SOCIETY OF ACTUARIES®

SOA Predictive Analytics Seminar – Malaysia

26 Aug. 2019 | Kuala Lumper, Malaysia

Session 6

Predictive Analytics in a Chaotic Data World

Wai Sum Chan, FSA, CERA, HonFIA, FRSS

SOA Predictive Analytics Seminar, Kuala Lumpur 26 August 2019, Session: 15:55-16:45

Predictive Analytics in a Chaotic Data World

Wai Sum Chan, PhD, FSA, HonFIA, CERA
Professor of Finance
The Chinese University of Hong Kong

Introduction

- Often actuarial practitioners are faced with working with data that is less than ideal.
- The data may be observed with gaps in it, a model may suggest variables that are observed at different frequencies, and sometimes predictive analytic results are very fragile to the inclusion or omission of just a few observations in the sample.
- Data, particularly big data, are often messy and something must be done about it.
- What is the actuary to do about these very practical matters?

What is the meaning of messy data?

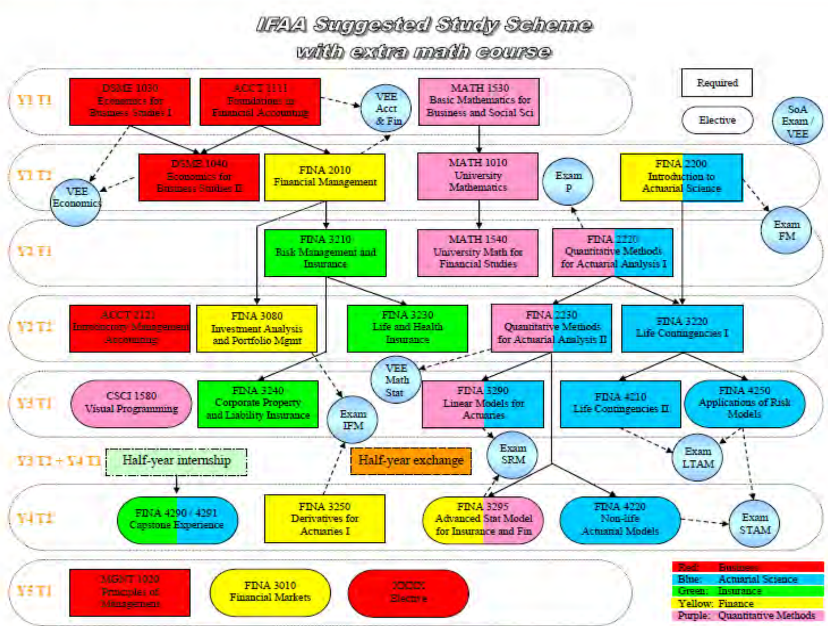
- Data sets large and small are rarely ready to use.
- There are many problems that associated with messy data:
  - missing values
  - outliers
  - structural changes
  - abridged and censoring data
  - lack of data and messy data
  - ... and many more
- We should perform cleansing and validating data before any predictive modeling
- garbage in, garbage out (GIGO)

My lovely data generator

SOA exam curriculum in 1990s

100 Calculus and Linear Algebra
110 Probability and Statistics
120 Applied Statistical Methods
130 Operations Research
135 Numerical Methods
150 Actuarial Mathematics
151 Risk Theory
160 Survival Models
162 Construction of Actuarial Table
165 Mathematics of Graduation

SOA exam curriculum in 2020s



(A) Missing Data

Missing Data: A climate change data case study

Missing Data: A climate change data case study

- The National Data Buoy Center (NDBC) is a part of the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) of the US government.

- NDBC deploys weather buoys which are instruments which collect weather and ocean data within the world's oceans.

Missing Data: A climate change data case study

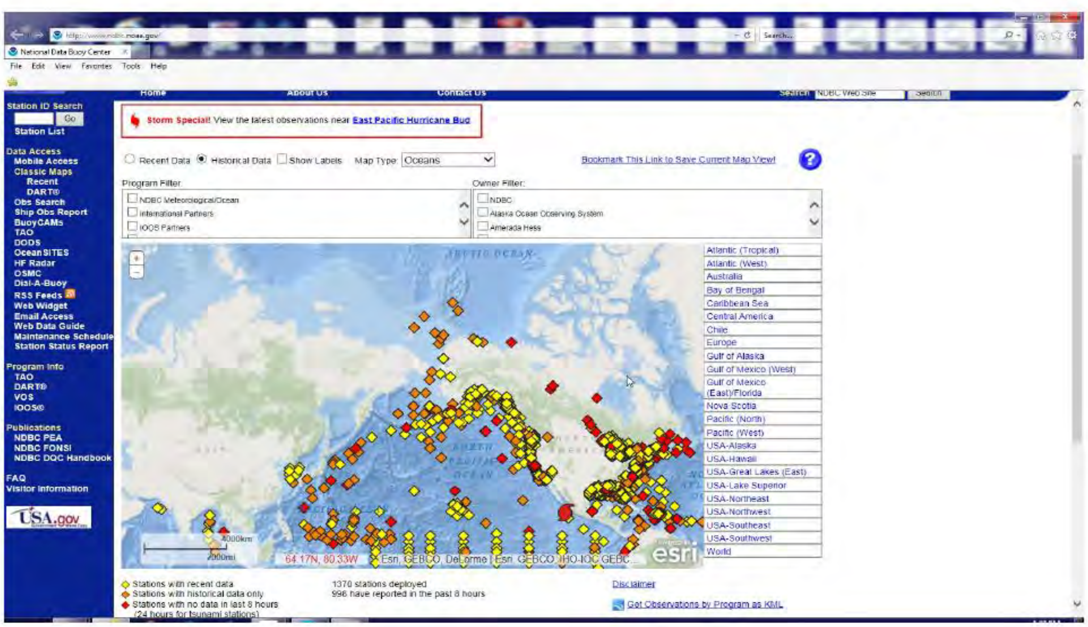
- The time-series weather data for each buoy are publicly available from the NDBC website (www.ndbc.noaa.gov).

- These data have been used for research and teaching purposes. I used this data set in my class "FINA3295 Predictive Analytics for Actuarial Science".

