Institute and Faculty of Actuaries

Model Telematics Data and its Application

- An Update from Advanced Pricing Methods GIRO Working Party

Dr. Ji Yao [a,b]

The views and opinions expressed in this presentation are independent of employer of the presenter or the Actuarial Profession

[a.] RPC Consulting
[b.] University of Kent

Institute and Faculty of Actuaries

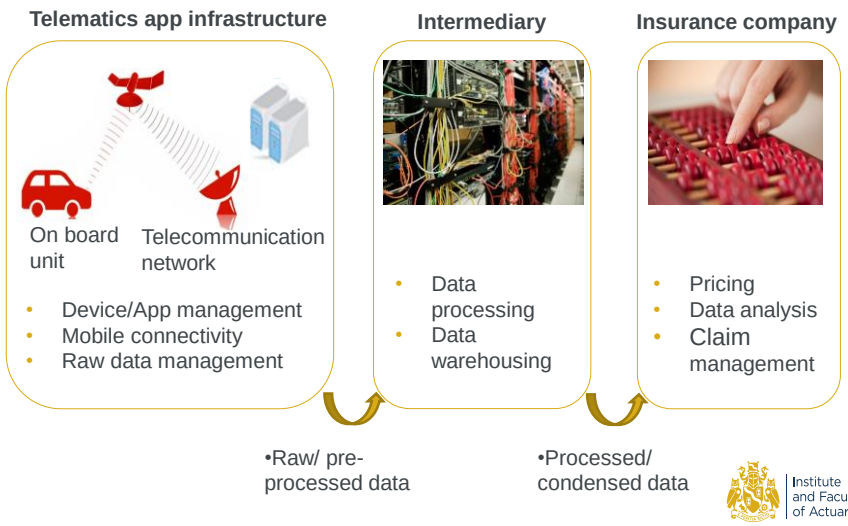
Agenda

- **Introduction**
- **Telematics Data**
- **Feature extraction**
- **Statistical techniques to model telematics data**
- **Summary and Q & A**

Introduction

- Advanced Pricing Methods (APM) GIRO WP was created in 2012
- Focus was on use of GLM in pricing during 2012-2014
  - Limitation of GLM
  - Possible solution: GLMM, credibility theory with GLM
- This year, focus is on telematics pricing
  - Telematics data
  - Feature extraction
  - Statistical models

The journey of telematics data



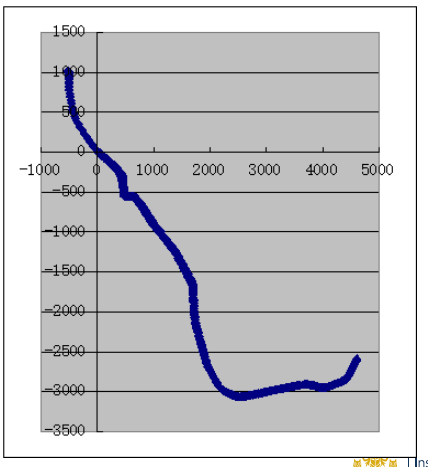
Institute and Faculty of Actuaries

Agenda

- **Introduction**
- **Telematics Data**
- **Feature extraction**
- **Statistical techniques to model telematics data**
- **Q & A**

Raw data is longitude and latitude (and/or altitude)

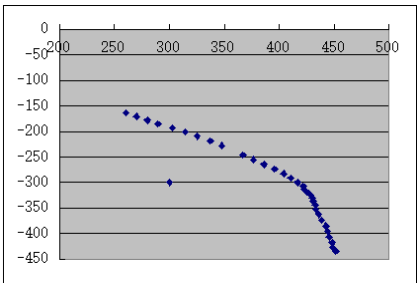
x	y
0	0
18.6	-11.1
36.1	-21.9
53.7	-32.6
70.1	-42.8
86.5	-52.6
101.7	-62.3
117	-71.6
131.2	-80.4
145.5	-88.7
159.7	-96.8
171.7	-104
182.5	-111
193.4	-117.7
202	-124
211.8	-130.4
221.5	-137
231.3	-143.6
239.8	-150.3
250.6	-157.1

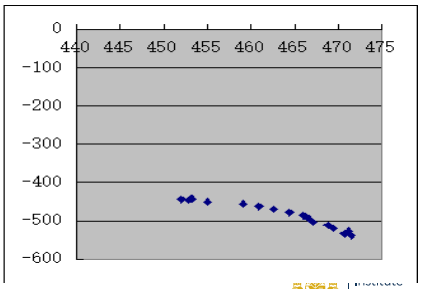


Institute and Faculty of Actuaries

Useful information could be derived from raw data, but...

