1 Background on statistical calculations, EAS 471

1.1 Introduction

In your Computing Lab 1, you will be provided time series of velocity (components u, v, w) and temperature T measured by fast response sonic anemometers on a tower standing on a salt flat in Utah. The sampling rate was 20 Hz, ie. the time series have sampling interval $\Delta t = 0.05$ s, and the sampling duration was one hour. Your assignment is to compute the micro-meteorological statistics from these data, and this document provides some background.

1.2 Signal decomposition

It is an almost universal approach in the treatment of signals to separate any time series (or any other type of series) into its mean and fluctuating parts. I shall demonstrate this in the case of a time series, but the idea applies equally well to a spatial series. Let us take the velocity component $u(t)$ which we know to the extent that we have a time series $u^n = u(n\Delta t),\ n = 1...N$. There are many conventions to represent the mean of the series, the most common being U, $\overline{u}$, $\langle u \rangle$ and E[u] where E[u] is termed the "expectation value". The decomposition may be written in various terminologies as

$$\begin{aligned} u(t) &= U + u'(t) \\ u(t) &= \overline{u} + u'(t) \\ u(t) &= \mathrm{E}[u] + u'(t) \end{aligned} \tag{1}$$

where u' is the deviation from the mean, also called the fluctuation[1]. Note that there will in general be other independent variables (here I only show time, t, because we have had time series dumped on our lap, with no reference to location). And note too that notwithstanding that we have time-averaged, the average U may still be a function of time, but it will be a function of time that has little or no energy on the rapid timescales, for averaging amounts to filtering out rapid variation. Here it would be appropriate to talk about the variance spectrum $S_u(f)$ of $u(t)$, but let's not be distracted.

[1]Sometimes people call U the mean and u the fluctuation, in which case a separate symbol such as u_t is needed for the total signal, thus: $u_t = U + u$.

Note that the fluctuation is by definition

$$u' = u - U \tag{2}$$

and that *it has zero mean*, ie. $\overline{u'} \equiv 0$. This is something it is always interesting to check, in your calculations. Machine roundoff errors will result in non-zero values, but they should be small relative to the mean U (or $\overline{u}$, depending on your choice of terminology) - if not, you have made an error. It is important to understand that in the statistical description of a signal we mostly cite statistics of the fluctuation... ie. we will talk about the mean of the signal and properties of the fluctuations, most saliently the standard deviation

$$\sigma_u = \sqrt{\overline{u'^2}} \tag{3}$$

(the root-mean-square deviation of the signal from its mean).

The above formulae are silent as to the exact mechanism by which we form our averages. If we had a continuous signal, we might define

$$U(t) = \frac{1}{T_a} \int_{t-T_a/2}^{t+T_a/2} u(t')\, dt' \tag{4}$$

where t' is a dummy variable and T_a is the averaging interval. On the other hand if we have a discrete time series $u(n\Delta t) \equiv u^1, u^2, u^3, ...u^N$ on intervals Δt then

$$U = \frac{1}{N} \sum_{n=1}^{N} u(n\Delta t) \tag{5}$$

I assume I don't need to elaborate on the utility of the statistical analysis and description of signals. In a nutshell, signal statistics amount to a *compact* and unambiguous descriptor of a signal... this is much easier than saying, "well, the temperature went up, and down, but then up again, and at one point it was over 30 C, and then suddenly...", if you see what I mean. A complete statistical description of a single signal $u(t)$ would provide

- the probability density function (pdf), eg. for a Gaussian variable with mean μ_u and variance σ_u

$$f(u) = \frac{1}{\sqrt{2\pi}\ \sigma_u}\, e^{-\frac{(u-\mu_u)^2}{2\,\sigma_u^2}} \tag{6}$$

or, at a lower level of completeness, all the moments of the pdf: the first moment is the mean $\overline{u}$, the second moment about the mean is the variance $\sigma_u^2 \equiv \overline{u'^2}$ (whose square root σ_u is the standard deviation), the third moment $\overline{u'^3}$ is the mean of the cube of the fluctuation, and when normalized it gives us the "skewness" $Sk_u = \overline{u'^3}/\sigma_u^3$... and so on to fourth and higher moments

- the auto-covariance function $C_u(\xi) = \overline{u'(t)u'(t+\xi)}$ and the power spectral density $S_u(f)$, which convey information about the longevity of fluctuations or equivalently the rapidity of the variations in the fluctuation $u'(t)$

If we have more than one signal and the signals may be related, we will want to look at cross-statistics, such as $\overline{u'w'}$ which is called the "covariance" of the signals u, w and the correlation coefficient

$$R_{uw} = \frac{\overline{u'w'}}{\sigma_u \, \sigma_w} \tag{7}$$

Covariances of the form $\overline{u_i' T'}$ are kinematic heat fluxes, and covariances of form $\overline{u_i' u_j'}$ are kinematic momentum fluxes (if $i \neq j$) or velocity variances (if $i = j$).

Now with a view to how best we may calculate statistics, let's look at different expressions for the variance of a signal... we may write

$$\begin{aligned}
\sigma_u^2 &= \overline{u'^2} = \overline{(u - U)^2} \\
&= \overline{u^2 + U^2 - 2\,u\,U} \\
&= \overline{u^2} + \overline{U^2} - 2\,\overline{u\,U} \\
&= \overline{u^2} + U^2 - 2U^2 \\
&= \overline{u^2} - U^2 \equiv \overline{u^2} - (\overline{u})^2
\end{aligned} \tag{8}$$

The result of this manipulation is that we can say "the variance equals the mean square minus the square of the mean", and this is the most convenient way to compute the variance, ie.

$$\sigma_u^2 = \frac{1}{N} \sum_n (u^n)^2 - \left(\frac{1}{N} \sum_n u^n \right)^2 \tag{9}$$

which implies we just need to sum up the individual members of the series (u^n) and also sum

up the squares (one's computer program to do this will use an "accumulator" for the u^n and an accumulator for the $(u^n)^2$).

Now in the above manipulation the first two equalities are definitional (the first follows from the definition of σ_u, ie. the meaning I define for the symbol σ_u; the second follows from my definition of the meaning of the prime, ie. definition of a fluctuation). Further on I used rules that may be obscure to you, viz. for any two variables f, g and any constant α

1. $\overline{f \pm g} = \overline{f} \pm \overline{g}$
2. $\overline{U} \equiv U$
3. $\overline{\alpha\, f} \equiv \alpha\, \overline{f}$

Note that from the third rule that $\overline{uU} = U^2$, because the mean, U, is effectively a constant.

The first of these rules follows from the fact that the averaging operation is either an integration (case of continuous variables) or a summation (discrete variables): in either case, it should be familiar to you that the operation is a linear operation, and the property I claim applies, eg.

$$\int [f(x') + g(x')]\ dx' \equiv \int f(x')\ dx' + \int g(x')\ dx' \tag{10}$$

The second rule (mean of the mean equals the mean) is exact, since

$$\overline{U} \equiv \frac{1}{N}\ (U + U + U + ...) = \frac{N\,U}{U} \tag{11}$$

The third rule can also easily be proven, by thinking about sums or integrals.

Note that the last line of eqn(8) can be rearranged to read

$$\overline{u^2} = (\overline{u})^2 + \overline{u'^2} \equiv (\overline{u})^2 + \sigma_u^2 \tag{12}$$

and you can easily prove that more generally for any variables f, g

$$\overline{f\,g} \equiv \overline{f}\,\overline{g} + \overline{f'g'} \tag{13}$$

or (restating it the other way around)

$$\overline{f'g'} \equiv \overline{f\,g} - \overline{f}\,\overline{g} \tag{14}$$

which are results that are useful for the efficient computation of cross statistics.

1.3 More on terminology, and the sample/population distinction

At this point I should note that I have been a little sloppy in my terminology. Strictly speaking, Greek letters are reserved for the "parameters" of a population, whereas in practise we only have a sample of a given population, and "statistics" are properties of a sample and (as such) only *estimators* of the corresponding parameters. Thus the standard deviation of a random variable x (properly denoted something like s_x) is only an estimator of the population property σ_x. In some contexts it is vital to make these distinctions of notation and interpretation. For example we can ask well defined questions like, what is the expected error of $|s_x - \sigma_x|$, or in words, how good is the estimate?

1.4 Shannon's Sampling Theorem and the Nyquist frequency

According to Shannon's Sampling theorem, if we sample a continuous signal $x(t)$ at discrete intervals Δt then we loose all information content at frequencies that exceed the Nyquist frequency

$$f_N = \frac{1}{2\,\Delta t} \tag{15}$$

Interpreting this in the context of the 20 Hz measurements with the sonic anemometers, the implied Nyquist frequency is 10 Hz, so (loosely speaking) our time series would not reveal fluctuations on frequencies exceeding about 10 Hz.