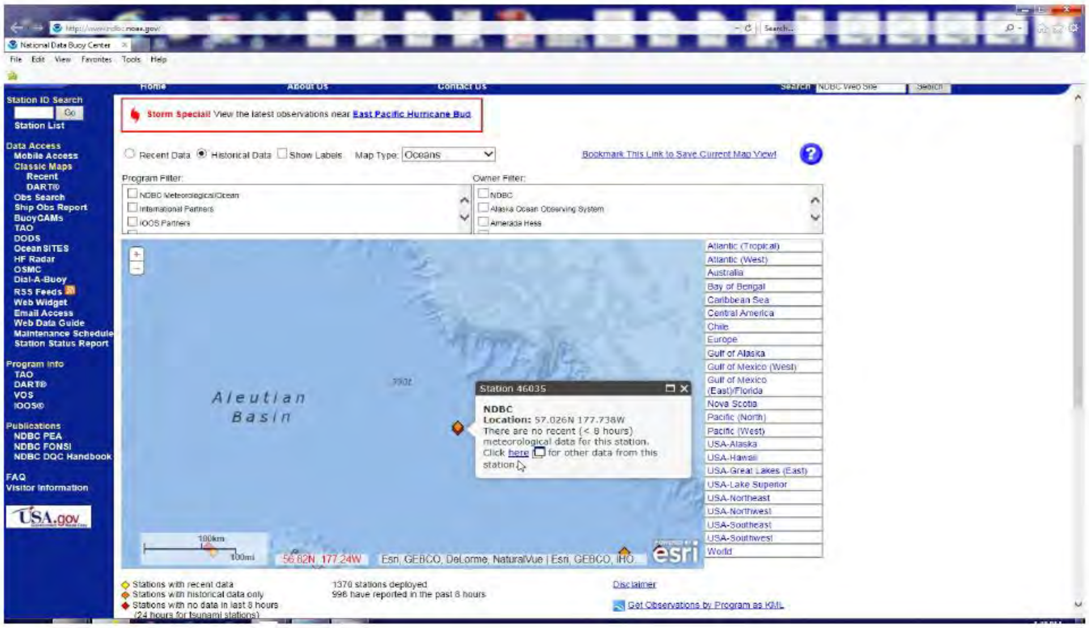
Part (A) - constructing the dataset

- Students are asked to locate the data webpage of the Weather Station buoy 46035 at 57.026 N 177.738 W from NDBC.
- Examine the data format for each yearly data file.
- Write an R program to extract and patch the data into two time-series of daily **Air Temperature** and **Sea Temperature** readings recorded at noon.

