Analysis Write-up

Gleb Furman[1]

[1] Who Kneads a PH.D. Bakery

Analysis Write-up

Cluster Analysis

Population Frame

The population frame is composed of data from three sources: (1) the Common Core of Data (CCD), (2) publically available accountability data, and (3) the U.S. Census. The CCD is a comprehensive database housing anually collected national statistics of all public schools and districts. Accountability data was used to calculate the proportion of students within each school performing at or above proficiency in Math and ELA. Finally, local median income was obtained from the U.S. Census and was matched to each school by zipcode. School level data was aggregated to get district level variables. These are reported in Table @ref(tab:tbl_desc)

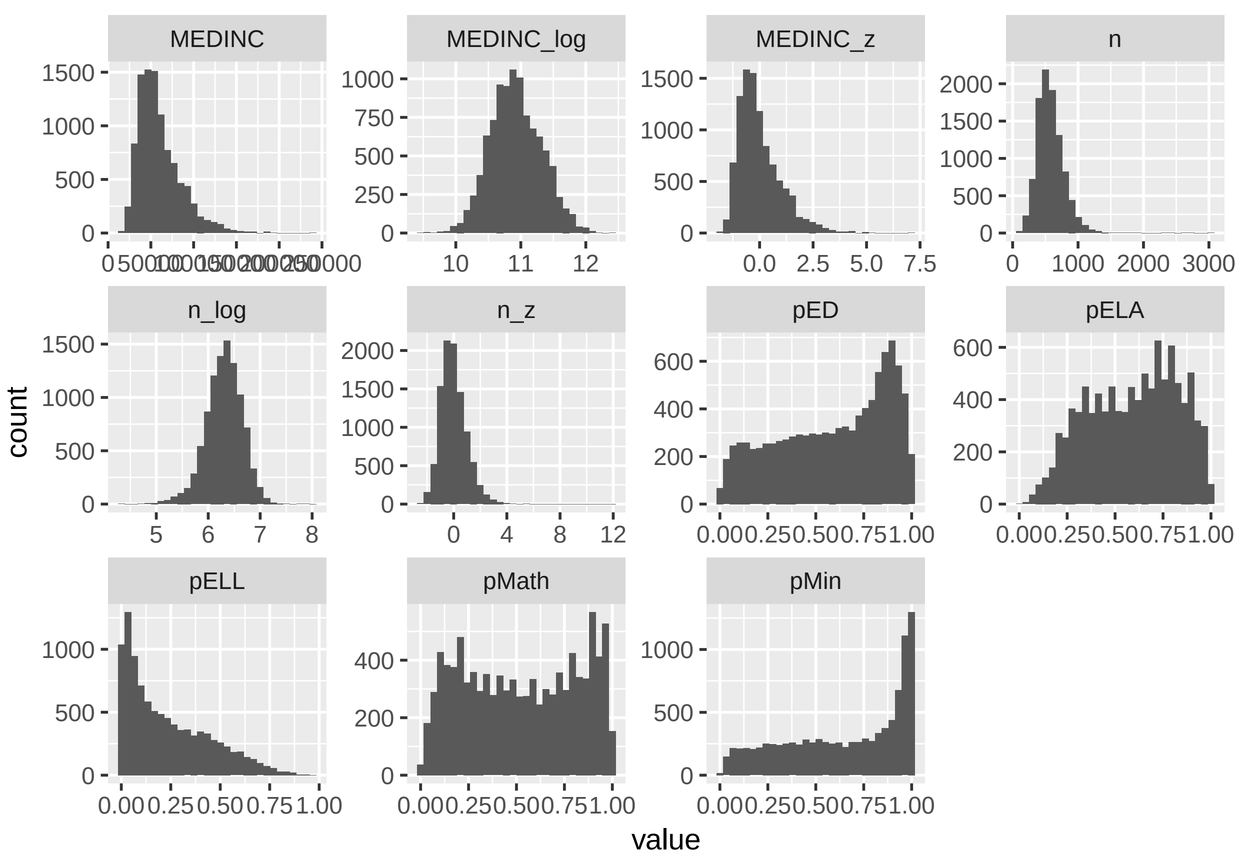
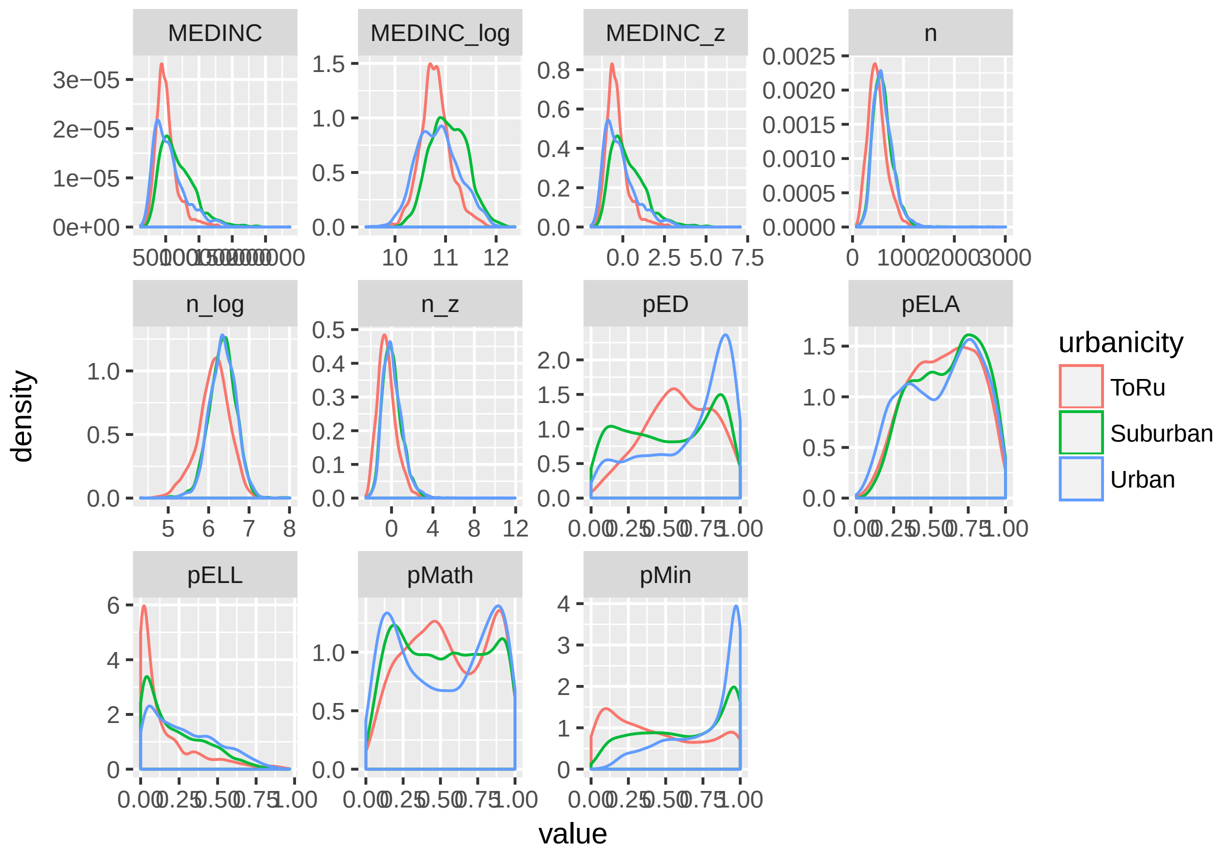


