Syllabus
EES 5891-03: Bayesian Statistical Methods

Jonathan Gilligan
Vanderbilt University

Fall 2022

Contents

1 Nuts & Bolts **2**
1.1 Class Meetings 2
1.2 Professor 2
1.3 Email 2
1.4 Course web site 2

2 Description **3**
2.1 Catalog Description 3
2.2 Prerequisites 3
2.3 Narrative Description 3

3 Goals **4**

4 Structure **5**
4.1 Reading Material 5
Required Reading 5
Optional Books 5
Overview of Reading Materials 6
4.2 Graded Work 6
Grading 6
Homework 6
Projects 6
Tests and Examinations 6

5 Honor Code: **6**
5.1 Research Integrity 7

6 Final Note: **8**

7 Meet Your Professor **9**

Class Schedule **11**

1 Nuts and Bolts

1.1 Class Meetings

TR 11:00–12:15 Stevenson 6740

1.2 Professor

Jonathan Gilligan
Associate Professor of Earth & Environmental Sciences
Associate Professor of Civil & Environmental Engineering
jonathan.gilligan@vanderbilt.edu
www.jonathangilligan.org
Office Hours: TBA, or by appointment.

1.3 Email

If you want to communicate with Professor Gilligan be sure to begin the subject line of your email with "EES 5891" This helps assure that we will see your message quickly and respond to it.

I have set my email reader to flag all messages like this as important, so I will read them first. This also assures that I do not mistake your email for spam. I typically receive over 100 emails per day, so if you do not follow these instructions, I may not notice your email.

1.4 Course web site

In addition to Brightspace, I have set up a companion web site for this course at https://ees5891.jgilligan.org, where I post the reading and homework assignments, my slides from class, and other useful material. That web site will be the central place to keep up with material for the course during the semester. This web site will direct you to Brightspace if there is anything you need to find there.

2 Course Description

2.1 Catalog Description

The class will begin with an introduction to Bayesian statistics and then focus on practical application of regression methods to data. We will use R together with the Stan software package (`https://mc-stan.org`) for Hamiltonian Monte Carlo methods and the R-INLA software package for Integrated Nested Laplace Approximation (INLA) analysis (`https://www.r-inla.org/`). The course will combine practical applications of Bayesian methods to real (often messy) data with more philosophical discussions of Bayesian approaches to statistics and how to interpret results of statistical analyses. We will focus on regression methods, including hierarchical or multilevel regression modeling methods, which can be very powerful when you have data that has a nested structure (e.g., cities and counties within states or species within genera). Students will do projects applying Bayesian methods to their own data sets.

2.2 Prerequisites

You should be comfortable with differential and integral calculus and have some previous experience with standard statistics.

This course will be very mathematical and will make extensive use of the R software system, but I do not assume that you already know R or advanced mathematics beyond calculus.

2.3 Narrative Description

Bayesian statistics is a branch of statistics that has been around for almost 300 years, but for most of that time, it was very difficult to apply to practical problems because the mathematical equations were too difficult to solve. In the last 30 years, as computers have become much faster and more powerful, new computational methods have emerged that make Bayesian statistics practical for research and applications.

Bayesian analysis is widely used across a wide variety of research as well as practical applications. It is used to analyze results from high-energy particle physics experiments to discover new subatomic particles. There are many other applications in a wide variety of domains. It's used by geologists to improve estimates of mineral distributions and radon hazards. It's used by biologists to identify and categorize variations in the genomes of humans and other species. It's used extensively in medicine to analyze the results of clinical trials, to determine the pharmacokinetics of drug metabolism, and to assess the predictive value of tests for diseases such as cancer or COVID infection. It's used in political science and sociology to improve the accuracy of public opinion surveys and to understand patterns of voting. It's widely used in marketing to identify consumer preferences and improve the effectiveness of advertising. If you use Google, Amazon, Netflix, Stitchfix, or practically any large online platform for shopping or entertainment, advanced Bayesian methods form the basis of their recommendations. Bayesian analysis has also been applied effectively to law and criminology to assess the value of evidence in proving guilt or innocence. It has been applied to public health to estimate the prevalence of dieseases and tomake more effective treatment decisions when medical tests are uncertain. It is widely used in meteorology to make weather forecasts and in climate science to combine data from many different sources and come up with quantitative predictions and detailed understanding of their associated uncertainties. Bayesian methods are also widely used in computational applications, such as image analysis and reconstruction, computational text analysis, and natural language processing. One of the earliest practical applications of Bayesian textual analysis, in 1964, identified the anonymous authors of the Federalist Papers. More recent applications of Bayesian textual analysis are used to separate desired email from spam.

