Teaching

Overall Teaching Philosophy: My teaching philosophy is appropriately summarized by an old Chinese proverb which states, “Give a man a fish and you feed him for a day. Teach a man to fish and you feed him for a lifetime.” Great teachers from history such as Jesus Christ, Charles Darwin, Martin Luther King Jr. and others are revered because the impact of their teachings on actions, understanding, and outlooks spanned lifetimes. In my opinion, taking a course at a university is a microcosm of lifetime learning. Students are either intellectually fed for a semester or students are given proper tools which facilitate an enriched lifetime of learning. My goal as a teacher is not to stimulate a student’s mind for a lecture, day, or semester, but to instill knowledge, appreciation, understanding, and confidence that can last a lifetime.

Self Assessment:

- Strengths:

  1. My greatest strength as a teacher stems from my love of the subject matter. I love statistics and I try to radiate that enthusiasm in my teaching. The majority of students who take statistics are required to do so (rather than by choice). Hence, attitudes towards the subject matter range from strongly apprehensive to curious. Exuding my personal excitement for statistics, however, allows me to capture the interest of all students and help them appreciate the value of a statistical education.

  2. I try to engage students during the lecture. On the first day of each class, I have students fill of a 3×5 card detailing their background, interest level, anxiety level, and expectations for the course. Then, rather than simply ask for volunteers to answer a question I pose in class, I will call on students by name in class using these cards. This use of index cards allows me to not only learn students names (a difficult task for large class sizes) but also ensures that the students are paying attention because they don’t know when they’ll be called on to answer a question.

  3. I design my courses to mimic job-like circumstances. Academia can be a different world from industry. Oftentimes, the problems posed in an academic setting are overly simplified and don’t reflect a “real-life” circumstance. For my classes, I strive to bridge the gap between academia and industry by centering my course around real applications. For example, each of my lectures and homework assignments in my Statistics 536 course centered around case studies where advanced statistical methods were required to answer complex questions using real data.

  4. I provide students specific feedback on their learning. I strongly believe that getting the “right” answer is not the goal of an education (oftentimes, there isn’t a “right” answer). Rather, I believe that the goal of an education is to learn how to use the mind
and spirit to guide action. To support this philosophy, I not only require students to show/justify their answers but I also strive to provide justification for deducting points on assignments, exams, and course projects. This allows students the opportunity to express their thinking and receive constructive feedback on where their understanding was incorrect. To this end, I always have a grading rubric that I supply to my students so they know exactly what was wrong and why it was wrong.

- Weaknesses

1. One common theme from the teacher ratings for the 2013-2014 academic year was that I didn’t incorporate gospel topics into my lectures. Admittedly, as a teacher of statistics, this is a difficult task to accomplish without getting sidetracked from the course content. I always started my lectures with a prayer but, to the students credit, they expect more of the gospel to be incorporated into lecture than this. This weakness will be a focal point for the 2014-2015 academic year (see goals below).

2. I teach at a fast pace. Admittedly, this can be both a strength and a weakness. It’s a strength in that it stretches students to keep up but is also a weakness is that I often gloss over material on which I should spend more time. I need to learn what material students find difficult (and which they find simple) to appropriately pace my teaching.

Report on 2013-2014 Academic Year Teaching Goals: The 2013-2014 year was my first year and, as such, I did not have any specific goals related to teaching. For this academic year, I primarily focused on learning how to teach by attending other professors lectures, attending the faculty development luncheons and discussion teaching strategies with Dr. Kenneth Plummer in the Center for Teaching and Learning (CTL).

2014-2015 Academic Year Teaching Goals: My teaching goals and plans for achieving these goals are as follows:

1. Attend and complete the faculty development series. During Fall semester 2013, I was involved in the bi-weekly luncheons held by the Faculty Center. Currently, I am registered to attend the Spring 2014 development series.

2. Invite my faculty mentor to review my student ratings and attend lectures to evaluate my teaching. After the Fall 2013 semester, I reviewed my student rating comments with Dr. Plummer. I found his feedback to be very constructive in helping me find patterns in the various comments. I plan to continue this method of feedback by reviewing my ratings with the faculty mentor.

3. Develop and implement new ideas for bringing gospel topics into my lecture. The weak point from my Fall 2013 students ratings was that I didn’t frequently incorporate principles from the gospel into my lecture. I plan to discuss ways to do this with my fellow faculty members and chair. Hopefully, these discussions will lead to new ways/ideas that will spiritually uplift the students in a non-religious class.

4. Engage in one-on-one learning through mentoring. I plan to mentor a graduate student on his/her Master’s thesis. At the time of writing, I have already solicited research opportunities and am currently looking for interested to students to mentor.
Scholarship

Research Overview: In the environmental and atmospheric sciences, epidemiology, ecology and many other fields of study, data are often collected at specific spatio-temporal coordinates (e.g., from the environmental sciences, pollution concentrations are measured at several monitoring stations over a specific time domain). Data collected in this manner often exhibit strong dependence in space and time. Thus, proper analysis of such data requires that this dependence be accounted for in the statistical modeling framework. While accounting for such dependence is often challenging, my primary research interests are in exploiting this dependence to enhance statistical inferences of multiple variables. For example, utilizing spatio-temporal dependence to enhance predictive accuracy or uncover complex non-linear spatio-temporal relationships between spatial variables.

My research is primarily published in statistics journal including Journal of the Royal Statistical Society Series C, Journal of Agricultural, Biological and Environmental Sciences, Environmetrics and Biostatistics, to name a few. I also have applied research published in Spatial and Spatio-temporal Epidemiology and Food Chemistry.

Self Assessment:

- Strengths:
  1. I feel my greatest strength as a researcher is my desire to collaborate with others. This desire has led me to reach out to various professors, both within my department and elsewhere, for research collaboration opportunities. I currently have several projects with professors in my department as well as collaborations with faculty members in the Departments of Health Sciences, Life Sciences and Plant and Wildlife Science.
  2. I have lots of ideas. Since graduating with my Ph.D. I have kept a “Next Steps” document detailing research ideas and next steps on various research projects. This document allows me to always have ideas to work on.