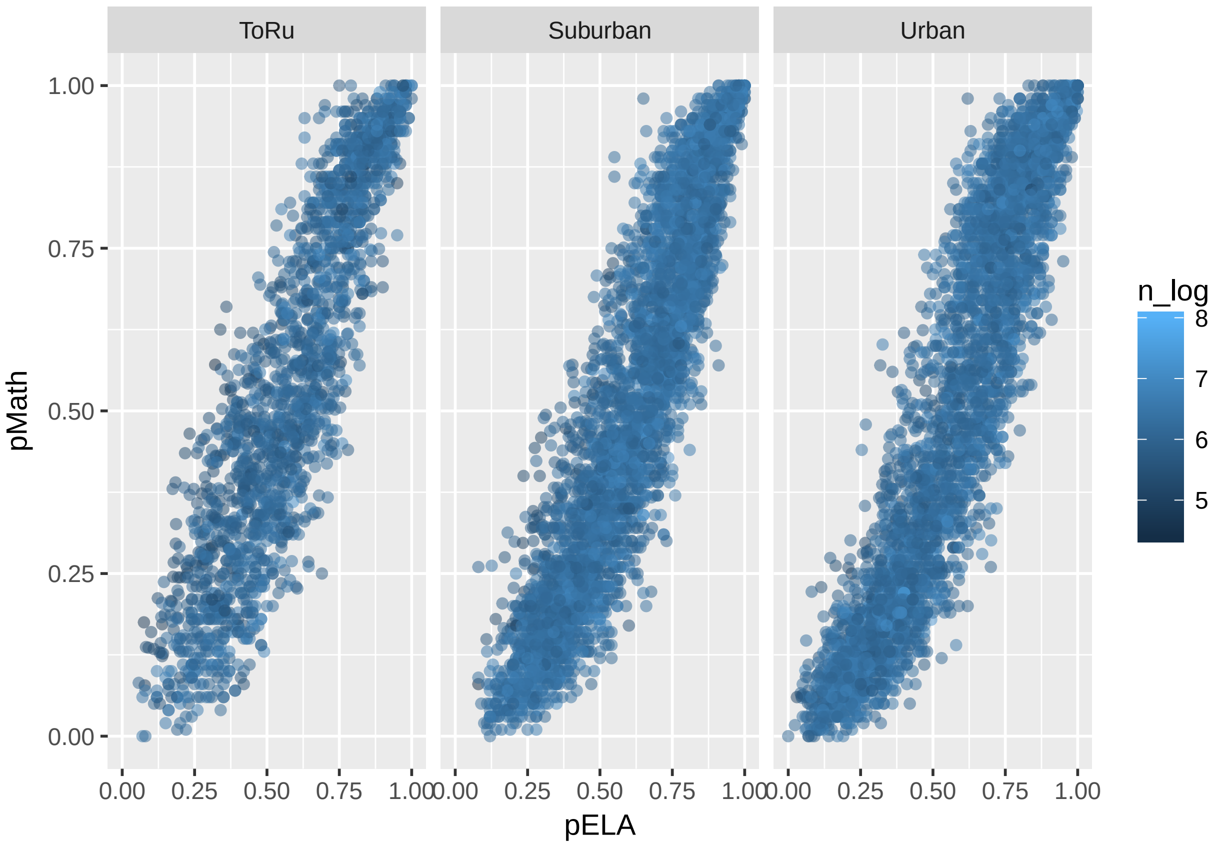
Table 1

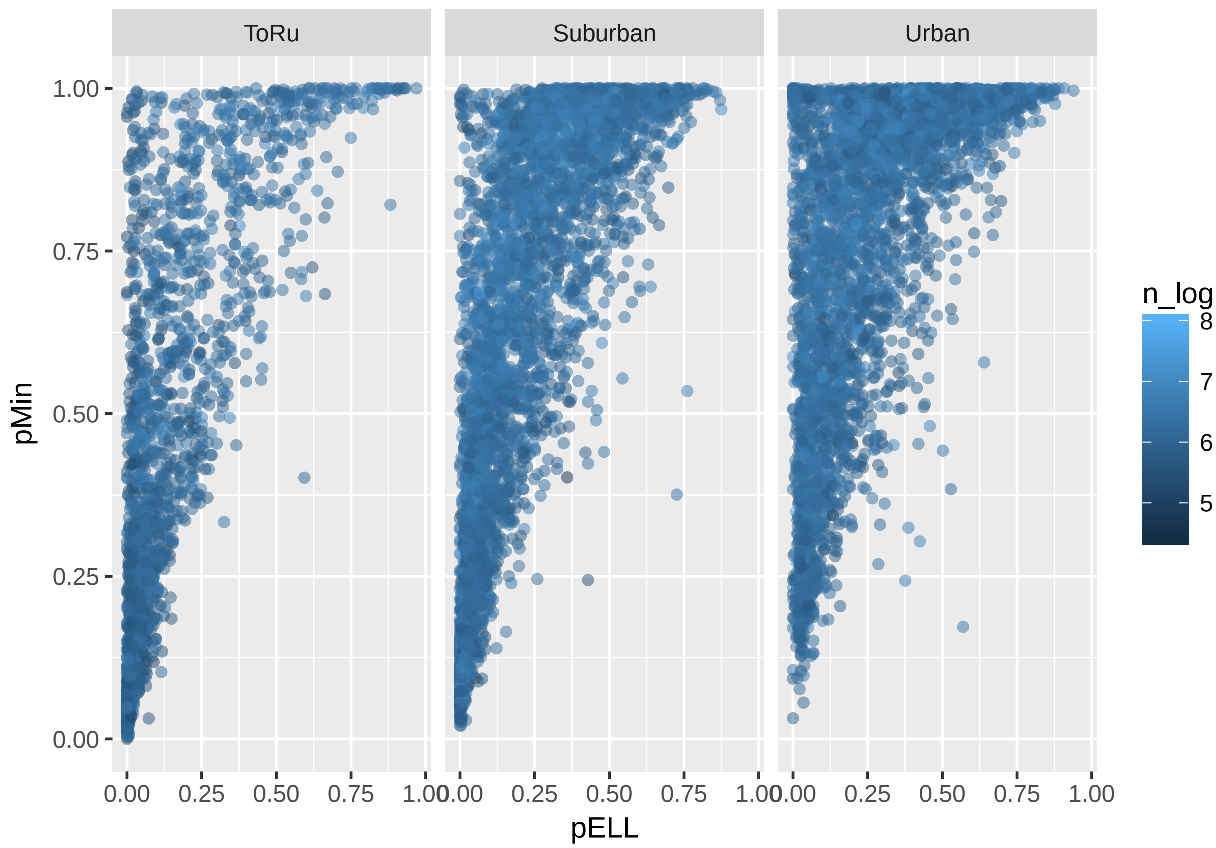
Descriptives of variables

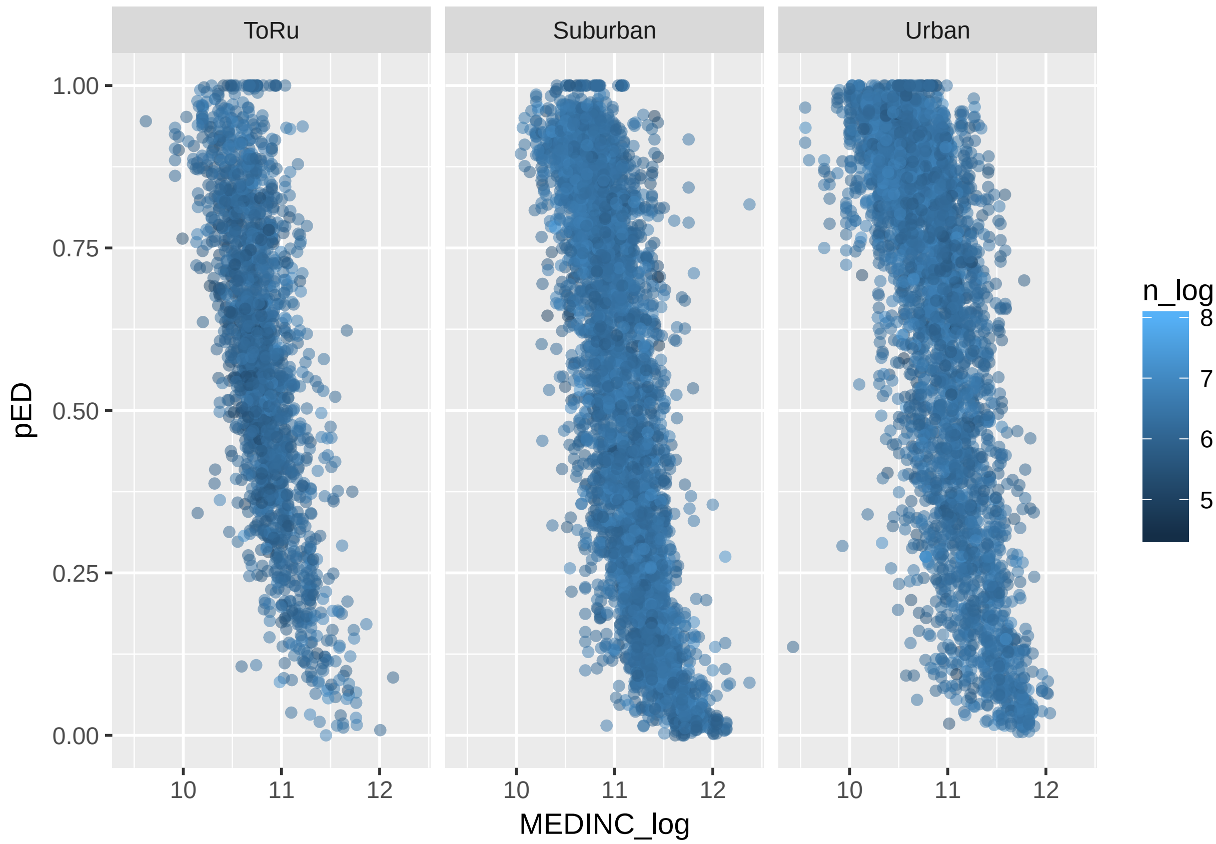
	School		District Weighted		District Unweighted	
Variables	Mean	SD	Mean	SD	Mean	SD
Number of Schools	NA	NA	9,875.00	0.00	4.84	12.72
School Size	579.07	203.19	534.77	225.34	534.77	225.34
Median Income	60,084.98	25,007.61	56,710.63	20,804.82	56,648.44	20,750.06
Average Proportions						
ELA Proficiency	0.59	0.23	0.60	0.20	0.60	0.20
Math Proficiency	0.53	0.29	0.54	0.26	0.54	0.26
Economically Disadvantaged	0.59	0.29	0.54	0.24	0.54	0.24
English Language Learners	0.23	0.21	0.14	0.17	0.14	0.17
Minority Status	0.65	0.30	0.46	0.32	0.46	0.32
Total/Free/Reduced Lunch	0.59	0.29	0.53	0.24	0.53	0.23
Indicators						
Urban	0.40	0.49	0.15	0.33	0.15	0.33
Suburban	0.41	0.49	0.33	0.44	0.33	0.44
Town or Rural	0.19	0.39	0.51	0.48	0.51	0.48

Note. District variables are derived as aggregate means of school variables


MEDINC
MEDINC_log
MEDINC_z
n
n_log
n_z
pED
pELA
pELL
pMath
pMin
density
value
urbanicity
ToRu
Suburban
Urban


