

Deterministic Anonymisation and Privacy Protected Data Visualisations

Demetris Avraam, Paul Burton

Data to Knowledge (D2K) Research Group

School of Social and Community Medicine

University of Bristol



Software
Sustainability
Institute



Deterministic Anonymisation



The challenge: ~~Anonymise the data to prevent re-identification~~

Var1	Var2	Var3	Var4
1	1	11.7165	7.380501
0	1	12.02422	8.125242
1	1	7.70104	7.327802
0	1	7.813473	8.126139
1	0	7.742192	8.266226
0	1	6.439849	7.960921
1	0	8.091229	7.853049
1	1	6.006454	7.513198
1	1	6.894984	7.167675
0	0	10.25784	8.636956
1	0	13.83798	8.016772
0	1	9.558858	8.623355
1	0	7.955962	7.066784
1	0	6.355218	8.12677
0	1	6.651313	8.504875
0	0	19.66334	7.980293
1	1	10.53635	7.856456
1	0	8.037256	7.41357
0	0	8.237809	9.543621
0	1	11.83836	8.562876
1	0	7.225119	9.368156

Deterministic Anonymisation Algorithm: Step 1 - Stratification

Var1	Var2	Var3	Var4
1	1	11.7165	7.380501
0	1	12.02422	8.125242
1	1	7.70104	7.327802
0	1	7.813473	8.126139
1	0	7.742192	8.266226
0	1	6.439849	7.960921
1	0	8.091229	7.853049
1	1	6.006454	7.513198
1	1	6.894984	7.167675
0	0	10.25784	8.636956
1	0	13.83798	8.016772
0	1	9.558858	8.623355
1	0	7.955962	7.066784
1	0	6.355218	8.12677
0	1	6.651313	8.504875
0	0	19.66334	7.980293
1	1	10.53635	7.856456
1	0	8.037256	7.41357
0	0	8.237809	9.543621
0	1	11.83836	8.562876
1	0	7.225119	9.368156

Deterministic Anonymisation Algorithm: Step 1 - Stratification

Var1	Var2	Var3	Var4
0	0	10.25784	8.636956
0	0	19.66334	7.980293
0	0	8.237809	9.543621

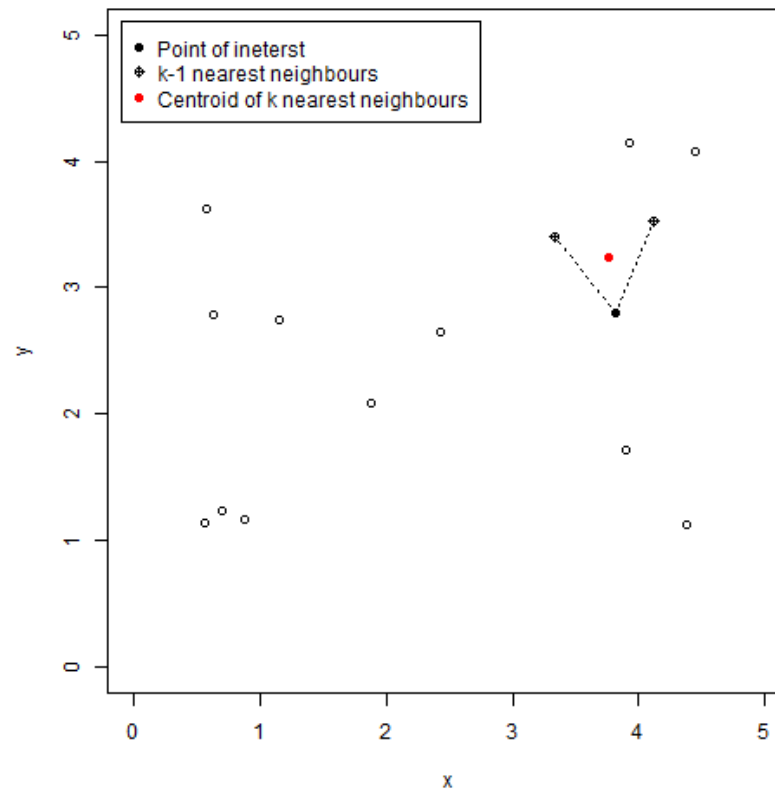
Var1	Var2	Var3	Var4
0	1	12.02422	8.125242
0	1	7.813473	8.126139
0	1	6.439849	7.960921
0	1	9.558858	8.623355
0	1	6.651313	8.504875
0	1	11.83836	8.562876

Var1	Var2	Var3	Var4
1	0	7.742192	8.266226
1	0	8.091229	7.853049
1	0	13.83798	8.016772
1	0	7.955962	7.066784
1	0	6.355218	8.12677
1	0	8.037256	7.41357
1	0	7.225119	9.368156

