



**SOCIETY OF
ACTUARIES®**

SOA Predictive Analytics Seminar – Malaysia

26 Aug. 2019 | Kuala Lumpur, 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



1 of 60

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?

2 of 60

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)

3 of 60

My lovely data generator



4 of 60

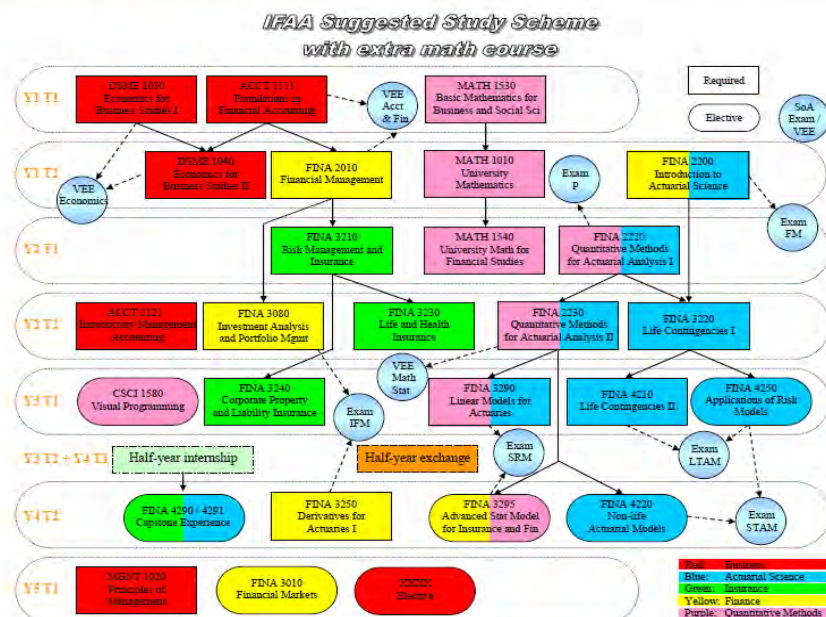
SOA exam curriculum in 1990s



- 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

5 of 60

SOA exam curriculum in 2020s



6 of 60

(A) Missing Data

7 of 60

Missing Data: A climate change data case study

The screenshot shows the homepage of 'THE Actuary' website. At the top, there is a navigation bar with a menu icon, a search bar, and social media icons for Facebook, LinkedIn, Twitter, and YouTube. Below the navigation bar is a horizontal menu with links for Home, Features, Current Issue, Departments, Web Exclusives, Archives, Annual Report, Advertising, About Us, and SOA.org. The main content area features a large image of a hand pulling back a white sheet to reveal a dark, stormy landscape with lightning bolts. Below the image, the article title 'The Challenges of Climate Change' is displayed in bold, followed by the subtitle 'The global actuarial response to the difficulties of climate change', the author's name 'YVES GUERARD', and the date 'APRIL/MAY 2018'. To the right of the article is a 'Related Posts' section with three links: 'An International Career', 'Double Threat', and 'Moving Past the Debate'.

8 of 60

Missing Data: A climate change data case study

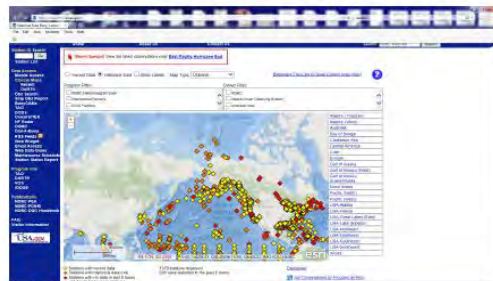
- The National Data Buoy Center (NDBC) is a part of the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) of the US government.
- NDBC deploys weather **buoys** which are instruments which **collect weather and ocean data** within the world's oceans.



9 of 60

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".

