MONT 103N – Analyzing Environmental Data
Final Projects
February 24, 2012

General Information

As announced in the course syllabus, one of the assignments for the seminar this semester will be a final project. You will be working on this project in *teams of 2 or 3* and the goals will be to prepare a roughly 15-page research paper and an oral presentation of approximately 25 minutes to the class on your project.

Schedule and Deadlines

- **On or before Friday, March 16** – Inform me *by email* which general topic you want to work on and with whom you will be working. Since each project area has several different directions that might be pursued (see the descriptions below), write up a short paper of two or three paragraphs giving a description of the aspect(s) you would like to work on and what you hope to do with the project. If you need assistance in forming "teams," I will be happy to help with that. As noted below the areas are large enough that if more than one group wants to try that, 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.
- **Friday, April 13** – Each group will submit, *by email*, a 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). The project descriptions below contain some first places to look, but you should plan to spend some time searching for additional sources.
- **During the week of April 16** – Each team 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 30, and May 2, 4, 7** – Each group will give a presentation to the class. We will have two presentations each of those days; the exact scheduling will be determined later.
- **Monday, May 7** – 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 overhead projector 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

This writing assignment will be different from the "opinion paper" on the "Connections" event that we did earlier this semester, and also different from the smaller-scale research paper from last semester (on *Collapse*). 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.
- Since our course has focused on techniques for understanding the patterns in data, *your topic should have some significant statistical component.* This could come from analysis of data that you collect as part of your project, or from learning about, explaining, and assessing 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 6*(!) Be sure you get started early enough so that the input can be put to productive use.

Project Ideas

Area 1 – A Holy Cross Campus Project

This area would be good for students looking for a project with a strong community action focus. There is enough to do here so that one or two groups could work on aspects of this. Collaboration between the groups would be fine too.

When people think of water pollution, they often envision big factories dumping toxic waste or raw sewage into a stream, lake or ocean. Although industrial polluters certainly can and do damage water quality, surface water run-off from agricultural fields, parking lots, construction sites, city streets and lawns probably contributes more to the nation's water pollution problems (because so much more area and water is involved).

In urban or suburban settings, water flows across impervious surfaces (surfaces that cannot be penetrated by water) such as streets, parking lots, and hard-packed soils. As it does so, it picks up contaminants like oil, sediment, excess fertilizer and pesticides, other wastes, etc. along the way. This polluted "urban run-off" flows into the storm-water system (via grates in the pavement) and then to waste-water treatment plants, or directly into water bodies. Pervious surfaces (surfaces that allow water to penetrate) allow rain and melting snow to *infiltrate* into the ground, which helps recharge groundwater and remove or break down pollutants. As background, you should look up sources about these issues, find estimates for how much pollution can be attributed to these sources, and ideas that others have generated for how run-off problems can be controlled.

A good project in this area would focus on types of land-use in the semi-urban setting of the Holy Cross campus. Although you may never have thought about this, there *is a fairly serious run-off problem here.* A large amount of run-off water flows from the Holy Cross campus into the nearby Blackstone River at the bottom of College Hill. This is aggravated by the steep slopes of our hillside campus. (If you have been out on College Street during a heavy rainstorm, you know what I am talking about!) Features of our land-use contribute to water pollution through run-off or help prevent water pollution through infiltration.

For this project, one or possibly two groups would begin by measuring categories of land-use for different areas of the campus. Possible categories are

1) ground area lying under roofs of buildings ("building footprint")
2) surface area of impervious (concrete, asphalt, closely packed brick) pedestrian pavement
3) surface area of vehicle pavement
4) surface area of gravel or more pervious brick walkways
5) surface area of grass
6) surface area of shrubs
7) number of large trees

Categories 1-3 indicate the amount of impervious surface, whereas Categories 4-6 indicate the amount of (more or less) pervious surface. Large trees (Category 7) reduce surface water run-off by soaking up water and releasing it slowly into the ground and air.

This project would involve quite a bit of actual data collection as well as some statistical analysis. The data collection phase of the project would involve first deciding exactly which areas of the campus will be surveyed – will you take a sample of areas, or try to survey the whole campus? Depending on how many people are involved, you might be able to make a complete survey. (Note: Quite detailed aerial maps of the campus are available in Google Maps, and these will be helpful in deciding how to split things up.) The first goal would be to get a good estimate of the exact area in each of Categories 1 - 6 within the part of the campus you are looking at, and get a good count of the number of large trees (say over 15 feet tall). This will involve a fair amount of "field work" actually walking the campus, measuring lengths and widths of pavements, buildings, plantings of shrubs, counting trees, etc.

The intermediate goal would then be to develop a Runoff Index (RI) to give a quantitative "rank" or "score" for an area's land-use characteristics that contribute to surface water run-off. The RI should allow you to pinpoint problem areas that are producing more run-off than others. This would involve some fairly extensive statistical work with the data you collected in the first phase, and you will need to decide exactly how the RI is to be defined.

Then, the final goal would be to use your RI scores to identify areas on the campus that are contributing especially heavily to run-off. Where are those areas? Are there things the College could do to improve the current situation? Which of those measures would likely be the most expensive? Which would be most affordable?

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 highest for these topics.

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 (a chapter in a larger source book of methods published by the USGS) that gives 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. 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 con- clusions 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 even be carried out 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. 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

1. 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.

2. 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 two group 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.)

3. 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 Carbon Dioxide Information Analysis Center (CDIAC) at Oak Ridge National Labs. 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 2 teams 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 – Perceptions of Environmental Issues Such As Climate Change

This project idea might be most interesting to those are studying social science and or psychology. There would be some statistical work involved, but that could be mostly devoted to understanding and explaining the statistical techniques used in a particular study mentioned in the next paragraph.

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? Some recent research by the Cultural Cognition Project at Yale Law School (in particular their paper titled "The Tragedy of the Risk Perception Commons") has thrown doubt on that. The first goal of this project would be to read that paper and other articles that led up to it to understand what the authors are saying about these questions. Exactly what does this 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? If you wanted to, you could try administering a survey here at the College that was similar to the one they report, but this would require some advance planning and some very careful design of the questions. If you decided to try this, then there would be a fair amount of statistical work to do to analyze the results.

Area 4 – A Topic of Your Choice

If there is another topic you would prefer to work on, 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 16 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.