Why study data analysis?

In August 2016, the following message was posted to all-at-sio:

> Last Saturday morning, the surface water temperature off La Jolla Shores was 77°F.
>
> This morning, it's 63°F.
>
> These are dive computer temperatures, so I went to check the CORDC gauges online, but it looks like those temperatures gauges are offline? Same with SCCOOS... but maybe those are pulling data from the same gauges? Are there any SIO sea surface temperature gauges currently online-to show us the last week or two?
>
> My dive buddy's theory was that this morning we were diving in the middle of a big incoming tide (note the full moon) and that was causing a lot of mixing of deeper water and explained the huge drop in sea surface temperature. I was skeptical.
>
> So who's right? Do big tidal flows cause lower sea surface temperatures?
>
> And what in the world caused this 14°F drop in six days???

You might be wondering about a couple of acroyms. CORDC is the Coastal Observing Research and Development Center (based at SIO) and SCCOOS is the Southern California Coastal Observing System. And you might also feel like we should use the metric system for temperature. Here 77°F = 25°C, and 63°F = 17.2°C, so the temperature drop 7.8°C. is The post raises a number of issues for us as analyzers of physical oceanograpic data. Here are some possible questions?

1. What's going on with the data sources? What should they provide, and why did the posters find that the observations were not available?

2. What data do we trust? How reliable is a dive computer and how does dive computer accuracy compare with pier sensor accuracy?

3. What is the typical temperature at the Scripps pier? And what is a typical range of variability? Is this event unusual?

4. What are the mechanisms behind this change? How much confidence do we have in our assessment of the mechanisms?

So having posed a set of questions, what do we do now? Figuring out what to do next is the goal for this class, and in general for your courses in physical oceanographic data analysis.

Introduction to the class

Welcome to SIOC 221A. This is the first in a 3-quarter sequence about data analysis in physical oceanography. The course is appropriate for first-year grad students, and it serves as an introduction to the basics of data analysis. Because we can't cover all of data analysis in one quarter, we'll focus this quarter on time series analysis (more on that later). This course is also your introduction to the scientific computing resources that are available to you at SIO, and we'll spend some time getting everyone up to speed on computing resources. First some basics.

What happens in each of the three quarters of this course sequence?

SIOC 221A. Time series analysis. We'll build data analysis skills with an aim to understanding Fourier transforms, spectral analysis, and how to interpret data using these tools.

SIOC 221B. Everything else. Random variables, matrix inversion, the details of least-squares fitting, objective mapping, empirical orthogonal functions.

SIOC 221C. Laboratory class. Project-based examples to put the techniques of data analysis into practice.

What are the expectations for this course? See the syllabus.

1. Weekly problem sets, which you may do collaboratively.

2. Midterm and final problem sets, which you MAY NOT do collaboratively. The final problem set will be a bit broader than the others and will involve working through a data set that you choose (or one suggested by your advisor) to test out the techniques that we explore in this class. It will be somewhat open ended, and you'll do a brief in-class presentation during our final exam time slot.

3. Learning something new. Everyone starts this class with a different set of prior experiences. Some of you are experienced programmers; others are new to Matlab. Take advantage of this class to master a new skill (e.g. learn Matlab, learn python, learn version control throug git hub, learn shell scripting). Respect everyone's differing background.

4. Come to class and participate in discussion. Ask questions. Learn collaboratively.

5. Do assigned reading. Review course notes.

What will we learn in this class?

By the end of the course, you should come away with more confidence in your programming and a solid understanding of time series analysis, the Fourier transform, and spectral methods.

Data analysis is at the heart of evidence-based decision making. This is what defines the science that we do—all science in fact. Data analysis provides you with the tools to think about uncertainty, to allow you to be skeptical and to set limits on your skepticism.

What should I read? See the syllabus for some examples of excellent books that will help broaden your understanding. The books are available on reserve from the Eckart Building. Course notes are fairly practical, but you want to come away from this course with enough knowledge to understand the notation when you need to consult a statistics book for a slightly more complicated topic. Three of the titles are also available in electronic form. We'll rely quite a bit on Bendat and Piersol, and you can start by reading Ch. 1, with a focus on sections 1.1 and 1.3.

What is a time series?