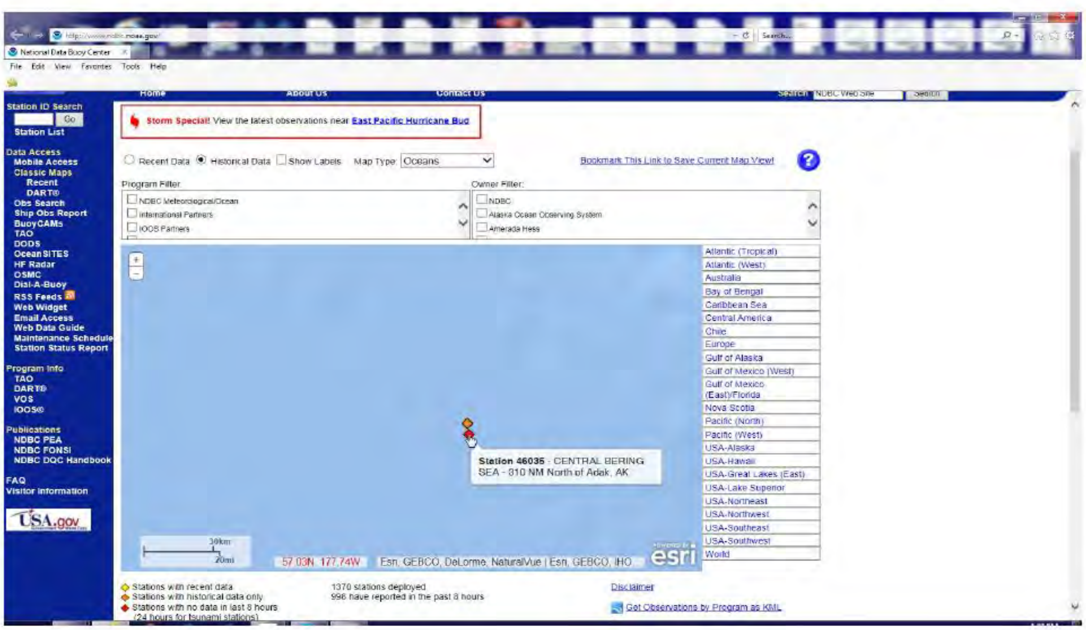
Part (A) - constructing the dataset



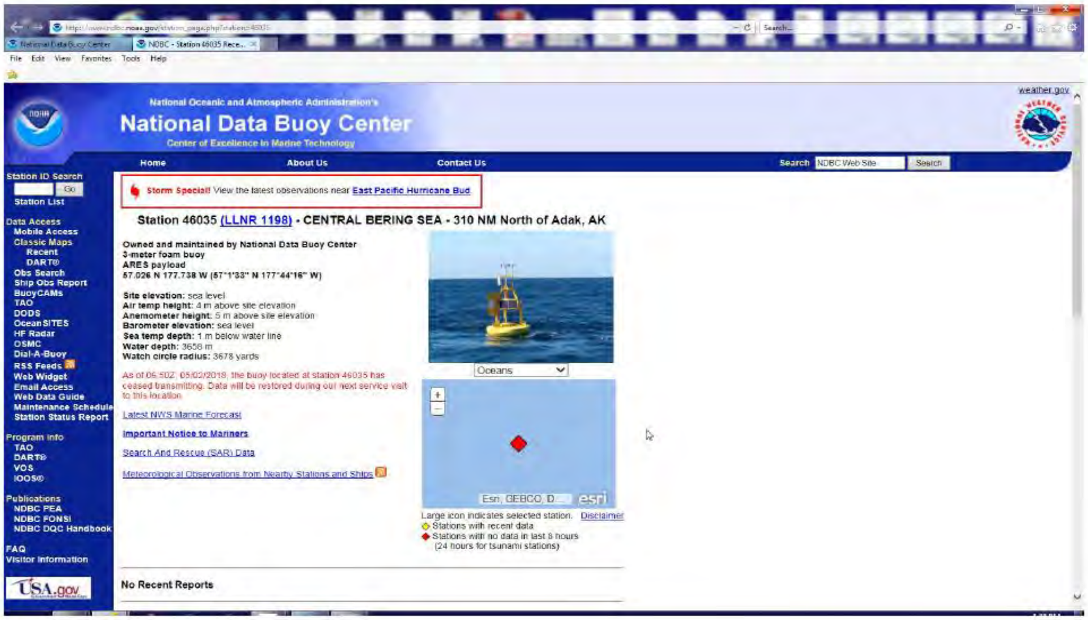
Part (A) - constructing the dataset



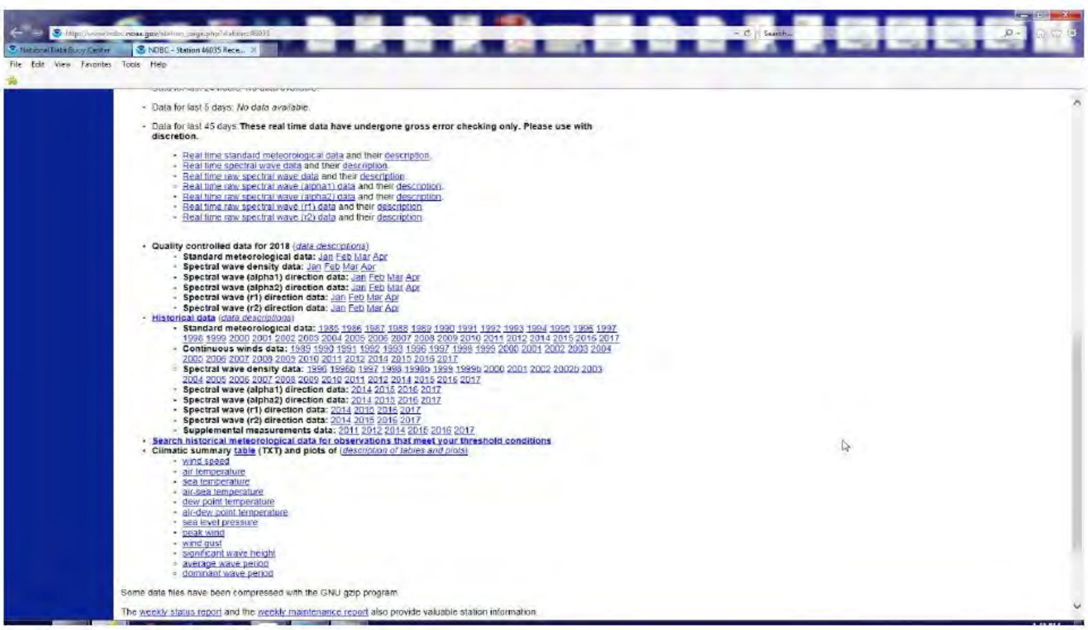
Part (A) - constructing the dataset



Part (A) - constructing the dataset

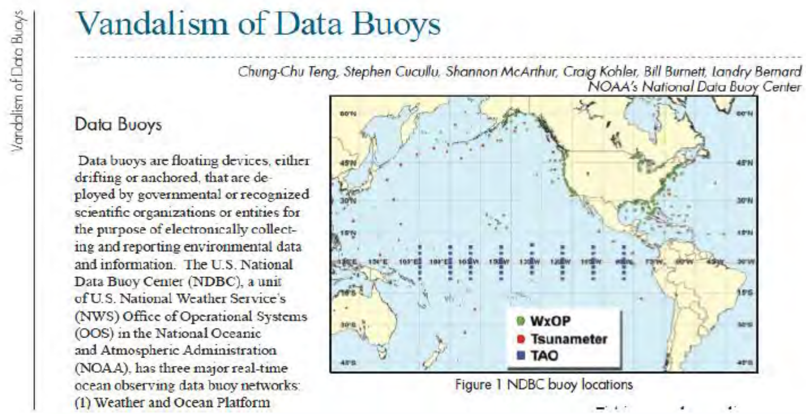


Part (A) - constructing the dataset



Part (B) - data cleansing

- Students are asked to plot and clean the data.
- Messy data: outliers, missing values, lost of data – due to vandalism/stolen of data buoys


Vandalism of Data Buoys

Vandalism of Data Buoys

Chung-Chu Teng, Stephen Cucullu, Shannon McArthur, Craig Kohler, Bill Burnett, Landry Bernard
NOAA's National Data Buoy Center

Data Buoys

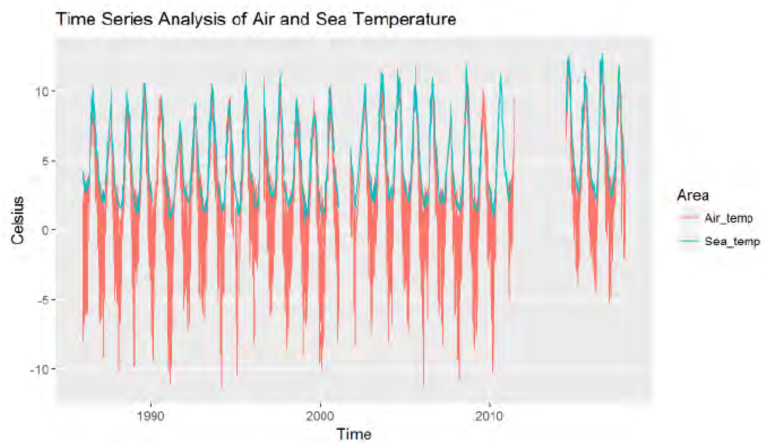
Data buoys are floating devices, either drifting or anchored, that are deployed by governmental or recognized scientific organizations or entities for the purpose of electronically collecting and reporting environmental data and information. The U.S. National Data Buoy Center (NDBC), a unit of U.S. National Weather Service's (NWS) Office of Operational Systems (OOS) in the National Oceanic and Atmospheric Administration (NOAA), has three major real-time ocean observing data buoy networks: (1) Weather and Ocean Platform

Figure 1 NDBC buoy locations

Part (B) - the research question

- Students are asked to answer the question: Global warming - have the temperatures (both sea and air) increased over the past 30 years?
- Students can use any statistical methods learned under the SOA new ASA exam curriculum .
- All computations have to be carried out in R.
- Two students form a team.
- Each team has to make a presentation and hand-in a final report (professionally written with proper conclusions and justifications).

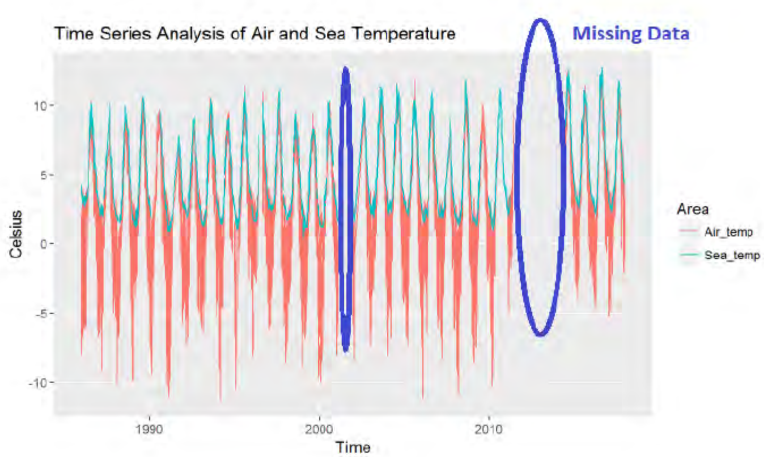
Part (C) - data cleaning



- Students have to research and decide on how to clean the data.
- If you were asked to analysing this data set, what would you do?

Part (C) - data cleaning

PROBLEM OF MISSING DATA!



How to deal with missing data?

- The first action, most of my students have done, is to ...
- Ask 'Goo-Goo'

How to deal with missing data?

- The followings are 'more reasonable' choices adopted by my students:
  - Replace the missing value with the historical average of that corresponding month
  - Replace the missing value with the corresponding observation obtained from a 'nearby' buoy
  - Fit a seasonal ARIMA model to the data and imput the missing values with the fitted value
  - Use an AI alogarithm to imput the missing value
  - Use Kalman Filter...... The R package `na.kalman()`
- There is no 'right' or 'wrong' answer in dealing with missing data...

Missing data in a BIG data set

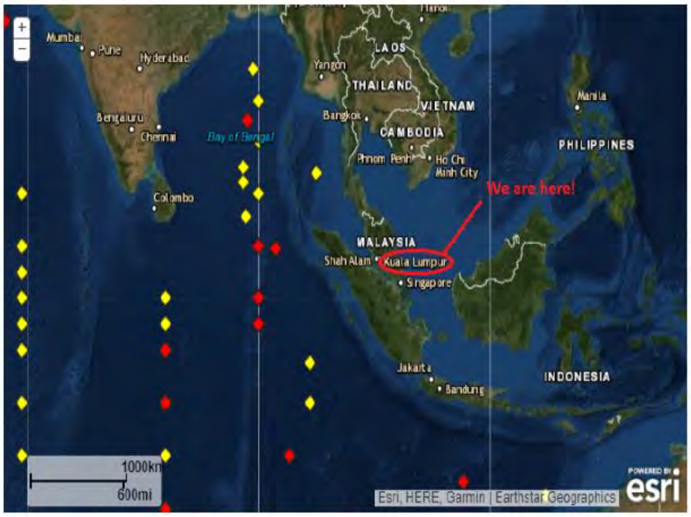
- In this climate study, we only use data from one buoy.
- In order to study the issue of gobal warming, we may use all the data in all buoys.
- It is a very BIG data set and each buoy may have different missing value problems.
- For missing value problems, we may not be able to deal with each buoy individually.
- A deep learning or AI alogarithm may help.

Missing data in a BIG data set

- There are many many buoys around the world:

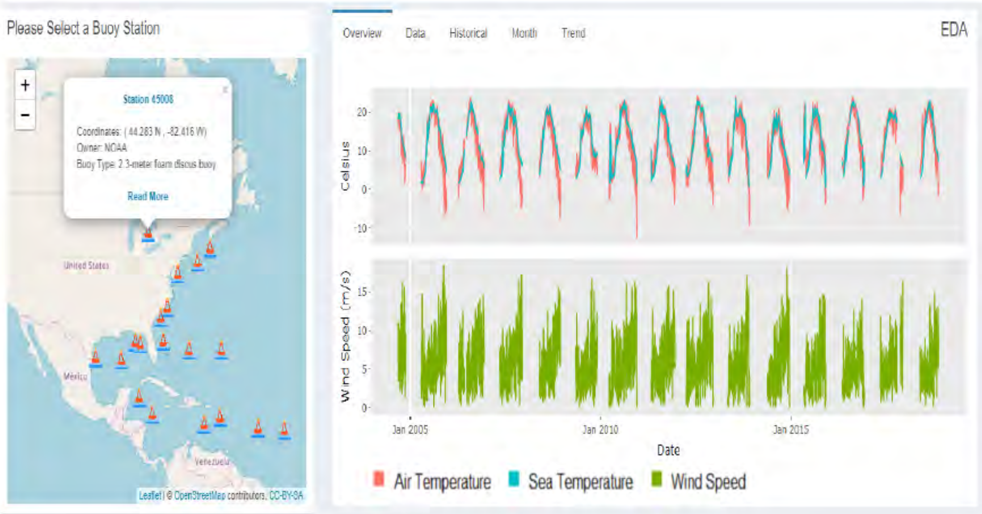
Missing data in a BIG data set

- There are no buoys near the Malaysia!