ToRu
Suburban
Urban
pMath
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
pELA
n_log
8
7
6
5


ToRu
Suburban
Urban
pMin
pELL
n_log
1.00
0.75
0.50
0.25
0.00
0.00 0.25 0.50 0.75 1.00
8
7
6
5



SUBS

Stratification using balanced sampling (SUBS) was performed prior to simulation because the group of schools in each strata would be static across conditions except where the balancing model is manipulated. The set of covariates in both the full model (SUBS-F) and the omitted variable model (SUBS-OV) include binary indicator variables

Number of Clusters. Selecting the number of clusters, k, is one of the most difficult problems in cluster analysis (Steinley, 2006). To date, the most extensive investigation of methods for determining k was conducted by Milligan and Cooper (1985) who analyzed 30 methods. However, aside from the limited generalizability of this study, many methods are also innapropriate in the context of non-hierarchical and thus do not support k-means clustering. Hennig and Liao (2013) argue that the method of selecting k should depend on the context of the clustering and frame the issue as one of obtaining an appropriate subject-matter-dependent definition of rather than a statistical estimation.

- Everitt (2011), p126
- clusterSim
- Continuous data?
  - Calinski and Harabasz (1974)
  - Duda and Hart (1973)
- Steinley, D. (2006a) K-means clustering: a half-century synthesis. British Journal of Mathematical & Statistical Psychology, 59, 1–34.
- Milligan and Cooper (1984)
- list 30

Subs-Full.

Subs-OV.

```
## # A tibble: 11 x 8
##    cluster_full_10   `1`   `2`   `3`   `4`   `5`   `6`  `20`
##              <dbl> <int> <int> <int> <int> <int> <int> <int>
##  1              1.   978    NA    NA   176    NA    NA  1154
##  2              2.    NA    NA    NA    NA    NA  1241  1241
##  3              3.    NA    NA    NA    NA   658    NA   658
##  4              4.    NA   957   193    NA    NA    NA  1150
##  5              5.    NA    NA     2    NA   667    NA   669
##  6              6.   283    59    NA    NA   217    NA   559
##  7              7.    NA     6  1368    NA    NA    NA  1374
##  8              8.    NA  1458    NA    NA    NA    NA  1458
```