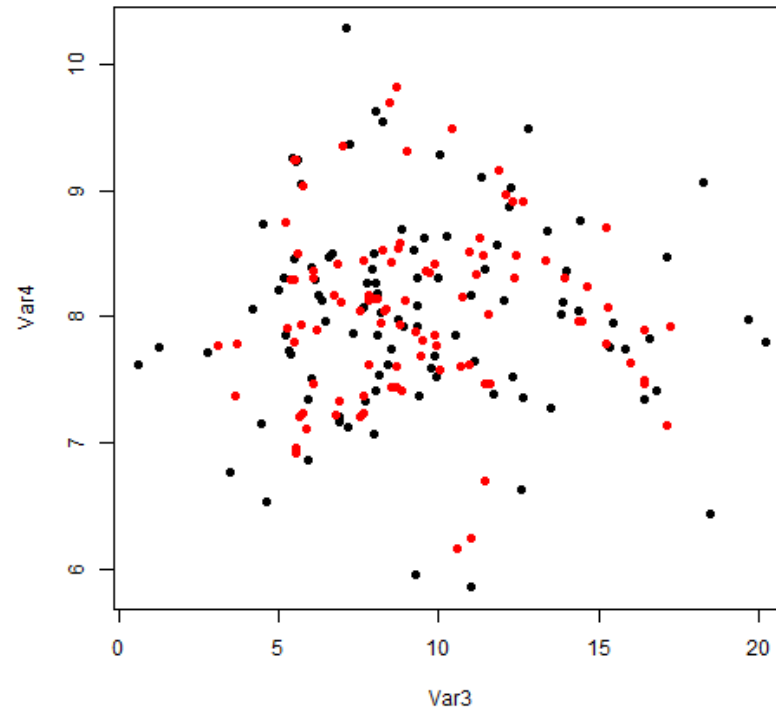
Var1	Var2	Var3	Var4
1	1	11.7165	7.380501
1	1	7.70104	7.327802
1	1	6.006454	7.513198
1	1	6.894984	7.167675
1	1	10.53635	7.856456

Deterministic Anonymisation Algorithm:

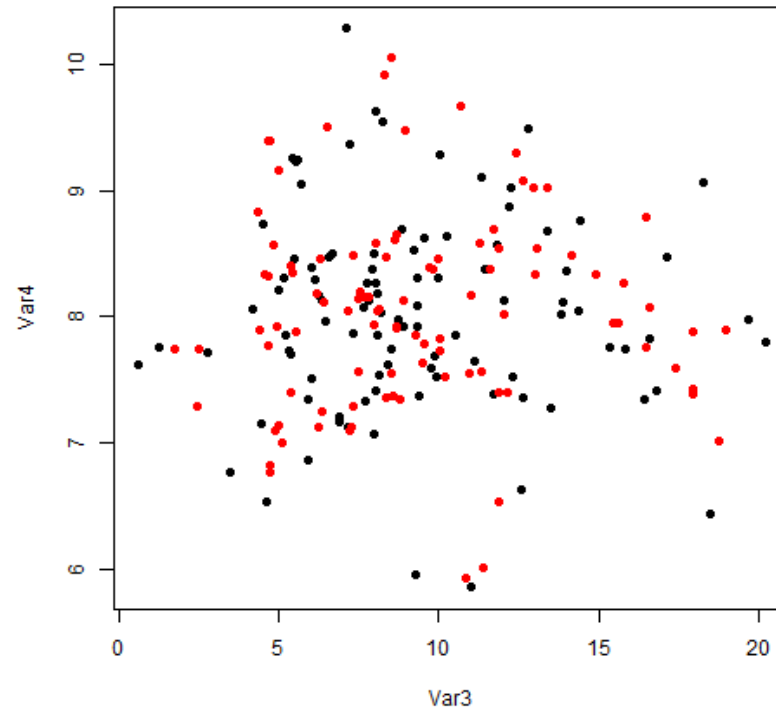
Step 2 – Find the centroids of each k -Nearest Neighbors



Deterministic Anonymisation Algorithm: Step 3 – Replace the data with the centroids



