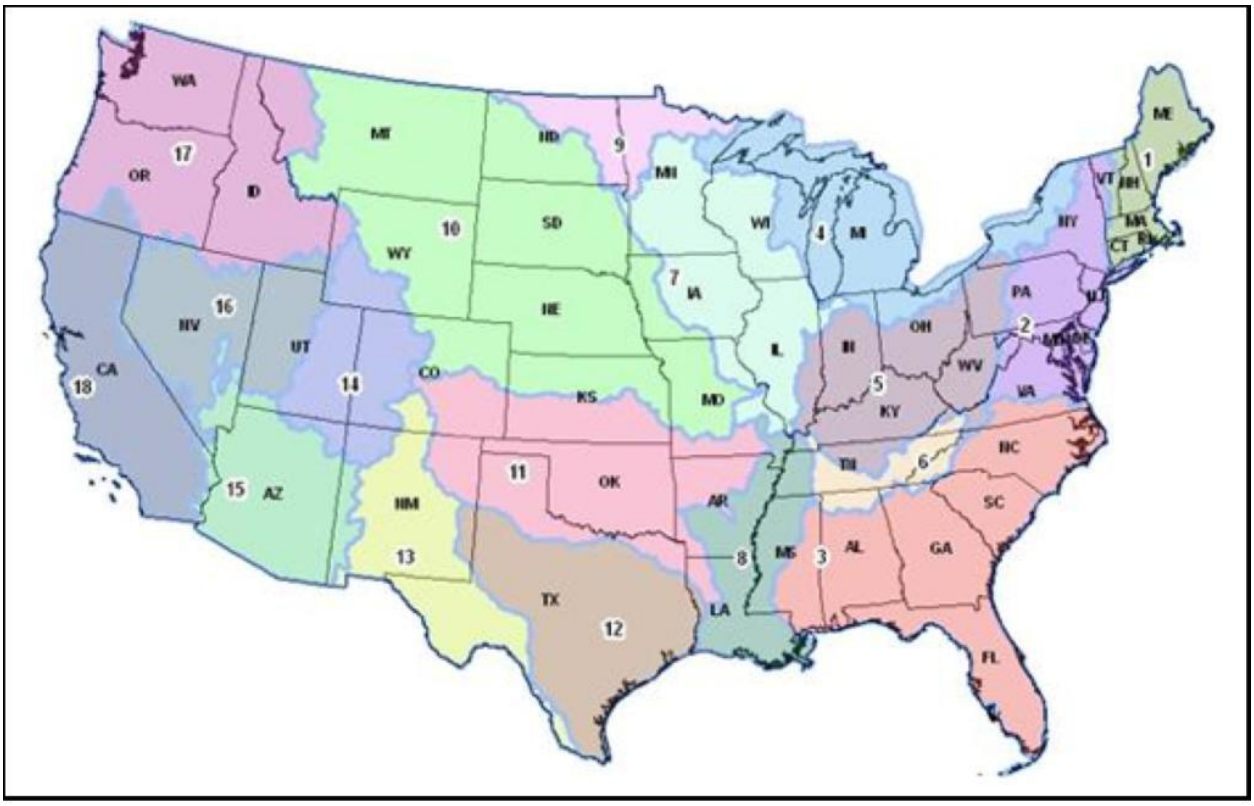

Figure 5. Hydrologic Unit Code (HUC2) Map (Jones et al., 2010). Each shading indicates one of the 18 regions.

4.4 Solution

The challenges presented in this project were considerable. The dataset was messy, including duplicate and conflicting observations that had to be resolved. There were also at least four critical complications that precluded a straightforward, standard analysis of sedimentation rate regressed on time, for each region.

First, the response was highly skewed due to the natural bound of 0 on the sedimentation rate. This issue was largely remediated by a log transformation (though, since the bound of 0 was not inviolable, nonpositive observations were omitted based on the argument made by Dr. Renwick that it is unusual for a reservoir to naturally exhibit a negative sedimentation rate and more likely to be the result of a man-

made intervention such as dredging). Scatterplots of the log-transformed sedimentation rates reveal many regions with apparently increasing trends (Figure 6).

