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)

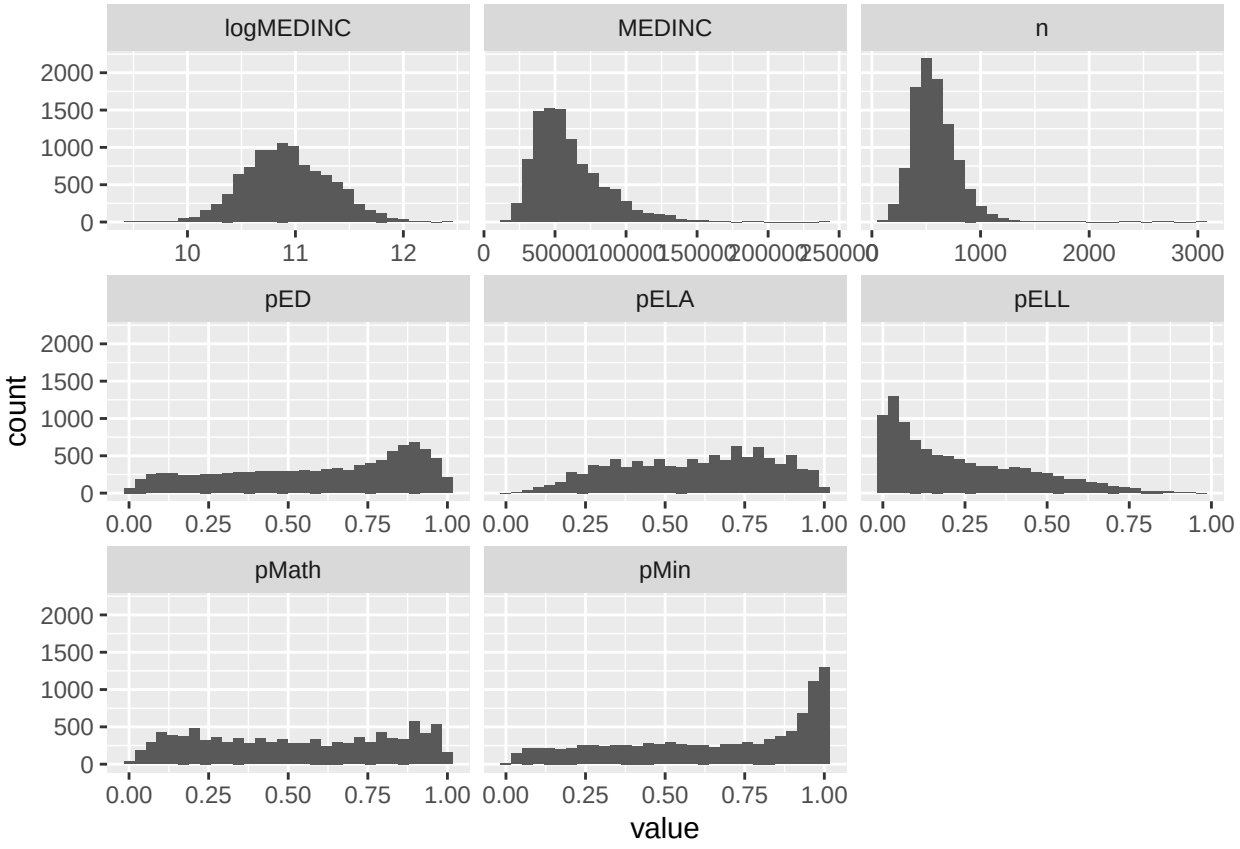
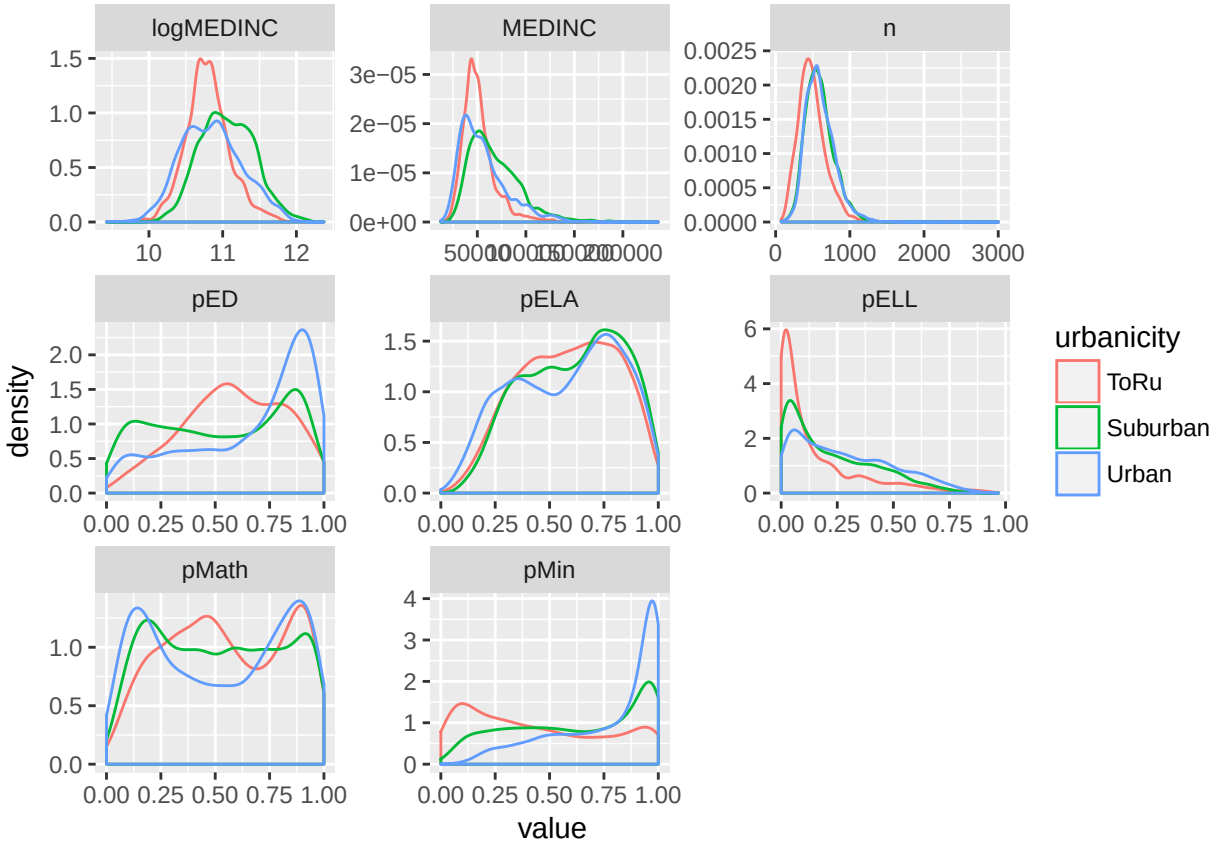