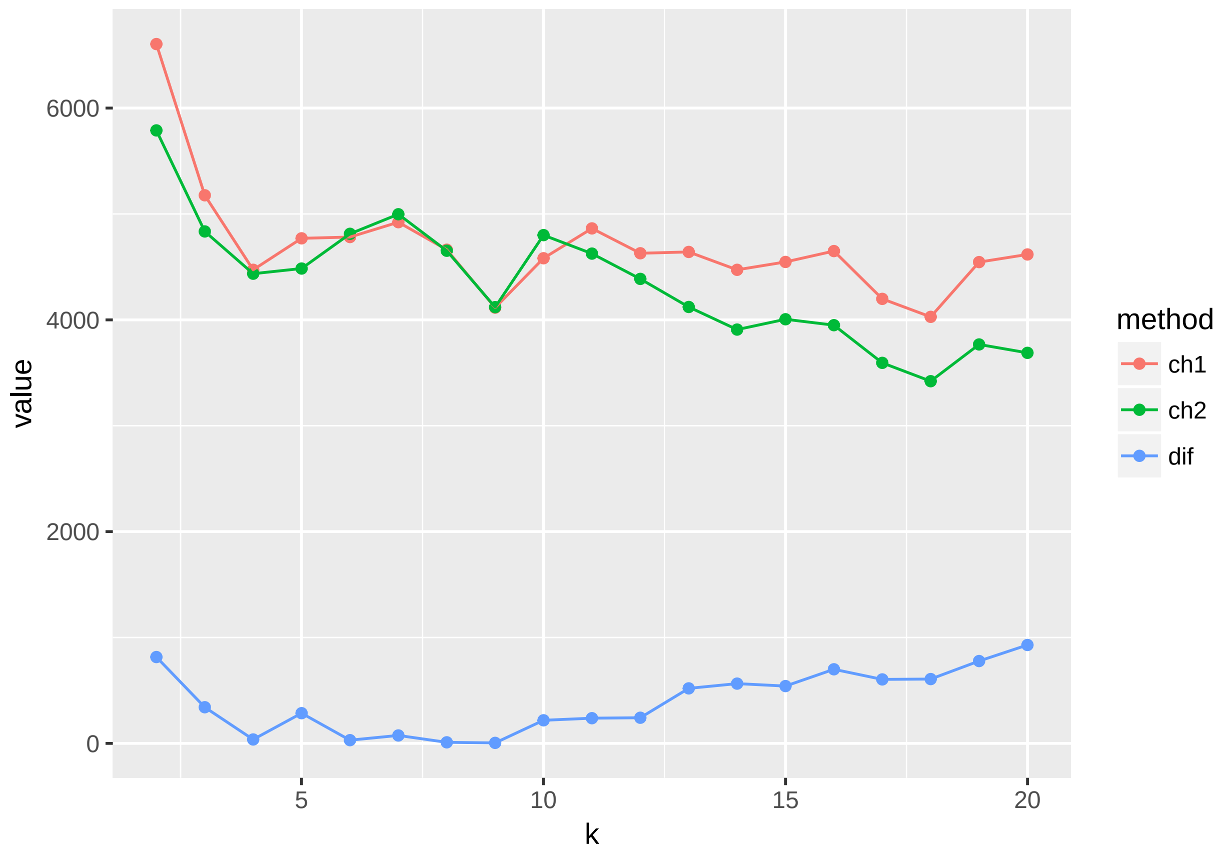


Figure 1

```
## 9                 9.    NA    NA    NA   749    NA   319  1068
## 10               10.    45    NA    NA   495    NA     4   544
## 11               20.  1306  2480  1563  1420  1542  1564    NA
```

```
## # A tibble: 11 x 8
##    cluster_OV_10   `1`   `2`   `3`   `4`   `5`   `6`  `20`
##            <dbl> <int> <int> <int> <int> <int> <int> <int>
## 1             1.    NA    NA  1372    NA    NA    NA  1372
## 2             2.    NA    36    NA    47    10   247   340
## 3             3.    NA    NA    NA    NA  1106    NA  1106
## 4             4.    NA   597    NA    54     5    NA   656
## 5             5.    NA   888    NA    NA    NA    NA   888
```