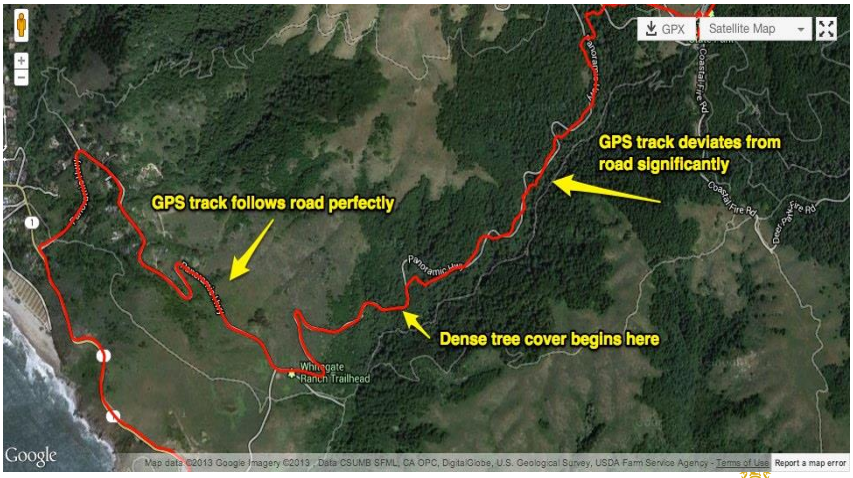
- Speed = s = ((x_t+1 - x_t)^2+(y_t+1 - y_t)^2)^0.5
- Acceleration = ((sx_t+1 - sx_t)^2+(sy_t+1 - sy_t)^2)^0.5
  - where sx_t+1 = x_t+1 - x_t and sy_t+1 = y_t+1 - y_t
- Journey length = sum of speed