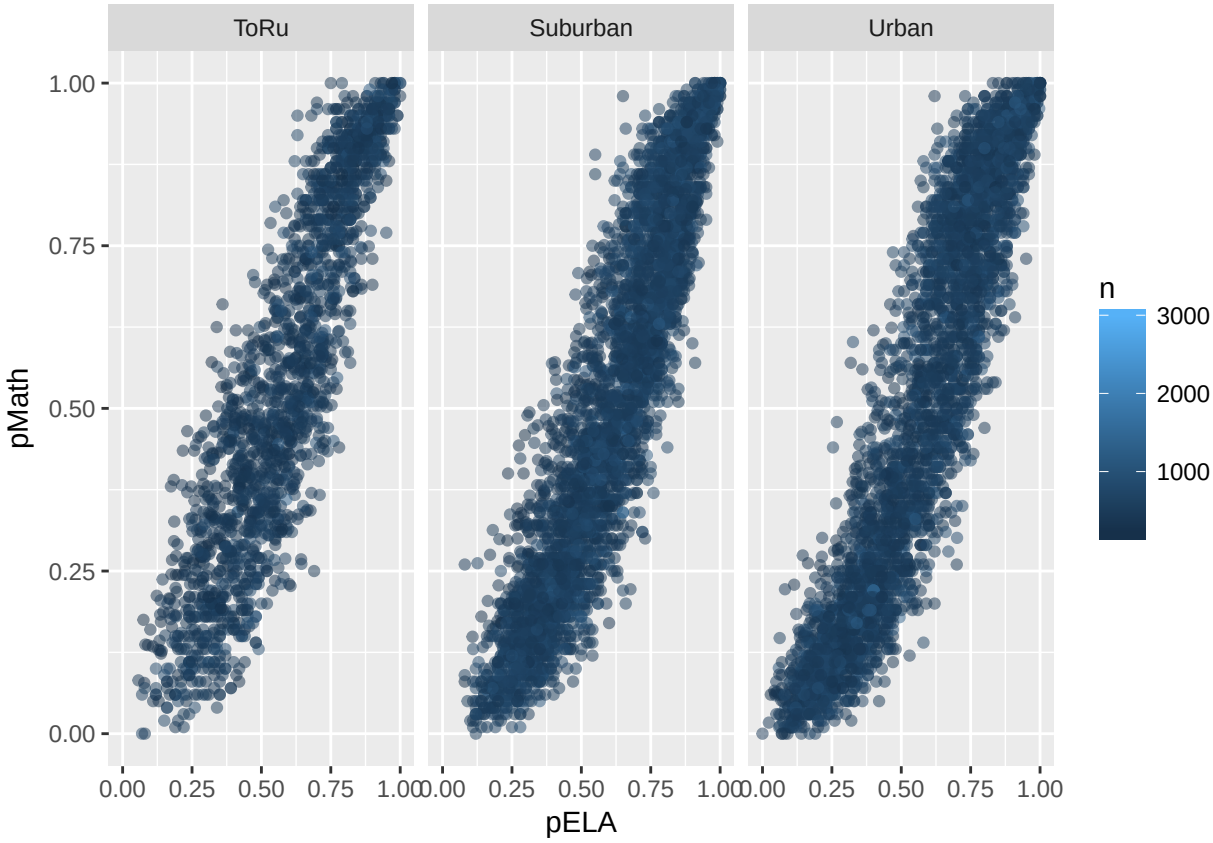
Table 1

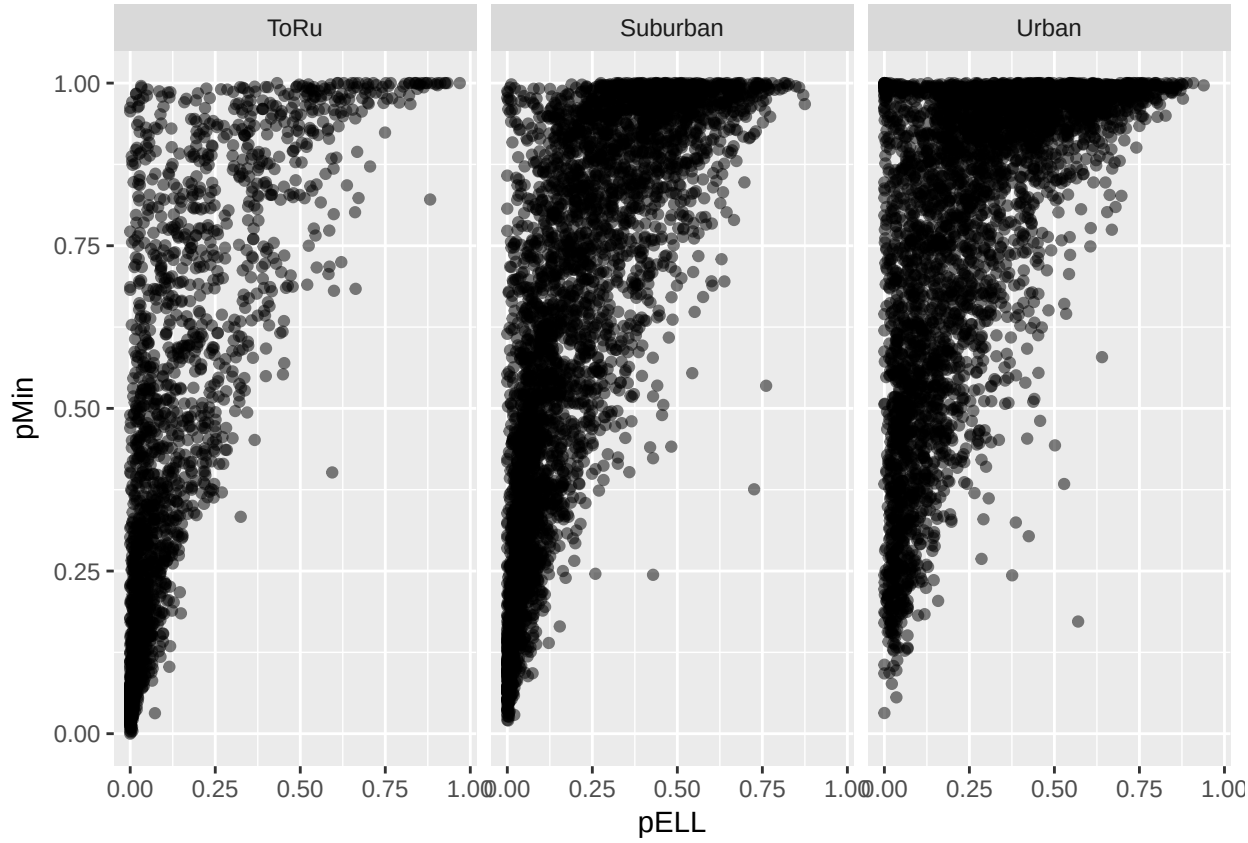
Descriptives of variables

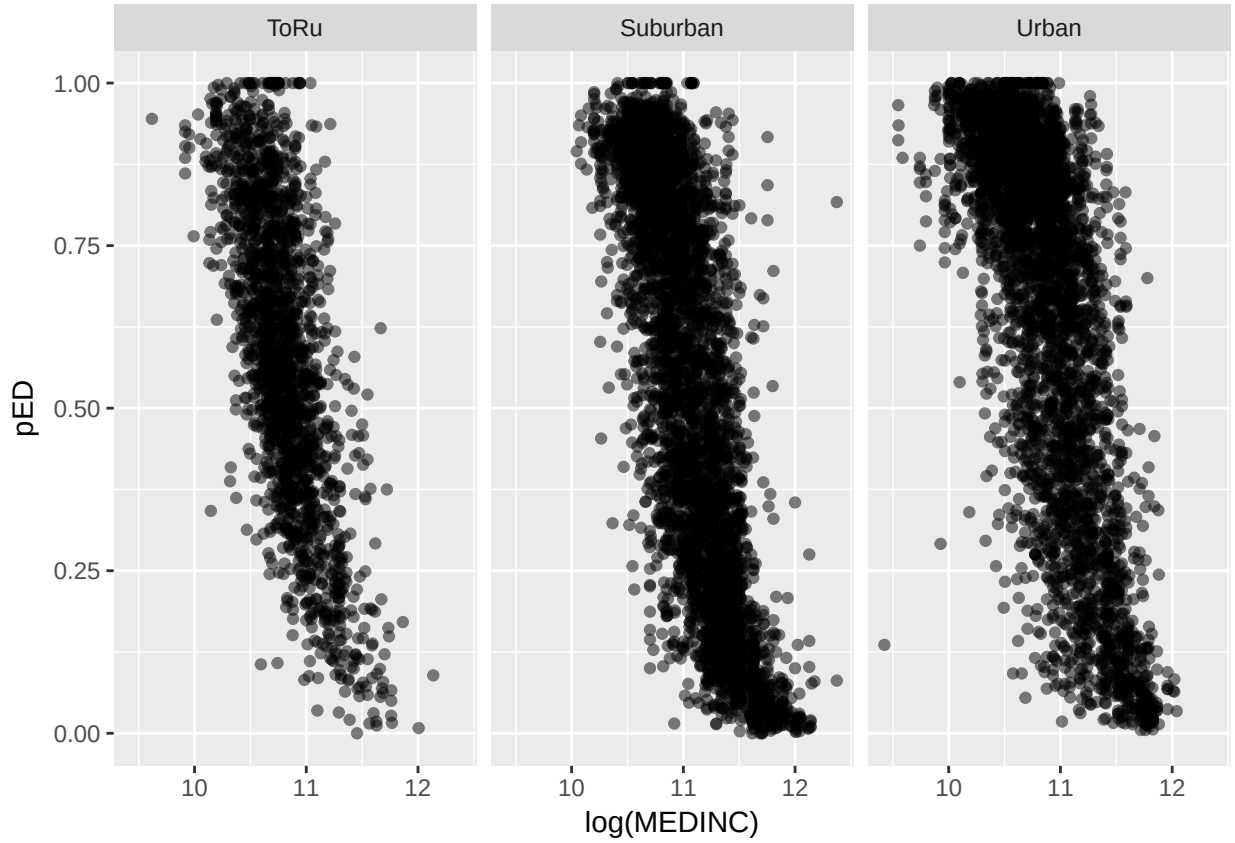
	School		District Weighted		District Unweighted	
Variables	Mean	SD	Mean	SD	Mean	SD
Number of Schools	NA	NA	9,875.00	0.00	4.84	12.72
School Size	579.07	203.19	534.77	225.34	534.77	225.34
Median Income	60,084.98	25,007.61	56,710.63	20,804.82	56,648.44	20,750.06
Average Proportions						
ELA Proficiency	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