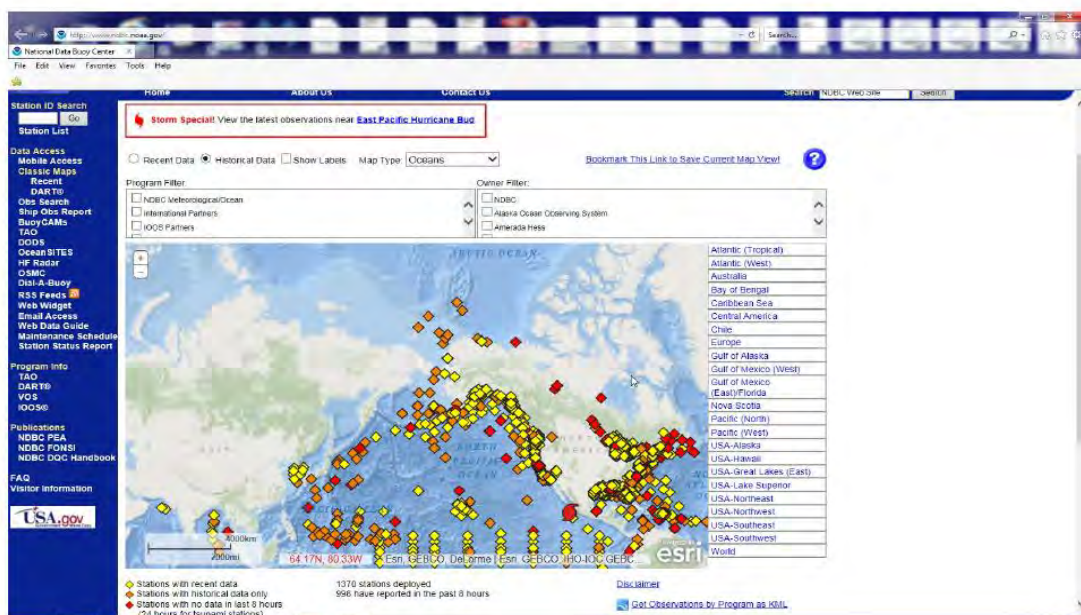
10 of 60

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**.

11 of 60

Part (A) - constructing the dataset



12 of 60

Part (A) - constructing the dataset

Station ID Search:

Station List

Data Access: Mobile Access, Classic Maps, Recent, DARTto, Obs Search, Ship Obs Report, BuoyCAMS, TAO, DQDS, Ocean SITES, HF Radar, OSMC, Dist-A-Buoy, RSS Feeds, Web Widget, Email Access, Web Data Guide, Maintenance Schedule, Station Status Report

Program Info: TAO, DARTto, VOS, IOOSp

Publications: NDBC PEA, NDBC FONSI, NDBC DQC Handbook

FAQ, Visitor Information

USA.GOV

Home | About Us | Contact Us

Storm Special! View the latest observations near **East Pacific Hurricane Bud**

Recent Data Historical Data Show Labels Map Type: Oceans

Program Filter: NDBC Meteorological/Ocean International Partners IOOS Partners

Owner Filter: NDBC Alsea Ocean Observing System Amerasia Press

Station 46035
NDBC
Locations: 57.026N 177.736W
There are no recent (< 8 hours) meteorological data for this station. Click [here](#) for other data from this station.

Aleutian Basin 57.02N 177.24W Esri GEBCO, Dalorme, NaturalVue | Esri, GEBCO, IHO

Atlantic (Tropical), Atlantic (West), Australia, Bay of Bengal, Caribbean Sea, Central America, Chile, Europe, Gulf of Alaska, Gulf of Mexico (West), Gulf of Mexico (East)/Florida, Nova Scotia, Pacific (North), Pacific (West), USA-Alaska, USA-Hawaii, USA-Great Lakes (East), USA-Lake Superior, USA-Northeast, USA-Northwest, USA-Southeast, USA-Southwest, World

1370 stations deployed
596 have reported in the past 8 hours

Stations with recent data
Stations with historical data only
Stations with no data in last 8 hours (24 hours for buoy stations)

Disclaimer
[Get Observations by Program as KML](#)

13 of 60

Part (A) - constructing the dataset

Station ID Search:

Station List

Data Access: Mobile Access, Classic Maps, Recent, DARTto, Obs Search, Ship Obs Report, BuoyCAMS, TAO, DQDS, Ocean SITES, HF Radar, OSMC, Dist-A-Buoy, RSS Feeds, Web Widget, Email Access, Web Data Guide, Maintenance Schedule, Station Status Report

Program Info: TAO, DARTto, VOS, IOOSp

Publications: NDBC PEA, NDBC FONSI, NDBC DQC Handbook

FAQ, Visitor Information

USA.GOV

Home | About Us | Contact Us

Storm Special! View the latest observations near **East Pacific Hurricane Bud**

Recent Data Historical Data Show Labels Map Type: Oceans

Program Filter: NDBC Meteorological/Ocean International Partners IOOS Partners

Owner Filter: NDBC Alsea Ocean Observing System Amerasia Press

Station 46035 - CENTRAL BERING SEA - 810 NM North of Adak, AK
NDBC
Locations: 57.03N 177.74W
There are no recent (< 8 hours) meteorological data for this station. Click [here](#) for other data from this station.

Atlantic (Tropical), Atlantic (West), Australia, Bay of Bengal, Caribbean Sea, Central America, Chile, Europe, Gulf of Alaska, Gulf of Mexico (West), Gulf of Mexico (East)/Florida, Nova Scotia, Pacific (North), Pacific (West), USA-Alaska, USA-Hawaii, USA-Great Lakes (East), USA-Lake Superior, USA-Northeast, USA-Northwest, USA-Southeast, USA-Southwest, World

1370 stations deployed
596 have reported in the past 8 hours

Stations with recent data
Stations with historical data only
Stations with no data in last 8 hours (24 hours for buoy stations)

Disclaimer
[Get Observations by Program as KML](#)

14 of 60

Part (A) - constructing the dataset

Station ID Search: [] [GO]
Station List

Data Access: Mobile Access, Classic Maps, Recent, DART, Ship Obs Report, BuoyCAMS, IAO, DODS, Ocean BITES, HF Radar, OSMC, Data Buoy, RSS Feeds, Web Widget, Email Access, Web Data Guide, Maintenance Schedule, Station Status Report

Program Info: IAO, DART, VOS, DODS

Publishers: NDBC PEA, NDBC PONS, NDBC DQC Handbook, FAQ, Vector Information

USA.gov

National Oceanic and Atmospheric Administration's
National Data Buoy Center
Center of Excellence in Marine Technology