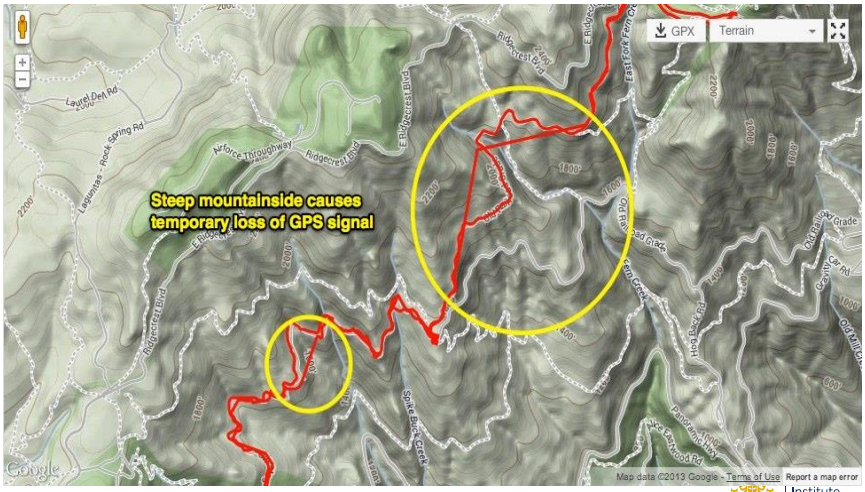
Institute and Faculty of Actuaries

Example of bad GPS data



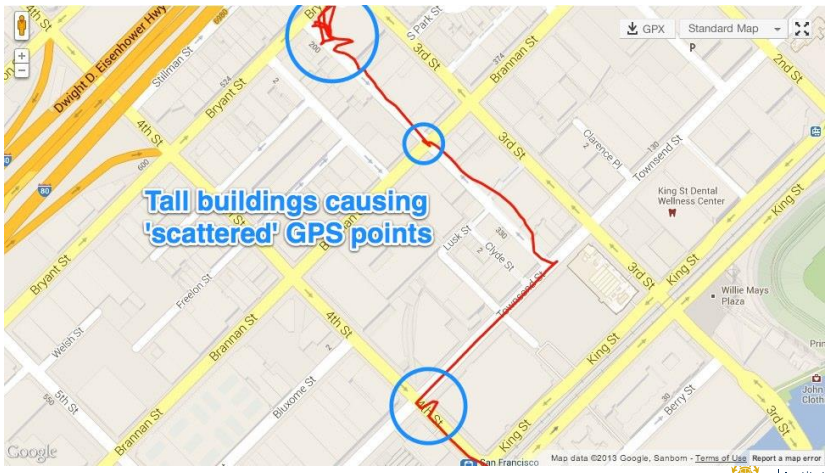
Institute and Faculty of Actuaries

Example of bad GPS data



Institute and Faculty of Actuaries

Example of bad GPS data



Institute and Faculty of Actuaries

GPS data requires cleansing and smoothing before processing

- **In particular, the quality of data is worse when there is an event**
- **Generally there is no unique 'correct' to do this**
- **Options are:**
  - **Delete suspicious data**
    - Exclude point where speed is too high
  - **Apply simply rules:**
    - Only re-calculate direction if speed is higher than 5m/s
    - Speed is capped
  - **Least square fit**
  - **Kalman filter**

Institute and Faculty of Actuaries

Agenda

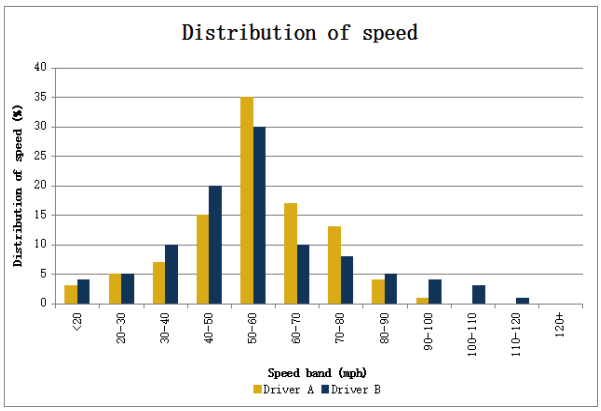
- **Introduction**
- **Telematics Data**
- **Feature extraction**
- **Statistical techniques to model telematics data**
- **Q & A**

Literatures show there are many driving behaviours possibly correlating with accident

- Mileage
- Speed
- Acceleration
- Deceleration
- Celeration
- Start behaviour
- Braking behaviour
- Behaviour at turn
- Behaviour at roundabout
- Sharp turn/U-turn

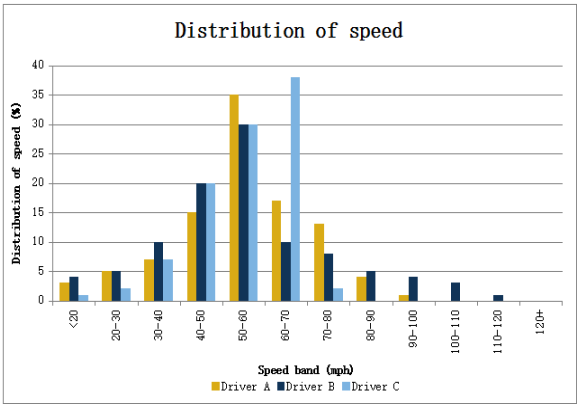
How to define and use them in a pricing is an art

There are many possible ways to define a feature



- Mean
- Standard deviation
- Skewness
- Extreme value

There are many possible ways to define a feature



- Middle value (median)
- Most frequency value (mode)
- Distribution band

Institute and Faculty of Actuaries

Layering context data can greatly enhance understanding of customer and driving behaviour

- To communicate with the business, the data could be displayed to a map
  - Any analysis produced can be applied back to a journey and mapped to aid understanding and communicate with others
- Other external data to overlaid
  - Type of road
  - Points of interest e.g School, pub, car park etc
  - Weather
  - Traffic density
  - Known road issues e.g. Potholes, broken traffic lights
- Some of the external data could be approximated by raw data (longtitude, latitiude)

Institute and Faculty of Actuaries

More features could be generated with interaction of context data

- Average speed on class of road
- Occurrences of speeding
- Occurrences of speeding while there is no traffic
- Occurrences of speeding in rain
- Purpose of journey
- And every conceivable permutations will need to be considered and/or modelled for consideration

