Analyzing Change Over Time

School of Information Studies
Syracuse University

Learning Topics for This Week

Cross sectional data vs. time sequenced data

Nonindependence between observations

Comparing observations at two points in time; individual differences; dependent samples t-test/one-sample t-test

Repeated measures ANOVA; balanced design;

Time series analysis; trend, seasonality, cyclicality; lags and auto-correlation

Learning Topics for This Week

The danger in time series: using differencing to eliminate trend

Diagnosing and testing stationarity

Change-point analysis

A word about ARIMA

Learning goal: recognize the difference between cross-sectional and non-cross-sectional data; recognize data suitable for repeated measures analysis vs. time series analysis; conduct basic repeated measures analysis; diagnose stationarity and remove trend; conduct a change-point analysis

Cross Sectional Data vs. Time Sequenced Data

Cross Sectional Data: No Implication of Time

In all of the previous chapters of the book, the data we used had no implication of time; for all we know, they could have been collected at the same moment

In the data set to the right, five people were measured with a test in the same class session yielding five grades

But what if we could learn something about the change in grades over time?

There are two essential ways. . .

Obs.	Person	Exam
1	Art	8
2	Bill	7
3	Cody	9
4	Deirdre	9
5	Ellen	7

Repeated Measures: Change Over One Interval (or Just a Few Intervals)

When you have two or more sets of measurements from the same "source" at different points in time, we call this a "repeated measures" design

With the data on the right, we now have two snapshots of each student at two different points in time

One benefit this affords is that we can now see what improvement (or decline) each student has achieved using the difference score

Obs.	Person	Exam	Exam	Difference
1	Art	8	9	1
2	Bill	7	8	1
3	Cody	9	9	0
4	Deirdre	9	7	-2
5	Ellen	7	9	2

Time Series: Many Observations From One Source

Art:	3	4	4	7	8	9	8	8	9

Here's another approach: nine consecutive weekly observations (quizzes?) from just one student. Using a series of measurements on the same variable, there are additional questions we can ask:

- This student seems to improve over nine weeks, is this a real trend?
- If we "decompose" these data to extract the trend, would there be any other time-based patterns (e.g., a cycle) in the data? Perhaps Art does better in even-numbered weeks than in odd-numbered weeks.
- Is there a certain point in time, a "change point," where Art's grades change?

Nonindependence

The Key Data Difference Is: Nonindependence

When two or more measurements of the same variable are obtained from the source, these data points are dependent upon one another because they share a source of variance

Dependency between observations breaks one of the key assumptions of the analytical techniques we have used so far:

- Independent-samples t-test
- Oneway ANOVA
- Linear Multiple Regression
- Logistic Regression

All of these assume that each measurement is independent from the others; a litmus test: Can you shuffle the order of rows or columns without changing the meaning of the data?

The Shuffle Test

Obs.	Person	Exam
1	Art	8
3	Cody	9
4	Deirdre	9
2	Bill	7
5	Ellen	7

Row order changed:
No analytical difference

Obs.	Person	Exam	Exam	Difference
1	Art	8	9	1
2	Bill	7	8	1
3	Cody	9	9	0
4	Deirdre	9	7	-2
5	Ellen	7	9	2

First and second exams
swapped: time is reversed

Art:	9	3	4	7	8	9	8	8	4

First and last observations
swapped: time is jumbled

Examining Two Intervals

Use Dependent-Samples T-Test for Two Intervals

Measuring at two points in time allows us to divide variance into an individual difference component and a change component

Who is the best student? How much improvement does that individual show over time?

Who is the worst student? How much improvement does that individual show over time?

We can use the dependent measures (or one-sample) t-test to explore this

Obs.	Person	Exam	Exam	Difference
1	Art	8	9	1
2	Bill	7	8	1
3	Cody	9	9	0
4	Deirdre	9	7	-2
5	Ellen	7	9	2

```
t.test(Quiz, Final, paired = TRUE) # Dep samples
t.test(Difference) # One sample t-test
BESTmcmc(Difference) # Bayesian one sample test
```

Dependent Samples Results

Paired t-test

data: Exam1 and Exam2

t = -1.633, df = 4, p-value = 0.1778

Alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.1601748 0.5601748

Sample estimates:

Mean of the differences: -0.8

Data: difference

t = 1.633, df = 4, p-value = 0.1778

Alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

-0.5601748 2.1601748

Sample estimates:

Mean of x: 0.8

```
MCMC fit results for best analysis: 100002 simulations saved
          mean      sd        median    HDIlo     HDIup     Rhat      n.eff
mu        0.8432    0.9338    0.8617    -0.9297   2.503     1.009     13779
```

Repeated Measures ANOVA With Balanced Design

Analyzing Chick Weights Over Time

The textbook describes the built-in R dataset "ChickWeight" with weight in grams of about 47 different birds measured 12 times after hatching

We can use the dependent samples t-test to compare any pair of groups; a comparison between time 16 and time 18 yields:

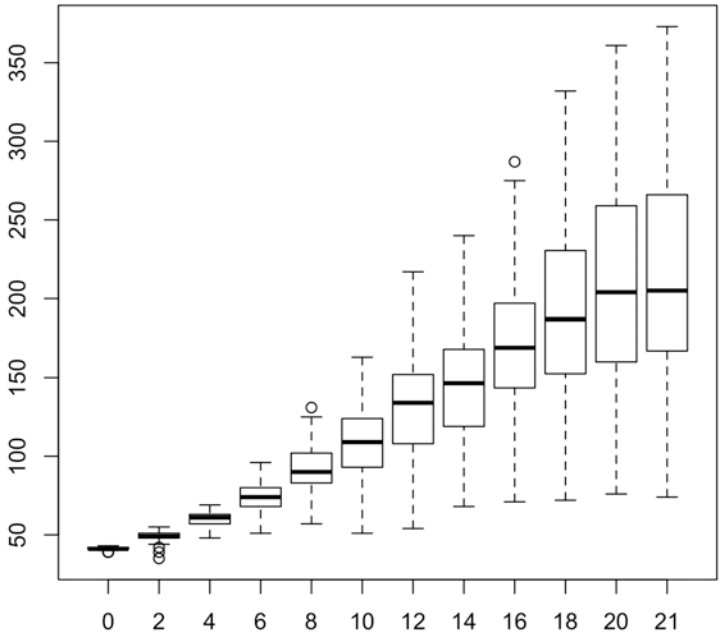
Paired t-test / data: time18weight and time16weight

t = 10.136, df = 46, p-value = 2.646e-13

Alternative hypothesis:
True difference in means is not equal to 0

95 percent confidence interval: 17.71618 26.49658

Mean of the differences: 22.10638



Balanced "Design": Same Number of Observations for Each Time Period

The aov() command for ANOVA requires a balanced design: the same number of subjects for each time period

```
# Using a copy of ChickWeights, keep only the birds that have a full list of 12 observations
list <- rowSums(table(chwBal$Chick, chwBal$TimeFact))==12  # Make a list of rows with 12
list <- list[list==TRUE]                                   # Keep only those with 12 observations
list <- as.numeric(names(list))                            # Extract the chick numbers
chwBal <- chwBal[chwBal$Chick %in% list, ]                 # Match against the data
table(chwBal$Time)                                         # Show a table of number of observations by time
```

Now there are 45 chicks at each time period:

```
 0  2  4  6  8 10 12 14 16 18 20 21
45 45 45 45 45 45 45 45 45 45 45 45
```

The ezANOVA() command warns about unbalanced data

Chick Weights Over Time With Repeated Measures ANOVA

To assess whether differences exist in weights across all time groups we need repeated measures ANOVA:

Error: Chick

	Df	Sum Sq	Mean Sq
Residuals	44	429899	9770

Individual differences

Error: within

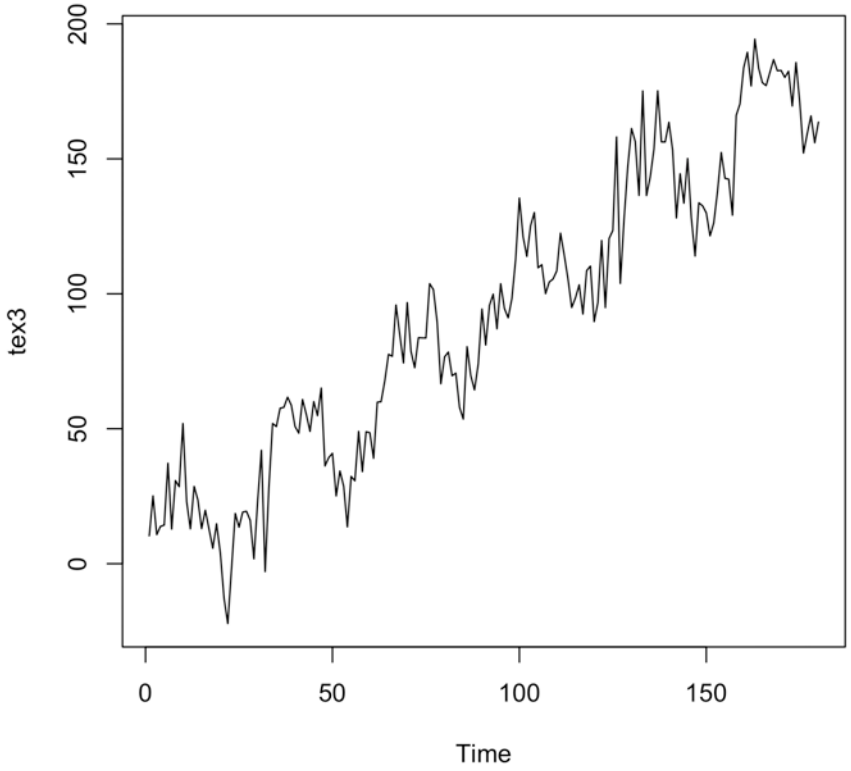
TimeFact	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TimeFact	11	1982388	180217	231.6	<2e-16 ***
Residuals	484	376698	778		