logMEDINC
MEDINC
n
pED
pELA
pELL
pMath
pMin
density
value
urbanicity
ToRu
Suburban
Urban


ToRu
Suburban
Urban
pMath
pELA
0.00
0.25
0.50
0.75
1.00
n
3000
2000
1000


ToRu
Suburban
Urban
pMin
pELL
0.00
0.25
0.50
0.75
1.00



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

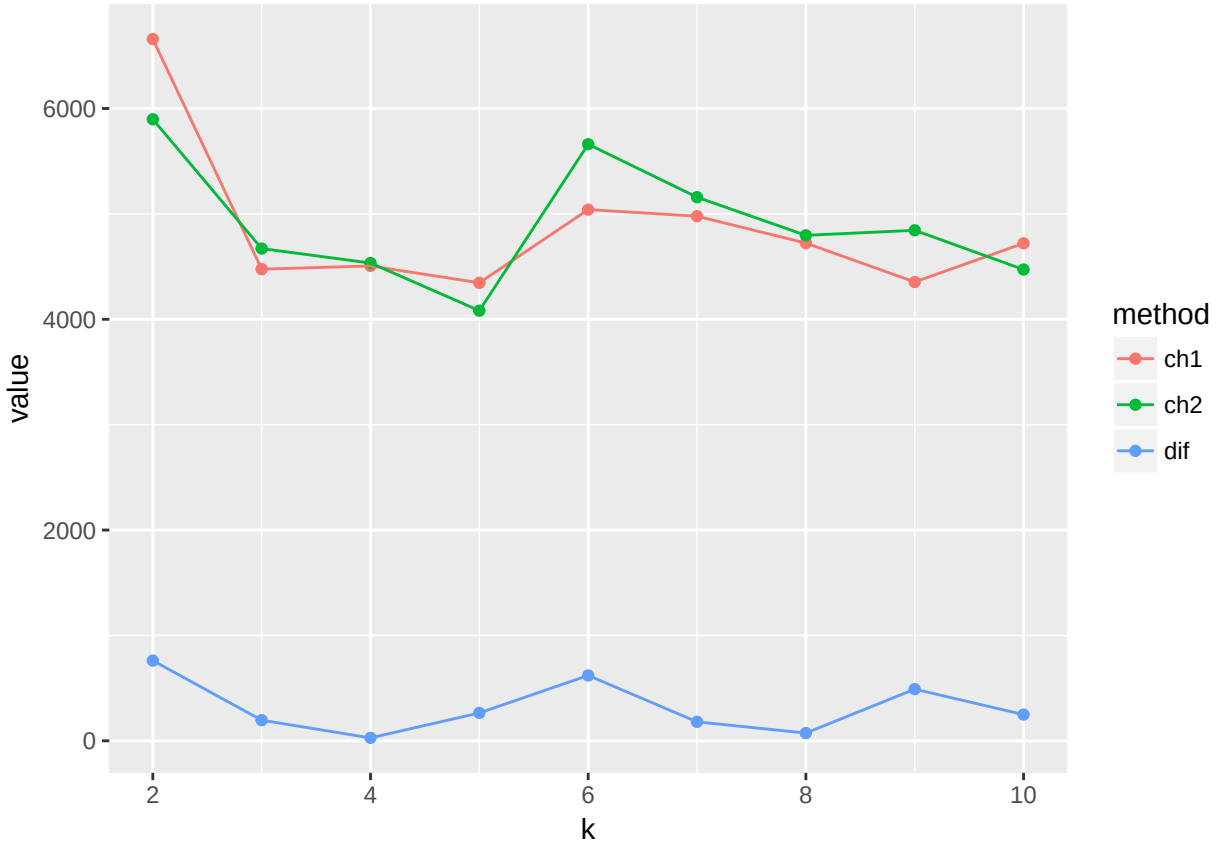


Figure 1

References

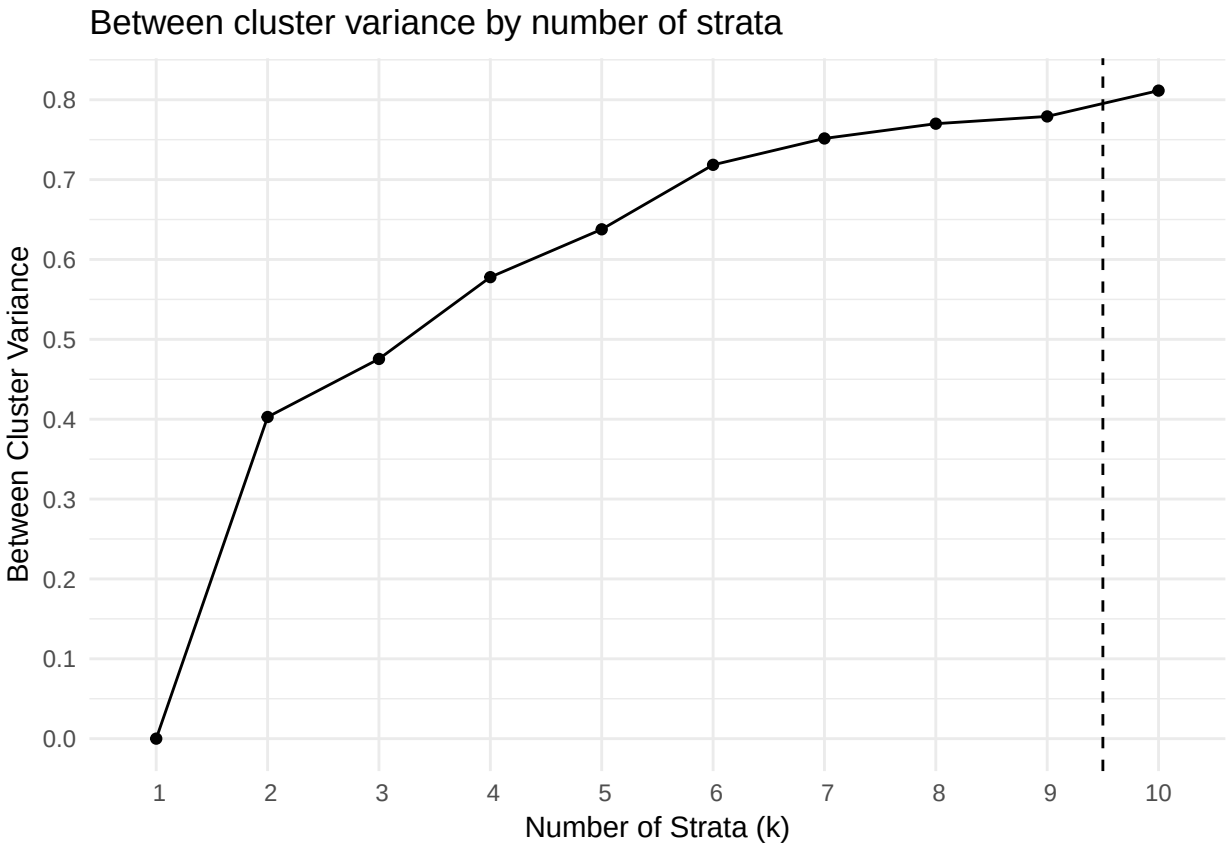

Between cluster variance by number of strata
Between Cluster Variance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
2
3
4
5
6
7
8
9
10
Number of Strata (k)

Figure 2