Undergraduates /35
Graduates /45

University of West Florida

GEO4990/5990 QUANTITATIVE METHODS

TIME SERIES ANALYSIS

A time series is a collection of observations made sequentially in time. Many types of time series occur in the physical sciences, particularly in meteorology, marine science and geophysics, including air temperature, rainfall or river discharge. In each case, the time series can be described in terms of three components:

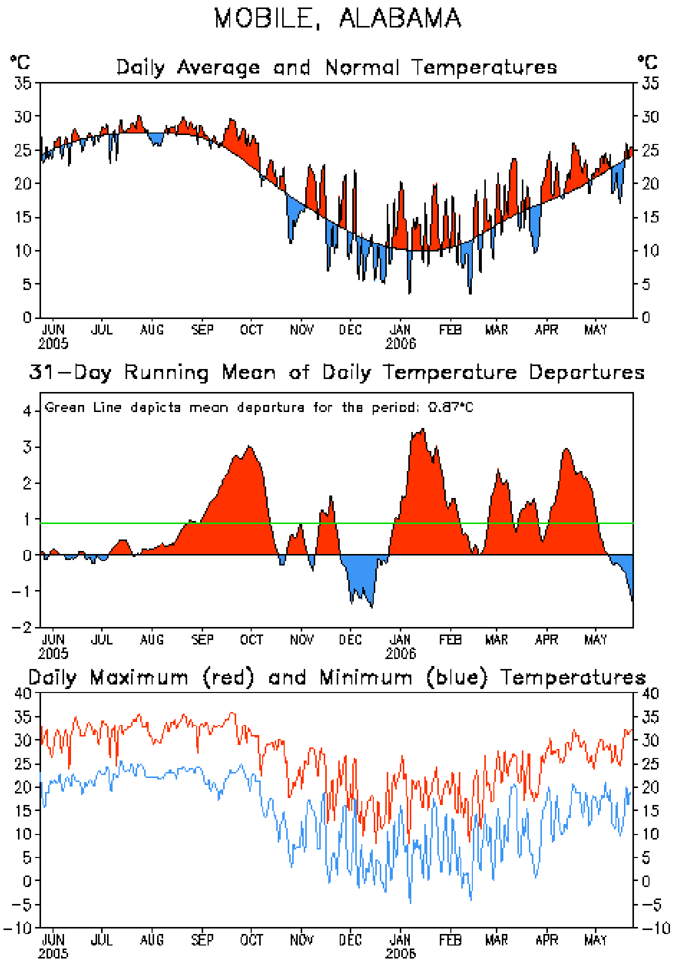
Time Series= Trend + Cycle + Residual (irregular variation)

Most time series exhibit a variation at a fixed period such as the seasonal variation in temperature. Beneath this cycle can be a long-term change in the mean (trend) that may be a true linear trend or a cycle in the data beyond the length of the time series. The shorter the time series the greater chance that the observed trends are due to low frequency (long) cycle. The residuals are components that are not associated with either the dominant cycles or trend. In this lecture we will be examining the residual, cycle and trend components of a time series.

RESIDUALS

In many climate studies, the deviation of a temperature from the climate normal (or historical average) is presented. In recent years, temperatures have been greater (positive residual) than the historical average, although the positive residuals are not present all of the time and everywhere.

The temperature variation for June 2005 to May 2006 is provided in the adjacent figure for Mobile, Alabama. The top image shows the historical daily average temperatures and the deviations (red or blue) from that average. The middle image shows the deviations from the average temperature (observed-average). The plot suggests that Mobile was warmer than the average (or expected) daily average temperature between June 2005 and May 2006.



Data updated through 23 MAY 2006

CLIMATE PREDICTION CENTER/NCEP

In the Excel Sheet labeled Time Series is the average monthly wind speed at Pensacola Airport from 1962 to 2005.

1. Plot the time series of monthly average wind speed (m s^{-1}) between 1962 and 2005. Make sure you fully label your graph and appropriately scale your x-axis. Being specific, describe the time series in terms of cycles, trend and residuals. /5

2. Compute the residuals (difference from the historical mean wind speeds) for each month in 1971, 1979, 1999 and 2005. Make a single plot of the residual wind speeds relative to month (1,2,3…..12) and describe the variation (ie. what years had slow wind speed and what years had fast wind speeds). /5

	Historical		Average			Residual			
	Average	1971	1979	1999	2005	1971	1979	1999	2005
January	8.31								
February	8.78								
March	8.97								
April	8.85								
May	8.06								
June	7.11								
July	6.35								
August	6.10								
September	6.99								
October	7.22								
November	7.79								
December	8.25								

CYCLES

Autocorrelation

Time series that exhibit cyclic variations are considered to be autocorrelated, because there is correlation between observations at different distances apart. The nature of the autocorrelation is examined by calculating a statistic similar to the Pearson correlation coefficient (Lecture 6) for the time series relative to itself with an increasing lag (k):

$$r = \frac{\sum (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum (x_t - \bar{x})^2}$$

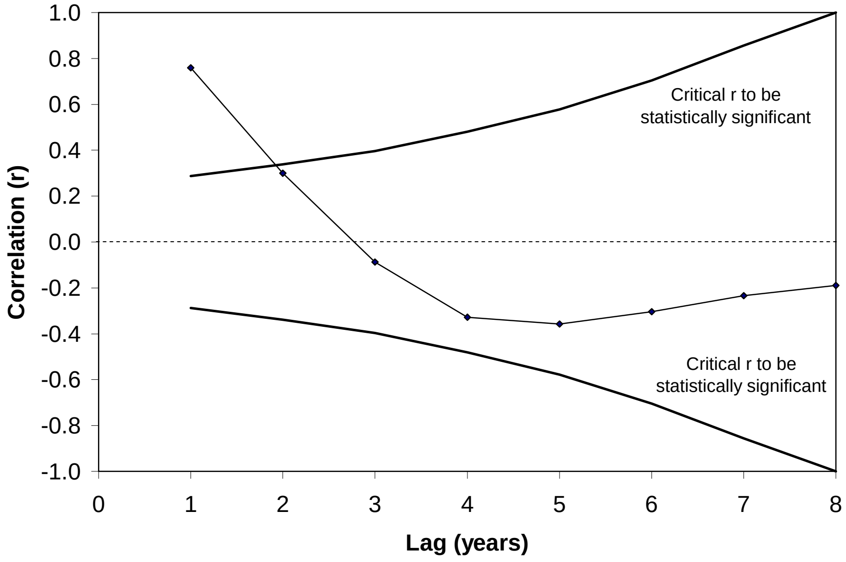
where x_t and x_{t+k} are the paired values. If there is a strong similarity (association) in deviations of the lagged data from the mean then there is a strong possibility of correlation. Consider the following time series (100,91,98….133) relating to sea-ice in the Hudson Strait. The correlation for a lag of 1 is 0.82 and the correlation for a lag of 2 is 0.53. As you will recall, the r value is not a measure of significance. Significance depends on the r relative to the number of samples and is calculated using the t statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

In this case the length of the time series (*n*) is the number of matched points: 10 for no lag, 9 for a lag of 1 year, 8 for a lag of 2 years, etc. As a consequence, the critical *t* value will get larger (and more difficult to meet) and the observed *t* value for the relationship will get smaller.

No Lag		**1 Year Lag**		**2 Year Lag**	
r=1.0		**r=0.76**		**r=0.53**	
100	100		100		100
91	91	100	91		91
98	98	91	98	100	98
119	119	98	119	91	119
133	133	119	133	98	133
140	140	133	140	119	140
144	144	140	144	133	144
127	127	144	127	140	127
133	133	127	133	144	133
133	133	133	133	127	133
		133		133	
				133	
		t	3.83	*t*	1.52
		df	9	*df*	8
		t_{crit}	2.25	t_{crit}	2.3
		Sign.?	Yes	Sign.?	No

The above table suggests that the time series is correlated at a lag of 1 year but that there are no statistically significant correlations over longer lags. The correlogram for this time series is provided in the following Figure showing the correlation at each lag and the associated critical value of r for the correlation to be significant.



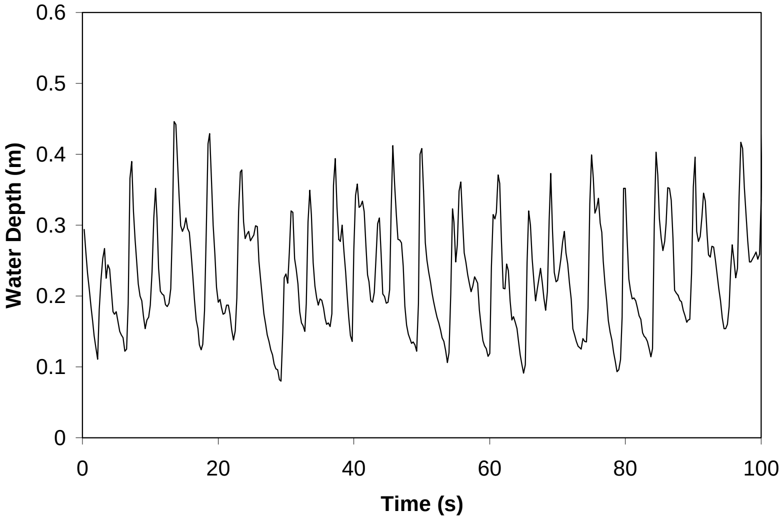
/10 3. Create a correlogram (to a lag of 5 years) for the data provided in the Excel sheet labeled Sea Ice Data. This data from 1990 to 1999 describes the length of the ice-free (summer) season in the Hudson Strait connecting the Atlantic Ocean with Hudson Bay.

/5 4. Create a table that summarizes the autocorrelation coefficients, t statistic and the critical t statistic at the 95% confidence level. Identify which lags exhibit a statistically significant correlation.

/5 5. It is believed (or hypothesized) that the length of the ice-free season in one year conditions the water temperatures (either warmer or colder) and affects the ice-free season in the following years. Based on your correlogram what would you tell the cryologist asking the question.

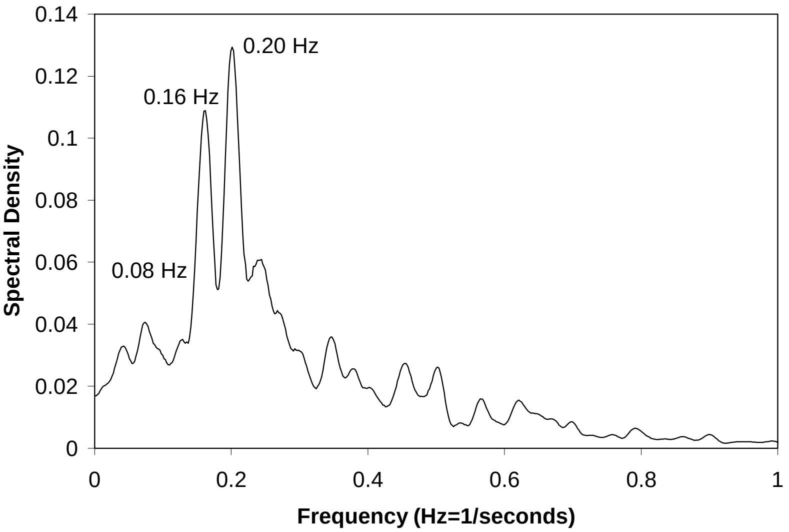
/10 **BONUS:** In the correlogram that I provided on you on the previous page I included the critical correlation (r) coefficients at each lag. Using the equation for the t-statistic calculate the critical r coefficients at each lag and plot on your correlogram from Question 3.

As a final note on correlograms, when you have data (such as daily average temperature) that has a strong seasonal cycle it is important to remove the seasonal component before calculating the autocorrelation. The correlogram will be dominated by the seasonal cycle and not provide useful information that you couldn't already get from the time series plot. For example, in the adjacent Figure the correlogram for raw monthly average temperatures is dominated by the seasonal cycle, while the correlogram for the time series with the seasonal cycle removed (not discussed in this class) provides more useful information.



Spectral Analysis

Many time series contain cycles at different frequencies, particularly wave data. The adjacent figure shows 100 seconds of wave data (as water

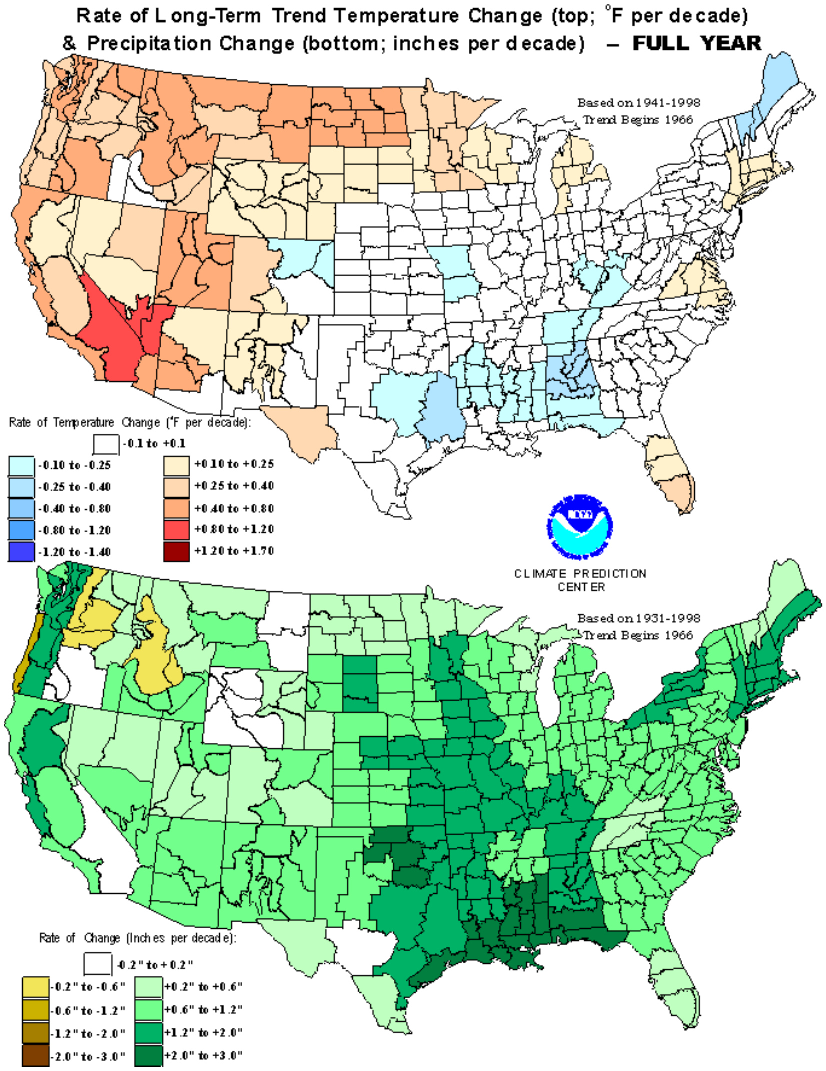


depth) from Pensacola Beach. Notice how there are different waves with different frequencies in the time series.

Spectral analysis is commonly used to identify the cycles in the data and their relative importance (energy). In the adjacent figure is the spectral density curve showing lots of energy at frequencies of 0.20 Hz, 0.16 Hz and even some at 0.08 Hz.

TREND

The debate over climate change is based on observed trends in time series of temperature and precipitation. The detection of a statistically significant trend is, however, not straightforward since the parametric regression test requires that the data be normal and independent. Also the test is very sensitive to outliers or extraordinary data, particularly at the start or beginning of the series.



The Mann-Kendall test provides a practical solution for trend analysis in data sets from environmental systems, since the test is non-parametric, does not require the dubious assumption of normality in the residuals, and is insensitive to outliers and low-order nonlinear trends. The test is applicable in cases when the data values *x* of a time series are assumed to be of a continuous monotonic increasing or decreasing function of time and residuals from the same distribution with zero mean. In other words, it is assumed that the variance of the distribution is constant in time (2nd order stationarity). A further assumption is that there is no serial correlation in the time series, such that a correlogram for the data shows no autocorrelation beyond a lag of zero.

This test considers whether the variable tends to increase or decrease with time, by computing a test statistic (*S*) calculated as the sum of the signs of the slopes for all possible combinations of two data points from the dataset:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}\left(x_j - x_k\right)$$

where x_j and x_k are the annual values in years *j* and *k* (*j*>*k*), respectively, and,

$$\operatorname{sgn}\left(x_j - x_k\right) = \begin{cases} 1 if x_j - x_k > 0 \\ 0 if x_j - x_k = 0 \\ -1 if x_j - x_k < 0 \end{cases}$$

If *n* is less than or equal to 9, the absolute value of *S* is compared directly to the theoretical distribution of *S* derived by Mann and Kendall (Gilbert, 1987), using four different significance levels α: 0.1, 0.05 and 0.01. If the absolute value of *S* equals or exceed a specified value of $S_{\alpha/2}$, the null hypothesis (*S=0*) is rejected in favor of the alternate hypothesis and a statistically significant trend is identified. A positive (negative) value of *S* indicates an increasing (decreasing) trend.

A test statistic (or *z* score) is used to check for the statistical significant of *S*:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\operatorname{var}(S)}} \\ 0 \\ \dfrac{S+1}{\sqrt{\operatorname{var}(S)}} \end{cases}$$

where *var(S)* is the variance of *S* and is computed as:

$$\operatorname{var}(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p (t_p - 1)(2t_p + 5) \right]$$

where *q* is the number of tied groups and t_p is the number of data values in the p^{th} group. For relatively large (*n*>10) samples a normally-distributed approximation Z_s can be constructed. However, for small data sets the absolute value of *S* is compared directly to the theoretical distribution of *S* derived by Mann and Kendall. At the 95% confidence level the significance threshold for the absolute value of *S* is 15 for 7 datapoints, 10 for 5 datapoints, and 7 for 4 datapoints. A trend will be found to be statistically significant when the magnitude of the change is large relative to the variation of the data around the trend line.

To account for seasonal cycles in the data series the **Seasonal Kendall test**, a variant of the Mann Kendall test, is used (Gilbert, 1987). In this test, the data are averaged by season (or month) and the trends for each season are examined independently. The test statistic is calculated as the sum of the test statistics from each season and the total variance is estimated as the sum of the seasonal estimates of variance. Statistical significance, which is obtained from a standard normal distribution for datasets larger than 10, is reported for the standardized test statistic.

The magnitude of the trend (rate of change over time) is computed according to Sen (1968). The trend slope is the median slope of all pairwise comparisons, with each pairwise difference is divided by the number of years separating the observations:

$$Q = \frac{x_j - x_k}{j - k}$$

If there are *n* values of x_j in the time series we get as many as *N=n(n-1)/2* slope estimates Q_i. The Sen's estimator of slope is the median of these *N* values of Q_i. The *N* values of Q_I are ranked from the smallest to the largest and the Sen's estimator is:

$$Q = Q_{\left(\frac{N+1}{2}\right)},$$

if *N* is odd, and:

$$Q = \frac{1}{2}\left(Q_{\left[\frac{N}{2}\right]} + Q_{\left[\frac{N+2}{2}\right]} \right)$$

if *N* is even. The trend is typically described as a percent of the mean water-quality concentration by dividing the slope by the mean and multiplying by 100.

While the test is relatively straightforward, you probably can tell that it would take some work to get it to work in Excel. Luckily someone has already developed an interactive Excel program to run the Mann-Kendall Test and the Sen Slope Estimator.

In the Excel file titled Sen Slope Estimator, I have entered the average monthly wind speed from Pensacola Airport (1962 to 2005) that we examined in Question 1.

/5

6. Run the Mann-Kendall test and fill out the following table using the results for each month. In which months has there been a statistically significant (at the 95% confidence level) increase in the average wind speed.

/10

7. **GRADUATE STUDENTS:** Copy the data to a separate Excel file and complete a regression analysis on the months you identified as having a statistically significant trend. How do the slopes for the trend compare? Is this unexpected?

	Test Z	Critical Z	Significant @ 95%	Slope Q	Constant B	Linear Slope Equation
January						
February						
March						
April						
May						
June						
July						
August						
September						
October						
November						
December						