Two former New Zealand Air Force photographic analysts described this photograph as "inscrutable". It was one of a series taken in the early morning on October 27, 1979 from a plane over Motonau, New Zealand.

Source: The Telegraph, UK

UFO sightings

Data Exploration and Forecast

Saurabh Gupta
sgup072@aucklanduni.ac.nz

The Dataset

The dataset is a CSV file where each row is a record of a UFO 'sighting', and each column is a piece of information about that sighting.

- Hence, each row is 1 sighting.
- In all, there were 80,331 rows and 16 columns
- After sub-setting for missing values, I used 80,061 sightings for analysis

Types of information contained

- Latitude and Longitude
- Location in different formats – city, country, code, etc.
- Time in different formats – full date.time, year, month, etc.
- Duration and Shape
- Comments

Data exploration and the code has been provided in R markdown files: UFOsightings_part1.Rmd and UFOsightings_part2.Rmd (for Time Series Model)

Missing Values

- Duration has 265 missing values
  - These were rows containing NA
  - They do not belong to any specific city, hence, not biased by location.
  - No rows had duration = 0
- Latitude and Longitude don't have any missing but name of the country (country_clean) has 3178 missing values.
  - Its probably because they may have been over international waters (including high skies) as they don't belong to a specific country. Ideally, I would check with Geonames api if there is any more info on those coordinates.
- Date time was missing in 5 rows

Sub-setting the data

- Assuming that Duration and DateTime is Missing Completely at Random and invalid entries, I'll subset the data to remove them.

Task 2

80,061 sightings

- Total **number of sightings** excluding missing values
- Assumptions explained on previous slide

157 countries

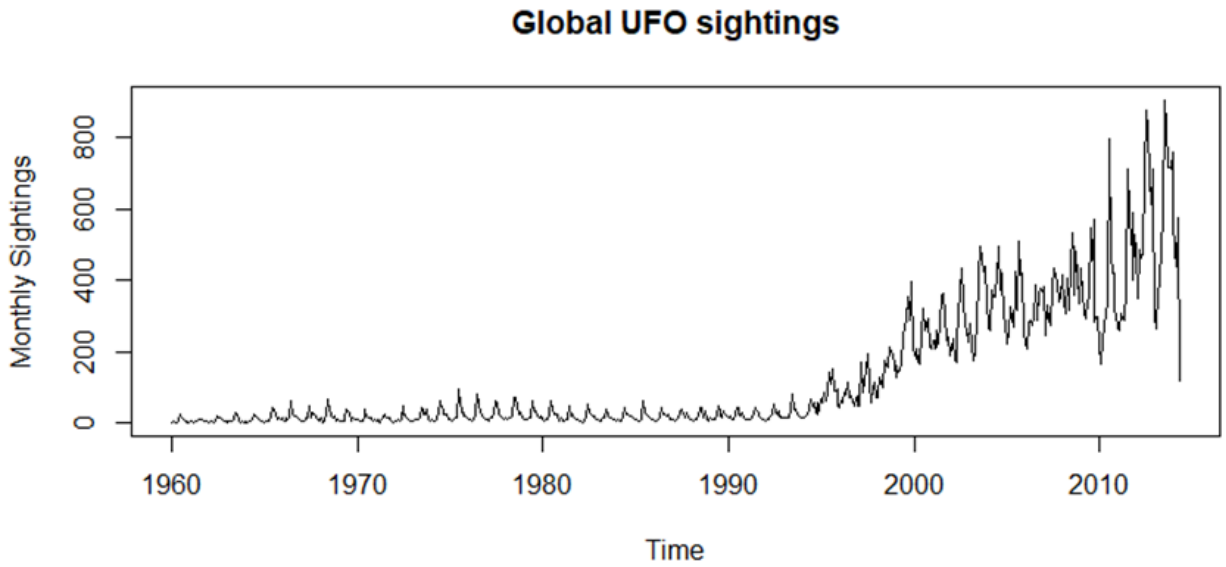
- **Number of different countries** from which the sightings originate
- Assumes that countries with 3,178 missing values (3175 after sub-setting) are international waters