- Weaknesses
  1. I am interested in many projects. As such, I often overbook myself in research so I don’t get enough time to sit and finish a project. I need to learn to discipline myself to finish projects and submit them for publication in a timely manner.
  2. I have little experience with grant writing. The funding landscape is vast and confusing. Having never been an active participant in grant writing, I desperately need more experience in this area.

Report on 2013-2014 Academic Year Scholarship Goals: Since arriving at BYU in the Fall of 2013, in the area of scholarship, I:

1. Submitted 4 articles for peer review to statistics and applied journals. Two of these were accepted for publication and two are still under review.

2. Submitted two grant proposals. The first was submitted to the BYU ORCA office and for which I was a Co-PI. The second was a collaborative research proposal submitted to the NSF for which I was the PI for the BYU portion. The ORCA grant was rejected in January 2014 and the NSF grant is still under review.
3. Gave four presentations on my statistical research. Three of these were in academic conferences or seminars (NOLTA13 in Santa Fe, University of Connecticut Department of Civil and Environmental Engineering Seminar, and BYU Department of Statistics Seminar) and one was a consulting presentation to the Federal Highway Administration.

2014-2015 Academic Year Scholarship Goals: My scholarship goals and plans for achieving these goals are as follows:

1. Maintain a “next steps” document with research ideas and directions. I have already created this document but, in order to accomplish this goal, need to continually update its content.

2. Submit three first-author papers for peer review. Currently, I have two well-defined projects that should near completion in 2014 leaving one additional “new” project that needs to be defined and carried-out during the course of the academic year. In order to accomplish this goal I need to schedule large blocks of time to research.

3. Submit a grant proposal to the NIH. Currently, the NIH has a FOA titled “Spatial Uncertainty: Data, Modeling and Communication” which is well suited to my research interests. I plan to write a proposal and submit it in June of 2014.

4. Give at least two presentation of my research in academic settings. I currently have 1 presentation schedule for August at the Joint Statistical Meetings. There are also several conferences during the summer in which I could present research.

5. Actively recruit a graduate student to participate in my research. To accomplish this goal I need to be proactive in announcing research opportunities to students.
Citizenship

Overall Citizenship Philosophy: I believe a great department is built by faculty who are willing to serve within the department, in their respective professional societies and across the academic environment. A willingness to serve shows a greater concern for the whole instead of the individual.

Self Assessment:

• Strengths:

1. I am actively involved in departmental service. This is reflected in my involvement in the department recruiting and undergraduate curriculum committees.
2. I am actively involved in professional service. I am frequently asked to review journal articles that are under consideration for publication. I have also been nominated to serve as a treasurer for the Environmental section if the International Society of Bayesian Analysis.

• Weaknesses

1. My willingness to serve can also be my weakness in that I rarely turn down opportunities to do peer review or serve on a committee. This, occasionally, results in me being burdened with other responsibilities which then hinders my research progress. I need to be disciplined to know when to accept and when to decline service opportunities.

Report on 2013-2014 Academic Year Citizenship Goals: Since arriving at BYU in the Fall of 2013, in the area of citizenship, I:

1. Have served on the recruiting and undergraduate curriculum committees.
2. Reviewed 3 articles for professional journals.
3. Served on the thesis committee for one Masters student.

2014-2015 Academic Year Citizenship Goals: My citizenship goals are as follows:

1. Act as a reviewer for 4 peer-reviewed articles.
2. Seek opportunities to serve as a committee member on a student thesis.
3. Be an active participant in my department committee assignments.
4. Attend the weekly collegiality lunch with faculty.
1 Course Background

Name: Statistics 536: Modern Regression Methods

Purpose: To develop and present necessary statistical tools to promote statistical learning from complex data sets.

Characteristics: This class is part of the core curriculum in the MS program in statistics. I am the only professor who teaches the course. While I took over the course from another professor, changes in focus and emphases of the statistical profession warranted that the course receive a major restructuring from previous methods used in the course. It was also the general opinion of the statistics department faculty that the course needed to have a shift in focus to realign it with current demands of employers and the statistical profession.

2 Learning Outcomes

Learning Outcomes (LOs): My goal for students upon completing this course is to have the skills to analyze nearly any dataset they may encounter on the job. To this end, I expect that students by the end of this class will be able to:

1. Appropriately explore data to determine an appropriate statistical model.
2. Posit & explain an appropriate statistical model that answers questions related to a dataset.
3. Fit the posited statistical model to the data using statistical software.
4. Appropriately present model results and conclusions from the statistical analysis.
5. Be comfortable submitting a written or oral report of a statistical analysis.

Approach to Teaching LOs: Each of the above learning outcomes is a necessary step to performing a proper statistical analysis. Therefore, this class will focus on developing the skills to analyze any dataset that the student might be given on the job. My approach to the course is to teach the learning outcomes in parallel (i.e. simultaneously) but with increasing depth as the semester progresses. That is, early semester assignments (see the Course Activities section below) will be simpler data analysis projects but allow students to focus on all the above learning outcomes simultaneously. Later semester assignments will still focus on all the above learning outcomes simultaneously but on far more complicated data analysis examples.

Connection to Program LOs: This course supports the MS program learning outcomes by providing students with experience as a statistical researcher or consultant. Namely, this course gives students experience working with real data and presenting results of a statistical analysis in written and oral form.
3 Course Activities

Case Studies. This course is built around case studies. Case studies are real-life data sets from various clients which require statistical expertise to analyze. Students are required to utilize the techniques learned in class to analyze the case studies. Students submit a written or oral report of their analysis in teams of up to 2 students. Case studies are of increasing in complexity as the semester progresses. Hence, students are required to demonstrate their ability to achieve the LOs with increasing depth and complexity.