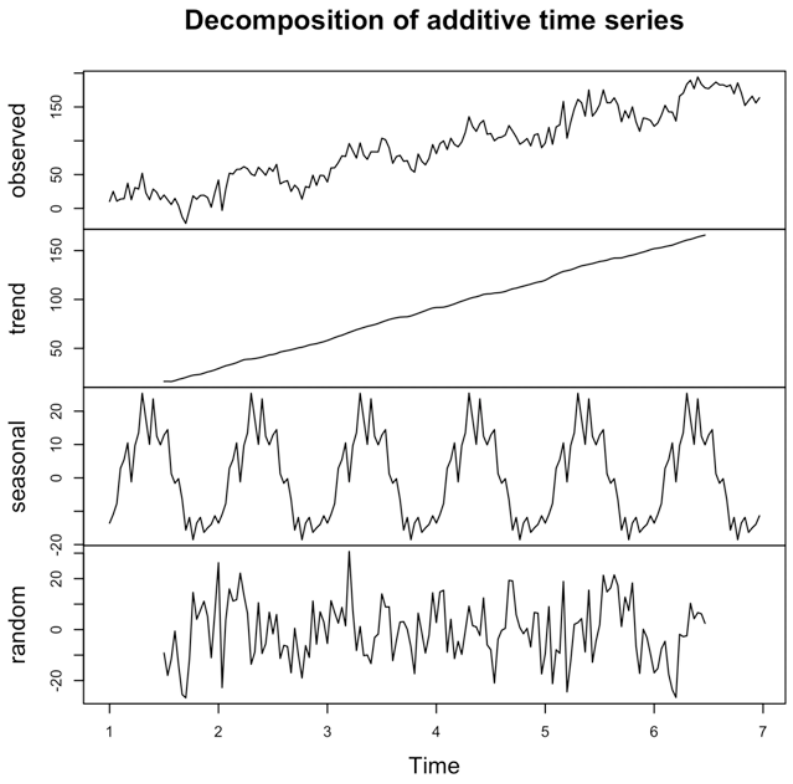
Effect

Random error

Time Series Analysis

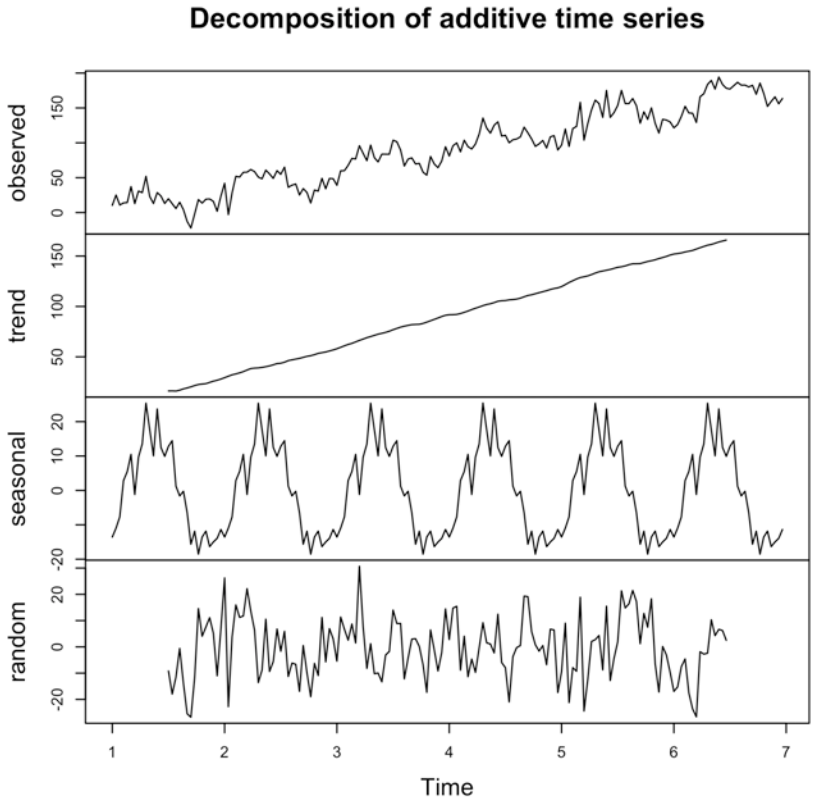
Time Series Analysis





Four Distinctive Components

1. **Trend** is one of the most noticeable and obvious aspects of a time series: growth or decline across time
2. **Seasonality**: regular fluctuations that occur over and over again across a period of time
3. **Cyclicality**: the idea that there may be repeating fluctuations that do not have a regular time period to them
4. **Irregular component**: what's left after trend, seasonality, and cyclicality are removed; some statisticians refer to this component as **noise, but this may be the most important component**



The Danger in Time Series and Using Differencing to Remove Trend

The Danger in Time Series

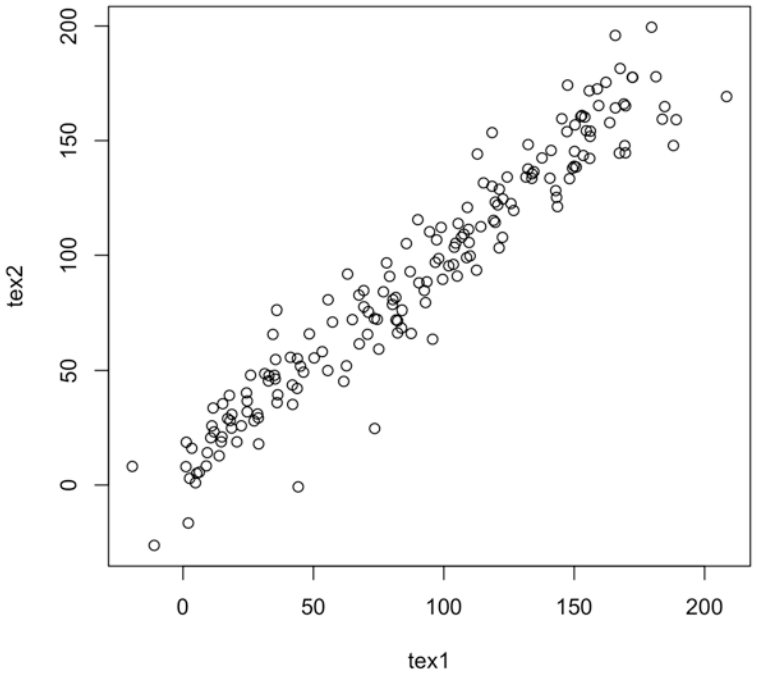
These two time series variables (tex1, tex2) look highly correlated

In fact, r=.96, indicating a very high correlation

But they both contain a strong positive/growth trend

That trend may result from a common, underlying growth pattern that obscures the true connection (or lack thereof) between the variables

