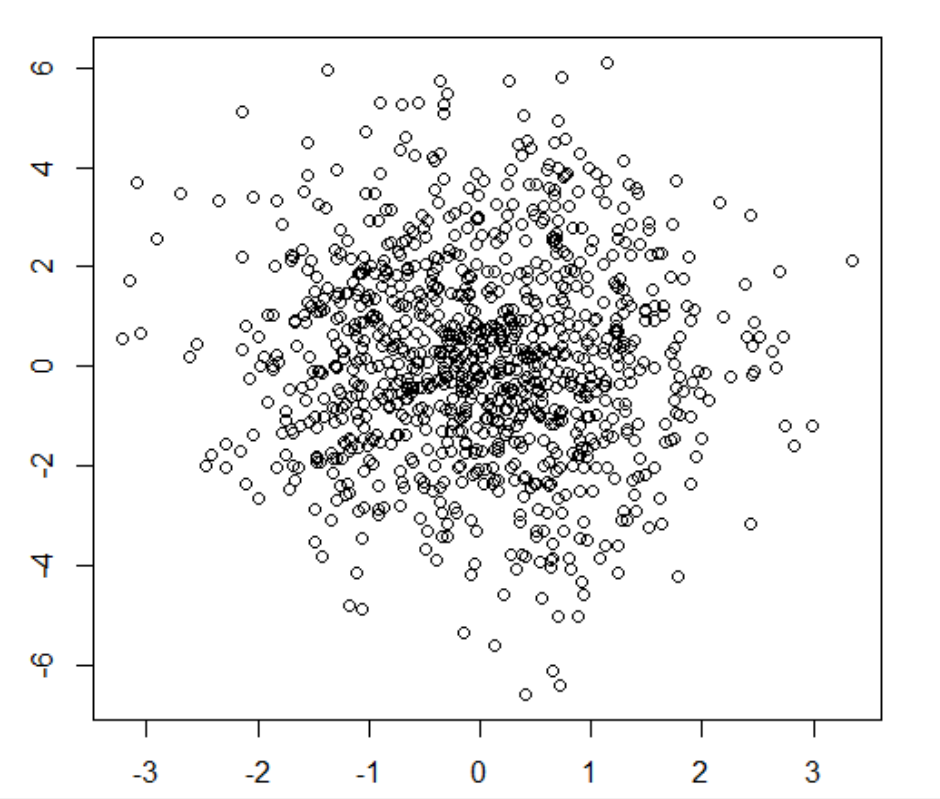
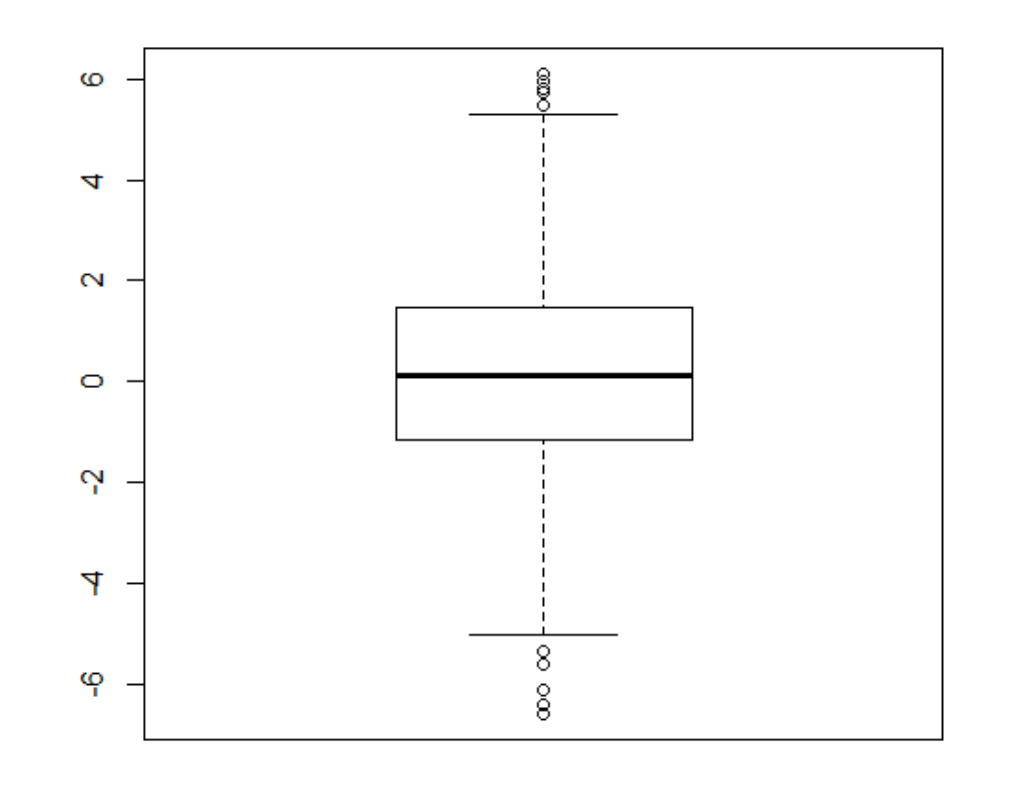
Privacy protected data visualisations in DataSHIELD

Demetris Avraam

Graphs are very informative but often very disclosive





Graphical functions in DataSHIELD

Released:

- ds.histogram
- ds.heatmapPlot
- ds.contourPlot

Developed:

- ds.scatterPlot
- ds.boxPlot

K-Anonymisation

A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.

There are two common methods for achieving *k*-anonymity*:* **suppression** and **generalisation**.

Low counts cell suppression: in tabular data, cells with small numbers (lower than a pre-specified threshold) are not displayed:

Actual 2D contingency table

		Gender		
		0	1	Total
Dis.	0	525	327	852
	1	3	18	21
	Total	528	345	873

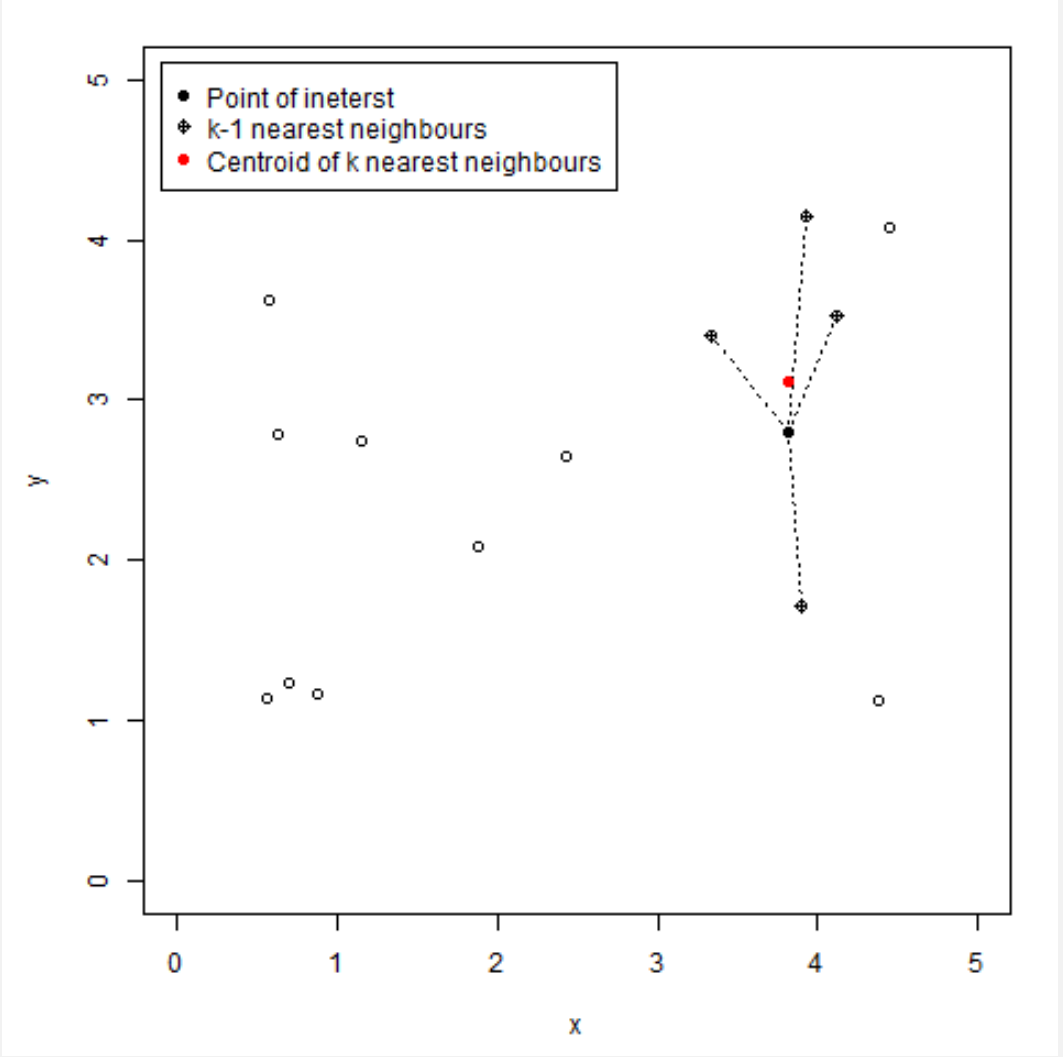
DataSHIELD output (with $Th = 5$)

		Gender		
		0	1	Total
Dis.	0	NA	NA	852
	1	NA	NA	21
	Total	528	345	873

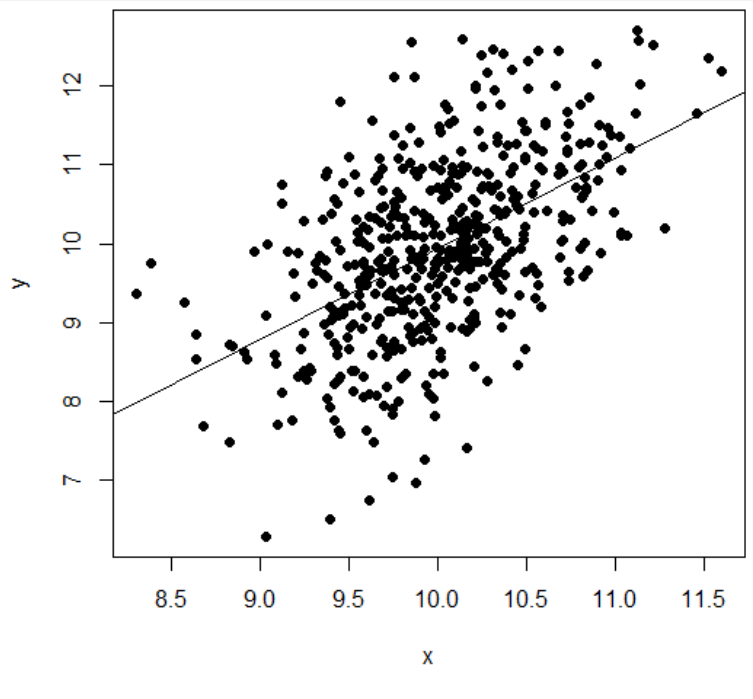
Deterministic Anonymisation

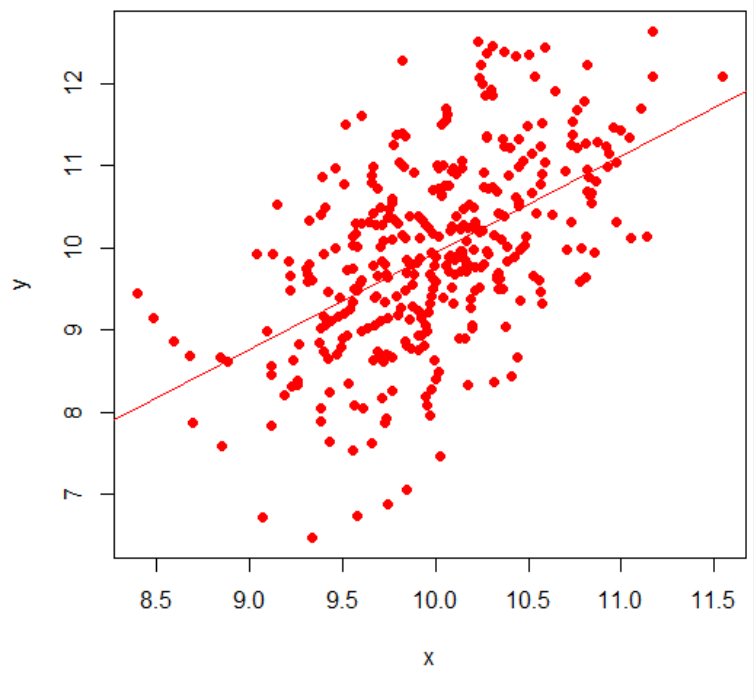
Algorithm:

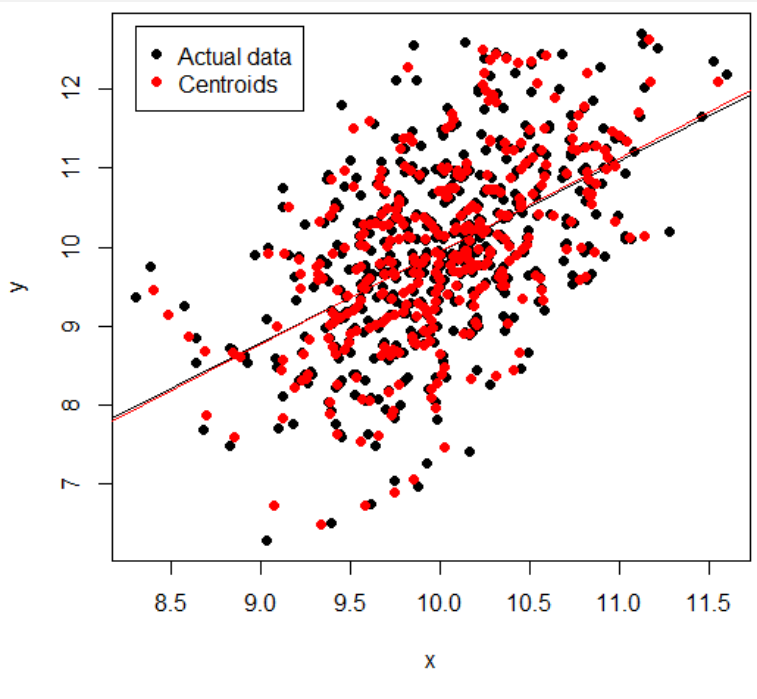
- Find the $k-1$ nearest neighbours of each data point
- Calculate the coordinates of the centroids of each k nearest neighbours
- Replace actual data with the centroids
- Apply scaling by multiplying the coordinates of the centroids with a scaling factor



Non-disclosive scatter plots





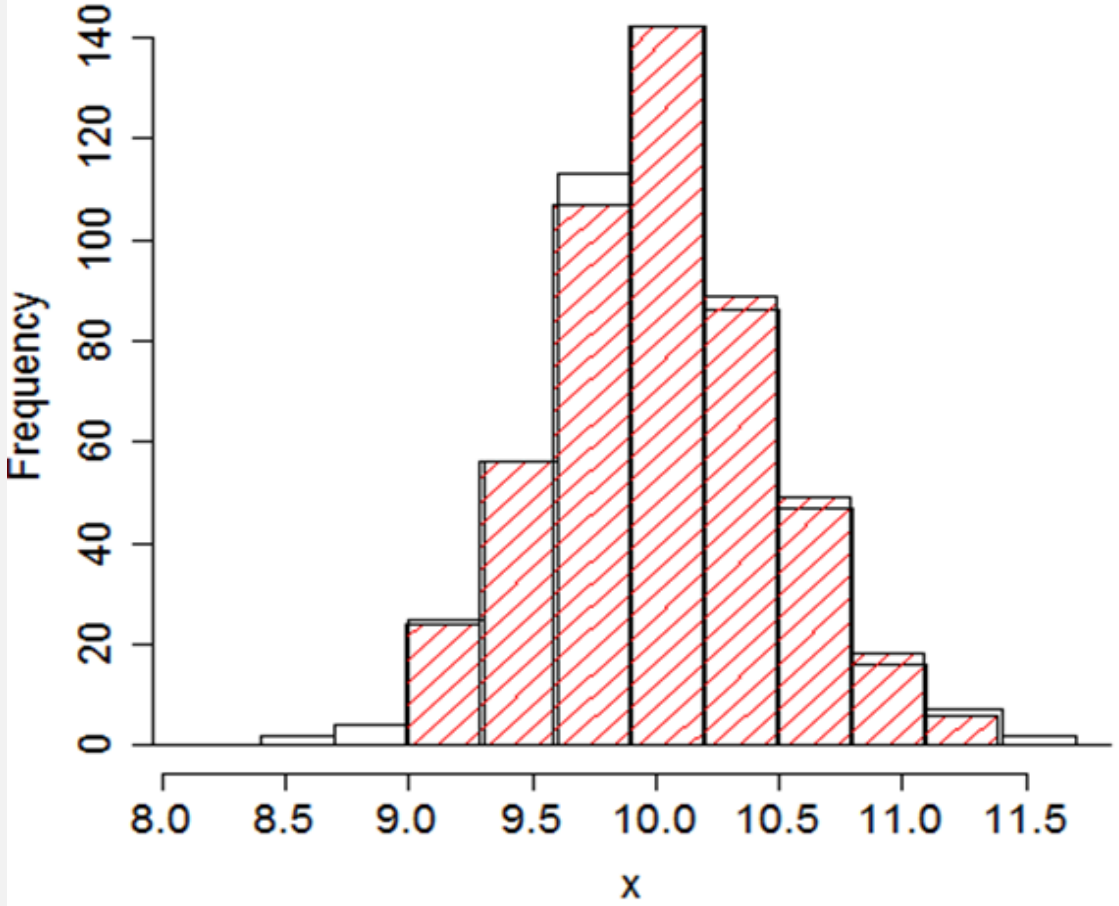


Advantages:

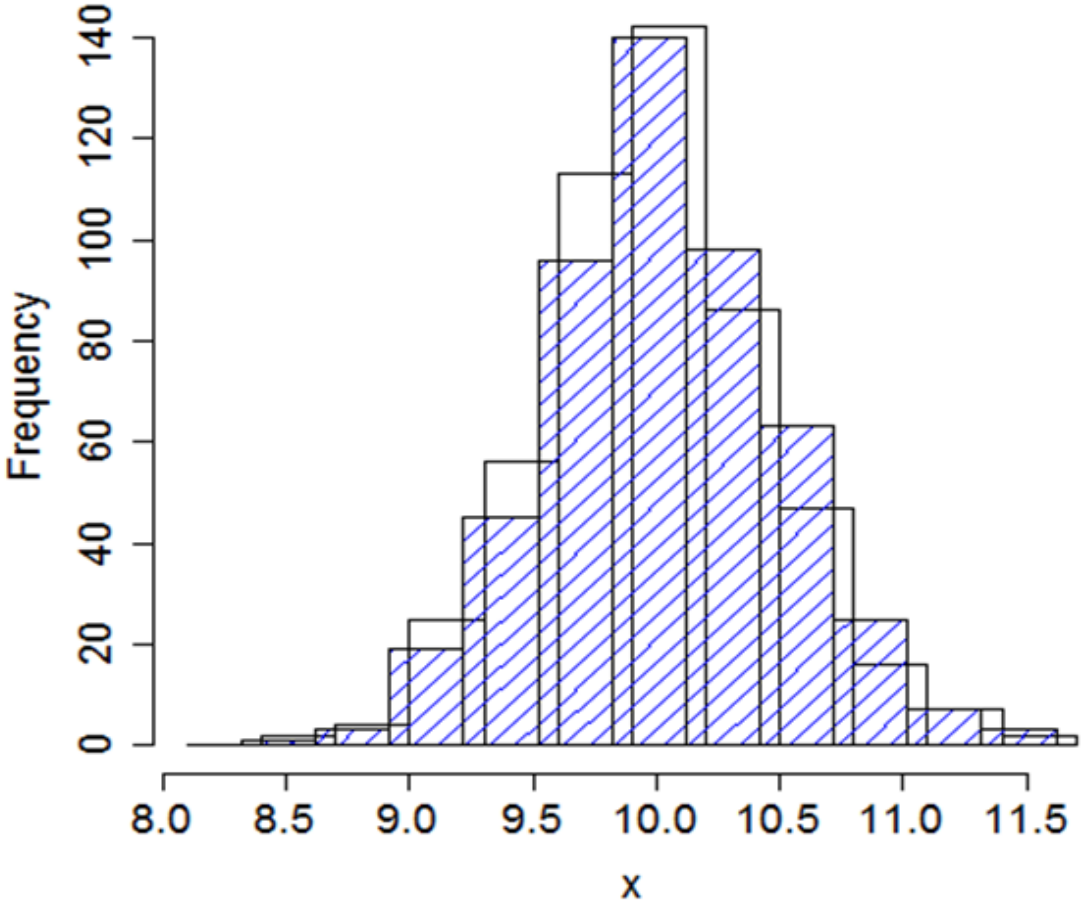
- The method is deterministic
- While k increases the disclosure risk decreases
- The information loss is minimum

Non-disclosive histograms

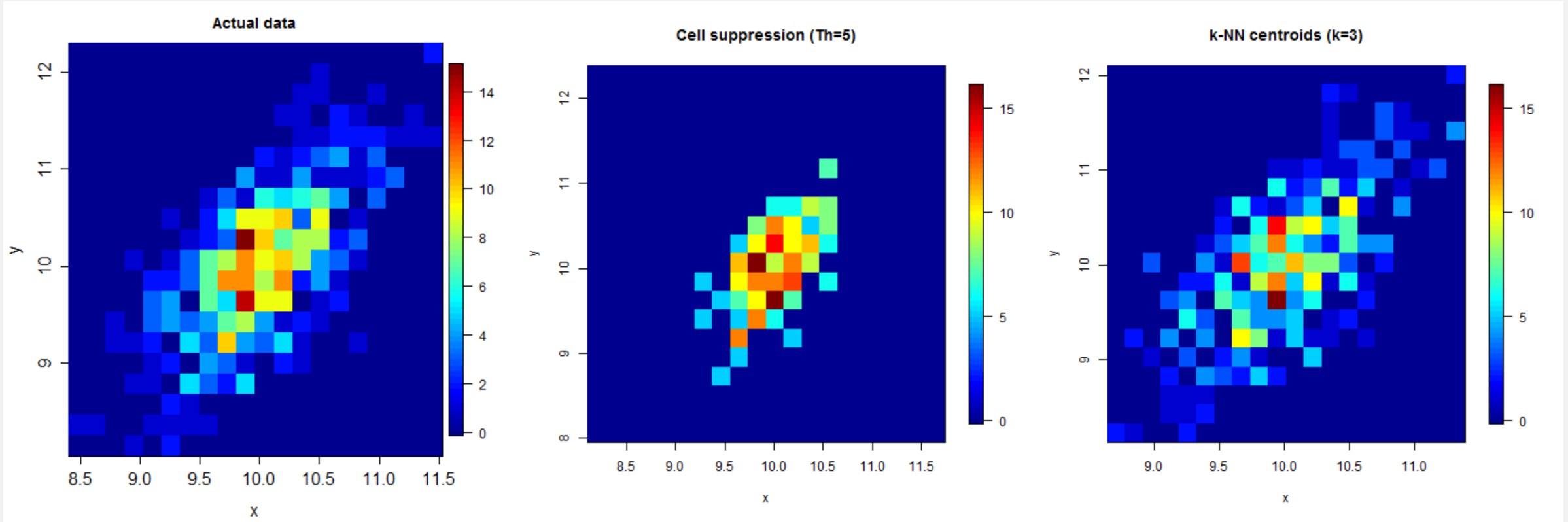
Cell suppression ($Th = 5$)



k-NN centroids ($k = 3$)

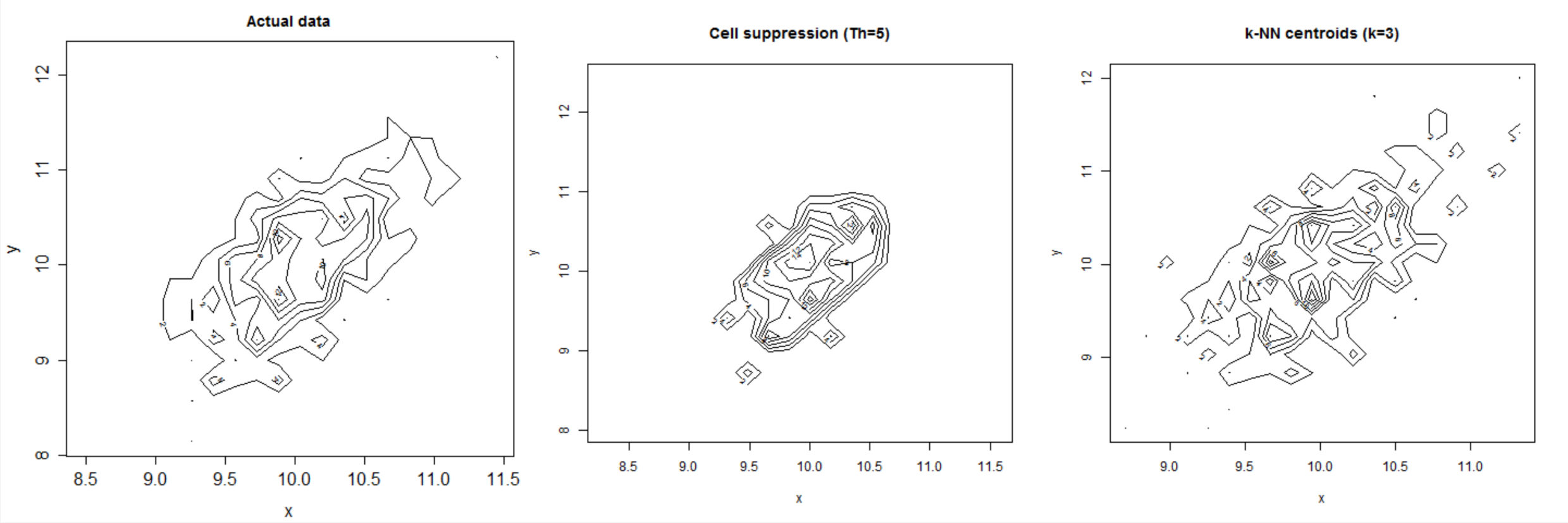


Non-disclosive heat map plots



The counts in 88 cells were > 0 and < 5 and they replaced by 0

Non-disclosive heat map plots



The counts in 88 cells were > 0 and < 5 and they replaced by 0

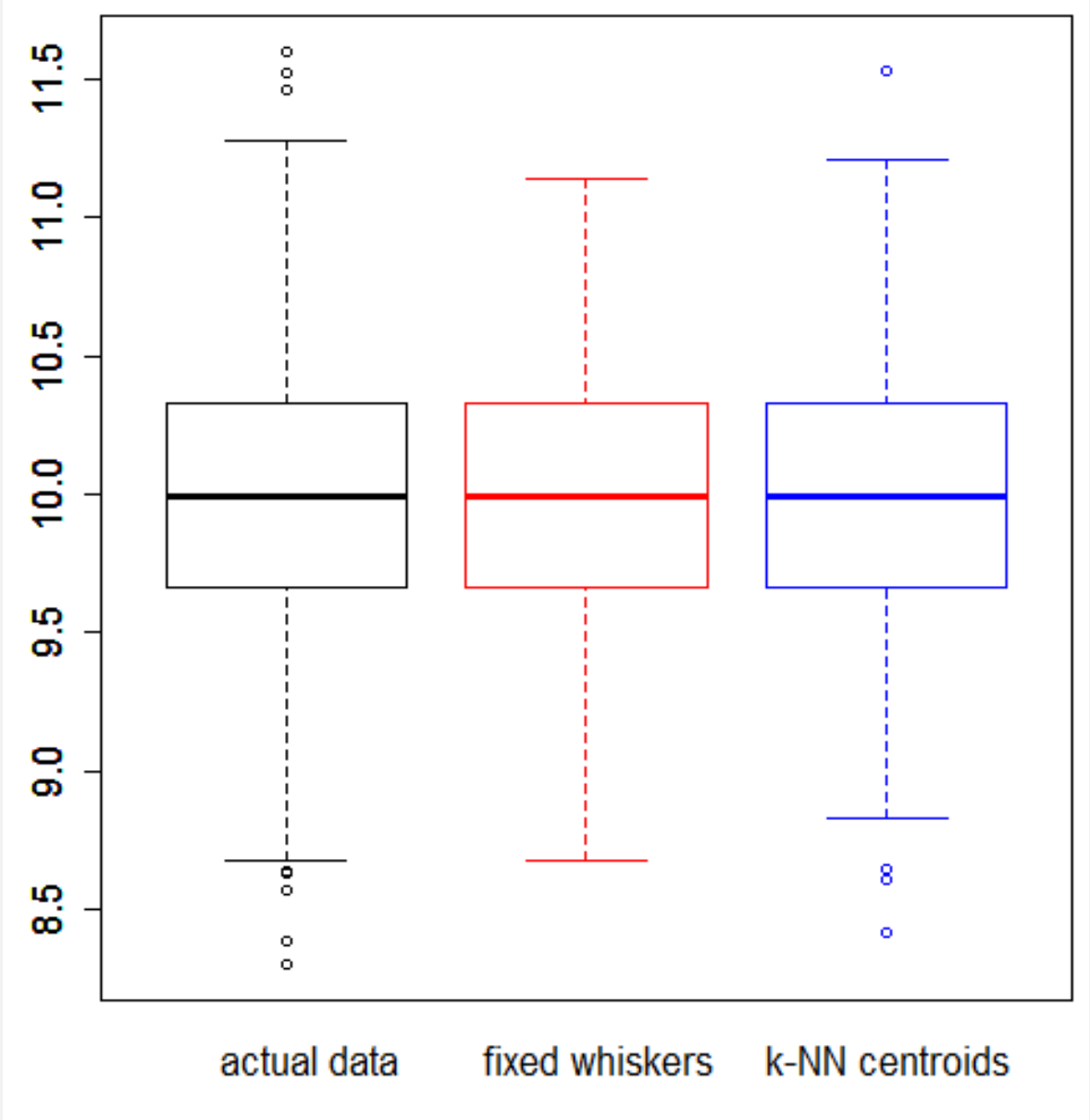
Prototyping non-disclosive box plots

Fixed extreme of the whiskers:

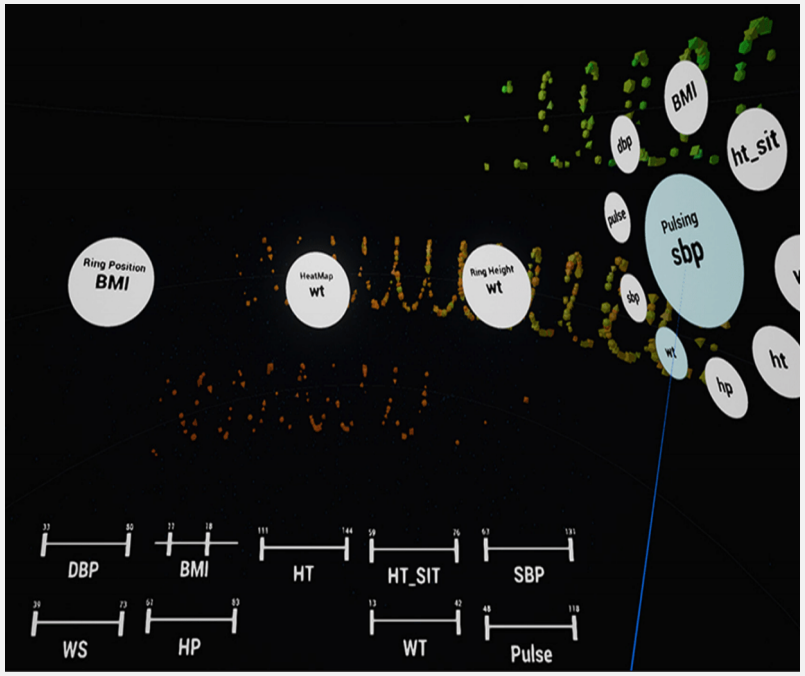
- Upper whisker = 95th percentile
- Lower whisker = 5th percentile
- The outliers are not displayed in the graph but their number is returned to the analyst

k-NN ($k = 3$):

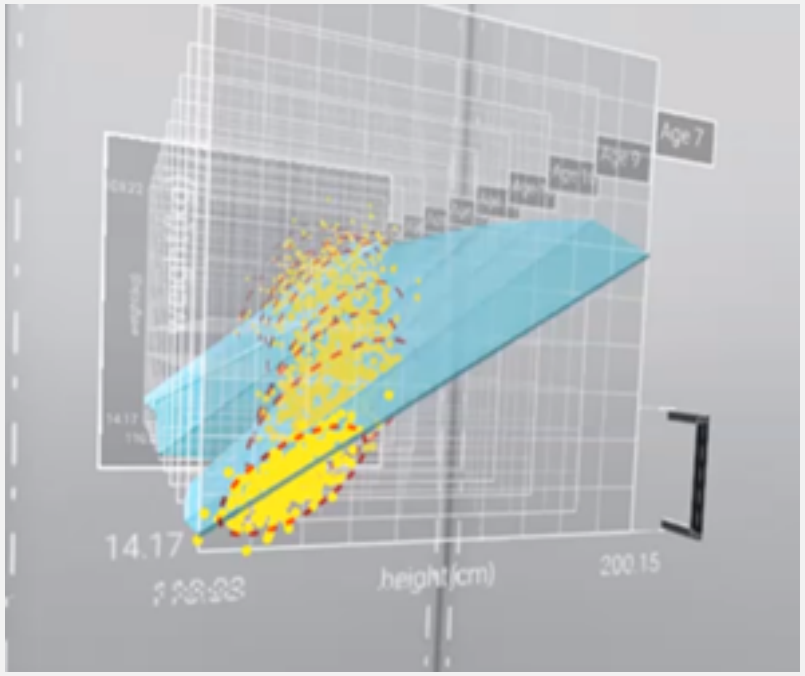
- Upper whisker = $Q_3 + 1.5\ IQR$
- Lower whisker = $Q_1 - 1.5\ IQR$
- The outliers are displayed in the graph



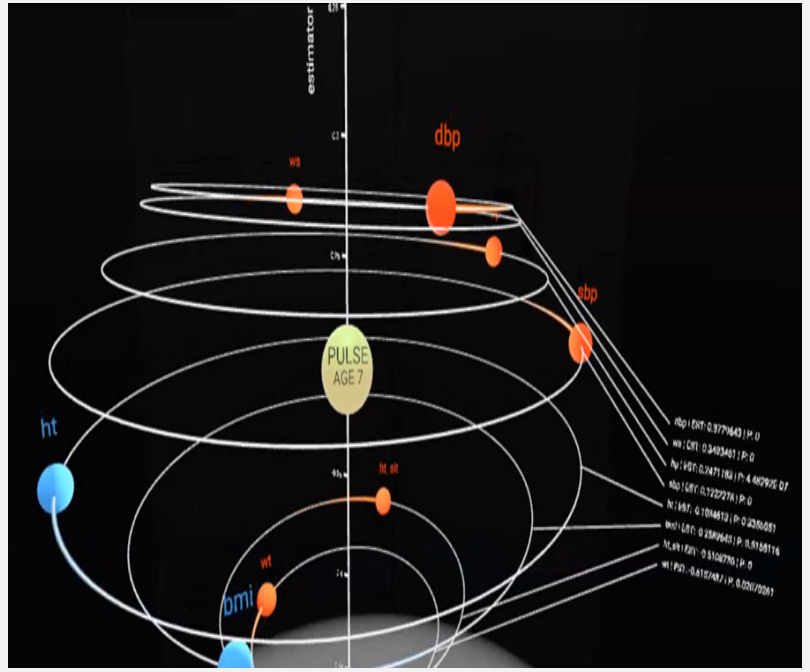
Exploring longitudinal data in Virtual Reality



Helix visualisation

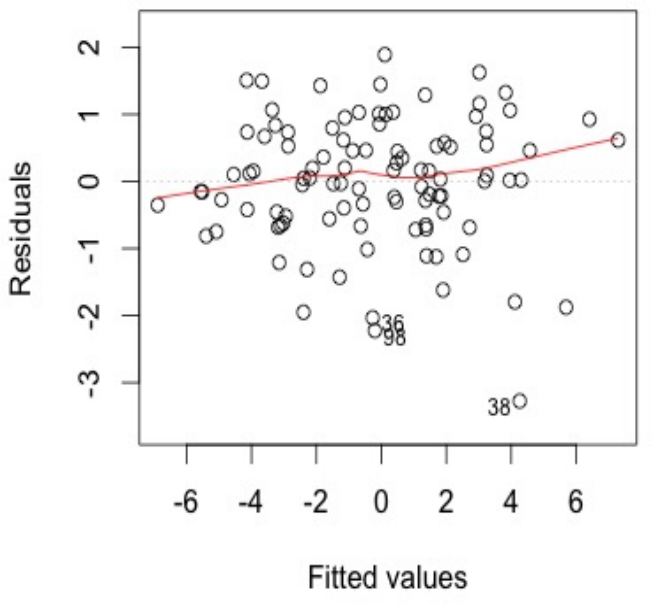


Ribbon-datashape visualisation

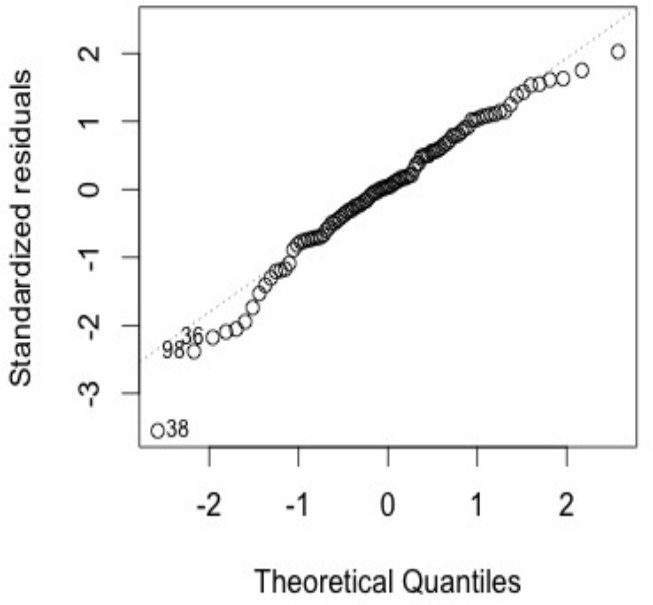


Orbit visualisation

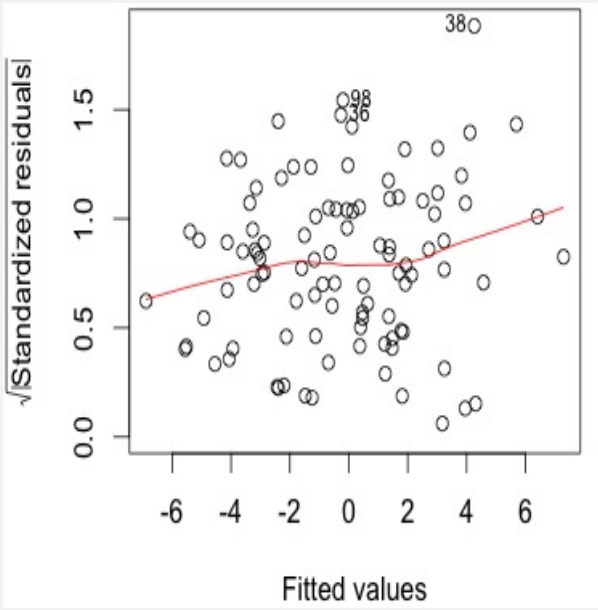
Plot diagnostics for regression models



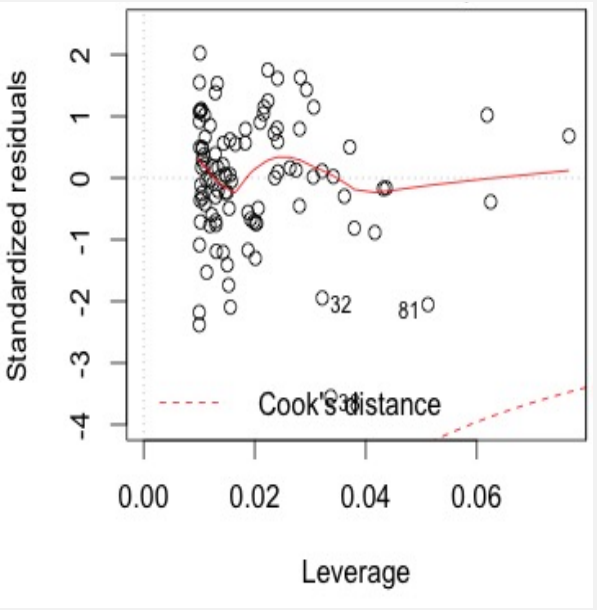
Residuals vs Fitted



Normal Q-Q



Scale-Location



Residuals vs Leverage

Anonymising and synthesising data

- Methods will include:
  - ‘weak’ and ‘strong’ k-anonymization (at different k thresholds)
  - probabilistic anonymization through addition of random noise (at different levels of variance)
  - deterministic anonymization through replacement of true records with the centroids of each k nearest neighbours (at different values of k)
  - simulating synthetic data using existing algorithms

- Disclosure Risk metrics
- Information Loss metrics

Thank you!