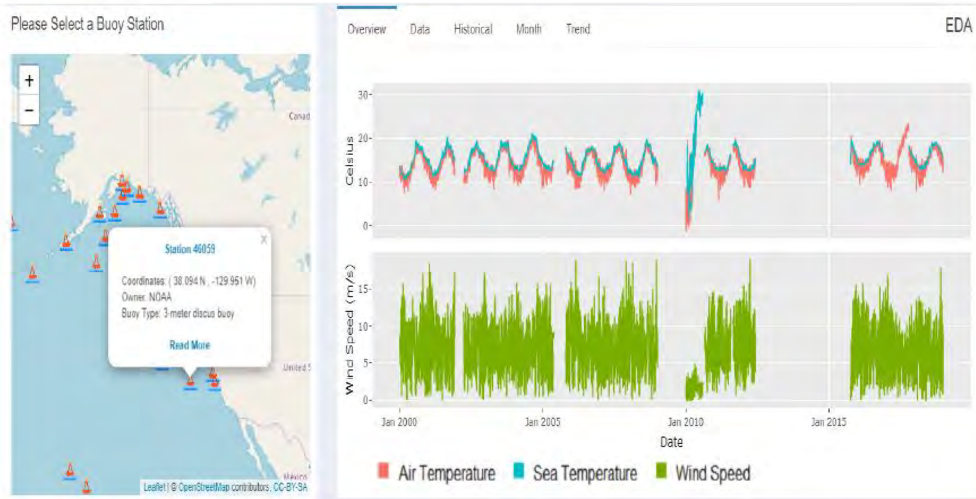
Missing data in a BIG data set

- Buoy 45008 :



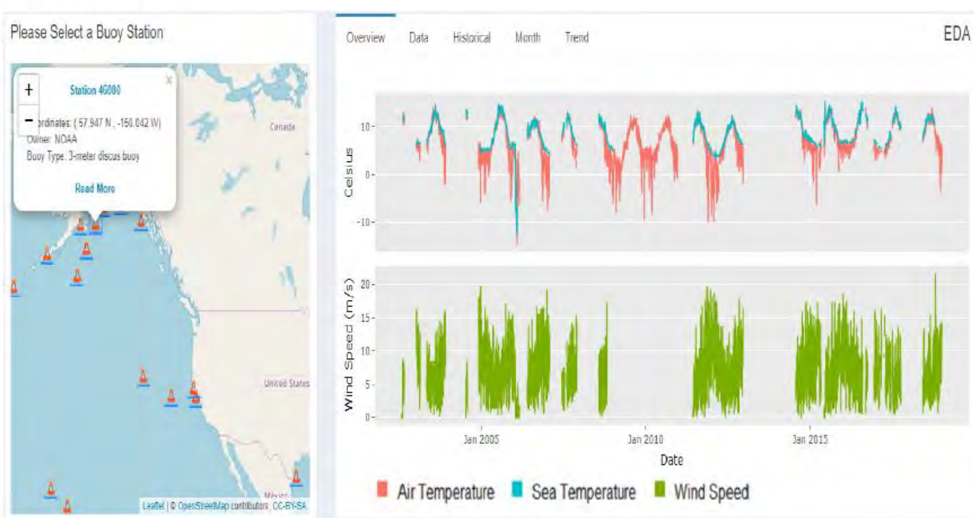
Missing data in a BIG data set

- Buoy 46059 :



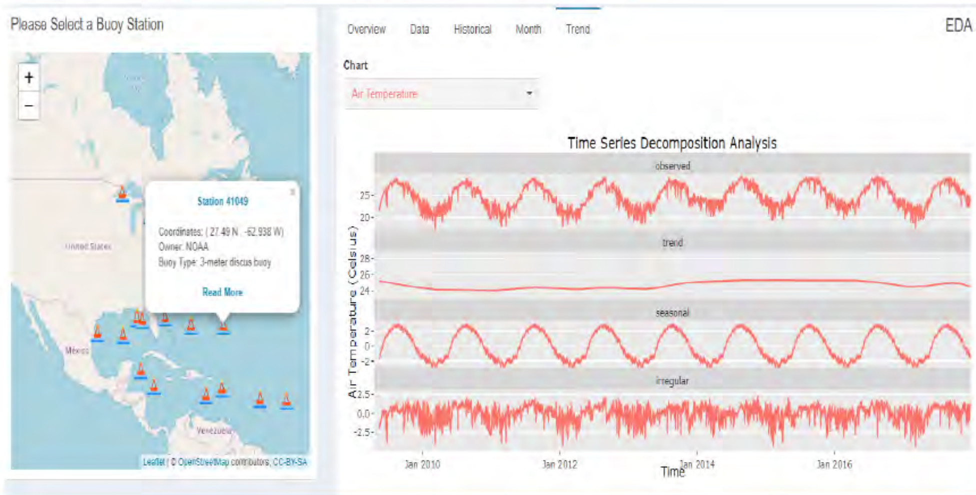
Missing data in a BIG data set

- Buoy 46059 :



Missing data in a BIG data set

- Buoy 41049 (with missing values imputed) :



(B) Outliers

What are the outliers?

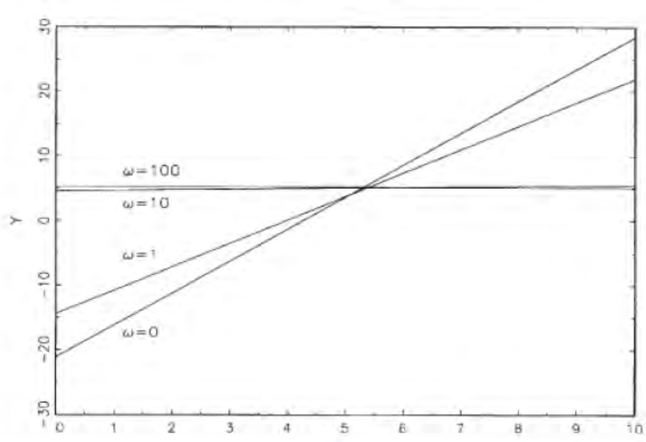
- In statistics, an outlier is a data point that differs significantly from other observations.
- differs significantly:
  - size
  - pattern (time-series)
  - catergory
  - influential
  - $\vdots \; \vdots \; \vdots$
- An outlier can cause serious problems in predicitive analyses.
- Here are some examples:

Impact of outliers on regression

- Consider a simple linear regression

$$y_i = \alpha + \beta x_i + e_i \quad \text{for } i = 1, \ldots 200.$$

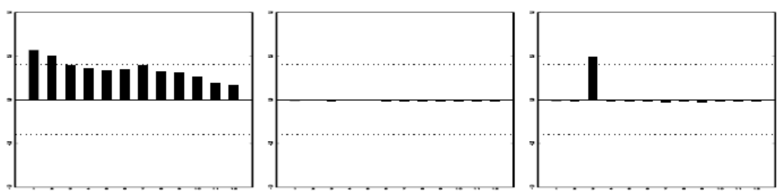
- An outlier with size ω is added to x_{100}



Impact of outliers on time-series autocorrelations

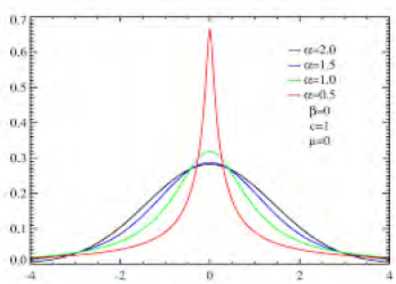
- Consider an ordinary time-series $(z_1, z_2, \ldots, Z_{200})$, according to the orthodox Box-Jenkins modelling approach, we examine the sample autocorrelation function (ACF)
- The following graphs show (a) no outlier, (b) one outlier at x_{100}, (c) two outliers at x_{100} & x_{103}

(a) (b) (c)



How to deal with outliers

- Three different philosophical approaches.
- The first one assumes that outliers occur by chances because the population has a **heavy-tailed distribution**.



- Under this approach, we can employ predicitive models which allows heavy-tailed distributions, e.g., GLM.

How to deal with outliers

- The second approach seeks to detect the outliers, provide plausible explanations, adjust the model (by dummy-variable regression or intevention method in time-series analysis) and perform prediction using the adjusted model.
- We shall briefly illustrate this approach using an actuarial example.
- Forcasting mortality rates using stochastic models has been becoming an important task for actuaries (pricing and reserving annuity products, reverse mortgages, social secutiry planning, among many others).
- We consider the classical Lee-Carter model for UK mortality data (See, Li and Chan, 2005, *Scandinavian Actuarial Journal*, 187-211).

Outliers in mortality data: an example

- **The data:** England and Wales (1841- 2000) from Human Mortality Database
- **The mortality model:** Lee-Carter (1992)

$$log(m_{x,t}) = a_x + b_x k_t + e_{x,t}$$

- where $log(m_{x,t})$ is central rate of death, a_x is a age-specific parameter, k_t is the time-varying mortality index parameter and b_x represents how rapidly or slowly mortality at each age varies when the mortality trend changes.
- **The time-series model on k_t:** ARIMA, Box and Jenkins (1976).

Outliers in mortality data: an example

- **The outlier model:**

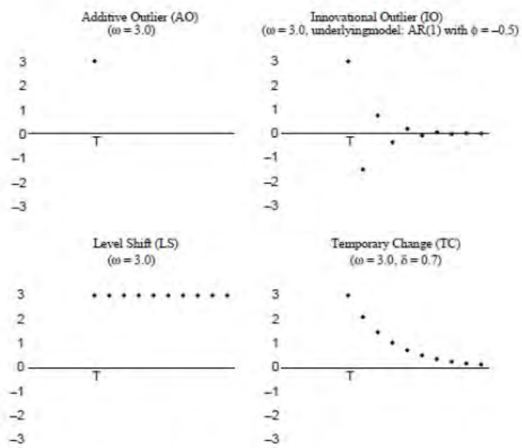


Figure 2. Different types of time-series outliers

Outliers in mortality data: an example

- **The Result:**

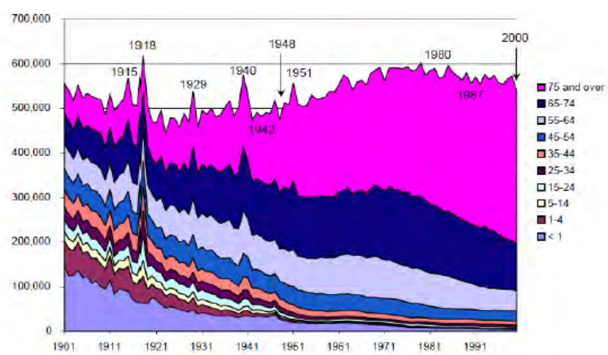
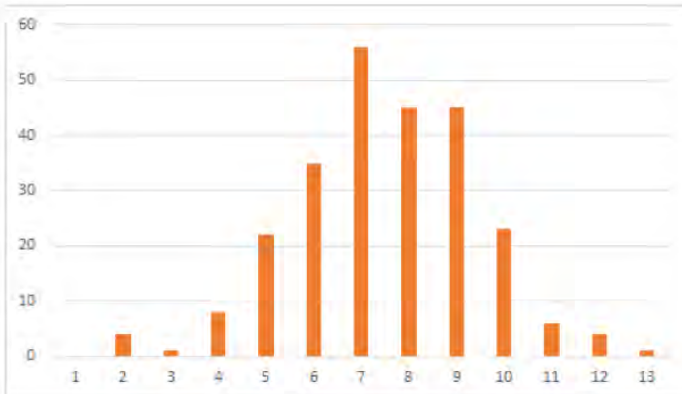


Fig. 4. Number of deaths per year (thousands), by age group, England and Wales, 1901-2000.

- **Remark:** The R package *tsoutliers* implements the above time series outlier detection procedures

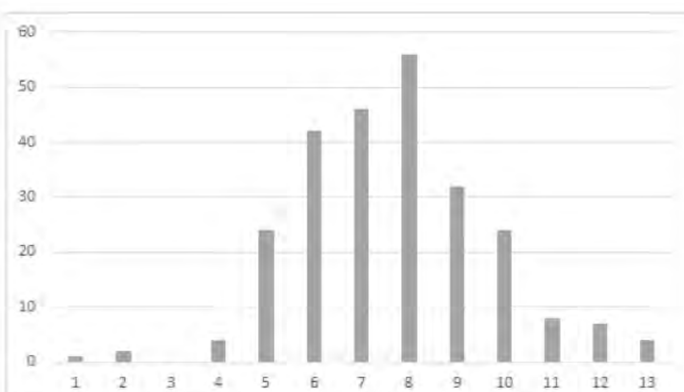
Outliers in two-dimensional data

- **Test 1:**



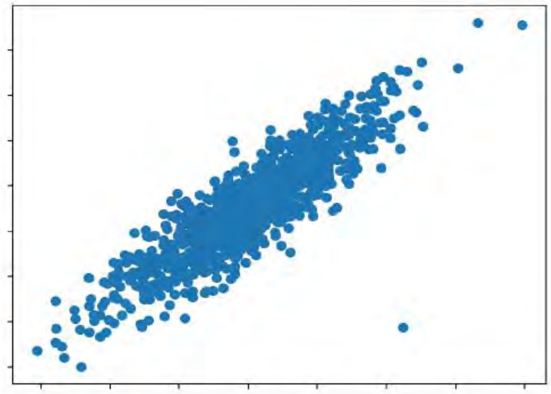
Outliers in two-dimensional data

- **Test 2:**



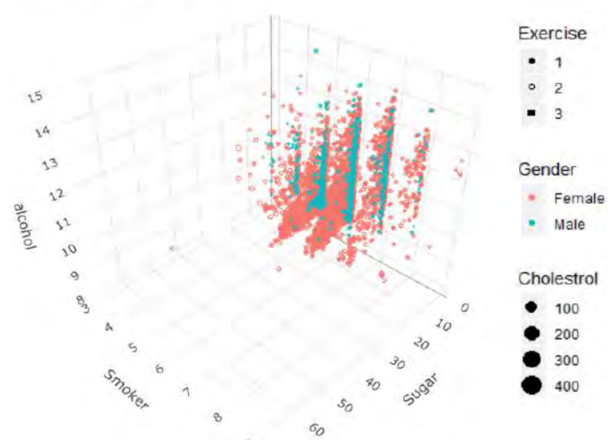
Outliers in two-dimensional data

- **Tests 1 and 2:**



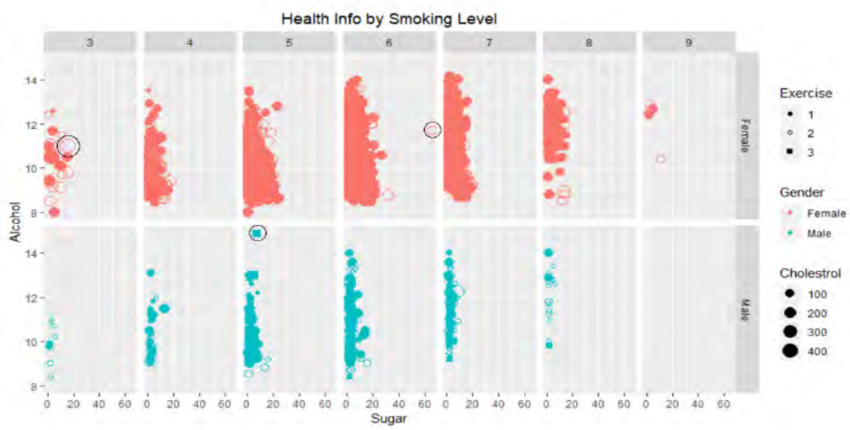
Outliers in high-dimensional big datasets

- **An Example - 6 variables: Gender, Alcohol, Smoking, Exercise, Cholestrol, Sugar**



Outliers in high-dimensional big datasets

- **An Example - 6 variables: Gender, Alcohol, Smoking, Exercise, Cholestrol, Sugar**



How to deal with outliers

- The third approach is to use **robust** and **resistant** methods for predicitive modelling.
- Robust statistical methods are expected with good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal.
- A resistant statistical method is relatively unaffected by unusual observations.
- Examples include:
  - robust regression analysis - R packages *MASS*, *robust*
  - robust time series analysis - R package *robts*
  - resistant lines - R packages *MASS*, *parody*

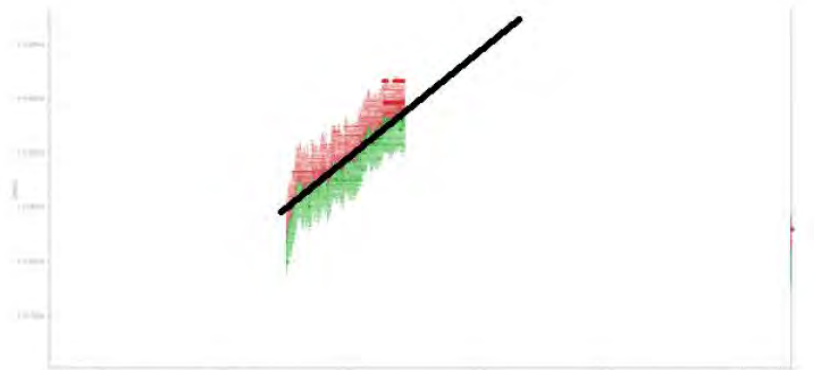
(C) Structural Changes

Structural changes

- In statistics, **structural change** is a shift or change in the basic ways the underlying mechanism functions or operates.
- For predictive modelling purpose, we may only consider the latest portion (or the most relevant portion) of the data set.
- Structural change tests are a type of statistical hypothesis test. They are used to verify the equality of coefficients across separate subsamples of a data set.
- Commonly used R packages include: *strucchange*, *segmented*, *breakpoints*
- This is particularly important for linear model analyses.

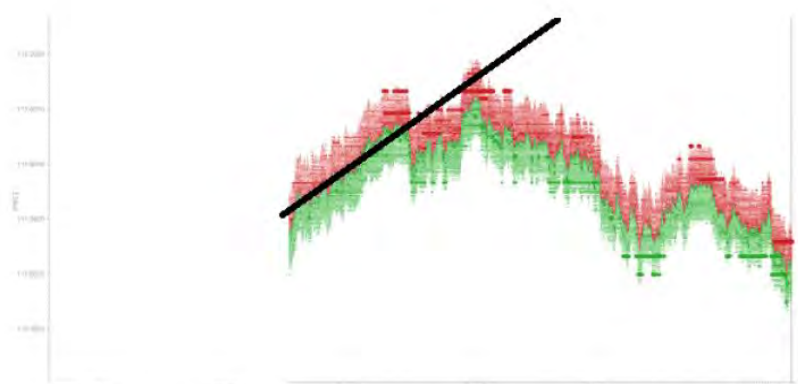
Structural changes

Structural changes



The End of the World

Structural changes



How to deal with structural changes

- One approach is to incorporate the structural changes into the predictive model.
- We shall briefly illustrate this approach using an actuarial example
- Forcasting mortality rates using stochastic models has been becoming an important task for actuaries (pricing and reserving annuity products, reverse mortgages, social secutiry planning, among many others).
- We consider the classical Lee-Carter model for US mortality data (See, Li, Chan, Cheung, 2011, *North American Actuarial Journal*, 13-31). Awarded the Edward A. Lew Research Award (Second Prize) - by SOA.

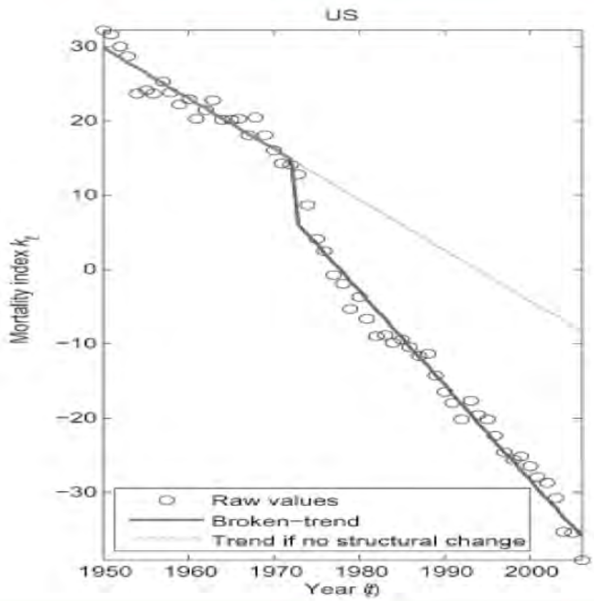
Structural changes in mortality data: an example

- **The data:** USA (1950- 2005) from Human Mortality Database
- **The mortality model:** Lee-Carter (1992)

$$log(m_{x,t}) = a_x + b_x k_t + e_{x,t}$$

- where $log(m_{x,t})$ is central rate of death, a_x is a age-specific parameter, k_t is the time-varying mortality index parameter and b_x represents how rapidly or slowly mortality at each age varies when the mortality trend changes.
- **The time-series model on k_t:** ARIMA, Box and Jenkins (1976).
- **Broken-Trend model**: R package: *ur.za*

Structural changes in mortality data: an example



(D) Abridged and Censoring Data

Abridged life tables, censoring data

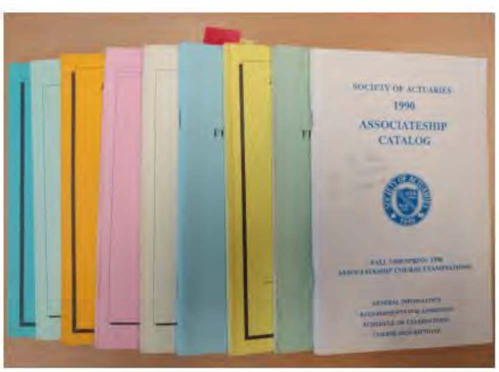
Table 1
Abridged Life Table
For Singaporeans (2001)
Age-Specific Death Rates

Age	$1000 \times {}_nM_x$ Male	$1000 \times {}_nM_x$ Female
0	2.4	2.1
$1-4$	0.3	0.3
$5-9$	0.1	0.1
$10-14$	0.1	0.1
$15-19$	0.4	0.3
$20-24$	0.7	0.2
$25-29$	0.7	0.2
$30-34$	0.7	0.5
$35-39$	1.0	0.6
$40-44$	1.6	0.9
$45-49$	2.5	1.5
$50-54$	4.6	2.6
$55-59$	8.1	4.6
$60-64$	13.2	7.2
$65-69$	23.2	12.8
$70+$	58.3	47.5

abridged

censoring

How to deal with abridged and censoring data

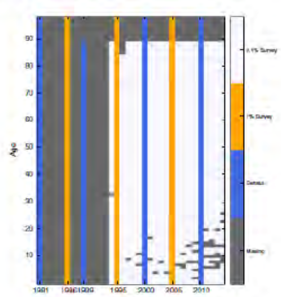


100 Calculus and Linear Algebra
110 Probability and Statistics
120 Applied Statistical Methods
130 Operations Research
135 Numerical Methods
150 Actuarial Mathematics
151 Risk Theory
160 Survival Models
162 Construction of Actuarial Table
165 Mathematics of Graduation

(E) Lack of Data, Messy Data

Lack of Data, Messy Data

- More than one problems exist in your data set
- Example: Chinese mortality data



(b)

Fig. 1. Lexis diagrams summarizing the availability of mortality data for (a) Chinese males and (b) Chinese females: ▮, data obtained from censuses; ▮, data obtained from 1% surveys; ▯, data obtained from 0.1% surveys; ▮, missing data

- Bayesian approach may be useful....

Summary

- There are many problems that associated with messy data:
  - missing values
  - outliers
  - structural changes
  - abridged and censoring data
  - lack of data and messy data
  - ... and many more
- The main purpose of this presentation is to draw audience's attention to this important topic in predicitive analytics

Thank You!

Q & A