Let's start by looking at some time series examples. The slides show a range of different time series, including winds in the region of Hurricane José, which bounced against the New England coastline in September 2017, Arctic ice cover, atmospheric concentrations of CO_2 from Mauna Loa, and temperature and pressure at the Scripps pier.

Simply put, a time series is a one-dimension data set that we can represent as a vector. For example, we can look at data collected at a single point and ask how they vary over time. Sometimes time series are not collected at a single point. We might examine the time series collected by an eXpendable BathyThermograph (XBT) as it free falls into the ocean. Officially an XBT makes measurements as a function of depth, but the engineering variables typically sample at a fixed sampling rate (e.g. 5 Hz), which means it can be useful to think of the data as a time series.

And sometimes we use the methods of time-series analysis to look at spatial records. Instead of asking about temporal variability (or variability in frequency space) we ask about spatial variability (or variability in wavenumber.)

Data vs models

What governs sea surface temperature at the Scripps pier. In an ideal and noise free world, we could hypothesize that upper ocean temperatures are controlled by seasonal changes in air-sea heat fluxes, Q_{net}. Thus

$$\frac{dT}{dt} = Q_{net} \tag{1}$$

If Q_{net} were sinusoidal (e.g. $Q_{net} = -q\cos(\omega t)$, where $\omega = 2\pi/365.25$, and t is in days, then we could solve for temperature:

$$T = T_o - \frac{q}{\omega}\sin(\omega t), \tag{2}$$

and we'd have a simple sinusoidal temperature. Real data are never so simple, and we inevitably have a signal that combines multiple factors:

1. Physical forcing we understand (e.g. the annual cycle in this case.)
2. Physical forcing that we might have hoped to ignore (e.g. the diurnal cycle, or tidal processes.)
3. Instrumental noise (e.g. instrumental inaccuracies.)

What can we learn from our time series?

So as an example, how can we characterize the Scripps pier sea surface temperature record? The pier system reports measurements every 5 minutes and 40 seconds. We can label each individual measurement as T_i to denote the ith realization of our temperature measurement. Each T_i can be referred to as a *random variable*, and for the moment we'll assume that each observation is independent.

If we go measure temperature at the pier, what do we expect to see? The mean or average of the temperature measurements is:

$$\text{mean}(T) = \overline{T} = \langle T \rangle = \frac{1}{N}\sum_{i=1}^{N} T_i. \tag{3}$$

This is the population mean. The distinction between overlines and angular brackets is a bit arbitrary, and it is up to you, the data analyzer, to explain how you are using the notation. For example, you might see angular brackets used for spatial averages and overbars for temporal averages (or vice versa).

Bendat and Piersol discuss the distinctions between the mean of an ensemble of realizations representing the same point in time, and the time average of a data value collected over time. Formally we like to think of the mean (the first moment) as the value that we would obtain if we sampled at time t_1 repeatedly:

$$\mu_T(t_1) = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} T_k(t_1), \tag{4}$$

but in reality we don't have an infinite number of samples at t_1, so we often rely on the *ergodic* assumption that the statistics are stationary so that variability in time and/or space is not significant (or that variability in time can be interchanged with variability in space, if we sampled one but not the other. Thus

$$\mu_T(k) = \lim_{N\to\infty} \frac{1}{T} \int_0^T T_k(t)\, dt. \tag{5}$$

Statisticians specify Greek letters for "parameters" determined from an entire population, and Roman letters for "statistics" determined from a finite-sized data sample. Compare the notation for the population mean $\overline{T}$ in (3) and the parameter μ in (4) and (5).

We can leave a detailed discussion of these distinctions for the future. But we'll want to remember that the expectation value (denoted with E) of the mean is defined to represent the value that we would expect if we could repeat our measurements many times:

$$E(T) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} T_i. \tag{6}$$

(Formally, the expectation value is weighted by the probability of occurrence, but the terminology is probably used most office for derived quantities, when a Greek letter would complicate the notation.)