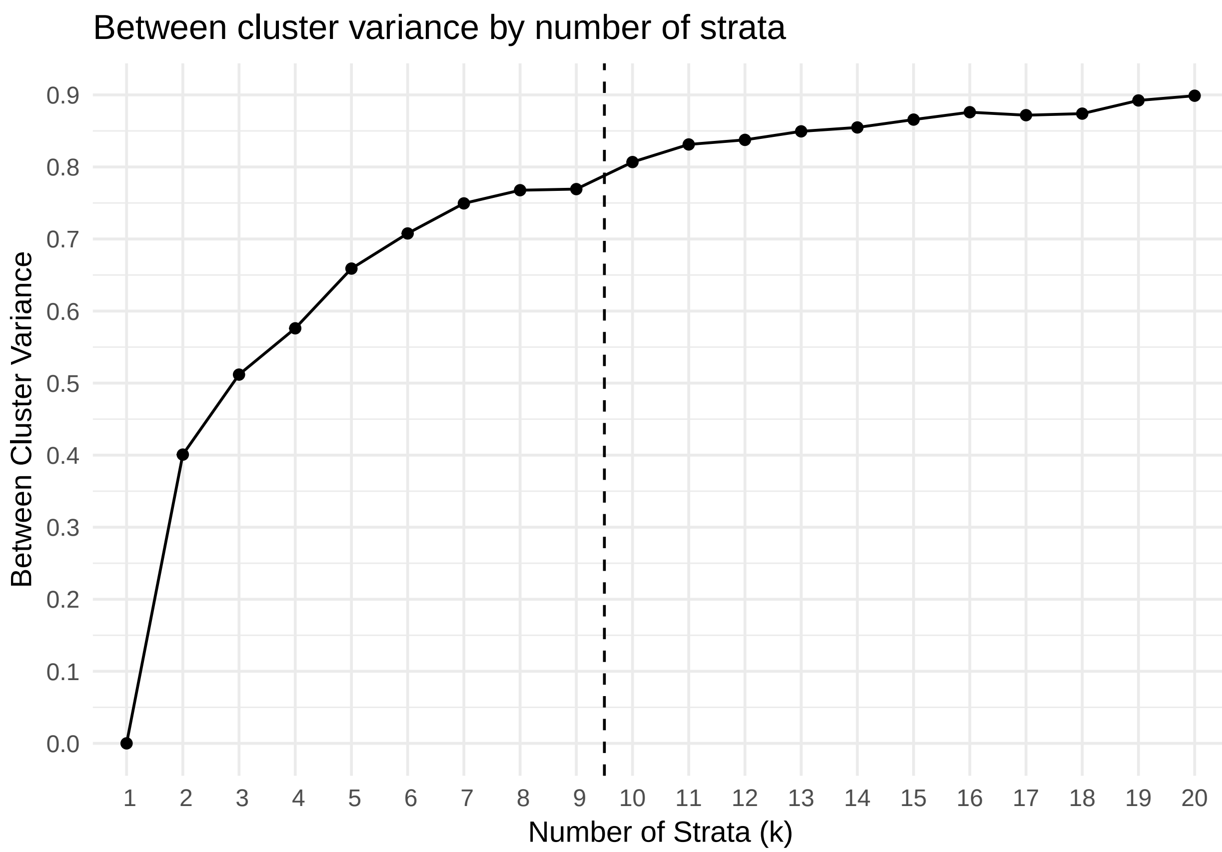


Figure 2

```
## 6             6.    NA    NA   573    NA   368    NA   941
## 7             7.    NA    NA    NA    10    NA  1455  1465
## 8             8.    12    NA    NA  1178    NA    NA  1190
## 9             9.  1328    NA    NA    NA    NA     1  1329
## 10           10.    NA    NA   185    NA   403    NA   588
## 11           20.  1340  1521  2130  1289  1892  1703    NA

## # A tibble: 7 x 8
##   cluster_OV_6   `1`   `2`   `3`   `4`   `5`   `6`  `20`
##          <dbl> <int> <int> <int> <int> <int> <int> <int>
## 1           1.    NA     8  1332    NA    NA    NA  1340
## 2           2.    20    NA    NA    NA  1501    NA  1521
```

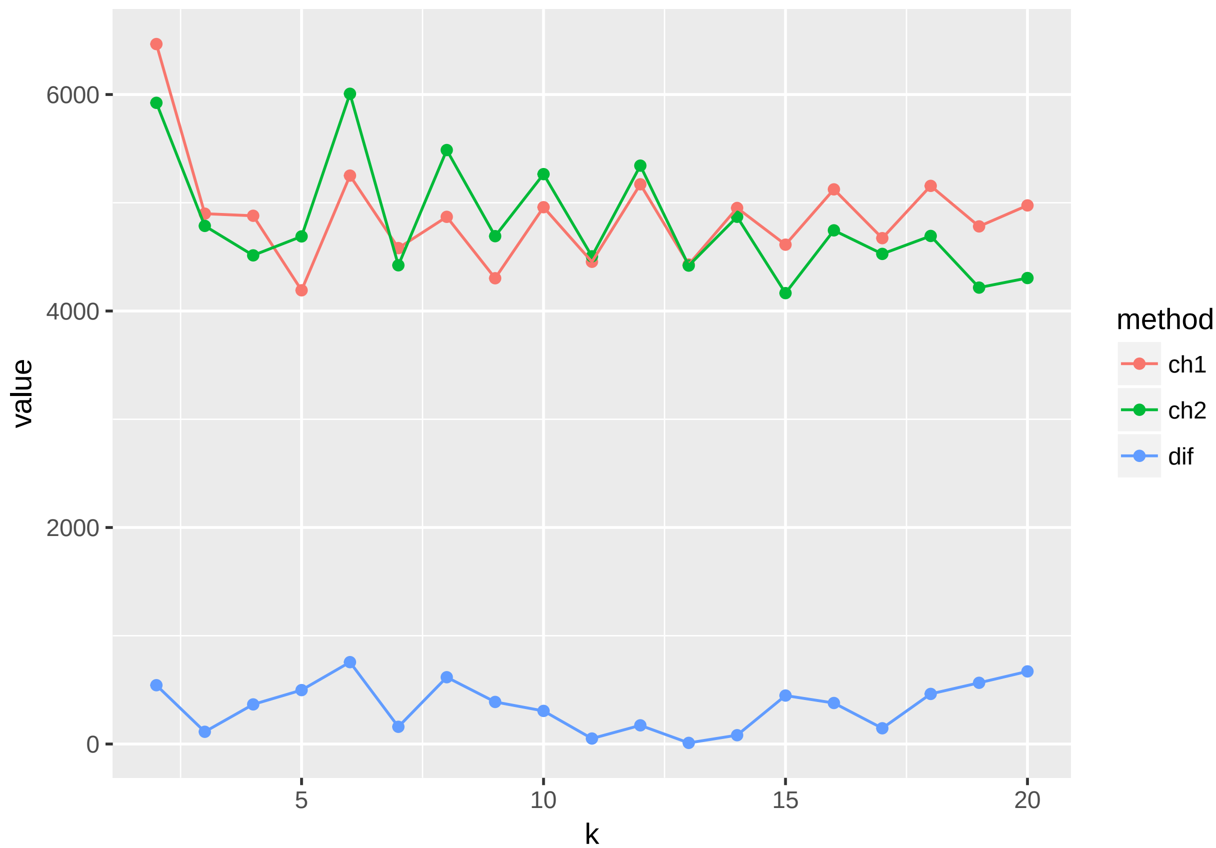


Figure 3

```
## 3            3.    NA    NA    NA   566    NA  1564  2130
## 4            4.     1  1019   228    NA    41    NA  1289
## 5            5.  1038    NA    NA   854    NA    NA  1892
## 6            6.   247  1453     3    NA    NA    NA  1703
## 7           20.  1306  2480  1563  1420  1542  1564    NA
```

```
## # A tibble: 11 x 12
##    cluster_OV_10   `1`   `2`   `3`   `4`   `5`   `6`   `7`   `8`   `9`
##            <dbl> <int> <int> <int> <int> <int> <int> <int> <int> <int>
##  1            1.    NA  1218    NA    NA    NA    NA    NA    NA   153
##  2            2.    NA    NA     1    NA    NA   339    NA    NA    NA
##  3            3.  1104    NA    NA    NA    NA    NA    NA    NA    NA
```