Home About Us Contact Us Search NDBC Web Site Search

Storm Special! View the latest observations near **East Pacific Hurricane Bud**

Station 46035 (LLNR 1198) - CENTRAL BERING SEA - 310 NM North of Adak, AK

Owned and maintained by National Data Buoy Center
3 meter beam buoy
ARE 5 payload
87.826 N 177.738 W (87°13'33" N 177°44'18" W)

Site elevation: sea level
Air temp height: 4 m above site elevation
Anemometer height: 5 m above site elevation
Barometer elevation: sea level
Sea temp depth: 1 m below water line
Water depth: 3658 m
Watch circle radius: 3675 yards

As of 06:00Z 05/02/2018, the buoy located at station 46035 has ceased transmitting. Data will be restored during our next service visit to this location.

Latest NWS Marine Forecast
Important Notice to Mariners
Search and Rescue (SAR) Data
Meteorological Observations from Nearby Stations and Ships

Oceans

Large icon indicates selected station. Disclaimer
Stations with recent data
Stations with no data in last 8 hours (24 hours for tsunami stations)

No Recent Reports

15 of 60

Part (A) - constructing the dataset

Data for last 5 days: No data available.

Data for last 45 days: These real time data have undergone gross error checking only. Please use with discretion.

- Real time standard meteorological data and their description
- Real time spectral wave data and their description
- Real time raw spectral wave data and their description
- Real time raw spectral wave (alpha1) data and their description
- Real time raw spectral wave (alpha2) data and their description
- Real time raw spectral wave (r1) data and their description
- Real time raw spectral wave (r2) data and their description
- Real time raw spectral wave (r3) data and their description

Quality controlled data for 2018 (data description)

- Standard meteorological data: Jan Feb Mar Apr
- Spectral wave density data: Jan Feb Mar Apr
- Spectral wave (alpha1) direction data: Jan Feb Mar Apr
- Spectral wave (alpha2) direction data: Jan Feb Mar Apr
- Spectral wave (r1) direction data: Jan Feb Mar Apr
- Spectral wave (r2) direction data: Jan Feb Mar Apr

Historical data (data description)

- Standard meteorological data: 1980 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
- Continuous winds data: 1980 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
- Spectral wave density data: 1980 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017
- Spectral wave (alpha1) direction data: 2014 2016 2016 2017
- Spectral wave (alpha2) direction data: 2014 2016 2016 2017
- Spectral wave (r1) direction data: 2014 2016 2016 2017
- Spectral wave (r2) direction data: 2014 2016 2016 2017
- Supplemental measurements data: 2014 2016 2016 2017

Search historical meteorological data for observations that meet your threshold conditions

Climatic summary table (TXT) and plots of (descriptions of tables and plots)

- wind speed
- air temperature
- sea temperature
- air-sea temperature
- sea-point temperature
- air-sea-point temperature
- sea level pressure
- total wind
- wind gust
- significant wave height
- average wave period
- dominant wave period

Some data files have been compressed with the GNU gzip program

The [recent status report](#) and the [recent maintenance report](#) also provide valuable station information.

16 of 60

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

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

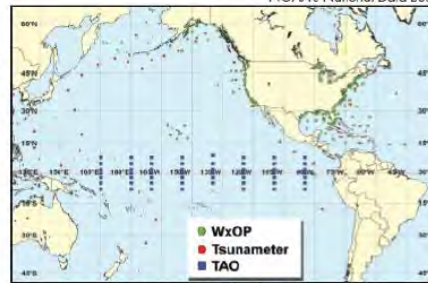
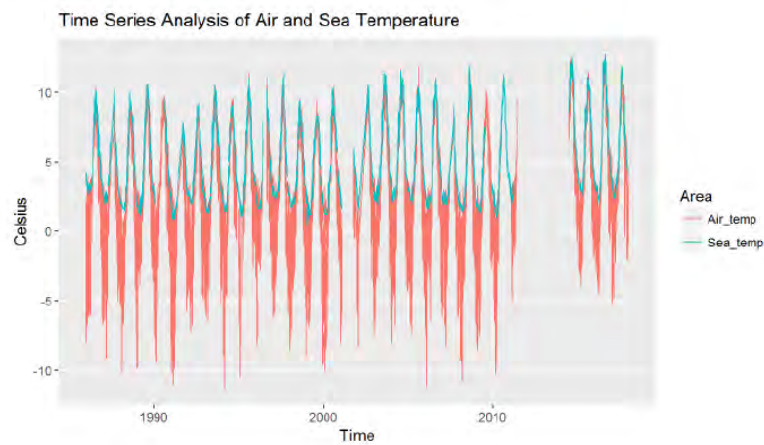