Knowing the mean temperature is great, but assuming that the temperature we measured today doesn't match the mean, how can we tell if the deviation from the mean is typical? As a first measure, we can compute the variance:

$$\text{var}(T) = E[(T - E(T))^2]. \tag{7}$$

When we deal with finite numbers of observations we need to be a bit careful:

$$\text{var}(T) = \sigma_T^2 = \frac{1}{N-1} \sum_{i=1}^{N} (T_i - \overline{T})^2, \tag{8}$$

and the standard deviation, which is the square-root of the variance:

$$\text{std}(T) = \sigma_T = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (T_i - E(T))^2}. \tag{9}$$

Why do we divide by $N-1$ instead of N? Here are a couple of ways to think about this:

1. In the limiting case in which we have only 1 sample, so $N = 1$, we can estimate the mean (badly) as our one measured value (T_1), but we don't know anything about the variability, so we need our standard deviation to be undefined. In essence, all our information has been allocated to estimate the mean, and we don't have enough degrees of freedom to estimate the standard deviation as well. Since one degree of freedom was needed for the mean, only $N - 1 = 0$ degrees of freedom are left for the standard deviation.

2. More formally, you can think about the problem this way:

$$\begin{aligned}
\text{var}(T) &= \frac{1}{M} \sum_{i=1}^{N} \left(T_i - \frac{1}{N} \sum_{j=1}^{N} T_j \right)^2 && (10)\\
&= \frac{1}{M} \sum_{i=1}^{N} \left(T_i^2 - \frac{2T_i}{N} \sum_{j=1}^{N} T_j + \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} T_j T_k \right) && (11)\\
&= \frac{1}{M} \sum_{i=1}^{N} \left(T_i^2 - \frac{T_i}{N} \sum_{j=1}^{N} T_j \right), && (12)
\end{aligned}$$

where we want to determine M. Rearranging,

$$\text{var}(T) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(T_i (T_i - T_j) \right). \tag{13}$$

For the summation over j, we'll end up using N values of T_j, but then when we start summing over i, every time $i = j$, $T_i - T_j = 0$, so we'll have only $N - 1$ non-zero terms in the sum. Thus we set $M = N - 1$ to avoid biasing our variance.

The mean and standard deviation give us quite a bit of information, but they don't tell us everything. What if temperature at the pier were generated by a coin toss, with only two values (e.g. 25°C or 17.2°C)? That would give us a bimodal distribution? With a hundred realizations, we could make a histogram of temperatures like Figure 1:

Histograms are useful, but they're hard to compare for different numbers of data points. To allow comparison, we normalize as a probability density function so that the area under the curve tells us the probability of finding an observed value within a given range. Formally, for a probability density function $p(x)$,

$$\int_a^b p(x)\, dx = \text{Prob}[a < x \leq b], \tag{14}$$

and the total area under the curve for $p(x)$ is fixed

$$\int_{-\infty}^{\infty} p(x)\, dx = \text{Prob}[-\infty < x \leq \infty] = 1. \tag{15}$$

If we use 1°C bins, then the pdf for our hypothetical bimodal data would be illustrated by Figure 2

On the other hand, we might find that pier temperatures are distributed uniformly around the mean. (In Matlab, the function 'rand' will generate a uniform distribution, which looks like a top hat.) A uniform distribution could have exactly the same mean and standard deviation as our bimodal distribution, so we wouldn't be able to tell them apart just by computing the mean and standard deviation.

More classically, we hypothesize that data have a Gaussian distribution, the classic bell-shaped curve, which can be represented as:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{16}$$

where μ is the mean and σ is the standard deviation. Again, we can define the Gaussian distribution to have the same mean and standard deviation as the bimodal distribution. Clearly we'll need some stronger tools to interpret our data.

So far we've talked about symmetric distributions, but data can have highly lopsided distributions. Think about rainfall in San Diego. There's no rain for months and months, and then suddenly we have 12 hours of continuous rain that deliver several mm of precipitation . A summer rainfall pdf would look like figure 3. In this case, the mean is non-zero, but the typical day has no rain. The median is useful for this. The median m is the value of x at the midpoint of the pdf:

$$\int_{-\infty}^{m} p(x)\, dx = \text{Prob}[-\infty < x \leq m] = 0.5 \tag{17}$$