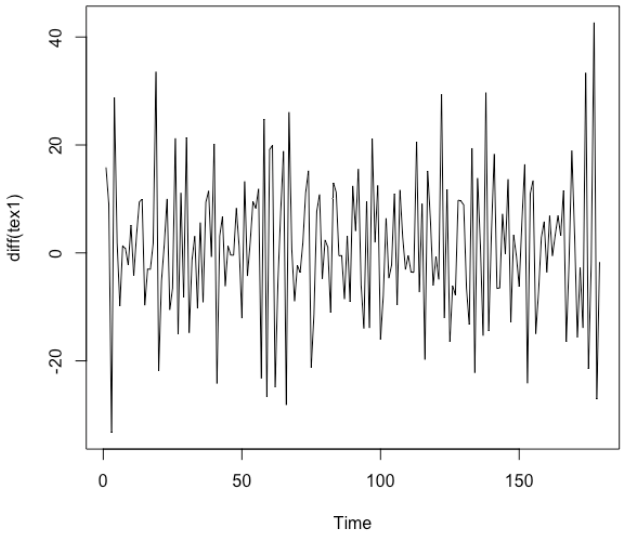
If we remove the trend from each series, we can reveal the "stationary" version of the data, which might be more useful

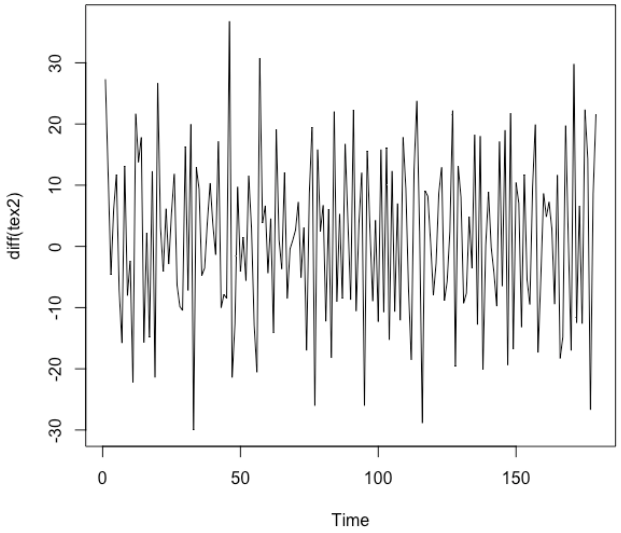


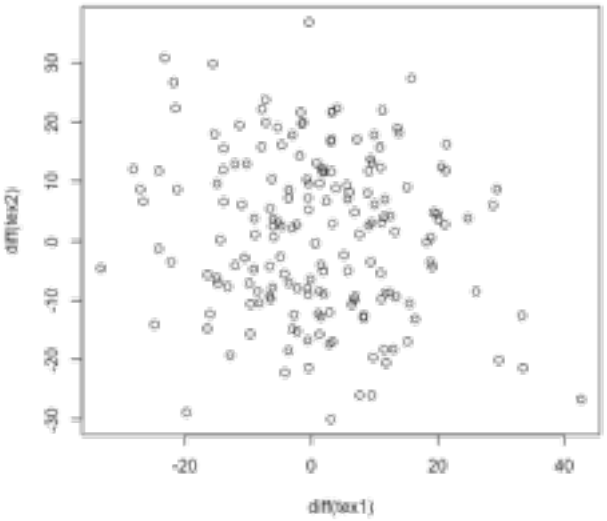
Removing Trend With Differencing

Consider the sequence: 1 2 3 4 5 6 7 8 9 10. If we take the difference between each pair of neighboring points we get: 1 1 1 1 1 1 1 1 1.

The plots below show what the data on the previous slide look like after differencing:







Diagnosing and Testing Stationarity

Measuring Stationarity With Lags

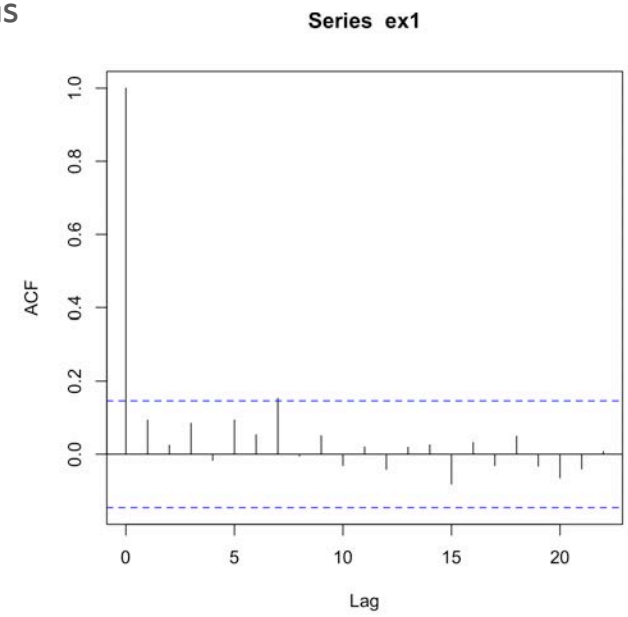
Consider the sequence: 2 1 4 1 5 9 2

The table below shows the data matched with itself at one lag and two lags

The correlation of a vector with itself at different lags provides valuable diagnostics

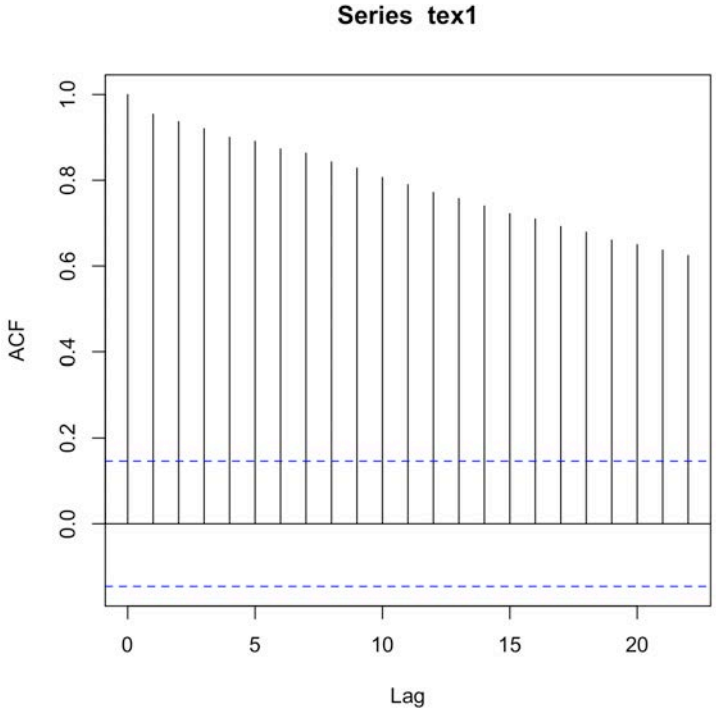
A time series is stationary when there are no lagged correlations

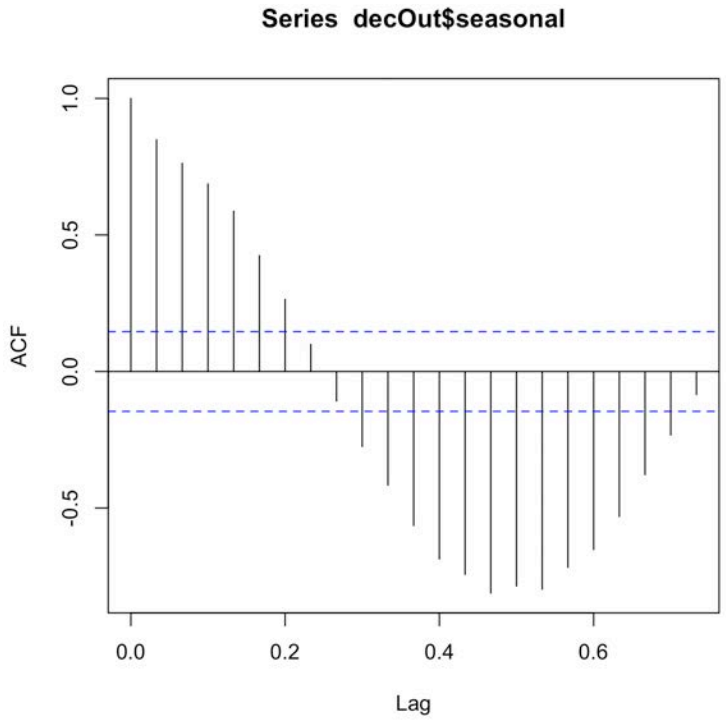
Observation Number	MyVar	lag(MyVar, k=1)	lag(MyVar, k=2)
1	2		
2	1	2	
3	4	1	2
4	1	4	1
5	5	1	4
6	9	5	1
7	2	9	2



Auto-Correlation Functions

On the left, an Auto-Correlation Function (ACF) showing a strong trend. On the right is an ACF showing a seasonal effect.