Figure 1 NDBC buoy locations

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

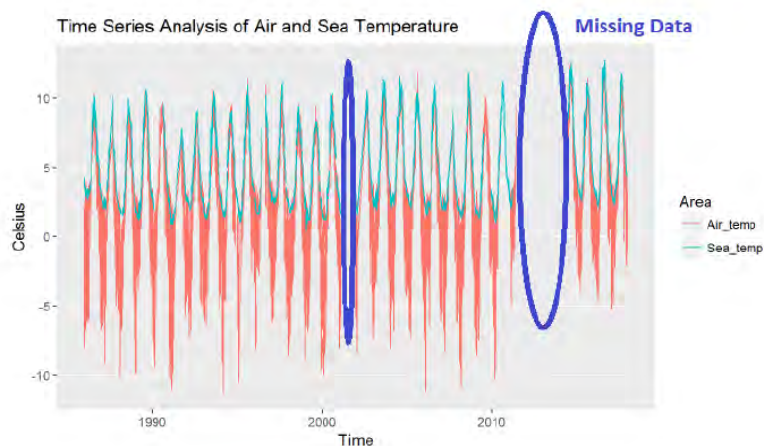


- Students have to research and decide on how to clean the data.
- If you were asked to analysing this data set, what would you do?

19 of 60

Part (C) - data cleaning

PROBLEM OF MISSING DATA!



20 of 60

How to deal with missing data?

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



21 of 60

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 impute the missing values with the fitted value
 - Use an AI algorithm to impute the missing value
 - Use Kalman Filter..... The R package `na.kalman()`
- There is no 'right' or 'wrong' answer in dealing with missing data...

22 of 60

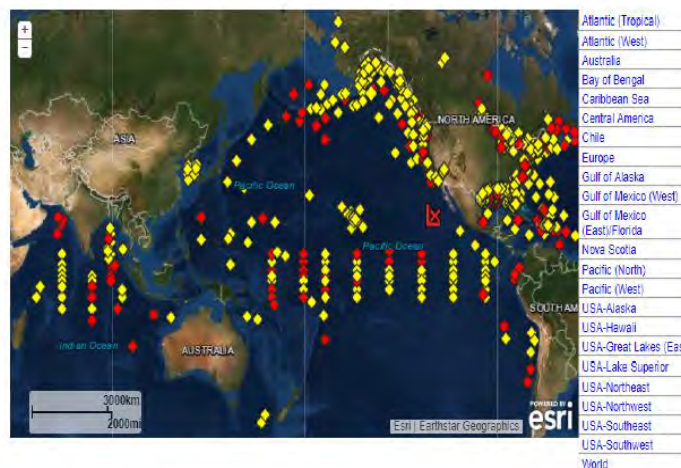
Missing data in a BIG data set

- In this climate study, we only use data from one buoy.
- In order to study the issue of global 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 algorithm may help.

23 of 60

Missing data in a BIG data set

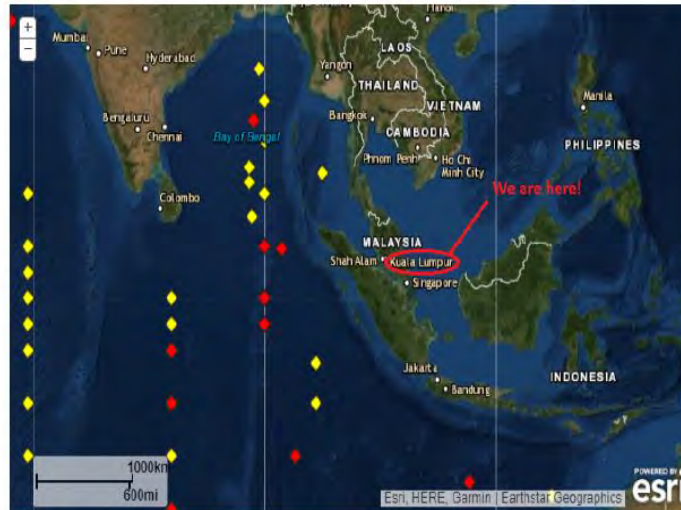
- There are many many buoys around the world:



24 of 60

Missing data in a BIG data set

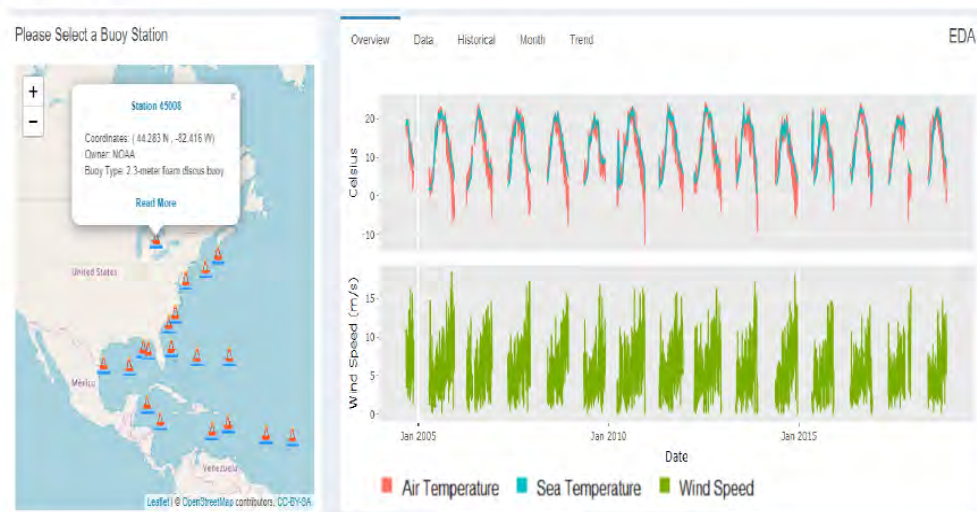
- There are no buoys near the Malaysia!