Each case study focuses on all of the 5 LOs simultaneously. That is, in successfully completing a case study, students gain experience in each of the LOs. Final written or oral reports will demonstrate a student’s understanding and ability to perform each of the LOs. Feedback on each case study will be based on the grading rubric which focuses on the learning outcomes.

Midterm and Final Case Study. The midterm and final case studies will be follow the same format as the homework case studies; however, students will be required to work alone.

As with the homework case studies, the midterm and final case study again focus on all 5 LOs simultaneously. In contrast, by working alone, students are able to gauge their personal understanding and ability to perform the LOs. The same grading rubric as was used on the homework case studies is used on the midterm and final case studies. The grading rubric focuses and grades students on their performance on each of the LOs.

4 Assessments of Student Learning

Case Study Reports: In an effort to simulate (as realistically as possible) real-life, job experiences as a statistician, the sole assessment used in this class will be case study reports. Reports can be either written or oral with the exception of the midterm and final case studies which are mandatory to be written and oral, respectively.

Case study reports are graded based on the provided grading rubric (see appendices below). Each point of the grading rubric is classified according to a learning outcome and students will receive a grade for each point. Hence, by constructing a final report according to the grading rubric, students will have the opportunity to demonstrate their ability in each of the LOs.

There are 8 case studies throughout the semester. By having each case study focus on all of the LOs simultaneously, students will have sufficient practice at each of the LOs. Additionally, students will have multiple opportunities to demonstrate their ability in each of the LOs. By focusing the grading rubric on the learning outcomes, students will be adequately assessed on each LO.
5 Student Achievement of Learning Outcomes

Cumulative Learning Because the grading rubric is LO-centric, student achievement all learning outcomes is evidenced by the quality of case study reports. The figure below displays the scores of one particular student on case studies 1-8. Notably, the trend in scores is positive denoting continued improvement in demonstrating the LOs.

Learning of Specific Outcomes As evidence of student learning of individual LOs, I highlight results from two case study reports (one written and the other oral) from two different students than was given above.

LO#1– One of the case studies involved spatially correlated data with the ultimate goal of utilizing the correlation to do prediction. In a case study involving spatially correlated data, a student wrote, “Like all spatial data, we do not have independent observations because of spatial correlation. To help EPA scientists, we will develop a spatial model that accounts for the spatial correlation in this problem. This model will model the relationship between CMAQ and ground ozone level and allow us to make ozone predictions at spatial locations of interest.” Clearly, this student was able to understand and explore the data to the point where he understood that a statistical model that accounts for spatial correlation was needed to achieve the goals of the analysis.
LO#2— The figure below displays one of the slides from a final case study oral presentation. In this slide, the student correctly wrote down a logistic regression model, included a description of the major components and orally described the model.

\[
\log \left( \frac{p_i}{1 - p_i} \right) = x_i \beta,
\]

where

- \( p_i \): probability of germination for the \( i^{th} \) bulb
- \( x_i \): the set of covariates for the \( i^{th} \) bulb
- \( \beta \): coefficients

LO#3— The following is a snippet of code written by a student demonstrating how to fit the statistical model in R.

```r
obs.data <- as.geodata(cbind(ozone$Ozone.8.hr.max., ozone$Longitude, ozone$Latitude), data.col=1, coords.col=2:3)
gp.fit <- likfit(obs.data, cov.model="matern", kappa=nu, fix.kappa=TRUE, ini.cov.pars=c(0.90*var(ozone$Ozone.8.hr.max()), Matern.cor.to.range(100, nu=.5, cor.target=0.5)), trend="ozone$cmaq_o")

phi <- 1/gp.fit$phi
s2 <- gp.fit$sigmasq
mu <- gp.fit$beta
tau2 <- gp.fit$tausq
K <- length(pred[,1])
D <- rdist(rbind(pred[,3:2], ozone[,5:4]))
V <- s2*Matern(D, alpha=phi, nu=nu) ##V = Sigma_Y
X <- cbind(rep(1,length(ozone$cmaq_o)), ozone$cmaq_o)
Xp <- cbind(rep(1,length(pred$cmaq_p)), pred$cmaq_p)
mup <- Xp %*% mu
mud <- X %*% mu

EV <- mup + V[1:K,K+(1:N)]%*%solve(V[1:K,K+(1:N)] + tau2*diag(N))%*%(ozone$Ozone.8.hr.max.-mud)
```

4
upper <- qnorm(0.975, mean=EV, sd=sqrt(cond.Var))
lower <- qnorm(0.025, mean=EV, sd=sqrt(cond.Var))
marg <- mean(upper-lower)
bet <- solve( t(X) %*% solve(V[K+(1:N),K+(1:N)]+tau2*diag(N)) %*% t(X) ) %*% solve(V[K+(1:N),K+(1:N)]+tau2*diag(N)) %*% ozone$Ozone.8.hr.max.
SE <- sqrt(diag(solve( t(X) %*% solve(V[K+(1:N),K+(1:N)]+tau2*diag(N)) %*% t(X) )))
mu[1] + c(-1,1) * 1.96 * SE[1]

range <- c(min(lower), max(upper))

LO#4– The following is a slide taken from a student presentation demonstrating results of fitting a statistical model. The results correctly interpolate the points and is presented in using professional quality figures.

LO#5– This learning outcome is demonstrated by professional reports and presentations. The entire reports from these two students are included as appendices to this document and clearly demonstrate ability to construct a professional report of an analysis.

Results from Midterm Evaluation. On March 6, 2014, I invited Ken Plummer to conduct a midterm course evaluation. For each question, students were asked to respond on a scale of 1 to 5 where 1 is “very little” and 5 is “very much.” The following are highlights from that evaluation.
• When asked by Dr. Plummer, “What are the learning outcomes?” students could not remember the exact wording but came up with the following list based on experience in the class:
  – Analyze data when you get data – what to do with it.
  – Learn modern approaches to analyze data.
  – Learn how to write decent reports for stake holders.
  – Ensure students understand and justify their explanations and methods.
  – Learn how to work in teams.
  – Communicate effectively both orally and in writing.

