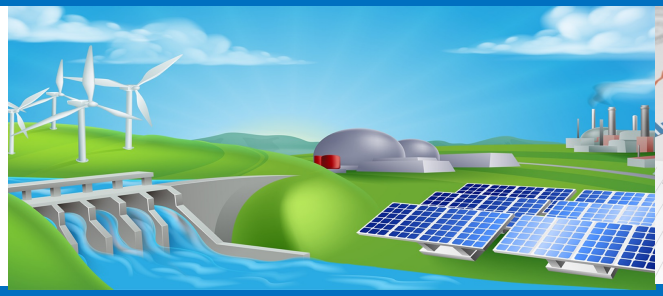



ENV797 - TIME SERIES ANALYSIS FOR ENERGY AND ENVIRONMENTAL APPLICATIONS

M4 - Missing Data and Outlier Detection

Prof. Luana Medeiros Marangon Lima, Ph.D.

Nicholas School of the Environment - Duke University

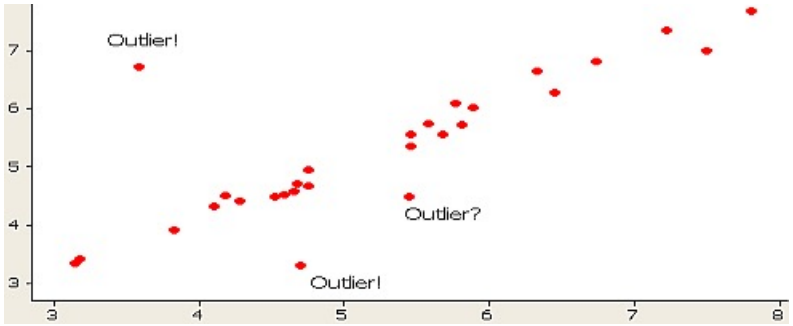
Learning Goals

- Missing Data
  - How to replace missing data
- Outlier Detection
  - Types of Outliers
  - Causes of Outliers
  - Outliers Detection and “Removal”
- Work on a R script for outlier/NAs detection and removal

Missing Data

Missing Data and Outliers

- Detecting missing data and outliers are considered the filtering step
- Missing data and outliers can occur for a variety of reasons, for example the breakdown of automatic counters



Missing Data

- Missing data can result in misinterpretation of the resulting statistics
- The following methods may be applied to replace the missing values in observed data set

Replaced by Local Mean

- Replace missing data in a variable by the average of the two adjacent observations

Replaced by Series mean

- Replaces missing values with the mean for the entire series

Missing Data in R

- How does R handle missing values?
- Missing data in R appears as NA
- NA is not a string or a numeric value, but an indicator of missingness
- Some useful functions associated with NAs from the "stats" package
  - *is.na()* is a tool for both finding and creating missing values
  - Other functions for NA are options for *na.action()*

Missing Data in R

- The possible *na.action()* settings within R include
  - **na.omit** and **na.exclude**: returns the object with observations removed if they contain any missing values
  - **na.pass**: returns the object unchanged
  - **na.fail**: returns the object only if it contains no missing values
- In some R functions, one of the arguments the user can provide is the *na.action()*.
- For example, if you look at the help for the **lm()** command, you can see that **na.action()** is one of the listed arguments – where user specify what to do in case of missing observations

Outliers

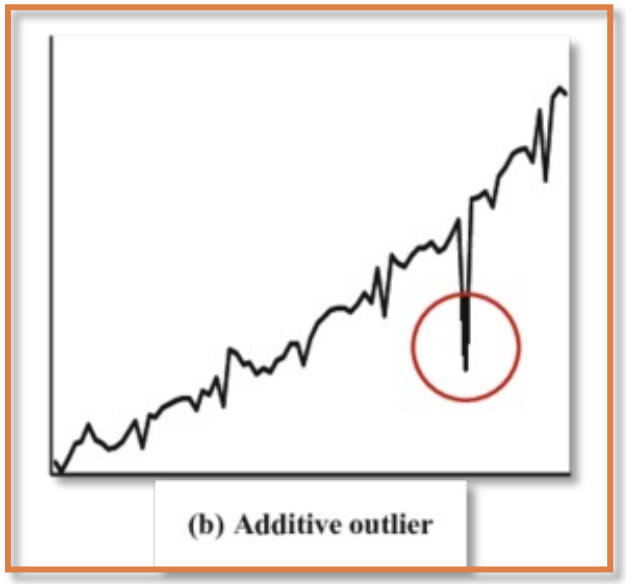
Outliers

- *"Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism"—Hawkins(1980)*
- It's is a simple assumption that the point is "wrong"
- Most common types of outliers

Additive Outlier (AO)	Level Shift (LS)	Temporary change (TC)

Types of Outliers: Additive Outlier (AO)

- An **additive outlier** appears as a **surprisingly large or small** value
- Affects a **single observation,** subsequent observations are unaffected by an additive outlier

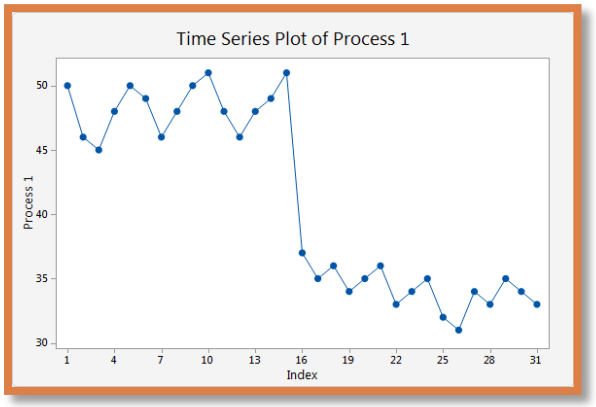


- It is the type of outliers that occurs as a result of a **mistake made by person in observation** or record
- Or may be caused by a random effect, a **strike or a short-term shock** in the system

Types of Outliers: Level Shift (LS)

- **Level shift** refers to a more **permanent change** in the time series level starting at a given time period
- All observations after the outlier move to a **new level**
- It may occur due to **changes in concepts and definitions** or compilation methods of the survey population
- Or as a result of **changes in economic behavior, in social traditions or in legislation**

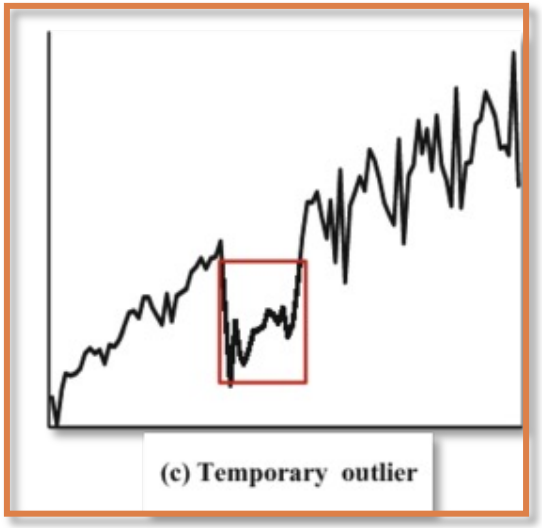
LS change the level of the time series, but do not modify seasonal behavior. Therefore, these outliers must be identified to avoid distorting the seasonal component.



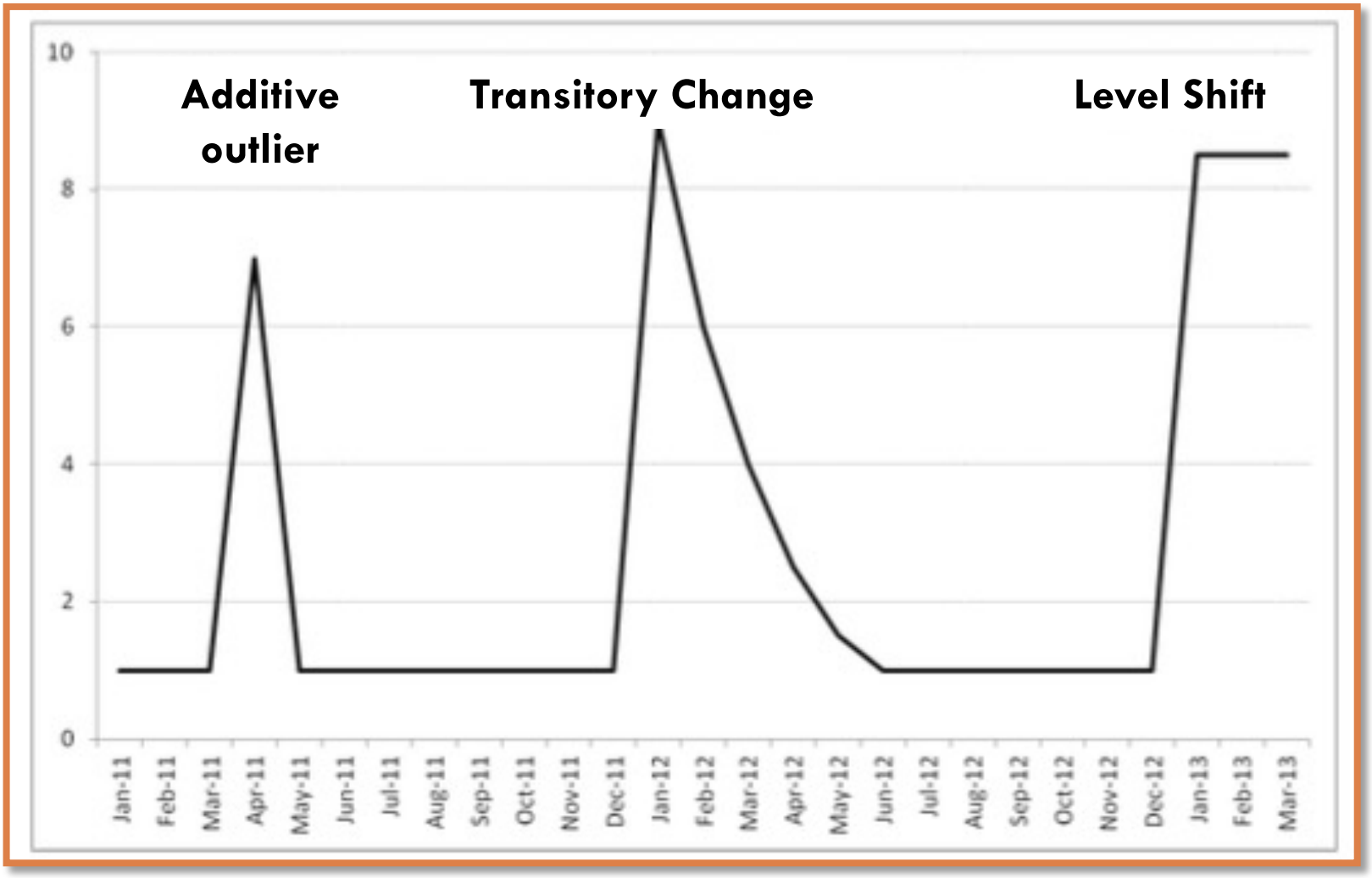
Types of Outliers: Transient Change (TC)

- **Transient change** outliers are similar to LS outliers, but the effect of the outlier **diminishes** over the subsequent observations
- Eventually, the series **returns to its normal level**
- TC may occur due to **deviations from average** monthly weather conditions

Ex: If the weather changes drastically, energy consumption may rise or fall. When the weather goes back to normal, energy consumption also returns to its usual level.

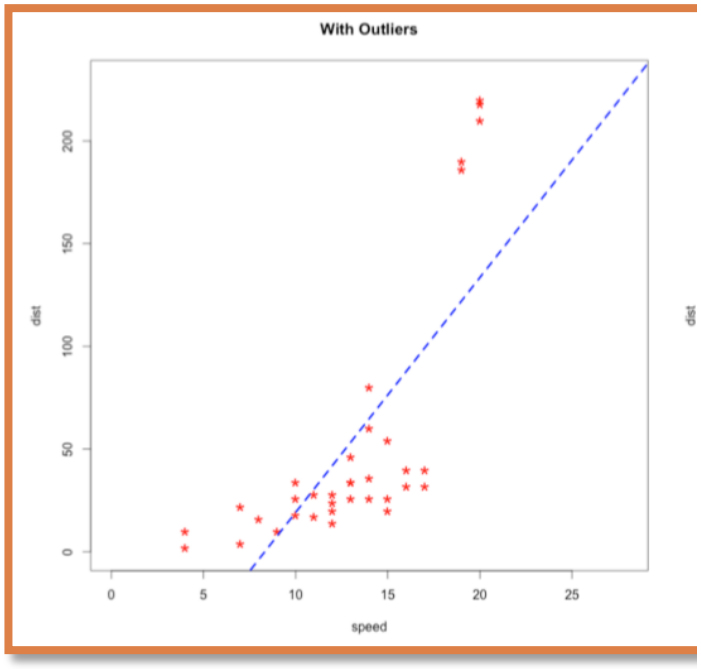


Frequent Types of Outliers



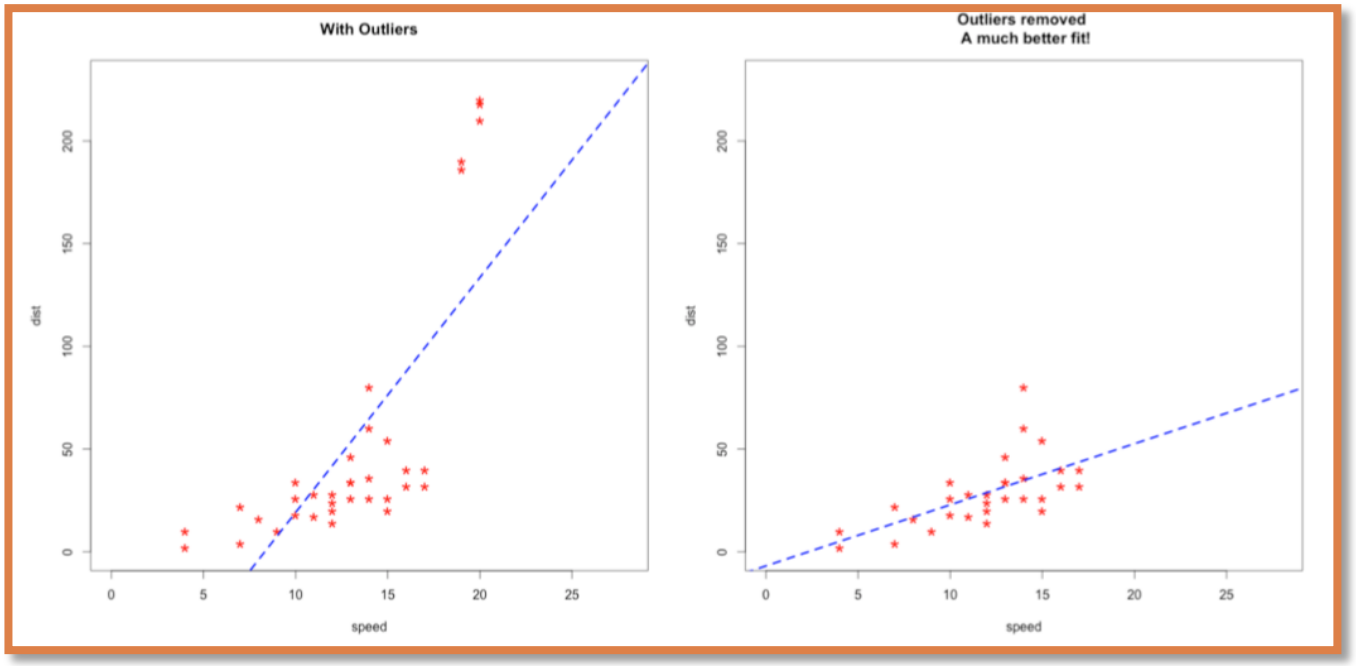
Modeling with Outliers

- Modeling and forecasting time series data in the presence of outliers is a difficult problem
- The presence of outliers can affect the **model identification** and **estimation**



Modeling with Outliers

- Modeling and forecasting time series data in the presence of outliers is a difficult problem
- The presence of outliers can affect the **model identification** and **estimation**



Modeling with Outliers (cont'd)

- Their presence close to the end of the observation period can have a serious impact on the forecasting performance of the model
- But keep in mind that…

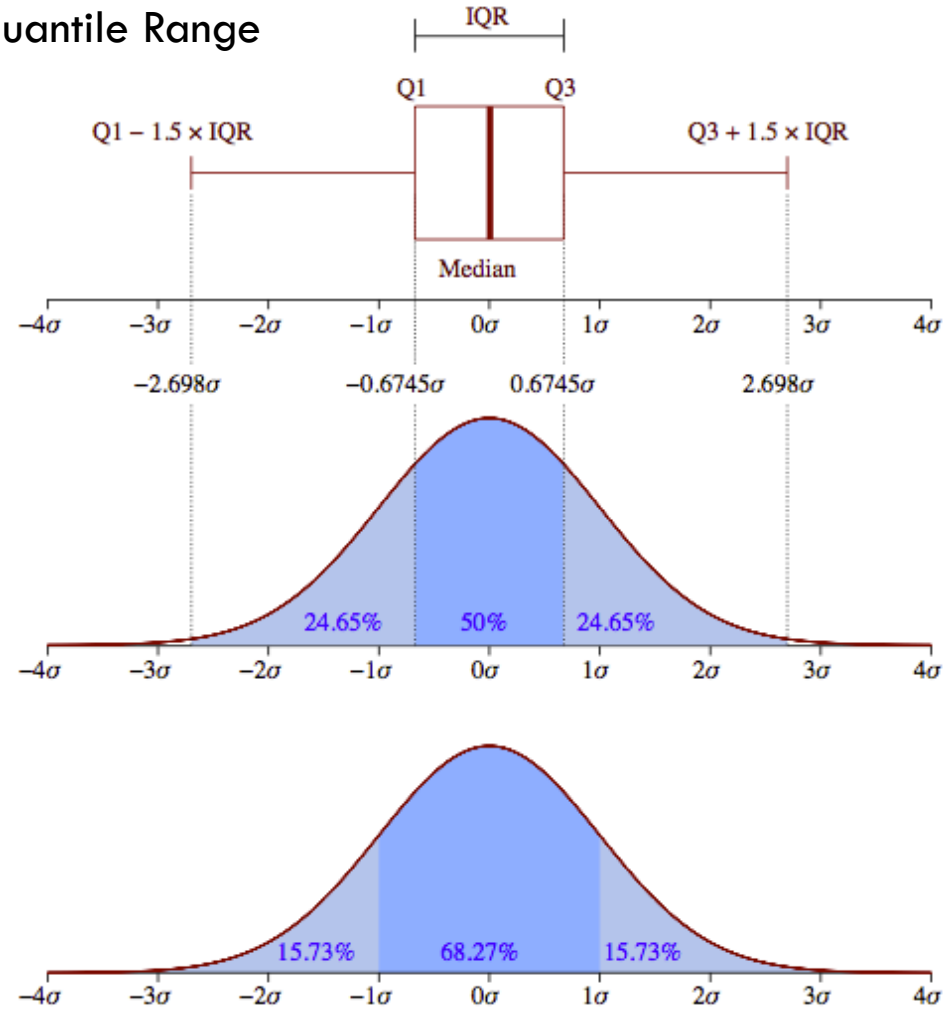
Outliers may contain information about unusual events, so they are an important part of the data!

Methods for detecting outliers

- Outlier tests are an iterative process.
  1. Check most extreme value for being an outlier
  2. If it is, remove it
  3. Check for the next extreme value using the new, smaller sample
  4. Repeat the process
- Once all outliers are removed the sample can be analyzed

Methods for Detecting Outliers

IQR - Interquantile Range

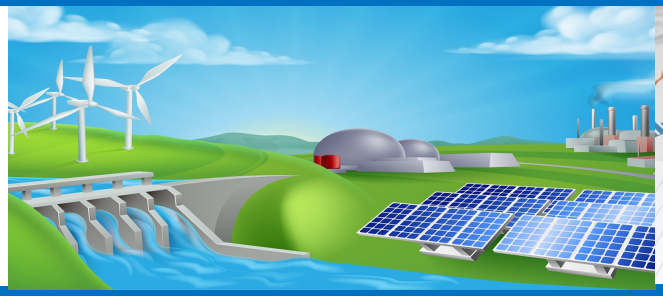


Box plots in R will identify outliers as black dots outside the interval $[-3\sigma, +3\sigma]$

Outliers with R

- Common functions for detecting and removing/replacing outliers from package "outliers"

outlier(x, opposite = FALSE, logical = FALSE)	Find value with largest difference from the mean
rm.outlier(x, fill = FALSE, median = FALSE, opposite = FALSE)	Remove the value(s) most differing from the mean
dixon.test(x, type = 0, opposite = FALSE, two.sided = TRUE)	Dixon tests for outlier
chisq.out.test(x, variance=var(x), opposite = FALSE)	Chi-squared test for outlier – assume known variance
grubbs.test(x, type = 10, opposite = FALSE, two.sided = FALSE)	Grubbs tests for outliers




THANK YOU !

luana.marangon.lima@duke.edu