25 of 60

Missing data in a BIG data set

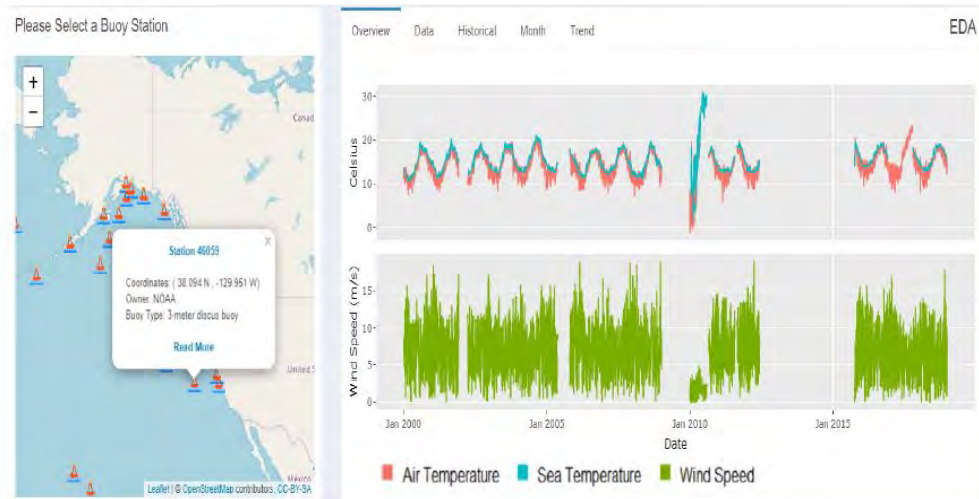
- Buoy 45008 :



26 of 60

Missing data in a BIG data set

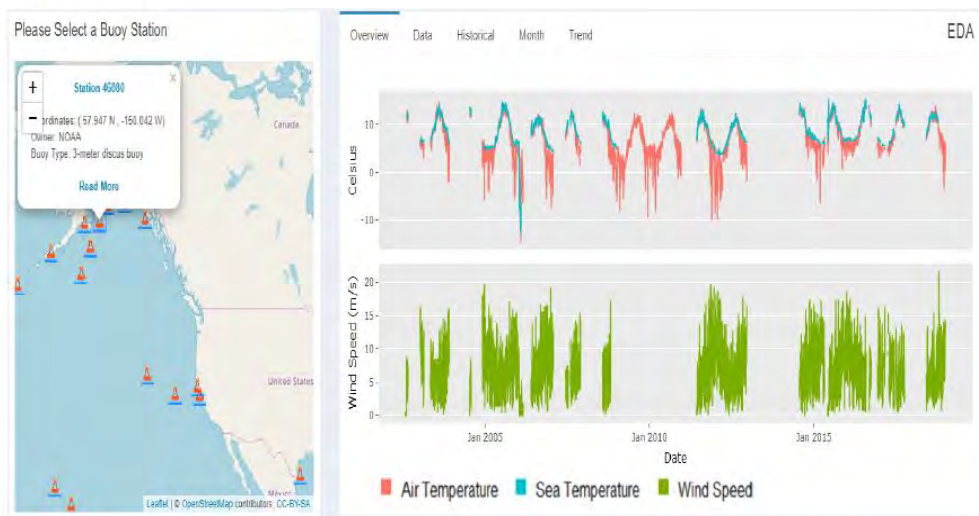
- Buoy 46059 :



27 of 60

Missing data in a BIG data set

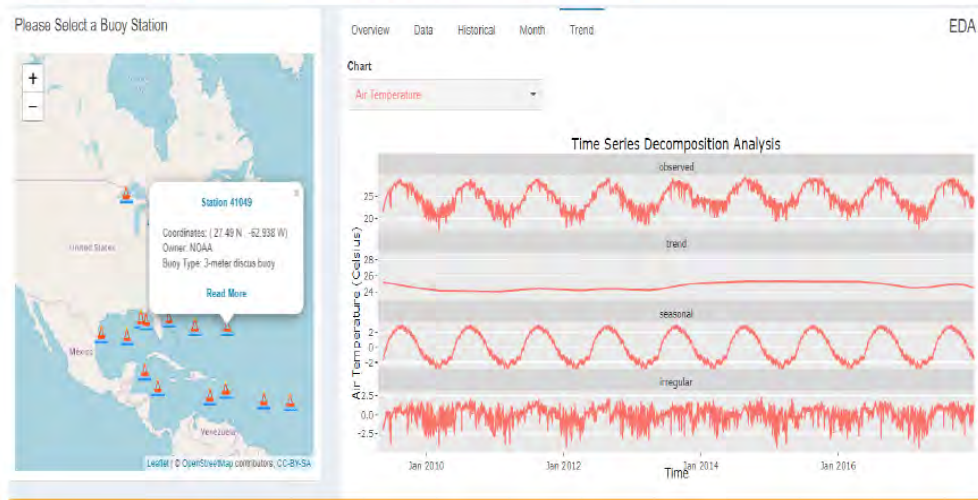
- Buoy 46059 :



28 of 60

Missing data in a BIG data set

- Buoy 41049 (with missing values imputed) :



29 of 60

(B) Outliers

30 of 60

What are the outliers?