125 teardrop shaped

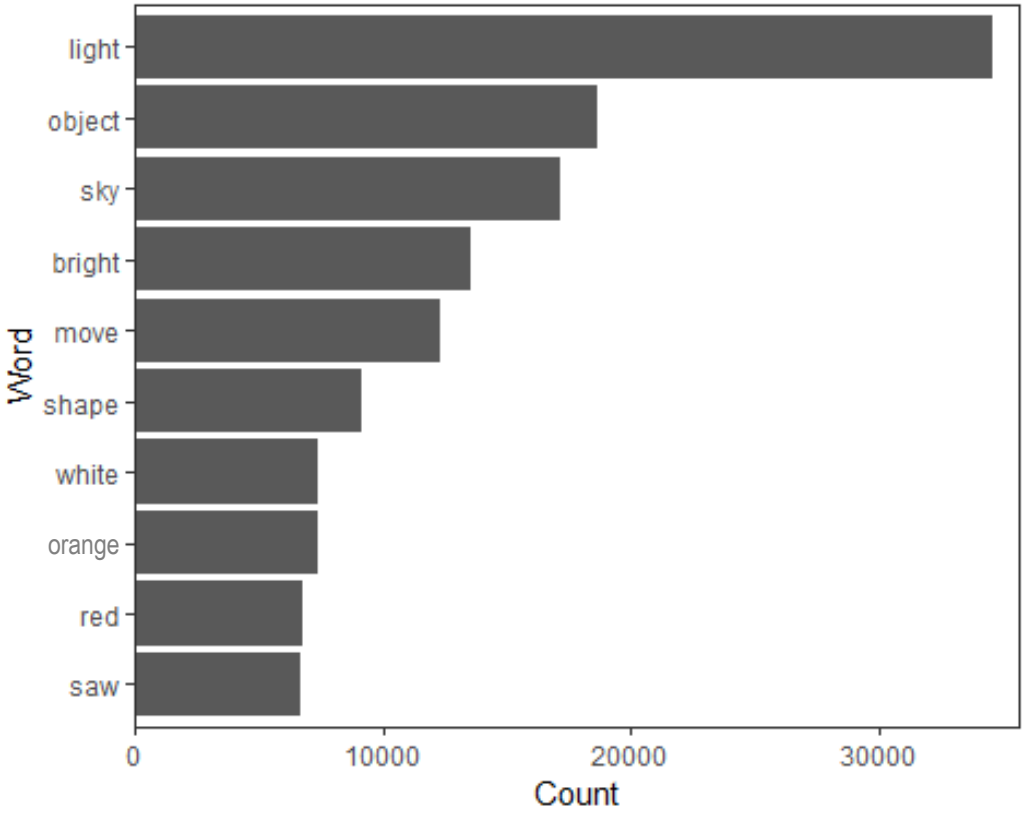
- Number of ‘teardrop’ shaped UFO sightings **between 1950 and 2000 (inclusive)**
- 1,932 have observations have missing values for shape i.e. 2.4% of the sightings.
- Assuming, they are MCAR, expect the actual number to be 128