Bayesian statistical methods are valuable because they provide a systematic way to combine what you already know about a problem with new data from experiments or observations, and the results of Bayesian analyses are more straightforward to interpret than conventional statistics.

This course will provide a general introduction to Bayesian statistics and will combine practical instruction in how to do Bayesian data analysis and philosophical discussions about how to think about the assumptions that go into a Bayesian analysis and how to interpret the results that it produces.

You do not need to have any prior knowledge of computer programming, but I do expect that you are familiar with basic statistics and calculus (both derivatives and integrals).

3 Goals for the Course

By the end of the semester, you will:

- Understand Bayes's theorem and how to apply it.
- Understand problems with the traditional statistical emphasis on null-hypothesis significance testing (NHST), why Bayesian approaches to NHST don't solve these problems, and how Bayeian statistics offers superior alternatives to NHST.
- Understand how think about statistical models, how to choose an appropriate model for your problems, and understand the tradeoffs between different kinds of models.
- Be able to design and conduct a comprehensive Bayesian analysis of data from start to finish.
- Understand how to choose appropriate priors for your Bayesian analyses and how to test whether your choice of priors is sound.
- Understand how to set up, perform, assess the validity of, and interpret the results of Bayesian regression analysis.
- Understand why Markov Chain Monte Carlo (MCMC) sampling is used in Bayesian analysis, what the limits of MCMC are, and how to test your MCMC analyses for validity.
- Understand and be able to perform analyses using more complex statistical models, such as interaction models, generalized linear models, models of discrete (categorical and count) data.
- Understand what multilevel or hierarchical models are, when to use them, and how to interpret the results of a multilevel analysis.
- Understand the Integrated Nested Laplace Approximation (INLA), why you might use INLA instead of MCMC analysis, and what the limits of INLA analysis are.
- Understand several types of Bayesian geospatial analysis, including Matern covariance models and conditional autoregressive (CAR) models.

4 Structure of the Course:

I divide the semester into three parts:

1. **Introduction to Bayes's Theorem and its Applications:** The first part of the course introduces the basic concepts of Bayesian statistics, using simplified approximations to calculate difficult equations. This section will focus on linear regression methods.

2. **Monte Carlo Methods:** Next, we study Monte Carlo methods, which help us solve more difficult problems that our earlier approximations are not powerful enough for. This section will introduce statistical models of discrete data (counts, categories, etc.), and generalized linear models. It will conclude with multilevel statistical models, which can be very powerful methods for working with large and complex data sets.

3. **Geospatial Modeling:** Finally, we will learn a different approach, called the Integrated Nested Laplace Approximation (INLA), which is very well suited for analyzing geospatial data that may be too difficult to analyze uding Monte Carlo methods.

4.1 Reading Material

There are two required textbooks and two optional books. Supplementary reading on the Internet or in handouts will also be assigned during the term and posted on Brightspace.

REQUIRED READING MATERIALS

- Richard McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan,* 2nd ed. (CRC Press, 2020; ISBN 978-0-367-13991-9). This will be the principal textbook for the semester. Be sure you get the second edition because it is significantly different from the first.

- Virgilio Gomez-Rubio, *Bayeian Inference with INLA* (CRC Press 2021; ISBN 978-1-032-17453-2). We will only use it for a few weeks in the third part of the semester when we study geospatial methods. The book is expensive, but there is a free web-based e-book version online at `becarioprecario.bitbucket.io/inla-gitbook/index.html`, which is identical to the printed book.