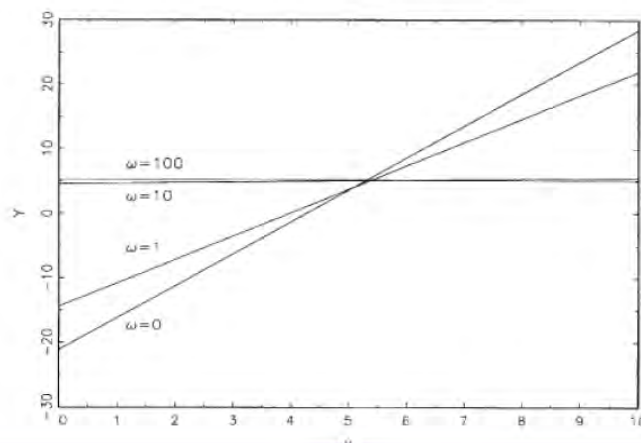
- In statistics, an outlier is a data point that **differs significantly** from other observations.
- **differs significantly:**
 - size
 - pattern (time-series)
 - category
 - influential
 - ⋮
- An outlier can cause serious problems in predictive analyses.
- Here are some examples:

31 of 60

Impact of outliers on regression

- Consider a simple linear regression

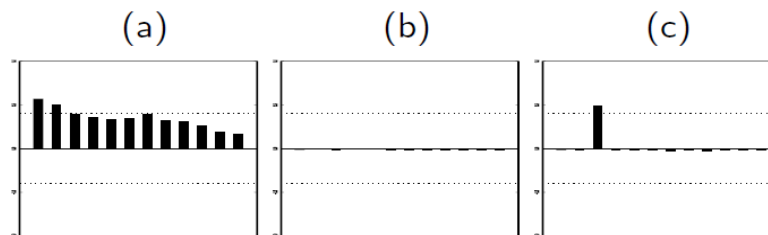
$$y_i = \alpha + \beta x_i + e_i \quad \text{for } i = 1, \dots, 200.$$
- An outlier with size ω is added to x_{100}



32 of 60

Impact of outliers on time-series autocorrelations

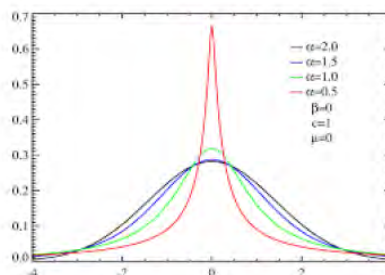
- Consider an ordinary time-series $(z_1, z_2, \dots, Z_{200})$, according to the orthodox Box-Jenkins modelling approach, we examine the sample autocorrelation function (ACF)
- The following graphs show (a) no outlier, (b) one outlier at x_{100} , (c) two outliers at x_{100} & x_{103}



33 of 60

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 predictive models which allows heavy-tailed distributions, e.g., GLM.

34 of 60

How to deal with outliers

- The second approach seeks to detect the outliers, provide plausible explanations, adjust the model (by dummy-variable regression or intervention method in time-series analysis) and perform prediction using the adjusted model.
- We shall briefly illustrate this approach using an actuarial example.
- Forecasting mortality rates using stochastic models has been becoming an important task for actuaries (pricing and reserving annuity products, reverse mortgages, social security 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).

35 of 60

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).

36 of 60

Outliers in mortality data: an example

- The outlier model:

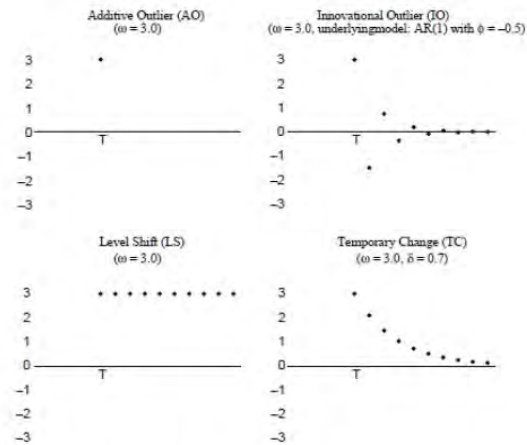


Figure 2. Different types of time-series outliers.

37 of 60

Outliers in mortality data: an example

- The Result:

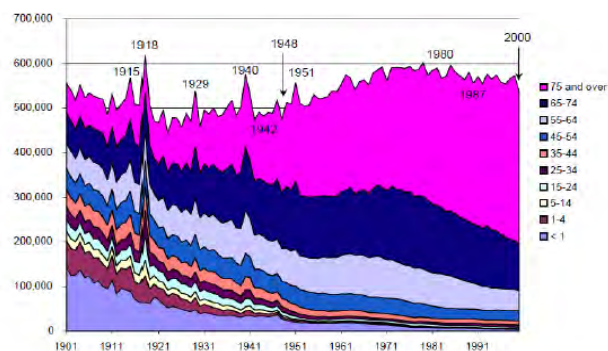


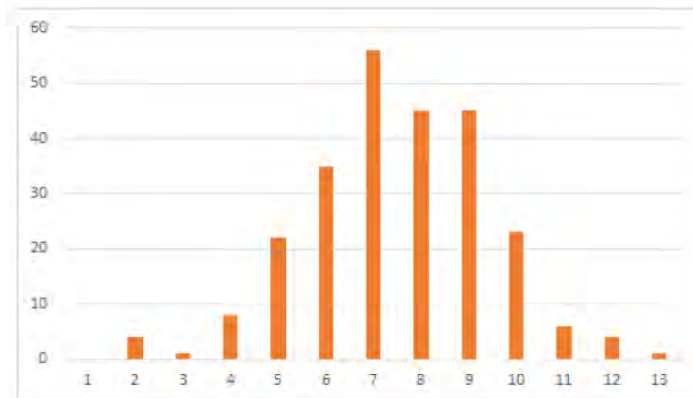
Fig. 4. Number of deaths per year (thousands), by age group, England and Wales, 1901-2000.

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

38 of 60

Outliers in two-dimensional data

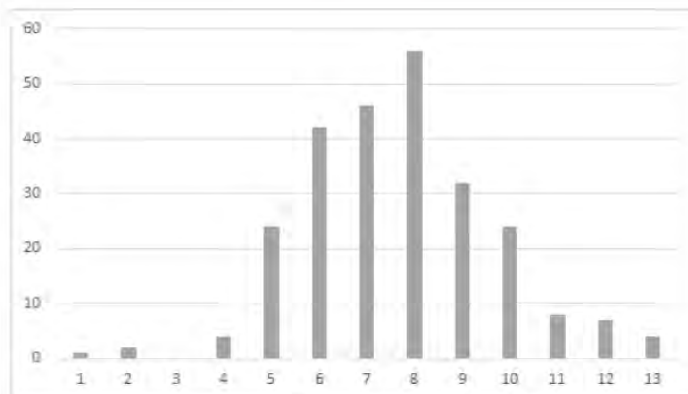
- **Test 1:**



39 of 60

Outliers in two-dimensional data

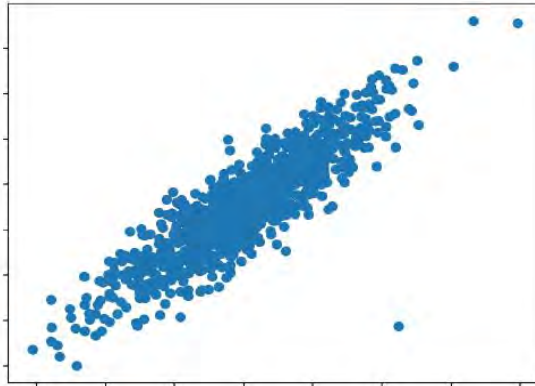
- **Test 2:**



40 of 60

Outliers in two-dimensional data

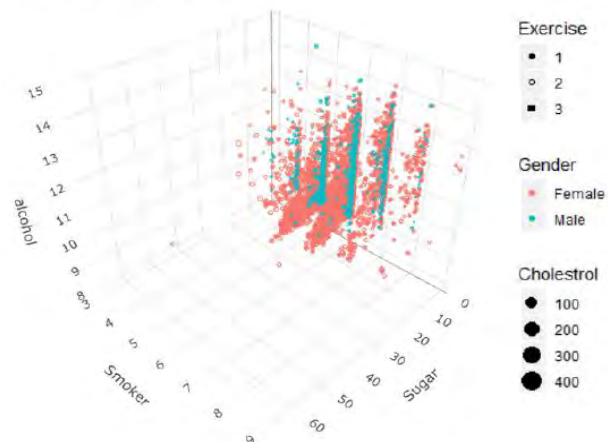
- Tests 1 and 2:



41 of 60

Outliers in high-dimensional big datasets

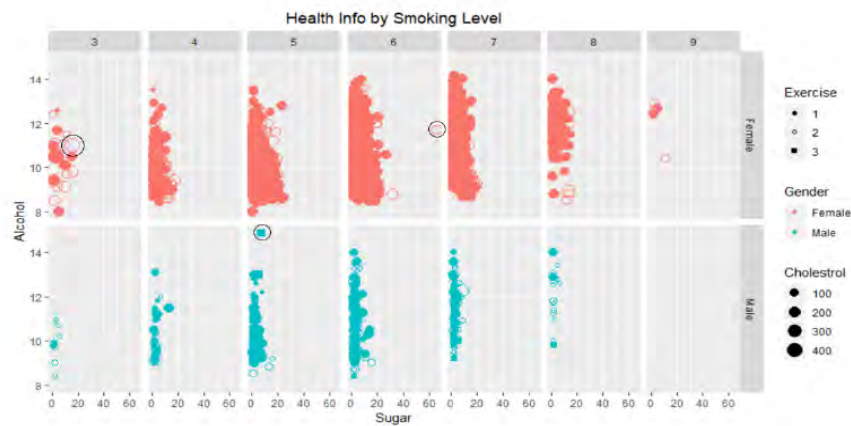
- An Example - 6 variables: Gender, Alcohol, Smoking, Exercise, Cholestrol, Sugar



42 of 60

Outliers in high-dimensional big datasets

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



43 of 60

How to deal with outliers

- The third approach is to use **robust** and **resistant** methods for predictive 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*

44 of 60

(C) Structural Changes

45 of 60

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.

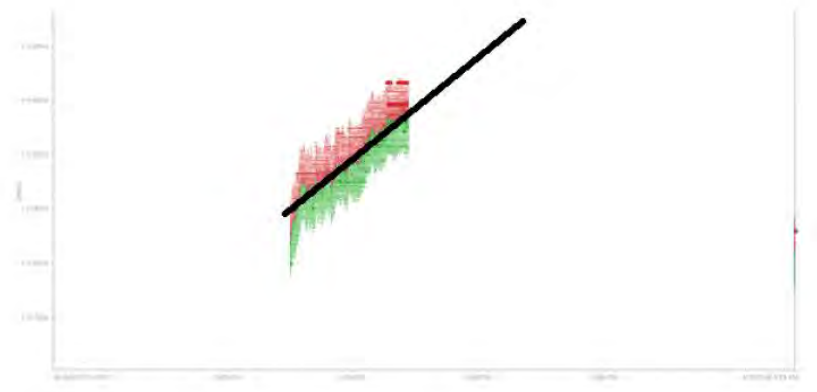
46 of 60

Structural changes



47 of 60

Structural changes



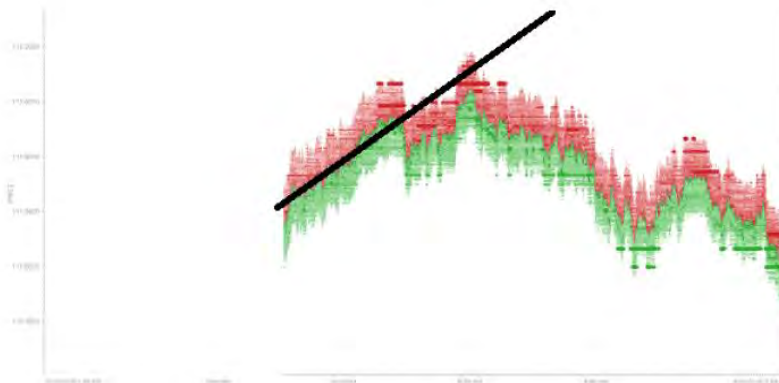
48 of 60

The End of the World



49 of 60

Structural changes



50 of 60

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
- Forecasting mortality rates using stochastic models has been becoming an important task for actuaries (pricing and reserving annuity products, reverse mortgages, social security 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.

51 of 60

Structural changes in mortality data: an example

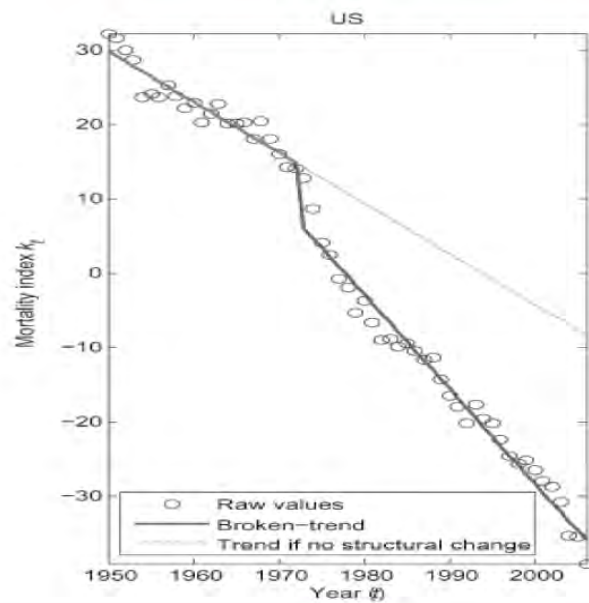
- **The data:** USA (1950- 2005) from Human Mortality Database
- **The mortality model:** Lee-Carter (1992)

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

- 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*

52 of 60

Structural changes in mortality data: an example



53 of 60

(D) Abridged and Censoring Data

54 of 60

Abridged life tables, censoring data

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

Age	$1000 \times {}_nM_x$	
	Male	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

censoring

55 of 60

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

56 of 60

(E) Lack of Data, Messy Data

57 of 60

Lack of Data, Messy Data

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

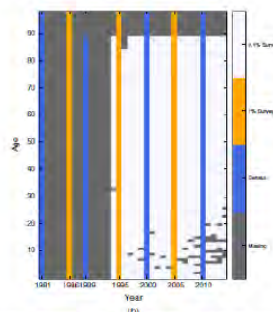


Fig. 1. Lexis diagrams summarizing the availability of mortality data for (a) Chinese males and (b) Chinese females. I, data obtained from censuses; II, data obtained from 1% surveys; III, data obtained from 0.1% surveys; IV, missing data.

- Bayesian approach may be useful....

58 of 60

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