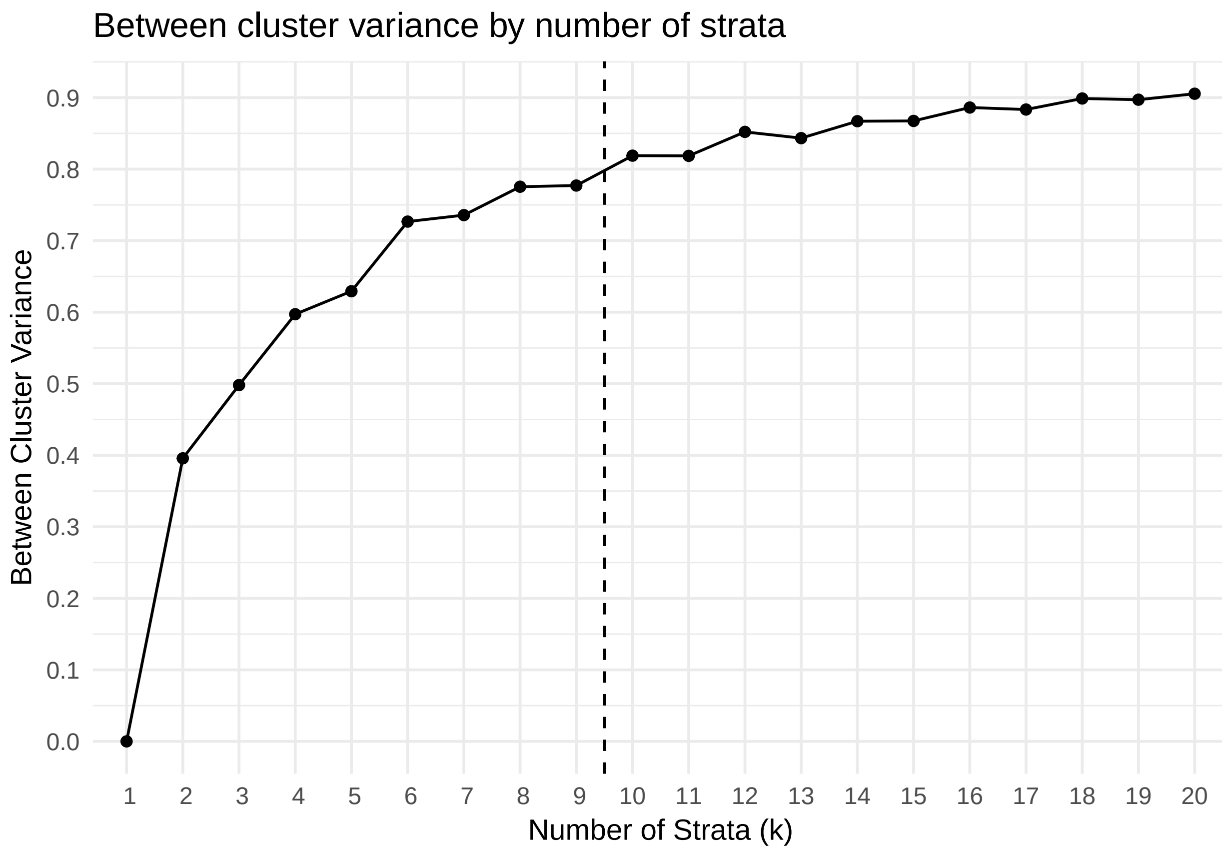


Figure 4

```
## 4              4.    NA    NA   300    NA   136   220    NA    NA    NA
## 5              5.    NA    NA   357    NA   531    NA    NA    NA    NA
## 6              6.    50    NA    NA    NA    NA    NA    NA    NA   878
## 7              7.    NA    NA    NA     8    NA    NA     5  1452    NA
## 8              8.    NA    NA    NA  1130    NA    NA    57     3    NA
## 9              9.    NA    NA    NA    12     2    NA  1312     3    NA
## 10            10.    NA    23    NA    NA    NA    NA    NA    NA    37
## 11            20.  1154  1241   658  1150   669   559  1374  1458  1068
## # ... with 2 more variables: `10` <int>, `20` <int>
```

References