While these are not the LOs exactly, they are highly similar suggesting that students can deduce the LOs simply based on class activities.

• When asked by Dr. Plummer, “To what degree are you accomplishing the learning outcomes?” students responded with mostly 4s and 5s.

• When asked by Dr. Plummer, “How helpful are the case studies in achieving the learning outcomes?” student responded with all 5s.

Results from Student Ratings. The following are excerpts of students comments about the class from their student ratings.

• “I feel like this class has helped allowed me to take everything I’ve learned so far about statistics and understand how to use it in ‘real’ settings. It has also helped me understand how to convey results and findings. If I am handed a data set and told to analyze it, I think that after taking this class I would be able to do it. (Within reason, of course.)”

• “This was an incredible coarse. Primarily because of this class, I now believe I could be an asset to a company in industry analyzing data. The applied nature of this class will be extremely valuable to my future.”

• “Great class. Definitely feel like (at this point) if I needed to use statistical methods in the future, it will be related to what I’ve done in this class.”

• “This class was perhaps the most useful stats class I’ve taken. It’s solidified my knowledge and understanding of statistics and working with data in ways that I hadn’t previously understood. it consolidated everything I was looking for in the masters program.”

6 Steps Planned to Improve Teaching and Student Learning

Based on the above results and student feedback in the teacher evaluations, I have decided to make the following changes:

1. Expand the case study reports to include a exploratory data analysis report that is due before each case study. This will require students to first explore the data prior to class discussion and lectures. I ultimately hope that this will increase class discussion regarding new statistical methodology.

2. Take steps to improve grading. While grading the case studies, I often felt that the order in which I graded them mattered. That is, I was often more strict with the first few reports. I realized this partway through the semester and started to randomize the order in which I graded. Next time around, I plan to read through all of the reports first then, on the second pass, determine how many points to take off. This will allow me to grade each report more evenly and fairly.
3. Give students direct feedback on their presentation skills. LO #5 suggests that students should be comfortable giving presentations. While I felt that this LO was achieved, I can give more feedback on their presentations skills.
Appendices
Statistics 536 - Case Study Grading Rubric  
Dr. Matthew Heaton, Winter 2014

/5pts EDA Score

/5pts Problem Statement and Understanding (Learning Outcome #1)
  - Does the report sufficiently describe the background of the problem?
  - Does the report sufficiently describe the data and point out any problems/issues?
  - Are the goals of the analysis clearly stated?

/15pts Describe the method/model(s) that are used (Learning Outcome #2)
  - Was a brief description of method/model used given in the report?
  - Were any greek letters used clearly defined?
  - Were any explicit or implicit assumptions needed to use the model adequately explained?

/10pts Model Justification (Learning Outcome #2)
  - Does the report give reasons for why the particular model was chosen?
  - Does the report describe why this model is appropriate for this data and how it solves the current problem?
  - Are the assumptions of the model justified (e.g. via exploratory analysis)?

/15pts Results (Learning Outcome #3 & #4)
  - Does the report adequately answer the questions posed in the case study?
  - Were estimates of the parameters and their uncertainties given?
  - Were the parameters interpreted in the context of the problem?
  - Did the report summarize the main points of the results in non-statistical terms?

/5pts Conclusions
  - Did the report summarize how the goals of the study were met?
  - Did the report discuss any shortcomings of the approach/model used?
  - Did the report provide suggestions for “next steps” in the analysis or further questions that may be of interest?

Teamwork
  - If the report is done as a team, please provide a description of each team member's responsibility on the project. Points may be deducted if I feel work was not equally distributed.
Midterm: Ozone Prediction

1 Problem Statement

Because ground level ozone is a main component of smog and causes breathing-related health issues, it is monitored by the Environmental Protection Agency (EPA). Community Multi-scale Air Quality Model (CMAQ) is used to simulate ozone creation based on population density, ground characteristics, temperature, etc. Problematically, CMAQ does not match observations exactly and returns average ozone level instead of point level predictions. EPA scientists want to understand the relationship between CMAQ and ozone station measurements. Furthermore, they want to predict ground level ozone at locations without measurements. Like all spatial data, we do not have independent observations because of spatial correlation. To help EPA scientists, we will develop a spatial model that accounts for the spatial correlation in this problem. This model will model the relationship between CMAQ and ground ozone level and allow us to make ozone predictions at spatial locations of interest.

2 Methods, Model, and Justification

2.1 Selecting Predictors

Since we do not have CMAQ calculations at ozone observation locations and prediction locations, we will estimate CMAQ using the calculations we have. However, because there are 66,960 calculated CMAQ values at locations $s_i \in S$, we have more predictors than we can possibly include in our model. To estimate CMAQ at locations of interest, we will take a weighted average of the four nearest CMAQ simulation locations. Using the four nearest CMAQ calculations effectively surrounds the point of interest, which will enable us to estimate CMAQ using the calculated CMAQ values that surround the point. I will weight the four CMAQ values according to how close they are to the point of interest (i.e. closest measurements will be most heavily weighted). Mathematically, I make weight $w_i \propto \frac{1}{d_i^2}$, where $d_i$ is the distance from the locations of interest to the measurement location. We will constrain the weights such that $\sum_{i=1}^{4} w_i = 1$. Thus, the estimated CMAQ is $CMAQ_{\text{pred}} = \sum_{i=1}^{4} w_i \cdot CMAQ_i$. This weighted average will enable us to estimate CMAQ at ozone observation locations and prediction locations. In our model, these estimated CMAQ values will be part of $X = (x(s_1), ..., x(s_n))$ at locations $s_i \in S$.

2.2 Spatial Model between CMAQ and Ozone

Because we expect spatial correlation in this problem, we will not assume that our observations are independent. Instead, we will create a Gaussian process spatial model that enforces smoothness using spatial correlation. Gaussian Processes are stochastic processes where any finite collection of observed random variables follow a multivariate normal distribution. In Gaussian process regression, we observe data $y(s_1), ..., y(s_n) \in Y$ and covariates $x(s_1), ..., x(s_n) \in X$ at spatial location $s_i \in S$. We want a function $w(s)$ so that $\sum_{i=1}^{n} [y(s_i) - w(s_i)]^2$ is small and $w(s)$ is smooth.

For our model, let $X(s) = (1, CMAQ_o)$ where $s$ is spatial location, 1 is an intercept, and $CMAQ_o$ is estimated CMAQ at ozone measurement locations; then, let $Y = (y(s_1), ..., y(s_n))$ be ground ozone measurements at the same locations. Since we have 800 ozone measurements, $n = 800$. Let $w(s_i) \sim GP(X \beta, \rho(\cdot))$, which implies $W = (w(s_1), ..., w(s_n)) \sim N(X \beta, \Sigma)$ where $ij$th element of $\Sigma$ is calculated using the Matérn covariance function

$$
\sigma^2 \cdot \frac{1}{2^{\nu-1} \Gamma(\nu)}(2d\phi \sqrt{\nu})^\nu K_\nu(2d\phi \sqrt{\nu})
$$

(1)
where \( \Gamma(\cdot) \) is the gamma function, \( K_\nu \) is the \( \nu^{th} \) order modified Bessel function of the second kind, \( d = \|s_i - s_j\| \), \( \sigma^2 \) is our estimated spatial variance, \( \phi \) is a range parameter, and \( \nu \) is a smoothness parameter. We assume that our observations \( Y \) are normally distributed about our Gaussian process \( W \), but we do not assume that our observations are independent; Thus, by our Gaussian process assumption,

\[
Y \mid W \sim \mathcal{N}(W, \sigma^2 I).
\]

Because \( [Y] = \int_W [Y \mid W][W] \, dW \),

\[
Y \sim \mathcal{N}(X\beta, \Sigma_Y + \sigma^2 I_n)
\]

where \( \Sigma_Y + \sigma^2 I_n \) denotes the covariance matrix for \( Y \).

With this closed from likelihood, I will use maximum likelihood estimation to estimate the parameters of the Matérn covariance function, \( \phi \) and \( \sigma^2 \), and of the distribution of \( Y \), \( \mu \) and \( \sigma^2 \). For the Matérn covariance function, I will fix \( \nu = 0.5 \) because this is computationally stable and the data contains no information about \( \nu \).

Within the Gaussian process model, we quantify the relationship between CMAQ and ground ozone. Because we do not assume independence, we will estimate \( \hat{\beta} \) for the intercept and CMAQ using generalized least squares. As noted above, \( Y \sim \mathcal{N}(X\beta, V) \), where \( V = \Sigma_Y + \sigma^2 I_n \). Thus,

\[
\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y
\]

\[
SE(\beta) = \sqrt{(X'V^{-1}X)^{-1}I}
\]

The result for the intercept will tell us expected ground ozone when CMAQ is 0. The estimate for CMAQ will tell us the expected change in ground ozone when CMAQ increase by one unit. These results also enable us to perform inference on \( \beta \).

2.3 Ozone Prediction Model

To use this model for prediction, I will create a vector \( Z = (z_1, ..., z_k) \) of prediction locations and \( X^*(z) = (1, CMAQ_p) \) where \( CMAQ_p \) is estimated CMAQ at prediction locations \( Z \) and 1 is an intercept. Let \( Y^* = (y(x_1^*), ..., y(x_k^*)) \) be the predicted ground ozone corresponding to \( Z \) and \( X^* \). It follows that \( Y^* \sim \mathcal{N}(X\beta, \Sigma_{Y*} + \sigma^2 I_k) \) where \( \Sigma_{Y*} + I_k \) denotes the covariance matrix for \( Y^* \). Thus, if we put \( Y \) and \( Y^* \) together so that

\[
\begin{pmatrix} Y^* \\ Y \end{pmatrix} \sim \mathcal{N} \left[ \begin{pmatrix} X^*\beta \\ X\beta \end{pmatrix}, \begin{pmatrix} \Sigma_{Y*} & \Sigma_{Y*}Y \\ \Sigma_{Y*}Y & \Sigma_Y \end{pmatrix} + \sigma^2 I \right]
\]

where \( \Sigma_{Y*Y} \) and \( \Sigma_{YY*} \) are the covariance matrices between \( Y \) and \( Y^* \), then, by our joint normal assumption, the conditional normal model gives,

\[
Y^* \mid Y \sim \mathcal{N}(\mu_{Y^*}, \Sigma_{Y^*})
\]

where

\[
\begin{align*}
\mu_{Y^*} &= X^*\beta + \Sigma_{Y*Y}(\Sigma_Y + \sigma^2 I_n)^{-1}(Y - X\beta) \\
\Sigma_{Y^*} &= (\Sigma_{Y*} + \sigma^2 I_k) - \Sigma_{Y*Y}(\Sigma_Y + \sigma^2 I_n)^{-1}\Sigma_{YY*}.
\end{align*}
\]

The corresponding 95% prediction interval about \( X^*\beta + \Sigma_{Y*Y}(\Sigma_Y + \sigma^2 I_n)^{-1}(Y - X\beta) \) is simply the 0.025 and 0.975 quantiles of \( Y^* \mid Y \). This simple confidence interval form is a major advantage to Gaussian process regression.

Because our model is based on a joint normal assumption, I will verify that this assumption is acceptable by demonstrating that the error terms \( \epsilon = Y - W \) are approximately normally distributed. Thus, I will fit the Gaussian process to the data so that we can calculate the residuals \( Y - W \). If we can show that the residuals are nearly normally distributed than our model will be justified.
2.4 Model Validation: 16-fold Cross Validation

We will verify this model using 16-fold cross validation where we take random hold-out test subsets of \( n = 50 \) and fit the model using the remaining 750 data to verify that the coverage of our 95% prediction interval is close to 0.95. K-fold cross validation does not allow overlapping test subsets, and since we have 800 data, we will repeat this process 16 times so that every data point is tested exactly once. We will calculate the mean coverage of all 16 validation steps and plot the results. Even though this step will be computationally expensive, it can help to verify that our model accurately predicts ground ozone levels.

3 Results

3.1 Model Fits

Because our model is based on a joint normal assumption, I will verify that this is acceptable demonstrating that the error terms \( \epsilon = Y - W \) are normally distributed (see Figure 1).

Figure 1: Note that the residuals are approximately normally distributed about the line \( y = 0 \). The residuals histogram also appears approximately normal. This implies that the normal assumption is reasonable.

Note that, while there are some large residuals, the residuals are approximately distributed as a normal about \( y = 0 \). Furthermore, the residual histogram appears approximately normal. For this reason, the normal assumption is justified; thus, the following results based on normal assumptions are also justified.

Using maximum likelihood estimation, we estimate \( \hat{\sigma}^2 = 1863.53, \hat{\phi} = 0.0069315 \), and \( \hat{\tau}^2 = 17.98387 \). We interpret \( \hat{\sigma}^2 \) as the estimated spatial variance, \( \hat{\tau}^2 \) is the estimated observation variance, and \( \hat{\phi} \) as the estimated decay parameter of the Matérn covariance function. We also report estimates and 95% confidence intervals on \( \hat{\beta} \) for CMAQ:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Lower Bound of 95% CI</th>
<th>Upper Bound of 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>10.80</td>
<td>-68.21</td>
<td>89.81</td>
</tr>
<tr>
<td>CMAQ</td>
<td>0.52</td>
<td>0.42</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 1: Estimates and Confidence Intervals on \( \hat{\beta} \). Note that 0 is the confidence interval for the intercept and is, thus, insignificant.
Our estimate for the intercept suggests that when CMAQ is 0, we expect that ground ozone will be 10.80 parts per billion (ppb). However, since 0 ppb is within the 95% confidence interval about the interval, it is not statistically significant. The estimate for CMAQ, on the other hand, is statistically significant and suggests, holding all else constant, that if we increase CMAQ by one unit, we expect a 0.52 ppb increase in ground ozone on average.

Using the predictive model, we predict ground ozone at the desired prediction locations. For the predictions, the average prediction interval margin was 23.94 ppb. I plot heat maps for the expected 8 hour ground ozone maximum as well as the lower and upper bounds on the 95% prediction intervals. Note that I have fixed the color scale so that the maps can be interpreted more easily (see Figure 2).

![Expected 8 hour Ozone Maximum](image1)

![Predicted 8 hour Ozone Lower Bound](image2)

![Predicted 8 hour Ozone Upper Bound](image3)

Figure 2: Plot of the expected ground ozone level, 95% prediction lower bound, and 95% prediction upper bound. Note that the minimum and maximum of the expected ground ozone level is 23.80 ppb and 92.97 ppb. The color legend, however, has minimum and maximum of 5.95 ppb and 102.59 ppb to capture the minimum of the 95% lower bound and maximum of the 95% upper bound. Thus, all plots are on the same color scale. The average prediction interval margin was 23.94 ppb, but note that at boundary locations that the prediction intervals widen.
3.2 Cross Validation Results

Using 16-fold cross validation, we found that the mean 95% prediction interval coverage was 0.94375 which suggests that our model predicts ground ozone levels accurately. Furthermore, since the mean 95% prediction interval margin is 23.94 ppb, our model is not only accurate but has small variance. I plot a histogram of the coverage of the 16 cross validation stages (see Figure 2).

![16-Fold CV Coverage](image)

Figure 3: Note that the mean 95% prediction interval coverage is 0.94375.

4 Conclusions

The Gaussian process model using spacial coordinates and CMAQ effectively modeled ground ozone as demonstrated by the results of our 95% prediction interval coverage (0.94375) and prediction interval margin (23.94 ppb). Furthermore, the normal assumption I made was justified by the distribution of residuals. While we could used other methods in this problem, the Gaussian process model performed well and was justified. However, my cross validation method was computationally expensive because I needed to fit the model 16 times.

While this model performed well with spatial coordinates and CMAQ, it would probably predict more accurately with more covariates. If there are covariates linked to ground ozone that are not used to calculate CMAQ, we could include these variables. Also, because CMAQ can be computed at any point, it would be helpful if CMAQ were computed specifically at ozone measurement locations. If we had CMAQ calculated at ozone measurement locations, we would not need to approximate it using a weighted average.

Because ground ozone is studied by the EPA and others, we have a lot of prior knowledge that we could incorporate into a Bayesian model. By using informed priors, the predictions on ground ozone levels would likely be improved. I did not pursue this because it is time intensive, but a fully Bayesian analysis would be interesting to carry out for this problem.
Problem: Climate changes in the Netherlands

- Increases in temperature, precipitation, risk of flood, and sea levels
  - Tulips (a huge agricultural export of this region) require a certain “chilling time.”
**Goals:** Understand the effect of chilling time on the germination of tulip bulbs

- Is the effect of chilling time the same across all populations? Which populations are the same/different?
- Is there an “ideal” chilling time? Does this ideal chilling time vary by population?
- What effect will a decrease from 10 to 8 weeks of winter/chilling time have for tulips?
Data Summary

**Data:** 2520 tulip bulbs collected from fields between 2005-2009

- Response Variable: 1 if the bulb germinated, 0 if it did not
- Explanatory Variables: Population (1-12), Chilling Time (0 - 12 weeks)
Data Decisions

How do we handle this data set?