There is a companion web site to *Statistical Rethinking* at `xcelab.net/rm/statistical-rethinking/`, which has links a number of resources, including videos of the author's own lectures on the material.

For people who are familiar with R and like to work in the `tidyverse` dialect, there is a free companion e-book on the web at `bookdown.org/content/4857/`, that has translated almost all the R code in the book into the `tidyverse` dialect of R.

OPTIONAL READING MATERIALS

- John Kruschke, *Doing Bayesian Data Analysis,* 2nd ed. (Academic Press, 2015; ISBN 978-0-12-405888-0). This book has a more elementary introduction to Bayesian statistics, at an undergraduate level. It is very clear, but it focuses almost entirely on Monte Carlo sampling and doesn't go as deeply into other important aspects of Bayesian statistics as we will do in this course. It is a very useful resource to check out if Monte Carlo sampling seems confusing. I have asked the Science & Engineering library to put a copy on reserve so you can read it there without needing to buy it.

- Hadley Wickham and Garrett Grolemund, *R for Data Science* (O'Reilly, 2017; ISBN 978-1-491-91039-9). This book is a great introduction to the R statistical programming language. It uses the `tidyverse` dialect of R, developed by Hadley Wickham. There is a free web-based ebook version at `r4ds.had.co.nz/`, so you won't need to buy a paper copy.

Overview of Reading Materials

I will give out detailed reading that give specific pages to read for each class and notes on important things you should understand. **I expect you to complete the reading before you come to class on the day for which the reading is assigned**, so you can participate in discussions of the assigned material and ask questions if there are things you don't understand.

4.2 Graded Work

Basis for Grading

Class participation	5%
Homework	45%
Laboratory & Homework	NA%
Final exam	NA%

Homework

Homework is due at the beginning of class on the day it is assigned.

Projects

You will do an extended research project in the second half of the semester, in which you will apply Bayesian methods to investigating a data set that you choose. This could be data from your dissertation research or another data set that interests you.

Tests and Examinations

There will be no tests or exams in this course.

5 Honor Code:

This course, like all courses at Vanderbilt, is conducted under the Honor Code.

Studying: As you study for this class, I encourage you to to seek help from me or from other classmates or friends.

Homework: I encourage working together. In most lab assignments you will be explicitly told to work with a partner. I also encourage you to talk with other classmates, as well as friends and acquaintances outside of class. You may discuss assignments, compare notes on how you are working a problem, and you may look at your classmates' work on homework assignments. But you must work through the problems yourself in the work you turn in: **Even if you have discussed the solution with others you must work through the steps yourself and express the answers in your own words. You may not simply copy someone else's answer.**

Research project: The research project will be conducted under the same ethical principals that apply to publishing papers in scientific journals. The work must be your own, but you may consult any other resources. If anyone else makes a substantial contribution, you must list them and their contributions in an Acknowledgements section.

If you ever have questions about how the Honor Code applies to your work in this course, please ask me. **Uncertainty about the Honor Code does not excuse a violation.**

5.1 Research Integrity

Beyond the University Honor Code, this course also emphasizes the scientific ethical principles of research integrity. Honesty is a very important part of research integrity, but it is only one part. Clearly, science cannot work if scientists are not scrupulously honest about the results of their research and there is no tolerance for scientists who lie. But research integrity goes much farther. Real science happens in the context of a scientific community and the integrity of this community is critical. The ethical principles of research integrity have grown over the centuries to protect the integrity of the scientific community. Indeed, the mathematician and poet Jacob Bronowski wrote, in his book, *Science and Human Values* (Harper & Row, 1956) that what makes science work and makes it great is much less about the intellectual brilliance and skills of individual scientists than about the ethical commitment to truth and human dignity by the community of scientists.