Deterministic Anonymisation Algorithm: Step 4 – Rescale the centroids



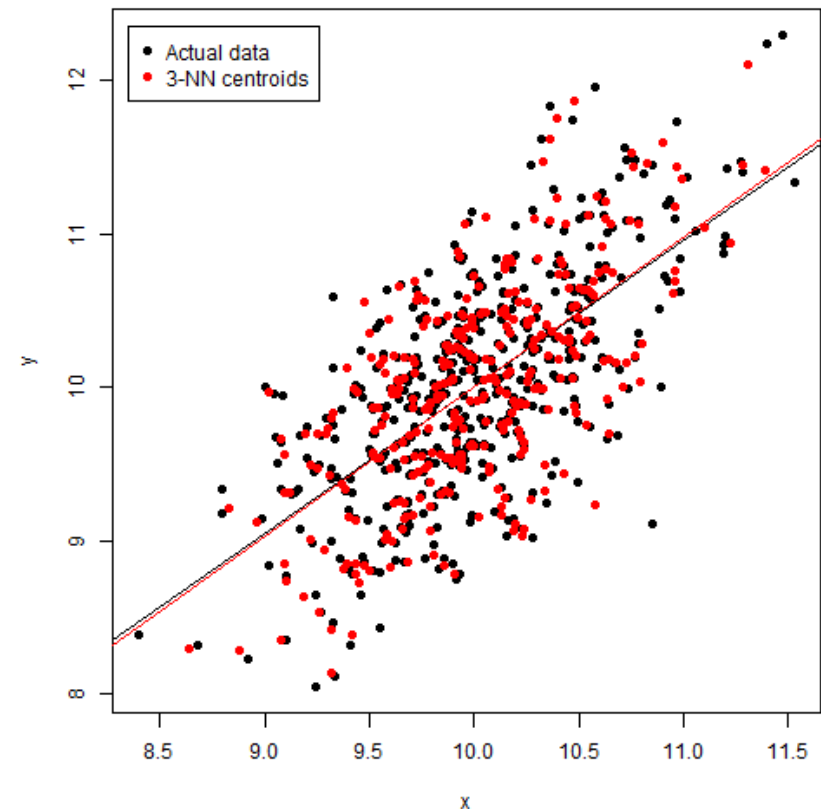
Privacy Protected Graphical Functionality in DataSHIELD



Non-disclosive scatter plots

Produce scatter plots using the algorithm of k -NN centroids

- The method is deterministic
- While k is increasing ($k \geq 2$) the disclosure risk is decreasing
- Information loss is minimum (for example when $k = 3$ the correlation line of actual data and the correlation line of the centroids are almost identical)



Low counts cell suppression

In tabular data, cells with small numbers (lower than a pre-specified threshold) are not displayed

Actual 2D contingency table

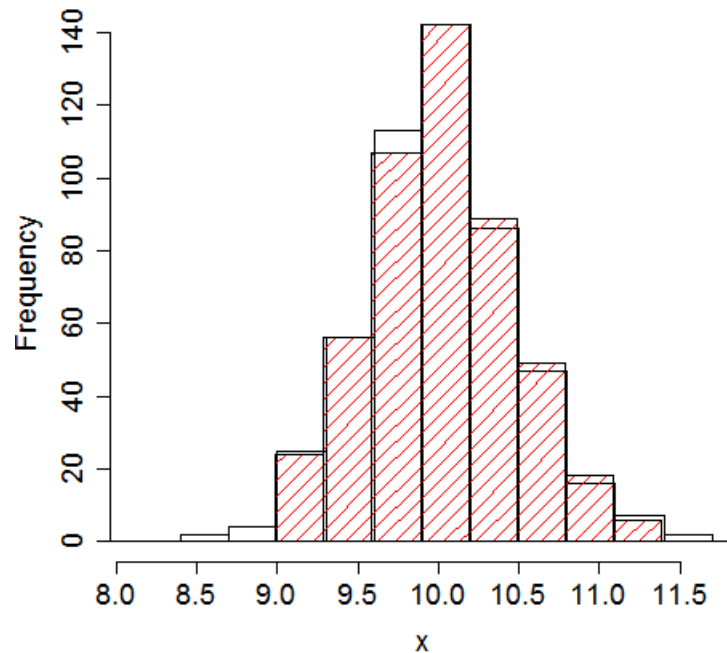
		Gender		
		0	1	Total
Dis.	0	525	327	852
	1	3	18	21
	Total	528	345	873

DataSHIELD output (with $Th = 5$)