So how typical is 2018 sea surface temperature data in La Jolla Shores? We can plot these data with the following procedure, and you can judge for yourself. First download the 2018 Scripps Pier data from the SCCOOS data server:

http://sccoos.org/thredds/catalog/autoss/catalog.html

Now load the data into Matlab:

```
% use ncdisp to survey the contents of the file:
% ncinfo is also useful
ncdisp('scripps_pier-2018.nc')
% or use the THREDDS server to download on the fly
ncdisp('http://sccoos.org/thredds/dodsC/autoss/scripps_pier-2018.nc');

%
% read the time and temperature variables
time=ncread('scripps_pier-2018.nc','time');
temperature=ncread('scripps_pier-2018.nc','temperature');
%
% note that time is measured in seconds since January 1, 1970
%  so define a reference date
date0=datenum(1970,1,1);
%
% plot the time series
plot(double(time/24/3600+date0),temperature,'LineWidth',3)
%  here we use double to force the time to be a real number
%  we divide the time by 24*3600 to convert seconds into days since 1970
% label the x-axis in months
datetick('x','mmm')
set(gca,'FontSize',16)
xlabel('months (of 2018)','FontSize',16)
ylabel('temperature (^oC)', 'FontSize',16)
```

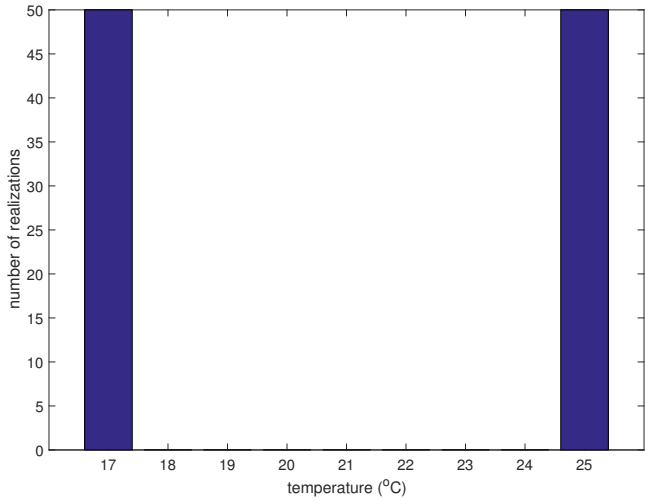

Figure 1: Bimodal histogram for hypothetical sea surface temperature distribution.

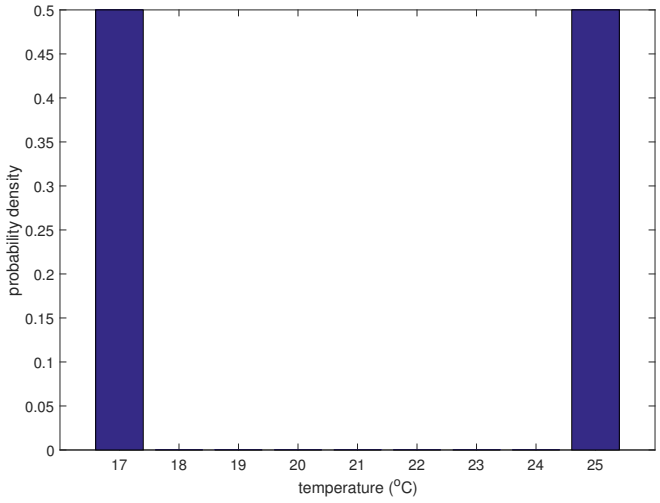

Figure 2: Probability density function for bimodal hypothetical sea surface temperature data shown in Figure 1.

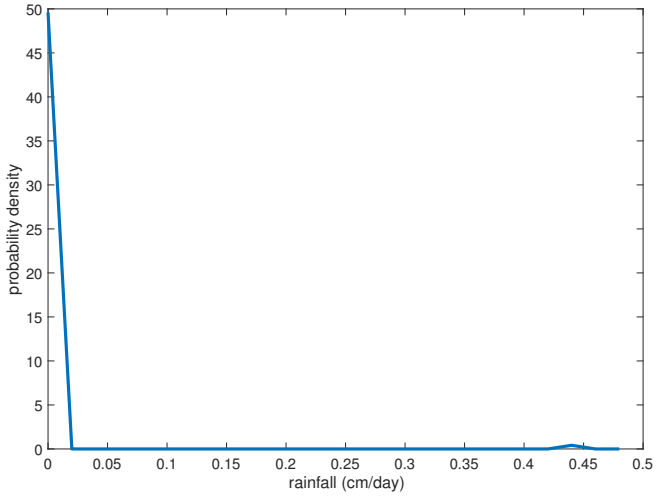

Figure 3: Probability density function for hypothetical southern California summertime rainfall. Most days have zero rainfall, and a small number of non-zero rainfall days occur.