MONT 105N – Analyzing Environmental Data
Information on Final Projects
February 26, 2019

General Information

As announced in the course syllabus, one of the major assignments for the seminar this semester will be a final project. You will be working on this project *either in pairs or individually, as you prefer* and the goals will be to prepare a roughly 15-page research paper and an oral presentation of approximately 15 minutes to the class on your project.

Schedule and Deadlines

- *On or before Wednesday, March 13* – Inform me *by email* which general topic you want to work on and whether you will work alone or who you will be working with (one email from each pair is OK, but please include both of your names so I know how the groups will break down). Since each project area has several different directions that might be pursued (see the descriptions below). To do this, write up a short paper of two or three paragraphs giving a description of the aspect(s) you would like to work on. As noted below several of the areas are large enough that if more than one person wants to try them, there will be ways to "split up" the topic into several parts. We can discuss the possibilities before this is due if you would like.
- *Monday, April 8* – All groups will submit, *by email*, an annotated bibliography of the sources to be used for your project. You should identify *at least six books, articles, or web sites, including no more than two web sites*, that will be relevant. For each of your sources, write up a short paragraph giving a rough description of how that source relates to your main topic, what kind of information you will take from it, and how you will be using it (including a preliminary estimate of how reliable you think the information there is). This part of the project will involve searching for sources, and this is where the library resources to be described by Ms. Merolli will be useful.
- *During the weeks of April 8 and April 15* – Everyone will meet with me during office hours (or at another time if that is not convenient) for a progress report and a chance to discuss any questions that have come up as you have started to work on the project. I will be happy to meet to talk over any aspect of the project at other times too, of course.
- *April 29, and May 1, 3, 6* – Each group will give a presentation to the class. We will have two or three presentations each of those days; the exact scheduling will be determined later.
- *Monday, May 6* – All final project papers will be due (via email) by 5:00pm. This assignment will take the place of a final examination for this course. Grading details and weights of each component of the final project are given in the course syllabus. Ask me if you have a question about those.

Other Information

- Ms. Merolli, our Science Librarian and a Montserrat Natural World Cluster member, will be visiting our class during the week after Spring Break to introduce herself and give some introductory information about using the library resources to identify sources for your project. She will be more than happy to assist you in the important process of assembling the resources you will use for your project.
- The presentations can be done either with Google Slides, or with PowerPoint. I will be happy to help out with technical details either way if needed.
- I will ask each group to do a "dry run" of their presentation with me at least one day before you go in front of the class. The purpose of this is to give you some feedback about what is working and what is not, and to give you some practice to minimize the effect of "nerves" when the time comes for the real thing.

The Final Project Paper

The goals of this assignment are for you to collect information about your topic from the various sources you find, and then present your analysis of that information. The evaluation of your project reports will be based on how well you have addressed the following guidelines and expectations:

- Distill your investigations into a central argument. A good research paper of this kind should be more than just a compilation of information from all the different sources you consulted. It should clearly show that you have thought independently about the information you found, that you have weighed the evidence for the various claims that were made in your sources, and that you have a central theme or argument about your topic that you want to present. It is certainly permissible to say you disagree with points of view presented in some sources, if you can explain why you think that and back up your opinions with appropriate evidence. (The templates from *They Say, I Say* that we discussed last semester can be a good way to think about structuring your argument.)
- Since our course has focused on techniques for understanding the patterns in data, *your topic should have some significant mathematical or statistical component.* This could come from analysis done as part of your project, or from learning about, explaining, and assessing mathematical or statistical work in some of your sources.
- The paper should be *well organized* and the writing should give the reader a clear indication where you are heading with your central argument at all times.
- Pay special attention to the first few paragraphs that will serve as an introduction. Catch the reader's attention, explain the significance of the topic or theme you will discuss. Say what you will do in general terms, without going into all the details from the start.
- Also pay special attention to the final few paragraphs, which will serve as a conclusion for your paper. Don't overstate the importance of your findings, and be honest if there are limitations. You might discuss how your investigations could be continued in further research.

- Give proper credit to sources you consulted that contributed to your ideas about the questions you studied. (In a longer thesis, it would be expected that another section reviewing the most relevant contributions of previous work on related subjects would be included – that kind of full literature review is *not expected for this assignment.*) Use footnotes or endnotes to identify direct quotations from your sources, and also to indicate the sources that contribute to specific points you are making.
- In a *References* section at the end, include all books, articles, websites you used in the preparation of the work. For books, give the author(s), title, publisher, place and year of publication. For articles, give the author(s), title, journal name, volume, year, and pages. For any websites, give the full URL, the author (if that can be determined), and the date you consulted.
- Be clear, concise, and correct in your writing. Aim for *no typos, misspellings, or grammatical problems.* But even more importantly, each paragraph should have a clearly evident purpose in relation to your main argument.
- Use figures, graphs, etc. sparingly in the main text. (If you want to include more of these, that can be done in an additional Appendix section at the end.)
- Proofread your work carefully and have an "impartial" reader or readers look at it and give you comments. This can be one of the other teams or me. Be prepared and willing to *revise* your work based on the comments you get. Of course, this means that *the writing must not be put off until the evening of May 5*(!) Be sure you get started early enough so that the input can be put to productive use.

Area 1 – A Systems Approach to Understanding Human Exploitation of Resources

Chapter 2 of Donella Meadows' book *Thinking in Systems* contains an analysis of how and why economic forces can lead to exhaustion of renewable and non-renewable resources and the ways that plays out. We mentioned some related topics at the start of the semester, but for this project, you would want to go much deeper. You would start out by researching the question: are there examples of this sort of using-up of resources that have occurred in the past or that are thought to be occurring now? (It might be interesting to look at the book *Collapse* by Jared Diamond.) What have been the effects on the human societies that did that? You will want to extend this to non-renewable resources too, but the two cases have some different properties. Then the second goal would be to develop and implement difference equation models exhibiting the results and patterns Meadows claims should be true in her book in Google Sheets or Excel.

Area 2 – Statistical Methods for Detecting Trends

This area would be good for students eager to explore additional statistical methods and their applications. The mathematical prerequisites are higher for these topics than for those in some of the other areas below.

Several possible project topics described below deal with using statistics to understand trends and patterns in data that change in some way over time. The technical name for this kind of thing is *time series data.* For example, a time series might represent the

concentration of a chemical in a water source measured in successive weeks, the temperature at a particular location at particular time of day over a range of days, the GNP of a national economy over a sequence of years, or even the batting average of a baseball player over the years of his career. We can think of breaking a time series up into several components:

$$\text{Time Series} = \text{Trend} + \text{Cycle} + \text{Residual},$$

where the Trend might be an upward or downward movement, the Cycle might represent a regular, repeating changing pattern (e.g. the normal temperature changes due to the change of the seasons), and the Residual represents random variation. One of the most important questions to ask in dealing with time series data is whether there is a way to identify whether there is some increasing or decreasing long-term trend involved in a given time series.

There are many different methods that statisticians use to address questions of this type. You should look for sources published by the U.S. Geological Survey on this. There is one in particular, titled "Statistical Methods in Water Resources" by D.R.Helsel and R.M.Hirsch (part of a larger source book of methods published by the USGS). The "Correlation" chapter (Chapter 8) and "Trend Analysis" chapter (Chapter 12) in this online book give a good (but perhaps somewhat technical) overview of several approaches (geared toward applications in water resources management).

One basic method would be to use a linear regression of the time series data against time as the independent variable. The slope of the regression line can then be used to decide if there is an upward or downward trend over time, as described above. However, if the trend is not linear, or if some of the necessary technical conditions necessary for hypothesis testing on the regresion coefficients are not met, then this approach is not appropriate.

There is an alternative class of methods called *nonparametric* statistical methods, including in particular a method called the *Mann-Kendall test* for trends, that do not require any of the assumptions needed for regression. As a result, the Mann-Kendall test has been very widely applied to study time series arising in areas such as pollution control, climate science, and other areas. The Mann-Kendall test works like this. Call the time series x_i, for $i = 1, \ldots, n$.

- Essentially the only assumptions necessary (for the basic version) are
  (1) if there is a trend, then it is *monotone* (either increasing for the whole time period, or decreasing for the whole time period),
  (2) there is no nonzero periodic Cycle term, and
  (3) the Residual term is purely random (not "autocorrelated").
- The test proceeds by computing for each time i and all later times $j > i$ the sign of the difference $x_j x_i$ ($+1$ if $x_j > x_i$, 1 if $x_j < x_i$, and 0 if $x_j = x_i$). Let S be the sum of all of these signs.
- A statistic called the *variance* V is computed by the following formula:

$$V = \frac{1}{18}\left(n(n-1)(2n+5) - \sum_{k=1}^{g} t_k(t_k-1)(2t_k+5)\right),$$

where g is the number of different groups of "ties" in the time series data, and t_k is the number of terms x_i in the kth group of ties for $1 \leq k \leq g$. (If there are no duplicate values in the time series, then this last term is zero.)

- Then the statistic

$$Z = \begin{cases} \frac{S-1}{\sqrt{V}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{V}} & \text{of } S < 0 \end{cases}$$

is computed and this is used to infer the presence or absence of trends as follows.

- Provided that $n > 10$, Z is approximately normally distributed, so we can set up confidence intervals or hypothesis tests with rejection regions specified by the percentage points of the standard normal curve.
- For instance, for a (upper-tail) test of the alternative hypothesis H_a : There is an increasing trend in the data, versus the null hypothesis H_0 : there is no upward trend, at the $\alpha = .05$ (Type I error probability) level, we would reject H_0 if $Z \geq 1.645$, and not reject H_0 if $Z < 1.645$. As usual with hypothesis testing, the rationale for why the test is set up this way is that the chance that Z has a value greater than 1.645 is only .05 if the null hypothesis is true. There are corresponding two-tail tests as well, where the alternative hypothesis would be that there is some trend (either upward or downward).
- For smaller length time series, $n \leq 10$, there are tables available in many books and online that replace the standard normal table for setting up the corresponding rejection regions.

Among the advantages of the Mann-Kendall test are that:

- it applies very widely,
- its conclusions are not greatly affected by gross errors or outliers (because it is only whether an increase or a decrease has occurred from one time period to another that matters, not the magnitude of the change), and
- the computations can be carried out easily in a spreadsheet, or even by hand if necessary.

One disadvantage is that it does not apply (and generates misleading results) when an upward or downward trend is combined with a cyclic (seasonal) variation. (You might experiment, for instance with the monthly Mauna Loa atmospheric CO_2 dataset to see what happens.) There are "souped-up" versions that deal with seasonal variations as well, though ("Seasonal Mann-Kendall" tests). Other "corrections" have been devised to deal with cases where the Residual term is not purely random.

As indicated above, it is certainly possible to compute the Mann-Kendall Z-statistic by hand if necessary. However, as you can probably guess, for long time series this can be somewhat tedious and prone to computational errors if you are not careful. For that reason, it is much more common to perform the calculations in software. Many of the

major commercial and research statistical software packages contain commands or have add-on packages to do this computation. There is also a commonly-used Excel template spreadsheet called MAKESENS (developed in Finland for environmental applications) that is set up to perform these calculations. See me early if you want to use this so that we can get technical issues about obtaining and using this resolved.

Topic ideas

(a) Implementing the Mann-Kendall and Seasonal Mann-Kendall Tests. Even though there are available Mann-Kendall spreadsheets and packages available for general use, I am still a firm believer that when you are learning a new computational process, then it can be very valuable to convince a computer to do it for you (;)) by programming the process (either in a spreadsheet, or in some other sort of programming environment). For this project, the goal would be to develop your own Mann-Kendall and Seasonal Mann-Kendall procedures and test them thoroughly on various inputs. Others doing projects in this area would be using the Mann-Kendall test more or less as a statistical "black box"– your presentation would involve digging a bit deeper into exactly how the tests work and explaining some of the fine points. Note: If you choose to work on this one, you should have a fairly high level of skill and experience in either spreadsheet macro programming, or in programming in some other environment (e.g. C++, Java, etc.) If you have never done anything like this, it might be better to consider a different part of this topic.

(b) Analyzing Trends in Water Resources Data. The USGS publication mentioned above has exercises that could form the basis for good more technical project topics. The data sets are either presented in the questions, or are available from the web page associated with the publication in an Excel spreadsheet or a text data file format. For this project, you would basically develop solutions for one or two of those exercises by working with the data and Excel. (You might copy the given data into the MAKESENS template, for instance, to do the calculations.) Then analyze and discuss your results. (Note: It would be OK if more several people wanted to work on these data sets. If so, we would just need to split things up so that you were looking at different questions. Also, when the questions say "use all the methods presented in this chapter, it would be OK just to do a regression and then a Mann-Kendall analysis, and compare and contrast the conclusions.)

(c) Analyzing Trends in Time Series on the Environment and Climate Change. This topic would be similar in spirit, but somewhat more open-ended than the previous one. There are a large collection of interesting data sets related to measurements of levels of various "greenhouse" gasses such as CO_2, NO_2 etc. in the atmosphere, temperatures, rainfall amounts, snow cover levels and durations, and many other things available for download from the web site of the U.S. Department of Energys (DOE) Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), a new data archive for Earth and environmental science data funded by the Data Management program within the Climate and Environmental Science Division – `https://data.ess-dive.lbl.gov`. (Use of some the data here requires a registration, but it is free and seems innocuous.) This contains all the data formerly housed in the CDIAC (Carbon Dioxide Information Analysis Center) archive. In a way, some of these data sets are too simple for the kind of trend analysis provided by regression or Mann-Kendall. For instance if you look at the time series of

atmospheric CO_2 levels measured at various locations on the Earth like the Mauna Loa Observatory data that we studied in the fall, there is a clear monotonically increasing trend in the yearly averages that requires no statistical analysis whatsoever(!) On the other hand, whether there are trends for some of the other trace gas amounts (e.g. carbon monoxide, chloroform, carbon tetrachloride, chlorofluorocarbons, etc.) is far less obvious. One very good project here would be to "pick your favorite trace gas(es)," look at the various data sets available here and perform trend analyses, comparing results from different locations. The measurements from the CSIRO Gaslab flask sampling network are especially good here because there are 9 different locations that measured various gas levels monthly over the period from 1992 to 2001. (Note: There is certainly enough to do for several people to work in this general area.) For climate data, perhaps the most interesting collections of data sets for US climate data are on the United States Historical Climatology Network. (There is a link to here from the CDIAC web site mentioned above.) One interesting question is how precipitation amounts have changed at the locations where temperatures have increased over the past 50 or so years. Do you find significant associations? (Note: Again by looking at different geographical areas, 2 teams could work on different aspects of this.)

Area 3 – Other More Advanced Mathematical and Statistical Topics

For those who would like to go farther with the mathematical and/or statistical ideas we have introduced in this course, here's one possible group of topics to consider.

(a) As we hinted in class, a normal distribution is not always appropriate as a model for the distribution of some quantity in a population. The mathematical theory of statistics has developed a whole repertoire of other model distributions for this reason. One of the most important for biological and environmental applications are the *lognormal* distributions. For this project, you would research these and some of their applications in biology. What does the probability density function of a lognormal random variable look like? What kinds of skewed distributions can be obtained in this way? What are the theoretical expected value and variance? What quantities are (empirically) well-described by lognormal distributions and what are some examples of data that can be analyzed using the properties of lognormal distributions? (Note: To choose this topic, you should have completed or be currently enrolled in Calculus II, since some familiarity with integration will be necessary.)

(b) Another important class of distributions are the so called χ^2 (chi-square) distributions. These are introduced briefly in Chapter 11 of our course textbook and we will discuss using them for testing independence of random variables in class. For this project topic, you would research another of their most important applications – the so called "goodness of fit" tests. What does the probability density function of a χ^2 random variable look like? What kinds of skewed distributions can be obtained in this way? What are the theoretical expected value and variance? What quantities are (empirically) well-described by χ^2 distributions? Then, how is the X^2 statistic used to measure the goodness of fit of a model? There is a very nice discussion of this in the book *Statistics* by Freedman, Pisani, and Purves (as well as in many other elementary statistics texts. The Freedman, Pisani, Purves text also includes a very interesting discussion of the way the statistician

R.A. Fisher used this test to question the results of Gregor Mendel's famous original genetics experiments on pea plants. Were Mendel's results "too good to be true?" Did he "cherry-pick" his data? If so, was that defensible on grounds of experimental ethics? Does your answer change knowing that later experiments showed his conclusions were correct even though his data analysis methods might have been questionable? (Note: To choose this topic, it would help to have completed or to be currently enrolled in Calculus II, since some familiarity with integration will be useful.)

(c) Last semester, recall that we used linear regression extensively to fit linear, exponential, and power law models to data. There are also strong statistical applications of this idea. In particular, it is possible to do hypothesis testing on the coefficients in a regression line $y = \beta_0 + \beta_1 x$ to determine, for example, whether it is reasonable to assume $\beta_1 > 0$ (or $\beta_1 < 0$, or β_1 is nonzero), under certain assumptions on the variability in the data. (This is the idea behind regression used as a trend-detection method for time series, as discussed above.) For this topic, you would research these methods, how they are applied, and discuss some examples. No calculus is really required for this one, although you should be confident about mathematical computations, etc. The results of doing these tests are generated automatically by Google Sheets (and Excel) in an alternate method for finding the regression line called the LINEST command. You would use that to compute several examples and explain the results. This is discussed in a general way in Chapter 12 of our class text. You will probably want to consult other statistics texts for more details and examples.

(d) One key method used in the generation of the "hockey stick" graph in Michael Mann's original work on rising global temperatures was a statistical technique known as *principal components analysis* (PCA). (See his book "The Hockey Stick and the Climate Wars" for more information.) For this project topic, you would research the basics of this method and try to understand in detail what it does and how it is applied. This would probably be the most challenging of all these topics for several reasons. First, you should probably only attempt this if you have completed or are currently taking Multivariable Calculus (MATH 241). In addition, some basic familiarity with matrices and general systems of linear equations would be important. Second, in order to use PCA on real data, you would need to use more advanced statistical software than Google Sheets or Excel. The open-source package `R` can be used for this, though. See me soon if you want to try this topic.

Area 4 – Environmental and Political Topics

(a) New York Times op-ed columnist Thomas Friedman has been a consistently interesting and provocative contributor to our public discourse over the past 20 years through his columns and books about the Arab-Israeli conflicts in the Middle East, the influence of globalization on the world and local economies, and the challenge of climate change. His book *Hot, Flat, and Crowded* (originally published in 2008, and updated in 2009) would be the basis for this project. The title refers to Friedman's characterization of our world as we move into the 21st century – "hot" because of climate change, "flat" because of the influence of globalization and the rise of western-style consumer cultures in formerly third-world areas of the world such as India and China, "crowded" because of the rising

human population. All this sounds dire, but Friedman tries to argue that the first challenge here (climate change) is actually an *economic opportunity* for any country able to muster the political will to face the reality and make real changes in how energy is produced and consumed. By addressing that challenge we could also start to address the others as well. For this project, you would read Friedman's book, summarize his arguments, then try to assess how realistic his ideas are and how influential this point of view has been (or not). For instance, it would be good to look for reviews of the book from major newspapers and periodicals to understand criticisms people have had.

(b) Harvard Forest is an ecological research station in Petersham, MA where much research on changes in the New England landscape over time (and also research on understanding forests' role in the global carbon cycle) has been done. The starting point for a good project would be to look at the article "Ecology and conservation in the cultural landscape of New England" by David Foster and Glenn Motzkin from the journal *Northeastern Naturalist* (available from JSTOR via the Holy Cross Libraries web site). What conclusions do Foster and Motzkin reach about how central Massachusetts landscapes have changed over the past 250 years or so? How do they illustrate these conclusions with statistics and graphical displays of quantitative information? Other work by the same authors addresses different aspects of these questions and they would be good to bring in as well.

(c) The release of the 2006 documentary film "An Inconvenient Truth" and the 2017 sequel "An Inconvenient Sequel" about former Vice President Al Gore's efforts to raise public awareness of issues of climate change was seen by some as a major turning point in the environmental movement. It was also one of the first efforts to bring Mann's "hockey stick" graph of average global temperature estimates to the attention of the general public. Others have harshly criticized it for various reasons on scientific and/or political grounds. For this topic, you would research the topics that were included in the movie and try to evaluate how well its predictions and the concerns it raised have been borne out since the original film from 2006 and the follow-up from 2017. You would also research some of the criticisms by reading the original sources and try to assess them. Are they reasonable criticisms or are they the same sort of propaganda efforts that Michael Mann describes in "The Hockey Stick and the Climate Wars."

Area 5 – "Cultural Cognition" and Perceptions of Risk

These project ideas might be most interesting to those are planning to major in a social science and or in psychology. There would be some statistical work involved, but that could be mostly devoted to understanding and explaining the statistical techniques used in some particular studies mentioned in the next paragraphs. I'm thinking that one project topic could be based around the issues discussed in each one of the following articles from the the Cultural Cognition Project at Yale Law School. This group has been pursuing a very interesting program investigating the ways individuals' general world views and cultural values influence their thinking about what are ostensibly purely scientific questions. This work is provocative because it provides a somewhat plausible explanation for the degree of polarization in our current politics over questions such as climate change, safety of nuclear energy, public health measures such as immunization, and others. One hope I had in designing this Montserrat course was similar to hopes that many other mathematics

and science educators have expressed over the past 10 or 20 years. Namely, by raising students' mathematical literacy, showing them how scientists use mathematical techniques to study the natural world, and looking at key aspects of the evidence we have for effects like climate change, students would come to realize that we are facing a number of serious environmental challenges and begin to work for changes in the way we use natural resources and influence the environment. But is that necessarily the case? Does a higher level of scientific and mathematical literacy necessarily correlate with increased tendency toward environmentalism?

(a) The paper "The tragedy of the risk perception commons" examines this idea and seeks to determine whether increasing scientific and mathematical literacy necessarily leads to a greater recognition of the risks involved in climate change.

(b) The paper "Cultural cognition of scientific consensus" investigates how cultural values influence the general public's estimates of the degree of scientific consensus on issues such as climate change. It would also be interesting to consider how their conclusions relate to the view expressed by Michael Mann in "The Hockey Stick and the Climate Wars" that tobacco companies, energy companies, and others have systematically sought to *create doubt* in the minds of the nonscientific general public by exaggerating the degree of scientific disagreement about whether tobacco use causes lung cancer, or whether human actions are changing global climate.

(c) Another paper by members of the same group entitled "Who Fears the HPV Vaccine, Who Doesn't, and Why?" examines the questions about the administration of the human papilloma virus vaccine (HPV – the virus responsible for many cases of cervical cancer in women) to young girls before they become sexually active and investigates how people's perceptions of the risks involved in vaccination are shaped by their other cultural values.

The goals of these projects would be to read the paper mentioned and other articles that led up to it to understand what the authors are saying about these questions. What is the "cultural cognition" theory? Exactly what does the study conclude? What are the factors that dispose people to be environmentalists or to ignore environmental issues or discount the evidence for problems such as climate change? What are the bases for the conclusions of the Cultural Cognition Group? How would you evaluate their reliability?

Area 6 – A Topic of Your Choice

If there is another topic you would prefer to work on – for instance a project analyzing data from another course, or something on other environmental topics you want to look at, perhaps from your current events journals – I am open to suggestions. If you want to propose a topic of your own, you *must get my approval* before starting to work. For the March 13 deadline above, write up a short but reasonably detailed description of the topic or questions you want to look at and how you want to try to address those questions. I will let you know as soon as possible whether you have my approval.