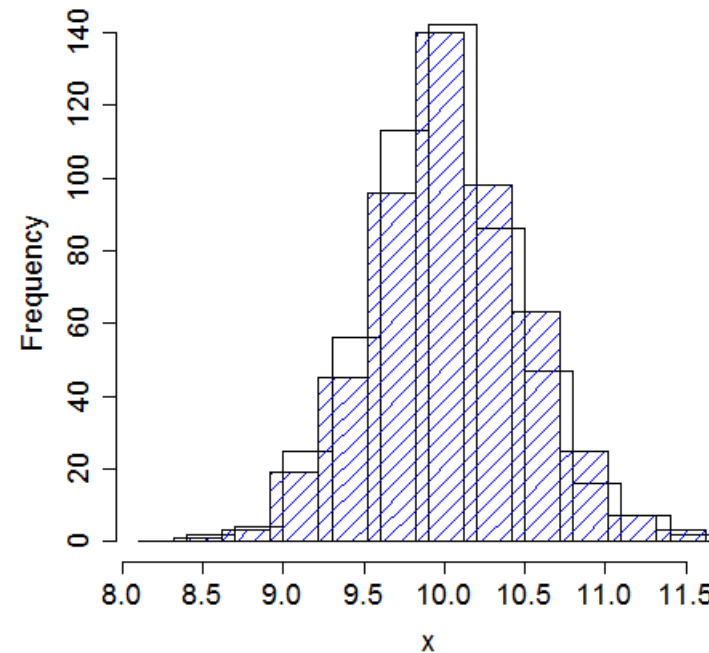
		Gender		
		0	1	Total
Dis.	0	NA	NA	852
	1	NA	NA	21
	Total	528	345	873

Non-disclosive histograms

Cell suppression ($Th = 5$)



k -NN centroids ($k = 3$)



Non-disclosive box plots

A – Actual box plot of a variable
(potentially disclosive)

B - Non-disclosive box plot of the
variable generated by suppressing the
five minimum and the five maximum
values of the variable and omitting
any outliers

C - Non-disclosive box plot generated
by the centroids of each three nearest
neighbours

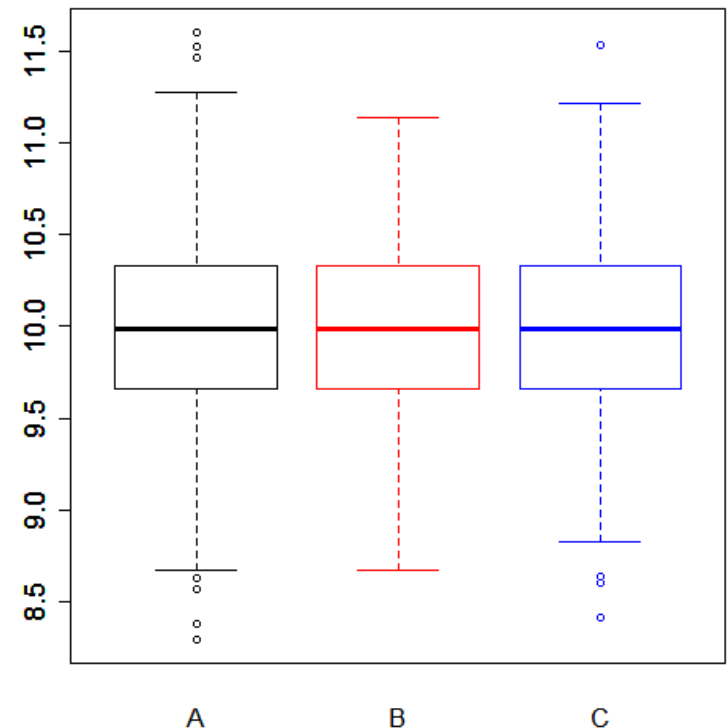


Figure 1 displays a comparison of three methods for identifying cell clusters, arranged in a 3x3 grid of plots. The rows represent different data processing methods, and the columns represent different visualization types.

Row 1: Disclusive Density Grid (actual data)

Row 2: Non-disclosive Density Grid (low counts cell suppression)

Row 3: Non-disclosive Density Grid (centroids of k nearest neighbours)

Column 1: Heatmap Plot (actual data)

Column 2: Heatmap Plot (low counts cell suppression)

Column 3: Heatmap Plot (centroids of k nearest neighbours)

Column 4: Contour Plot (actual data)

Column 5: Contour Plot (low counts cell suppression)

Column 6: Contour Plot (centroids of k nearest neighbours)

The plots show the spatial distribution of cell counts (x and y coordinates) and the resulting cluster identification. The color scale for the heatmaps ranges from 0 (blue) to 20 (red). The contour plots show the boundaries of the clusters.

Further analysis and developments

- Analyse the effect of outliers and influential points
- Analyse the effect of stratification and the effect of k value for the k -anonymity
- Compare the deterministic with probabilistic anonymization techniques in terms of disclosure risk and information loss
- Develop non-disclosive plots for regression diagnostics (e.g. graphs for residuals against fitted values, QQ-plots, etc.)
- Implement non-disclosive data visualisation into Virtual Reality technologies

Thank you!