Dickey Fuller Test for Stationarity

adf.test(decComplete) # shows significant, so it is stationary

The adf.test() procedure in the final line of code above yields the following output:

Augmented Dickey-Fuller Test

data: decComplete

Dickey-Fuller = -5.1302, lag order = 5, p-value = 0.01

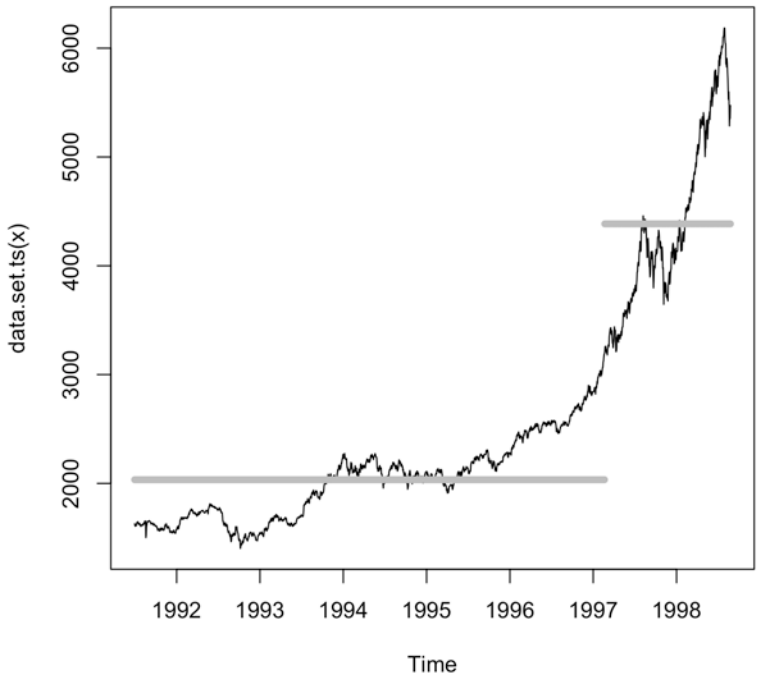
Alternative hypothesis: stationary

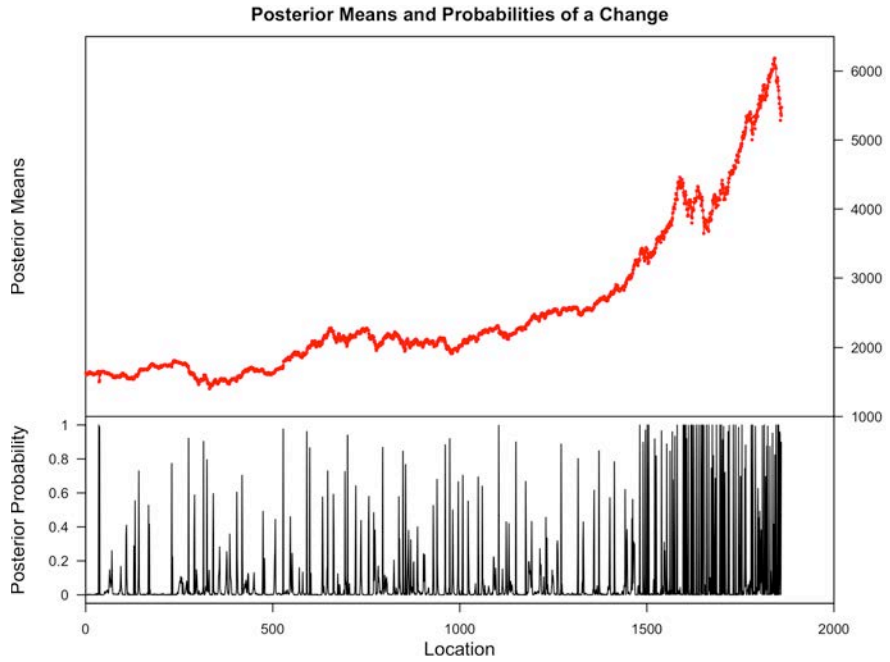
A significant Dickey-Fuller test means that a time series is stationary. Use the test after differencing or decomposition to confirm that trend, seasonality, and other time artifacts have been removed.

Change Point Analysis

When an Event or Intervention Occurs

There may be a meaningful shift in a time series, either a change in level or a change in volatility (variability). Change point analysis provides an analytical view of such changes:





Change in Variance

```
dEUstocks <- diff(EuStockMarkets)
dDAX <- dEUstocks[,1]
dDAXcp <- cpt.var(dDAX)
plot(dDAXcp,cpt.col="grey",cpt.width=5)
dDAXcp
```

Changepoint type: change in variance

Method of analysis: AMOC

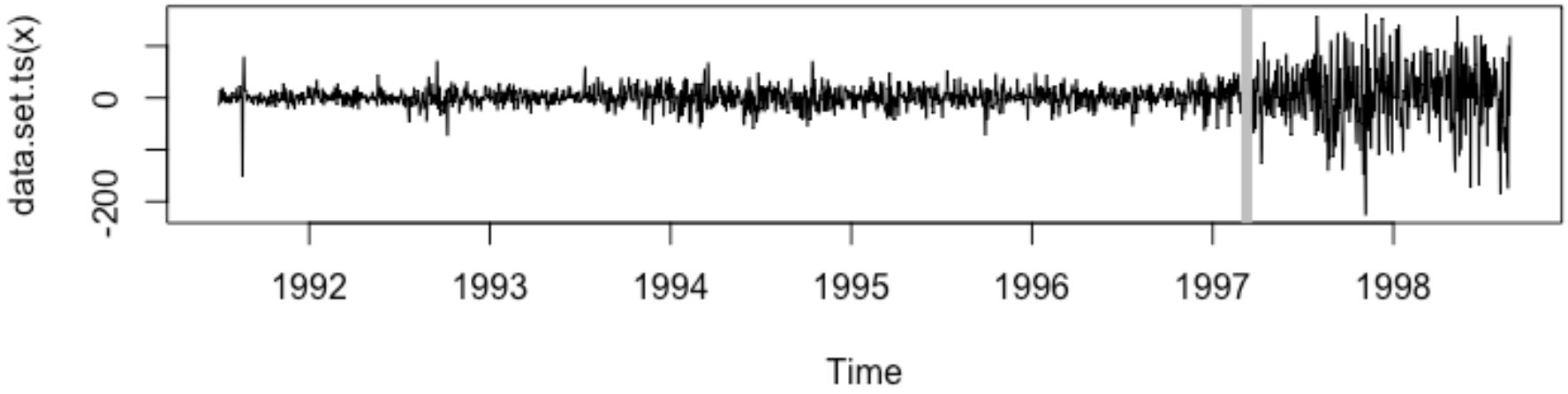
Test statistic: normal

Type of penalty: MBIC with value 22.58338

Minimum segment length: 2

Maximum no. of cpts: 1

Changepoint locations: 1480



Integrating the Results

We conducted an analysis of the European stock market index known as DAX, with daily time series data from 1991 to 1997. A strong positive trend was evident in the data. We conducted a change point analysis of means, using the “At Most One Change” (AMOC) search algorithm. A substantial change in the mean level of the index was detected in the early months of 1997. Likewise, after differencing to remove the trend, we conducted a change point analysis of variance with the AMOC algorithm and found a credible change in volatility at the same point in 1997. These changes in mean and variance appeared to coincide with several shocks to world markets.

A Word About ARIMA

A Word About ARIMA

```
tsFit <- arima(LakeHuron[1:90], order=c(1,0,1))   # Fit the model
lhPred <- predict(tsFit,n.ahead=8)                # Show the next few predicted values
LakeHuron[91:98]                                  # Compare with the actual values
plot(LakeHuron[91:98], lhPred$pred)               # Plot the actual vs. predicted
abline(a=0,b=1)                                   # Orient to unit line
```