Note: Code and data exploration provided in UFOsightings_part1.Rmd

Task 3 - Plot number of sightings over time



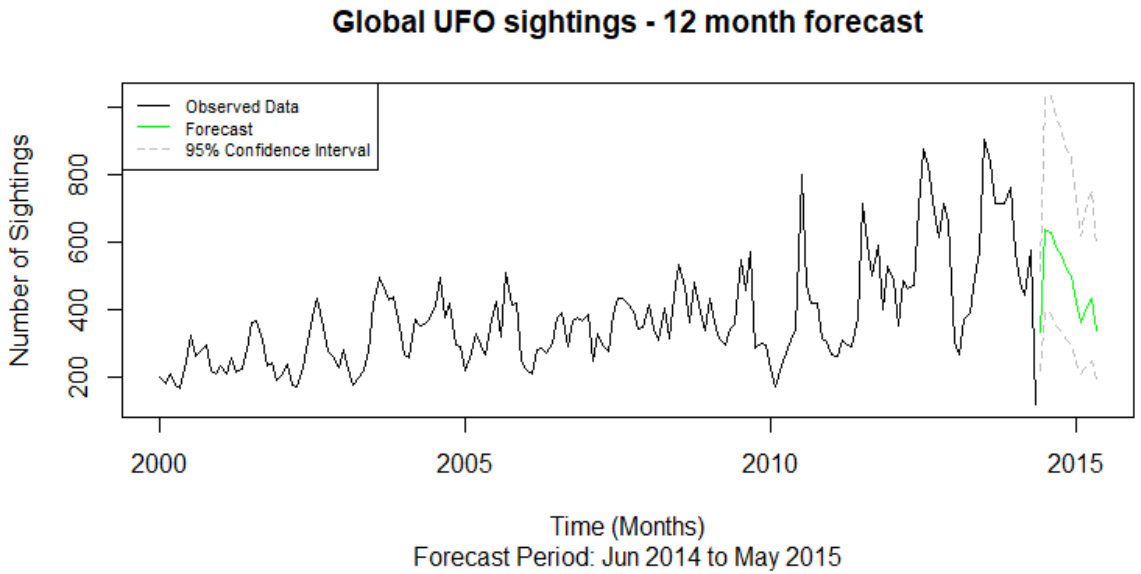
Note: Series includes months with no sightings. First, a vector of months was created. Sightings per month were aggregated using mySQL through R. Finally they were assigned to matching months.



Task 4 - The 10 most frequently used words in comments

- Indicates that most people saw bright lights in the sky – mostly orange or red in colour
- Others saw a moving shape in the sky
- Methodology:
  - Libraries used were `tm` and `qdap`
  - white spaces, punctuations, brackets and abbreviations were removed.
  - Words have been truncated to their roots. Hence, orange changed to orang but has been corrected on this slide.

Task 5: One year forecast of UFO sightings



Based on ARIMA(1,1,1)(1,1,1)[12] model. Forecast indicates median values.
Details provided on the next few slides. Code in UFOsightings_part2.Rmd

Task 5: One year forecast of UFO sightings (contd.)

Month	Forecast	95% Confidence Limit	
	Median	Lower	Upper
Jun-14	332	215	513
Jul-14	639	395	1032
Aug-14	630	383	1035
Sep-14	586	353	973
Oct-14	565	337	946
Nov-14	520	308	876
Dec-14	501	295	852
Jan-15	423	247	723
Feb-15	360	209	620
Mar-15	409	236	710
Apr-15	430	246	752
May-15	336	191	591

- The model explained 74% variability on test data
- The time series had an **upward trend** from mid-90s. Hence the data from January 2000 was used for accuracy.
- **Variability** is increasing with time hence it was log transformed. Consequently, the forecast is of the median values.
- **Seasonality** is quite conspicuous from the plot. pacf and differencing indicated seasonality of 1 year was significant.
- There also seems to be some **cyclicity** in the data.

The Model

```
Call:
arima(x = log_train, order = c(1, 1, 1), seasonal =
list(order = c(1, 1, 1),
    period = 12), include.mean = F)

Coefficients:
         ar1      ma1    sar1     sma1
      0.2080  -0.7829  0.0928  -0.9996
s.e.  0.1414   0.0972  0.0921   0.1949

sigma^2 estimated as 0.03141:  log likelihood =
30.14,  aic = -50.28
```

Methodology

Test and Train data

- Last 12 months of data were set aside as test data
- Data before that upto Jan 1995 and Jan 2000 were used as training data
- As mentioned before months with 0 sightings were included but they were all before 1995.

Transformation and differencing

- Log transformation was applied to reduce variability.
- Differencing of 12 and 1 months was applied to make it stationary.

ACF and PACF plots were compared to determine the time window of data to be used, differencing to be applied, terms and seasonality.

Lowest AIC and highest R2 on test data were used to determine the best model.

Output was back transformed to original scale. Hence, it reflects quantiles rather than the means.

Thank You