Science does not proceed only by making leaps of discovery but also by making useful mistakes and then discovering and correcting the errors in those mistakes. Because of this, scientific integrity requires scientists to be extremely transparent and forthcoming about all the details of their research. It is not enough to sincerely report a discovery or an idea in good faith, but one must also provide others with the tools to critically examine that discovery and idea, and if a scientist learns, even many years later, that a report or discovery contained an error, they must correct the error and actively inform other scientists about it.

When a scientist discovers an error in their past work and does not promptly and actively correct it, other scientists may continue to rely on the truth of that result and thus waste time, effort, and money. Thus, both making one's own work available to others so they can have the opportunity to find errors, and also to promptly and publicly report any errors that one finds in one's own work are two critical pieces of research integrity.

Another aspect also involves the communal nature of science: None of us works in isolation, and every scientist's work builds on work by others. There are two reasons why it is critical to acknowledge the role of others' work in our own research reports: First, it is important to give others the credit for their contributions to our shared body of scientific knowledge. Secondly, it is important for others to know where the data and methods we use come from. If I use someone else's data or methods for an analysis and it later turns out that there were problems with their data or methods, then it is important for people reading my work to be able to examine my work and evaluate how those errors might affect my own results.

I want to emphasize that these considerations about research integrity are not just negative things. They are very positive, which is why so many researchers are embracing them. **By**

being transparent and forthcoming, and by encouraging others to reproduce your research results, you can enhance your reputation, both for honesty (you show that you have nothing to hide) and for being a good citizen of the scientific community by making it easy for other researchers to learn from your work and build on it to make new discoveries and build new and more powerful tools for analyzing data.

Where this is relevant to this course on Global Climate Change is in our practice of reproducible research in the laboratory portion of the course. Making our work, however humble, fully open and transparent so that others may examine it, criticize it, or build on it to develop new tools and make new discoveries is an essential part of research integrity.

In your lab reports it will be important for you to document where the data you worked with comes from (this will mostly be clearly spelled out in the assignments) and what methods you used to analyze it. Using the tools of R and RMarkdown will make it easy to almost automatically include this kind of transparency in your reports. As you do this throughout this course, you will learn the best practices adopted by the scientific community and develop habits of openness, transparency, and reproducibility for any research you do in the future in any area of society, whether in science, journalism, business, or other endeavors.

6 Final Note:

I have made every effort to plan a busy, exciting, and instructive semester. I may find during the term that I need to revise the syllabus to give more time to some subjects or to pass more quickly over others rather than covering them in depth. Many topics we will cover are frequently in the news. Breaking news may warrant a detour from the schedule presented on the following pages. Thus, while I will attempt to follow this syllabus as closely as I can, you should realize that it is subject to change during the semester.

7 Meet Your Professor

Jonathan Gilligan has worked in many areas of science and public policy. His past research includes work on laser physics, quantum optics, laser surgery, electrical properties of the heart, using modified spy planes to study the ozone layer in the stratosphere, and connections between religion and care for the environment.

Professor Gilligan is the Alexander Heard Distinguished Service Professor, Associate Professor of Earth & Environmental Sciences, Associate Professor of Civil & Environmental Engineering, and the director of the Vanderbilt Climate and Society Grand Challenge Initiative, which is working to integrate research, teaching, and public outreach about climate change across the natural sciences, social sciences, and humanities.

His current research investigates the role of individual and household behavior in greenhouse gas emissions in the United States; how “smart cities” can use technology to reduce environmental footprints and promote health and citizen empowerment; water conservation policies in American cities; vulnerability and resilience to environmental stress in South Asia; and developing new directions for climate policy in the US.

Professor Gilligan and Professor Michael Vandenbergh won the 2017 Morrison Prize for the highest-impact paper of the year on sustainability law and policy. Gilligan and Vandenbergh’s book, *Beyond Politics: The Private Governance Approach to Climate Change* (Cambridge University Press, 2017), was named by *Environmental Forum* as one of the most important books on the environment of the last 50 years.

Apart from his academic work, Professor Gilligan dabbles in writing for the theater. His stage adaptation of Nathaniel Hawthorne’s *The Scarlet Letter*, co-written with his mother Carol Gilligan, has been staged at The Culture Project in New York City, starring Marisa Tomei, Ron Cephas Jones, and Bobby Cannavale, and was later performed at Prime Stage Theatre, Pittsburgh and in a touring production by The National Players. Most recently, it was performed as the principal fall 2019 production of the Fullerton College Classic Dramatic Series in Fullerton CA, directed by Michael Mueller, and was also chosen by the Classic Repertory Company in Watertown, MA, for its 2019–2020 repertory season.

Prof. Gilligan and Carol Gilligan also wrote the libretto for an opera, *Pearl*, in collaboration composer Amy Scurria, and producer/conductor Sara Jobin, which was performed at Shakespeare & Company in Lenox MA, starring Maureen O’Flynn, John Bellemer, Marnie Breckenridge, John Cheek, and Michael Corvino, and in Shanghai China, starring Li Xin, Wang Yang, John Bellemer, and Lin Shu.

Schedule of Classes (Subject to Change)

Important Note: This schedule gives a rough indication of the reading for each day. See the detailed daily assignments on the course web site at `https://ees5891.jgilligan.org`.

Date	Topic	Reading
Thu., Aug. 25	Introduction	No reading
Tue., Aug. 30	Rethinking statistics	*McElreath* Ch. 1–2 ("The Golem of Prague" and "Small Worlds and Large Worlds")
Thu., Sep. 1	Sampling	*McElreath* Ch. 3 ("Sampling the Imaginary")
Tue., Sep. 6	Geocentric Models	*McElreath* Ch. 4 ("Geocentric Models")
Thu., Sep. 8	Many variables	*McElreath* Ch. 5 ("The Many Variables and The Superfluous Waffles")
Tue., Sep. 13	Designing statistical models	*McElreath* Ch. 6 ("The Haunted DAG & The Causal Terror")
Thu., Sep. 15	Regularization	*McElreath* Ch. 7 ("Ulysses' Compass")
Tue., Sep. 20	Interactions	*McElreath* Ch. 8 ("Conditional Manatees")
Thu., Sep. 22	Monte Carlo sampling	*McElreath* Ch. 9 ("Markov Chain Monte Carlo")
Tue., Sep. 27	Generalized linear models	*McElreath* Ch. 9 ("Markov Chain Monte Carlo")
Thu., Sep. 29	Generalized linear models	*McElreath* Ch. 10 ("Big Entropy and the Generalized Linear Model")
Tue., Oct. 4	Discrete statistical models	*McElreath* Ch. 11 ("God Spiked the Integers")
Thu., Oct. 6	Mixture models	*McElreath* Ch. 12 ("Monsters and Mixtures")
Tue., Oct. 11	Discussion of student projects	No reading
Thu., Oct. 13	Fall Break	No reading

Date	Topic	Reading
Tue., Oct. 18	Multilevel models	*McElreath* Ch. 13 ("Models with Memory")
Thu., Oct. 20	Multilevel models, part 2	*McElreath* Ch. 13 ("Models with Memory")
Tue., Oct. 25	More multilevel models	*McElreath* Ch. 14 ("Adventures in Covariance")
Thu., Oct. 27	Messy data	*McElreath* Ch. 15 ("Missing Data and Other Opportunities")
Tue., Nov. 1	Geospatial data analysis	*Gomez-Rubio*
Thu., Nov. 3	Laplace approximations	*Gomez-Rubio*
Tue., Nov. 8	Nested Laplace approximations	*Gomez-Rubio*
Thu., Nov. 10	Matern models	*Gomez-Rubio*
Tue., Nov. 15	Conditional autoregressive models	*Gomez-Rubio*
Thu., Nov. 17	STAN Hamiltonian Monte Carlo sampler	No reading
Tue., Nov. 22 Thu., Nov. 24	**Thanksgiving Break**	No reading
Thu., Dec. 1	Diagnosing model output	No reading
Tue., Dec. 6	Project presentations	No reading
Thu., Dec. 8	Project presentations	No reading
Sat., Dec. 10		No reading