- Nonlinear effects across populations
- Removal of population 12
<table>
<thead>
<tr>
<th>Population</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.40</td>
<td>0.97</td>
<td>0.83</td>
<td>0.87</td>
<td>0.87</td>
<td>0.97</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>0.13</td>
<td>0.53</td>
<td>0.73</td>
<td>0.73</td>
<td>0.83</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.53</td>
<td>0.80</td>
<td>0.83</td>
<td>0.97</td>
<td>0.90</td>
<td>0.87</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.17</td>
<td>0.53</td>
<td>0.60</td>
<td>0.73</td>
<td>0.90</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>0.33</td>
<td>0.87</td>
<td>0.67</td>
<td>0.73</td>
<td>0.70</td>
<td>0.57</td>
<td>0.50</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
<td>0.40</td>
<td>0.43</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>0.00</td>
<td>0.10</td>
<td>0.33</td>
<td>0.47</td>
<td>0.83</td>
<td>0.67</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.03</td>
<td>0.27</td>
<td>0.33</td>
<td>0.33</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>9</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>10</td>
<td>0.00</td>
<td>0.17</td>
<td>0.10</td>
<td>0.53</td>
<td>0.87</td>
<td>0.87</td>
<td>0.83</td>
</tr>
<tr>
<td>11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.20</td>
<td>0.23</td>
<td>0.67</td>
<td>0.83</td>
<td>0.47</td>
</tr>
<tr>
<td>12</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Response is dichotomous: 1 if germination occurred, 0 if it did not.

- Basic linear regression will not work!

Classification Model Options:
- Decision Trees, Random Forests
- Discriminant Analysis
- Support Vector Machines
- K-Nearest Neighbors
- Probit Regression
- **Logistic Regression**
Logistic Regression Background

- Used for binary responses ($Y \in \{0, 1\}$)
- Consider the Bernoulli distribution:

$$f(y_i) = p^{y_i}(1 - p)^{(1-y_i)}$$

What is $p$, the probability of a 1 (tulip germination)?
Logistic Regression Model

\[
\log \left( \frac{p_i}{1 - p_i} \right) = x_i' \beta,
\]

where

- \( p_i \): probability of germination for the \( i^{th} \) bulb

\[
p_i = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}
\]

- \( x_i \): the set of covariates for the \( i^{th} \) bulb
- \( \beta \): coefficients
Logistic Regression and Non-Monotonic Probabilities

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \sum_{p=1}^{P} f_p(x_{ip}) + \epsilon_i,
\]

where \( f_p(x_{ip}) \) is the function for the \( p^{th} \) variable.

- Basis Function for Population: Categorical (1 in population \( m \), 0 otherwise)
- Basis Function for Chilling Time: Natural Spline
Non-Monotonic Probabilities & Natural Splines

- Natural Spline: cubic spline with linear behavior outside of range of data
Non-Monotonic Probabilities & Natural Splines

- What is the appropriate threshold?
  - Chosen based on AUC: If $\hat{p}_i > 0.55$, $\hat{y}_i = 1$, 0 otherwise

- What is the appropriate number of knots?
  - Chosen through cross validation: $df = 4$ (3 knots)
Model Assumption: Independence

- Justification: Difficult to prove – assume independence to move forward
Why Logistic Regression?

- Interpretability
- Ability to address nonlinear effects
- Straightforward likelihood ratio tests
- Prediction of probability of a success
- **Uncertainty of Estimates**
Why Our Model?

- The model we constructed that accounted for population, chilling time, and the interaction between these two variables provided an AUC higher than that of the model that just accounted for chilling time (0.7626 vs. 0.6577).
Results

Is the effect of chilling time the same across populations? Which populations are the same/different?
Is the effect of chilling time the same across populations?

- Likelihood ratio test:
  - $H_0$: The reduced model is sufficient; the effect of chilling time is the same across populations.
  - $H_1$: The full model is required; the effect of chilling time is not the same across populations.
  - p-value $< 2.2e-16$

- Specifically, which populations are the same/different?
Which populations are the same/different?

- Likelihood ratio test for pairwise comparisons: For each pair of populations A and B,
  - \( H_0 \): The reduced model is sufficient; the effect of chilling time is the same for populations A and B.
  - \( H_1 \): The full model is required; the effect of chilling time is not the same for populations A and B.
- Bonferroni correction for multiple comparisons: compare to 0.00091
Most populations show different responses to changes in chilling time.
Is there an “ideal” chilling time? Does this chilling time vary by population?
Because the effect of chilling time is not the same across population, the ideal chilling time will vary.

- Obtained predictions, then calculated uncertainties using bootstrap sampling.

\[ CI : (2.5\% \text{ quantile}, 97.5\% \text{ quantile}) \]
<table>
<thead>
<tr>
<th>Population</th>
<th>Estimated Ideal Time</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.018</td>
<td>2.607</td>
<td>12.000</td>
</tr>
<tr>
<td>2</td>
<td>9.946</td>
<td>7.507</td>
<td>12.000</td>
</tr>
<tr>
<td>3</td>
<td>9.634</td>
<td>4.829</td>
<td>12.000</td>
</tr>
<tr>
<td>4</td>
<td>9.898</td>
<td>9.381</td>
<td>12.000</td>
</tr>
<tr>
<td>5</td>
<td>3.147</td>
<td>2.619</td>
<td>9.009</td>
</tr>
<tr>
<td>6</td>
<td>10.547</td>
<td>4.937</td>
<td>12.000</td>
</tr>
<tr>
<td>7</td>
<td>10.258</td>
<td>4.877</td>
<td>12.000</td>
</tr>
<tr>
<td>8</td>
<td>5.790</td>
<td>4.312</td>
<td>12.000</td>
</tr>
<tr>
<td>9</td>
<td>10.847</td>
<td>9.849</td>
<td>12.000</td>
</tr>
<tr>
<td>10</td>
<td>9.922</td>
<td>9.272</td>
<td>12.000</td>
</tr>
<tr>
<td>11</td>
<td>9.813814</td>
<td>9.526</td>
<td>10.066</td>
</tr>
</tbody>
</table>
Results

Curve for Population 11

Chilling Time

Predicted Probability of Germination

Curve for Population 1

Chilling Time

Predicted Probability of Germination

Curve for Population 7

Chilling Time

Predicted Probability of Germination

Curve for Population 5

Chilling Time

Predicted Probability of Germination
What effect will a decrease from 10 to 8 weeks of winter/chilling time have on tulips?
<table>
<thead>
<tr>
<th>Population</th>
<th>Estimated Diff.</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.055</td>
<td>0.010</td>
<td>0.130</td>
</tr>
<tr>
<td>2</td>
<td>0.049</td>
<td>-0.027</td>
<td>0.120</td>
</tr>
<tr>
<td>3</td>
<td>0.025</td>
<td>-0.032</td>
<td>0.070</td>
</tr>
<tr>
<td>4</td>
<td>0.099</td>
<td>0.019</td>
<td>0.187</td>
</tr>
<tr>
<td>5</td>
<td>-0.029</td>
<td>-0.148</td>
<td>0.077</td>
</tr>
<tr>
<td>6</td>
<td>0.146</td>
<td>0.036</td>
<td>0.533</td>
</tr>
<tr>
<td>7</td>
<td>0.367</td>
<td>0.156</td>
<td>0.584</td>
</tr>
<tr>
<td>8</td>
<td>-0.018</td>
<td>-0.129</td>
<td>0.167</td>
</tr>
<tr>
<td>9</td>
<td>0.533</td>
<td>0.349</td>
<td>0.726</td>
</tr>
<tr>
<td>10</td>
<td>0.086</td>
<td>0.018</td>
<td>0.158</td>
</tr>
<tr>
<td>11</td>
<td>0.210</td>
<td>0.093</td>
<td>0.390</td>
</tr>
</tbody>
</table>
Conclusions

Through a logistic generalized additive model, we

- Identified differences in the effect of chilling time across tulip populations
- Estimated the ideal chilling time for each tulip population
- Estimated the effect of reducing the chilling time from 10 to 8 weeks for each tulip population
Conclusions

Other Potential Approaches

- Probit Regression
- Decision Trees, Bagging, Random Forests
- Discriminant Analysis
- K-Nearest Neighbors
- Support Vector Machines
Conclusions

Shortcomings & Future Work

- Other models better geared for prediction
- Large uncertainties
- Detailed characteristics of tulip populations
- Potential incorporation of spatial covariate
Questions?
Report on 2014 Citizenship Goals:

1. Act as a reviewer for 4 peer-reviewed articles.
   • **Report:** I shared this goal with my faculty mentor and he advised me that every paper I submit is reviewed by at least 2 reviewers. Hence, to act as a gain on the profession, I should be reviewing about 2 papers for every paper I submit. With this goal in mind, I have reviewed 8 papers this year.

2. Seek opportunities to serve as a committee member on a student thesis.
   • **Report:** As mentioned in my scholarship goals, I am currently the thesis advisor for 3 MS students.

3. Be an active participant in my department committee assignments.
   • **Report:** Since setting this goal, I have not only been active in my committee assignments but I’ve been asked to serve on an additional committee.

4. Attend the weekly collegiality lunch with faculty.
   • **Report:** With few exceptions, I have attended each collegiality lunch.

5. Attend Dr. Blades 330 class in preparation for teaching 330 in Winter ’15.
   • **Report:** I found that attending every lecture was a less efficient use of my time than constantly talking with Dr. Blades about preparing to teach 330. I attended several lectures but only do so when I’m interested in how a particular topic is taught.

6. Continue research collaborations with Drs. Berrett, Christensen and Reese.
   • **Report:** I submitted a paper with Dr. Christensen, two grants with Dr. Berrett and one grant with Dr. Reese.
1. Maintain a “next steps” document with research ideas and directions.

- **Report:** Maintaining a “next steps” document has proven more useful than I could have imagined when I started it. I use this document as “scratch paper” for research ideas. I have found this document particularly useful for recruiting students to help me with my research, kick-starting a new project after submitting a paper, merging multiple “next steps” into a grant proposal and a location to keep ideas that come to me but I don’t have time to pursue at the time. I certainly plan on keeping this document to help streamline my research activities.

2. Submit three first-author papers for peer review.

- **Report:** The “next steps” document referred to in goal #1 has proven pivotal in accomplishing this goal. That is, the “next steps” document serves as a progress report on all of my current and future research activities. To evaluate my progress on this goal, I frequently referred to that document on the status of paper or project. To date, I have submitted three first-author papers, two of which have been accepted for publication and the third is still under review. I have made significant progress on a fourth first-author paper as well as two other papers in which I appear as a co-author.

3. Submit a grant proposal to the NIH.

- **Report:** The proposal was submitted in June 2014 as planned.

4. Give at least two presentations of my research in academic settings.

- **Report:** At the time this goal was set, I only had 1 scheduled presentation. Since then, I was invited to give 2 other presentations bringing my total to 3 professional presentations. I attribute the success of this goal to constant networking efforts to colleagues with similar research interests.

5. Actively recruit a graduate student to participate in my research.

- **Report:** In order to recruit students to help me with my research, I gave a department seminar on several research opportunities. From that seminar, I was able to recruit 3 2nd year MS students and 1 1st year MS student. These students are currently working with me to complete their thesis requirement for the MS program and publish their results.

**Scholarly Research Strategies**

1. As mentioned above, the “next steps” document has proven invaluable as a research journal for progress reports, general results and future research ideas. I feel this strategy has allowed me to streamline research productivity without a lull in progress because I always know the “next step.”
2. Blocks of research time has also been useful but I have not been very diligent in scheduling this time. I find time to do productive research but I need to improve at scheduling such time.

3. I once thought that networking ended once I landed a professorial job. However, I have now come to realize that networking is a constant requirement for success. Through networking, I have been able to get involved in new research projects, give more presentations and expand my knowledge of cutting edge research. I plan to continue efforts to network whenever possible and invite those with whom I network out to BYU.

What I Learned

1. The largest lesson I’ve learned is that research needs to be done on purpose. Research time will not just magically appear and success will not just magically happen. Success in research requires consistent, diligent effort to achieve success. This best effort is given when you develop a particular strategy for success and stick with it.