- **Could generate c.200 features from the longitude and latitude data**

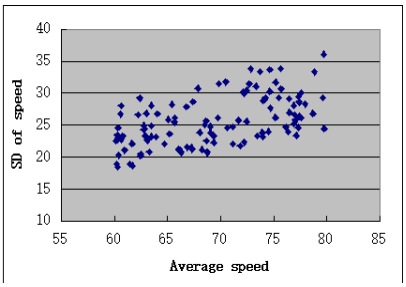
Institute and Faculty of Actuaries

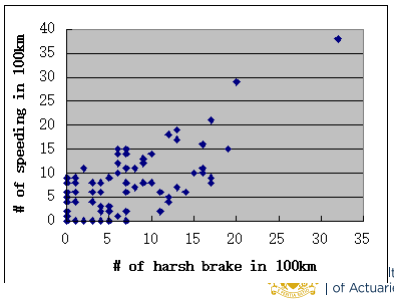
Agenda

- **Introduction**
- **Telematics Data**
- **Feature extraction**
- **Statistical techniques to model telematics data**
- **Q & A**

How to use telematics data in pricing

- Challenges are
  - Lack of attached claim experience
  - Large amount of correlated features





Variable clustering provides a possible approach to variable selection

- **Variable clustering is a method to detect subset of correlated variables. Variables which provide the same kind of information belong into the same group.**
- **More interpretable than PCA**

Variables
Average speed
Average acceleration
Average deceleration
Celeration
# of brake
# of speeding
Speed at right turn
Speed at left turn
Avg mileage per journey

Cluster	Variables
Cluster 1	Average speed # of speeding Speed at right turn Speed at left turn
Cluster 2	Average acceleration Average deceleration Celeration # of brake
Cluster 3	Avg mileage per journey

Variable clustering approach

- Define similaritiy
  - Correlation: Pearson correlation for continuous variable
  - Covert categorical variable to a set of 0/1 indicators

Journey	Type of road
1	Motorway
2	A road
3	A road
4	Motorway
5	Local minor road
6	B road

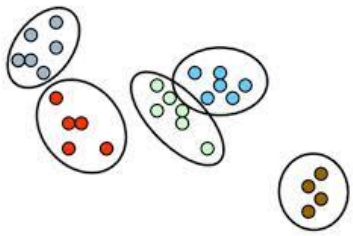
Motorway	A road	B road	Local minor road
1	0	0	0
0	1	0	0
0	1	0	0
1	0	0	0
0	0	0	1
0	0	1	0

  - Covariance is possible

Institute and Faculty of Actuaries

Variable clustering approach

- Apply clustering methods
  - Partitioniong
    - k-means
    - k-medoids
    - k-medians
    - k-means++
  - Hierarchical
    - Top-down appraoch
    - Bottom-up approach



Institute and Faculty of Actuaries

Algorithms are available in statistical packages

- **SAS: PROC VARCLUS**

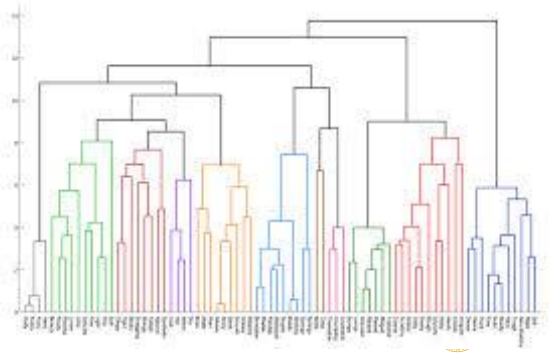
```
PROC VARCLUS < options > ;
   VAR variables ;
   SEED variables ;
   PARTIAL variables ;
   WEIGHT variables ;
   FREQ variables ;
   BY variables ;
```

- **R: ClustOfVar**

```
#mixture of quantitative and qualitative variables
data(wine)
X.quanti <- wine[,c(3:29)]
X.quali <- wine[,c(1,2)]
tree <- hclustvar(X.quanti,X.quali)
plot(tree)
```


Result of variable clustering is usually presented in a dendrogram

- **Number of cluster are judgementally chosen. Some criteria are**
  - **At a clear cut-off thredshed**
  - **Business requirement/contraint**
  - **Expert knowledge**
- **Fianlly, variables are chosen from each cluster**
  - Most representative of the cluster
    - Most orrelated with cluster component (cluster mean)
  - Most correlated with predicted variable



Summary

- The method of process of telematics data for pricing is discussed
- Use of telematics data for pricing is a multi-disciplinary topic that requires skills and knowledge of mathematics, physics, data mining, computer science, signal process, transportation, driving behaviour etc.

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter.