# “On repairing certain big data sets using KNN"

Shaila Sharmeen, PhD Intern, Australian Bureau of Statistics

Under the Supervision of Dr Siu- Ming Tam,

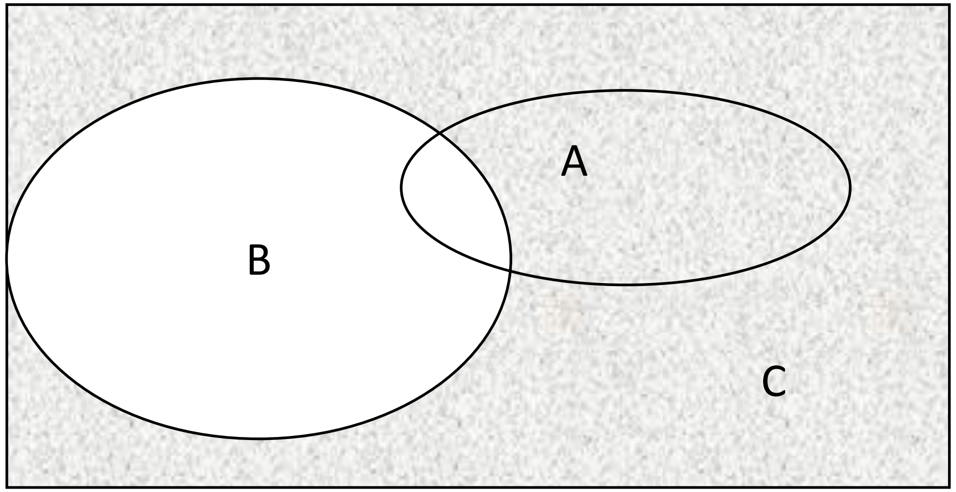
Ex-Chief Methodologist, Australian Bureau of Statistics

Honorary Professorial Fellow, University of Wollongong, Australia

# Motivation

- To address the representativeness of Big Data by using mass imputation
- To improve small area estimation using Big data
- To allow for optimum survey estimation in multi-purpose surveys by
  - applying the machine learning Algorithm for the prediction (imputation)
  - by minimising the prediction error and
  - by keeping the relationship between response variables intact

# Problem Description




- let, U is the finite population
- B is the big data sample
- A is the survey data sample
- C is the set in U, but not in B
- Our aims are to:
  - predict the data points in set C in such a way that the totals of predicted values will match some pre-determined control total eg the Regression Data Integration (RDI) total as described in the "Mining for the New Oil for Official Statistics" paper by Siu-Ming Tam; and
  - maintain the relationship between the response variables.
- For this project, we assume B has no measurement errors, when compared with A
- We illustrate our methods using a simulated data set

# Simulating the Data

- 1000 data points in U with 6 auxiliary variables X and 6 Response variables Y as outlined below:

| Auxiliary variables |
|---|
| x1 ~ uniform (0,1) |
| x2 ~ normal (u2, var(x2)) |
| x3 ~ Bernoulli(p=0.5) |
| x4 ~ Bernoulli (p=0.25) |
| x5 ~ uniform (0,1) |
| x6 "geographic identifier" : split x1 into quartiles to create 4 "geographic areas" |

| Name of the Response variable | Description | Nature |
|---|---|---|
| y1 | Normal distribution | Continuous |
| y2 | More complex regression | Continuous |
| y3 | More complex regression with positive skew | Continuous |
| y4 | A Bernoulli Trail with probability of Success related to geographical area(i.e: whether pineapples are grown indicator) | Categorical |
| y5 | A Mixture model- (i.e: if the farm grows pineapples – the amount is given by a normal distribution) | Continuous |
| y6 | Coin Toss- | Categorical |

***Source: Feasibility Simulation Study of regression Data Integration and Constrained Imputation, Susan Shaw, Susan Fletcher, December 2019, ABS***

# Simulating the Data

- Set A: Sample A or the 'survey' - fixed at 25% of population size (250 data points)
- Set B: **not missing at random**, a true bias scenario →607 data points

| No units to sample | x1<0.5 | x1>=0.5 |
|---|---|---|
| x2<u2 | 60%<br><br>(156 data points) | 80%, if y1>mean y1, twice as likely to be sampled<br>Suppose, X data points satisfied this condition, we will take 80% of them—S<br>So, n1+n2=S and<br>n1/N1= 2. n2/N2<br>N1= no.of data points where y1>mean Y1<br>N2= number of data points where y1<= mean Y1<br><br>(199 data points)(n1=25,n2=174) |
| x2>=u2 | 70%<br><br>(158 data points) | 35% if y1>mean y1, twice as likely to be sampled<br><br>(94 data points)(n1=12,n2=82) |

***Source: Feasibility Simulation Study of regression Data Integration and Constrained Imputation, Susan Shaw, Susan Fletcher, December 2019, ABS***

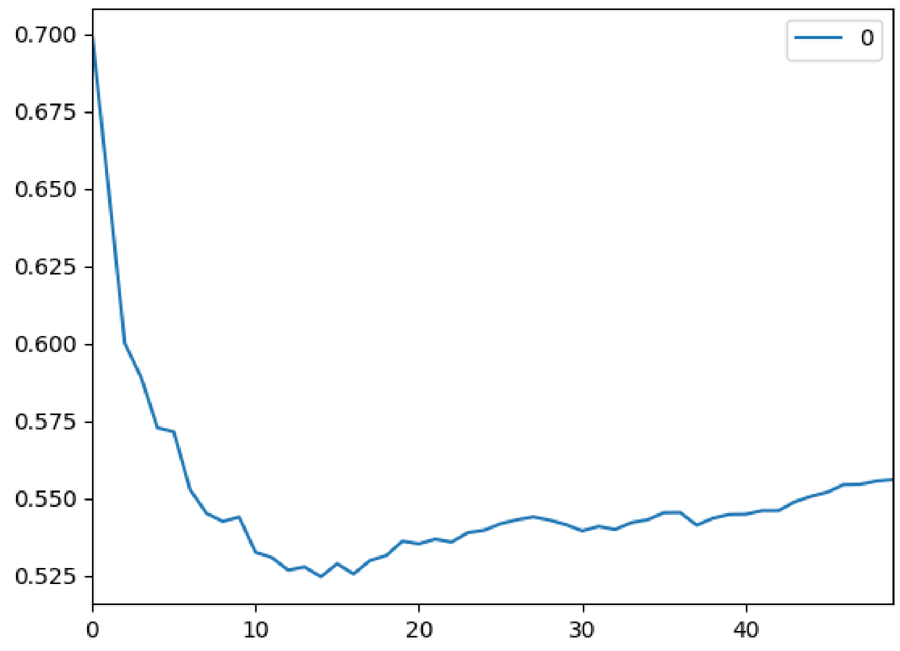
# KNN Algorithm steps

- We will use the KNN algorithm to predict the missing data points so that the control total and the predicted total will remain the same.
- Find K such that the prediction error for the 6 response variables as a set is smallest.
- Divide A into training and test data set
- Test performance of K based on accuracy of prediction measures
  - After applying feature selection where needed
- We use:
  - the HasD distance metric to find the NN
  - RMSE, f1 score for the accuracy metric for continuous and categorical data respectively.

# Optimum K Determination for continuous variables for individual response variables

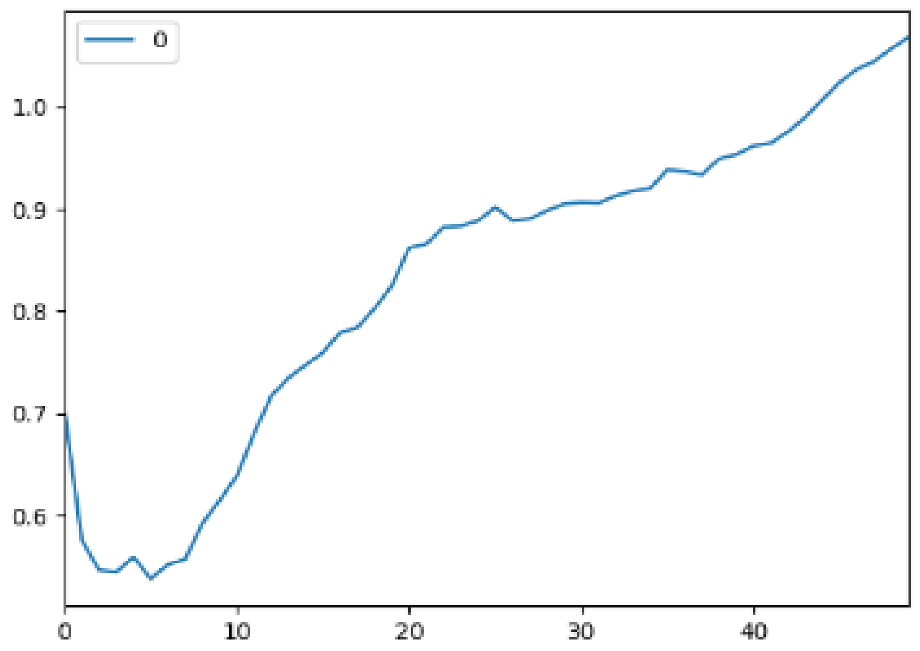
**Y1 : Optimum k: 15**

**RMSE :0.52455**




**Y2 : Optimum k: 6**

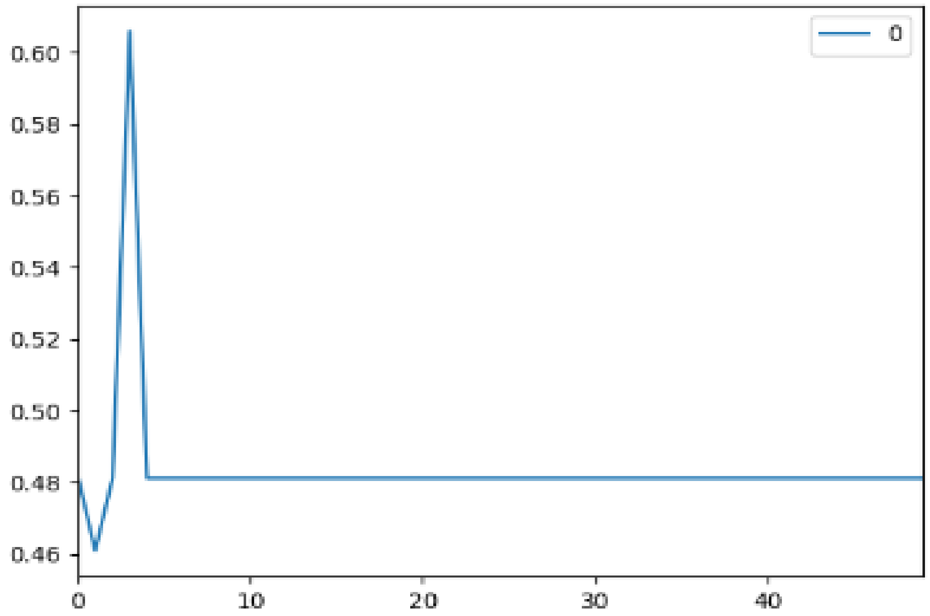
**RMSE :0.45977**



# Optimum K (Categorical Variables)

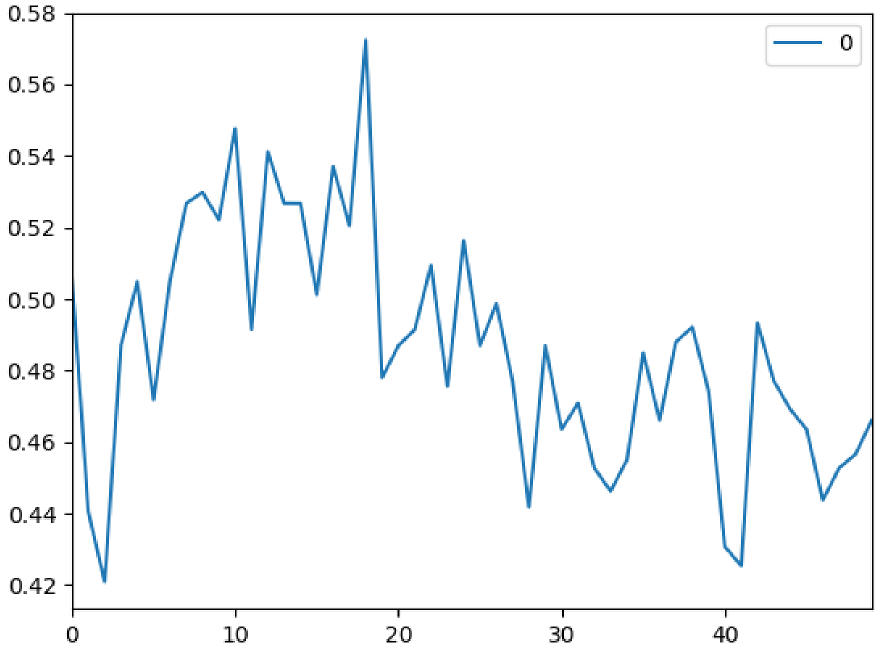
**Y4 :Optimum k=5**

**F1 Score 0.60**



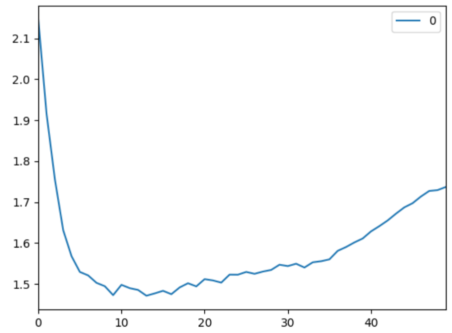

**Y6: Optimum k=19**

**F1 Score 0.57242**

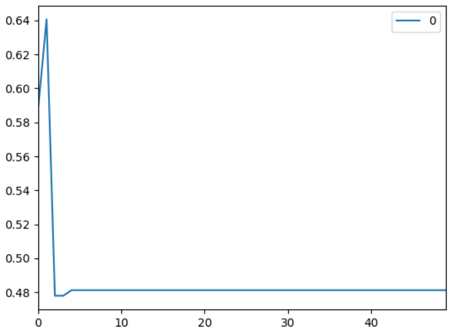


# Feature Selection

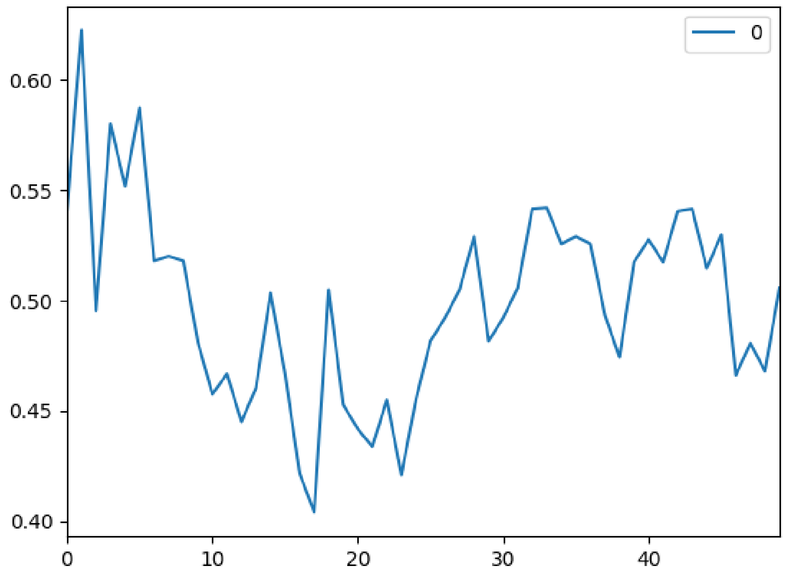
- Optimisation curves for Y3, Y4 and Y6 are not entirely satisfactory so feature selection was carried out
- Different set of features to obtain lowest RMSE or Highest f1 score
- Graphs are generated for these different set of features
- For Y3→ x1,x2,x3,x4,x6( RMSE- 1.47)
- For Y4→x2,x3,x4,x5(f1 score -0.64)
- For Y6→x1,x5 (f1 score -0.62)



Y3



Y4




Y6

# Why do we need Optimum K for the set of 6 response variables?

- Want same donors for all 6 response variables, so the relationship between the response variables are maintained – I call this “all in” donation
- We cannot have different K’s for different response variables, given the “all in” donation condition. Need to find one K that gives the highest prediction accuracy for all 6 variables – akin to local maxima and global maxima
- Assess the global prediction accuracy with the range of Ks for the local maxima
- Rescaled Error (RE)= $\frac{\sqrt{(y_m-\bar{y})^2}/\sqrt{n}}{\bar{y}}$ where $y_m$=$y_i - \widehat{y_i}$
- Loss from not using local optima = |kth rescaled RMSE- optimum rescaled RMSE|
- Total loss for the set of 6 response variables = y1 rescaled RMSE diff+ y2 rescaled RMSE diff + y3 rescaled RMSE diff - y4 f1 score diff+ y5 rescaled RMSE diff – y6 f1 score diff

# **Optimum K Determination:** Local loss = abs (kth rescaled RMSE- optimum rescaled RMSE)

| Variable Name RMSE/f1 score | K=6 diff | K=7 | K=8 | K=9 | K=10 | K=11 | K=12 | K=13 | K=14 | K=15 | K=16 | K=17 | K=18 | K=19 | K=20 | K=21 | K=22 | K=23 | K=24 | K=25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Y1 0.52455** | 0.0234 | 0.0141 | 0.010435 | 0.009125 | 0.009559 | 0.003945 | 0.003013 | 0.000917 | 0.001581 | 0 | 0.002162 | 0.000478 | 0.002544 | 0.003365 | 0.005675 | 0.005219 | 0.0057 | 0.0057 | 0.007228 | 0.007601 |
| **Y2 0.45977** | 0 | 0.002428 | 0.004838 | 0.012418 | 0.007474 | 0.008656 | 0.007958 | 0.010831 | 0.011247 | 0.020698 | 0.023038 | 0.031286 | 0.037003 | 0.043168 | 0.05259 | 0.057217 | 0.061105 | 0.070623 | 0.078454 | 0.083828 |
| **Y3 1.47136** | 0.00519 | 0.0045 | 0.00295 | 0.00207 | 0.000226 | 0.002746 | 0.001936 | 0.0014 | 0 | 0.000437 | 0.00231 | 0.003542 | 0.002313 | 0.003542 | 0.002567 | 0.004655 | 0.004166 | 0.003163 | 0.005267 | 0.005088 |
| **Y4 0.64069** | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 | 0.15944 |
| **Y5 1.73107** | 0.099439 | 0.155879 | 0.154329 | 0.130851 | 0.118109 | 0.108689 | 0.146031 | 0.171081 | 0.171923 | 0.16405 | 0.116489 | 0.130784 | 0.130778 | 0.080811 | 0.081093 | 0.040994 | 0.04246 | 0.011589 | 0.024759 | 0 |
| **Y6 0.62272** | 0.03529 | 0.10471 | 0.10263 | 0104718 | 0.14147 | 0.16520 | 0.15596 | 0.17766 | 0.16231 | 0.11929 | 0.15596 | 0.20111 | 0.21861 | 0.11784 | 0.16997 | 0.18090 | 0.18898 | 0.16773 | 0.20179 | 0.16773 |

# Optimum K Determination:

| | K=6 diff | K=7 | K=8 | K=9 | K=10 | K=11 | K=12 | K=13 | K=14 | K=15 | K=16 | K=17 | K=18 | K=19 | K=20 | K=21 | K=22 | K=23 | K=24 | K=25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Loss | -0.03134 | 0.017562 | 0.013112 | -0.00498 | -0.02407 | -0.0354 | -0.0005 | 0.024789 | 0.025311 | 0.025745 | -0.01544 | 0.00665 | 0.013198 | -0.02855 | -0.01751 | -0.05135 | -0.04601 | -0.06836 | -0.04373 | -0.06292 |

Total= y1 rescaled RMSE diff+ y2 rescaled RMSE diff + y3 rescaled RMSE diff - y4 f1 score diff+ y5 rescaled RMSE diff – y6 f1 score diff

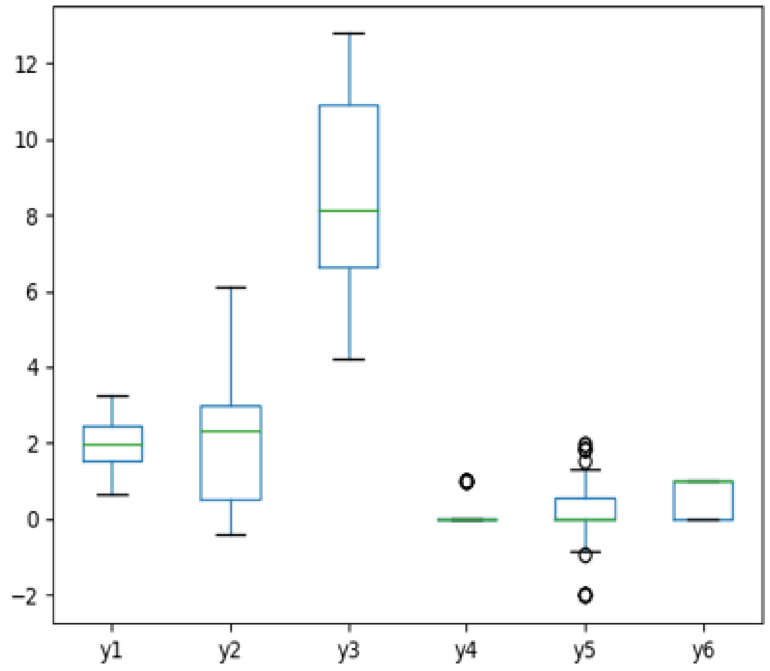
K=12 → Lowest Total

# We also use a procedure to align the predicted total to the RDI total, by using a weighted sum of NNs

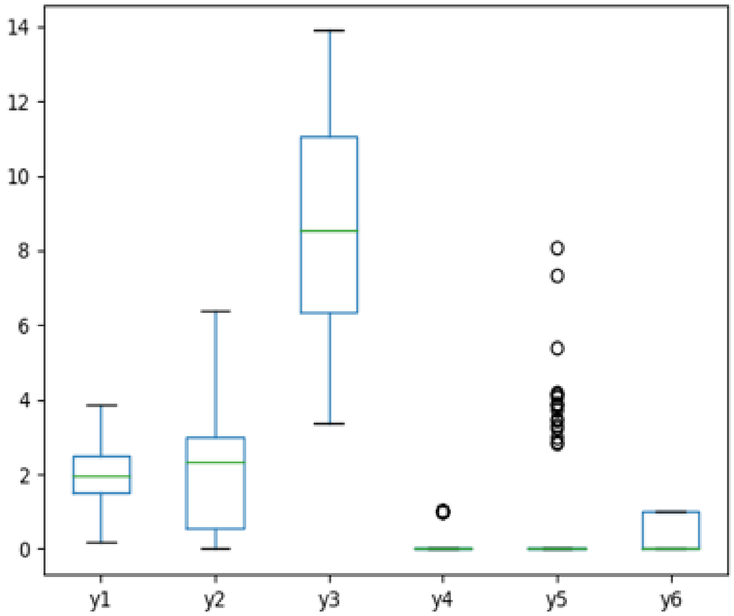
| | Old prediction using mean | New Prediction using wi |
|---|---|---|
| Difference of Y1 total(actual total-predicted total) | -3.73625431 | -6.02540240e-12 |
| Difference of Y2 total(actual total-predicted total) | 23.95503429 | -7.04858394e-12 |
| Difference of Y3 total(actual total-predicted total) | -1.72088161 | -4.04725142e-11 |
| Difference of Y4 total(actual total-predicted total) | 4.83333333 | -1.45661261e-13 |
| Difference of Y5 total(actual total-predicted total) | 13.27609791 | -6.25277607e-13 |
| Difference of Y6 total(actual total-predicted total) | -14.58333333 | -1.13686838e-12 |

# Box Plots

**RDI KNN predictors**




**Original data points**



# Thank You All.