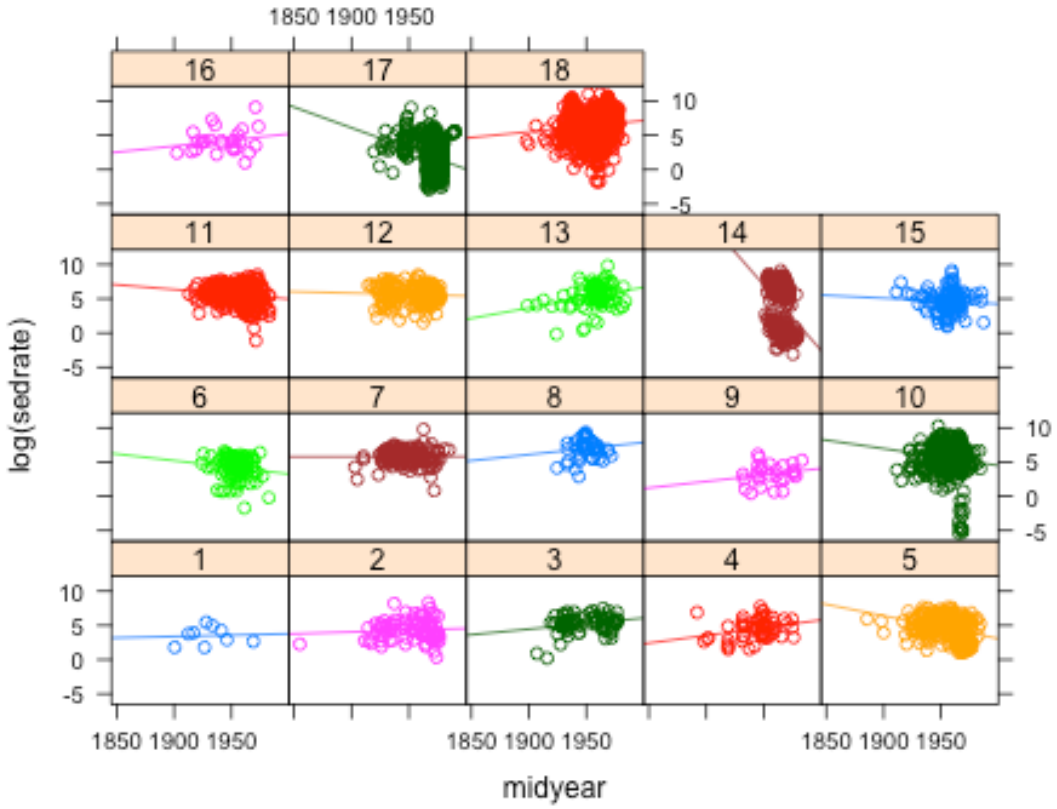

Figure 6. Scatterplots of the log of the sedimentation rates versus MidYear (with simple linear regression line superimposed), for regions 1-18. Taken from STA 475 written report to client in Spring 2013.

The second difficulty had to do with how the sedimentation rate of a reservoir was measured: At two different times, sometimes years apart, the volume of a particular reservoir was calculated and these two measurements used to estimate the sedimentation rate. As is clear from Figure 7, when this duration is smaller the response is much more variable (this pattern persists even when the sedimentation rate was log-transformed). This suggests that observations should be weighted as a function of the duration.

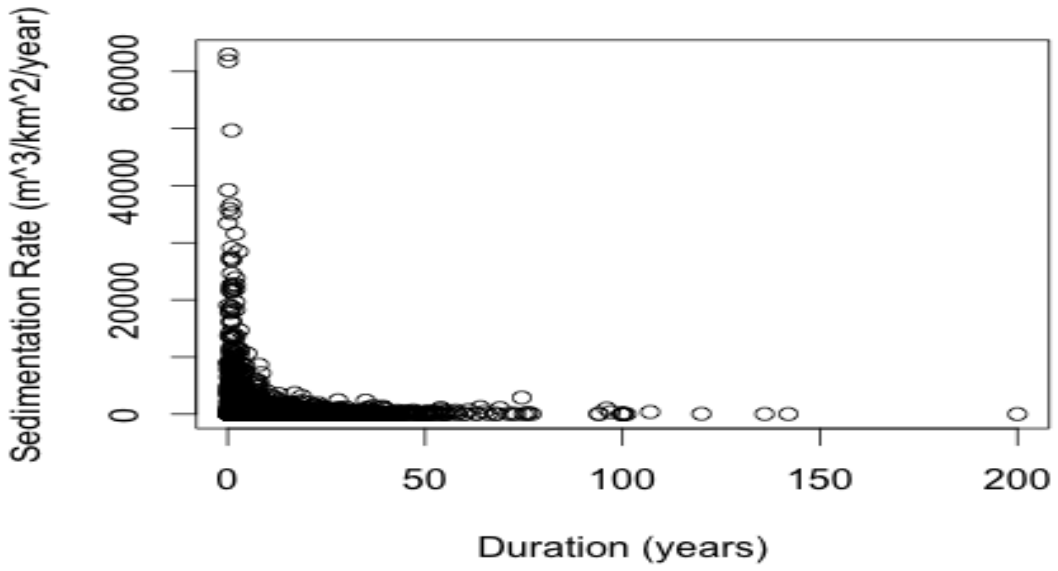

Figure 7. Sedimentation rate versus the duration used to calculate the sedimentation rate. Taken from STA 475 written report to client in Spring 2013.

The third important complication is that many reservoirs were measured multiple times, meaning that the correlation between observations on the same reservoir needed to be incorporated into the model. (The simplifying assumption was made, however, that observations from different reservoirs were independent.) None of the students in the class had ever fit a model with such structure.

Finally, multiple comparisons is an issue that should not be ignored, since there are 18 regions for which inference is desired.

The model fit was

$$\log(S_{jk}) = \sum_{i=1}^{18} \beta_i R_{ijk} + \sum_{i=1}^{18} \beta_{i+18} R_{ijk} M_{jk} + \varepsilon_{jk}$$

where

- j represents reservoir j and k represents the kth measurement taken on the jth reservoir;
- R_{ijk} is an indicator variable for Region i;
- errors are $\varepsilon_{jk} \sim N\left(0, \sigma^2 \left|D_{jk}\right|^{2\delta}\right)$, with D_{jk} the difference between the beginning and end of the sedimentation rate measurement (the "duration"), and σ^2 and δ are parameters estimated from the data.
- correlation between multiple measurements on reservoir j is $corr\left(\varepsilon_{jk}, \varepsilon_{jk'}\right) = \phi^s$ where ϕ is estimated from the data and s is the amount of time between measurements.

The model was parameterized to allow for direct hypothesis tests on each region slope parameter in the form of $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$ for $i=19,20,\ldots,36$. It was fit using the `gls` function in the `nlme` R package (Pinheiro et al., 2014).

The complications in this analysis posed significant challenges for the undergraduate consultants, and these issues required them to carefully learn several aspects of statistical modeling with which they were unfamiliar, while concurrently building up the model in an unfamiliar R function. Then, they had to demonstrate their adequate knowledge by writing clearly about these statistical issues.

4.5 Results

This model produced a standardized residuals vs. fitted plot (Figure 8) that gave a reasonable level of confidence in the subsequent inference. Upon fitting the model, there were eight regions with p-values less than $\alpha = 0.05$, indicating possible changes in sedimentation rates over time. However, once the p-values were adjusted to account for the multiple hypothesis tests, four regions (Mid-Atlantic; South-Atlantic Gulf; Rio Grande; California) exhibited a slope parameter significantly different from 0. The

procedure due to Holm (Holm, 1979), a method to control the family-wise error rate that is more powerful than the Bonferroni correction, was used to adjust for multiple tests.

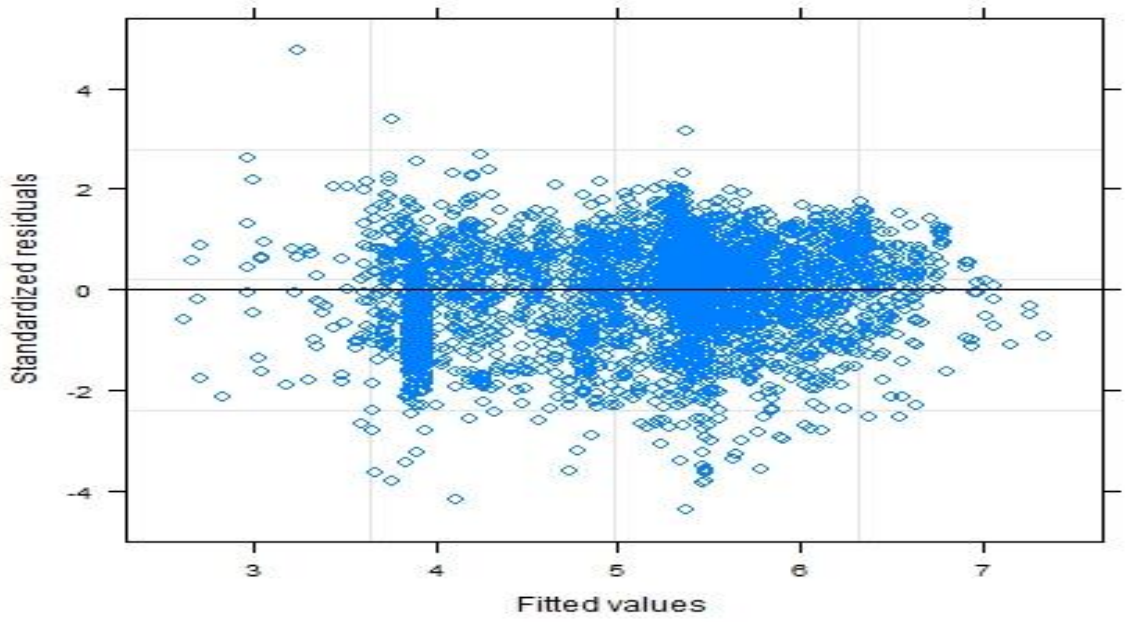

Figure 8. Standardized residuals vs. fitted values for the final model. Taken from STA 475 written report to client in Spring 2013.

Interestingly, each of the four significant slope parameters were positive, implying that the sedimentation rate was *increasing* across time, in contrast to the initial hypothesis. Though establishing which relationships exist and their direction was the primary objective of this analysis, the slope estimates can also be interpreted. For instance, the significant slope estimate for the Mid-Atlantic region was 0.017, which can be interpreted roughly as follows: For this region, a one year increase is associated with an average change in sedimentation rate by about a factor of $e^{0.017} = 1.017$, i.e. a 1.7% change.

4.6 Limitations

In most consulting projects, there are caveats that temper confidence in the results. In this case, one limitation was the elimination of the sedimentation rates less than or equal to 0, in order to facilitate the log transformation. Though there are scientific justifications for this move, it did reduce the sample size by several hundred. A possible work-around would be to add a constant to all responses in order to ensure that they all are positive. This would not affect the slope parameter estimates and would allow the use of the entire dataset.

Another limitation is that correlation between neighboring reservoirs was ignored. Though the HUC2 region predictor accounts in a very crude way for similarities in regions, it is possible that reservoirs in close proximity might be associated with each other. This could be accounted for by incorporating a spatial covariance component, but likely would have made the project much more challenging.

5. Discussion and Conclusions

In this article we have described the undergraduate consulting class at Miami University, and presented two case studies that illustrate the excellent work that can come from students even in the face of formidable projects that require the assimilation, execution, and presentation of statistical material with which they are unfamiliar. This class model has been successful in summoning the best from engaged students. There are, however, several recurring challenges.

First, not all students are engaged. There are many possible reasons for this, from senioritis to language barriers to personality. Since the class is based upon group work, a lack of engagement on the part of a minority of students can cause tensions within groups, and it is possible for some students to slip by without pulling their weight. This is ameliorated somewhat by measures like peer review of group

members that may affect the student's grade, but this remains a difficult, structural problem associated with any course that includes a substantial group component.

Another challenge is the fluid and overlapping nature of the projects. Though students routinely juggle assignments and material from different classes, many find it difficult to involve themselves deeply in more than one consulting project at a time. The incongruity might be handled by pointing out the multitasking they already do, and also comparing it to a professional environment that may require a similar balancing act.

Finally, the role of the instructor in this class is very different from traditional classes or even from the role played in a traditional canned lab consulting class. The instructor needs to solicit projects for the class and needs to gauge the difficulty of projects up front. For instance, the sedimentation rate was welcomed as the first project of the semester! There was no gentle ramp-up in this particular class, which was less than ideal. This suggests that the instructor might expend more effort to understand the projects ahead of time, though this militates somewhat against the philosophy of the class, which is to charge students address whatever project they are presented with. In addition to generating clients for the class, the instructor serves as research manager and senior consultant for the project. As part of this role, the instructor needs to provide just-in-time instruction for concepts that the students have not learned in previous classes (e.g. fitting regression models with correlated errors, producing faceted graphical displays, assistance with data processing to produce an analysis data set).

Employers consistently indicate that career success factor skills—such as leadership, teamwork, written and verbal communication competencies particularly aimed at a nontechnical audience—as promoted by recent ASA presidential initiatives and affirmed in curriculum guidelines, are those most in demand from students. They often hasten to add that technical skills are required as well, but their manner clearly indicates, perhaps because in their experience the technical skills are stronger than the nontechnical ones,

an emphasis upon those qualities that are more intangible and general-purpose. The practicum class plays an indispensable role in Miami's undergraduate program in developing these soft skills. Most of these characteristics could be fostered in the old-style consulting class environment, but the element that pushes this course to the next level—beyond normal—is the presentation of messy, unstructured, undefined statistical problems by content-specialist clients. Consequently, students report that potential employers are often more interested in the experience working on practicum projects than other course experience because it demonstrates an ability to perform useful statistical analysis in a real-world, non-academic setting. Furthermore, since the clients care about their projects, it imbues the students with satisfaction, because their work is relevant and valued.

References

Boomer, KB, Rogness, N., and Jersky, B. (2007), "Statistical Consulting Courses for Undergraduates: Fortune or Folly?" *Journal of Statistics Education*, 15(3). Available at www.amstat.org/publications/jse/v15n3/boomer.html

Holm, S. Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2) (1979), 65-70.

Hooks, T. and Malone, C. (2012), "Involving Undergraduates in Statistical Consulting," *The First Biennial Electronic Conference on Teaching Statistics (eCOTS 2012)*. Available at https://www.causeweb.org/ecots/ecots12/posters/16/.

Jersky, B. (2002), "Statistical Consulting with Undergraduates – a Community Outreach Approach," *International Conference on Teaching Statistics (ICOTS6)*. Available at http://icots6.haifa.ac.il/PAPERS/3A1_JERS.PDF.

Jones, R.D., Costello, K., Hetrick, J., Lin, J., Parker, R., Thurman, N., Peck, C., and Orrick, G. (2010), "Development and Use of Percent Cropped Area Adjustment Factors in Drinking Water Exposure Assessments," Office of Pesticide Programs, Environment Protection Agency. Available at http://www.epa.gov/pesticides/science/efed/models/water/archives/pca_adjustment_dwa_9sep2010.html.

Kim, H.-J., Alberts, K.S., and Thatcher, S. (2014), "Teaching Undergraduates Through Statistical Consulting," *International Conference on Teaching Statistics (ICOTS9)*. Available at http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C195_KIM.pdf.

Mackisack, M. and Petocz, P. (2002), "Projects for Advanced Undergraduates," *International Conference on Teaching Statistics (ICOTS6)*. Available at http://iase-web.org/documents/papers/icots6/3e4_peto.pdf.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014), *nlme: Linear and Nonlinear Mixed Effects Models*. Available at http://CRAN.R-project.org/package=nlme.

Spurrier, J (2001), "A Capstone Course for Undergraduate Statistics Majors," *Journal of Statistics Education*, 9(1). Available at http://www.amstat.org/publications/jse/v9n1/spurrier.html.

Taplin, R (2003), "Teaching Statistical Consulting Before Statistical Methodology," *Australian & New Zealand Journal of Statistics*, 45(2): 141-152.