An Alternative to the Carnegie Classifications:

Developing Tools to Identify Similar Doctoral Institutions

Paul Harmon
w/ Sarah McKnight, Laura Hildreth, Ian Godwin, & Mark Greenwood
Montana State University

June 28, 2018

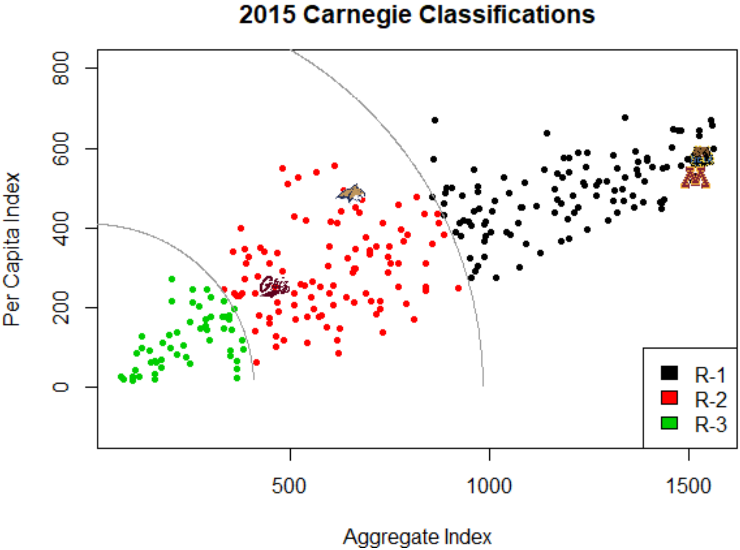
The Carnegie Classifications for Doctoral Institutions

The Carnegie Classifications are a widely-used tool for institutional classification. Some schools use these to inform strategic plans (e.g., Montana State University, University of MT, University of Idaho) and other policy decisions. For doctoral-granting schools, the Carnegie Classifications delineate three groups:

- **R1**: Highest Research
- **R2**: Higher Research
- **R3**: Moderate Research

For teachers, these data constitute a rich multivariate dataset that is publicly available, easy to understand, and often a hot topic on campuses. Data are available at:
http://carnegieclassifications.iu.edu/downloads.php

The Carnegie Classifications: 2015 Update



What Data are used?

Data used in the Carnegie Classifications are based on a snapshot from the Integrated Postsecondary Education Data System (IPEDS), NSF HERD Survey, NSF GSS, etc. The variables used in the classifications are:

- **STEM expenditures** (in thousands of dollars)
- **Non-STEM expenditures** (in thousands of dollars)
- **Postdocs/Nonfaculty PhD Research Staff size**
- STEM PhD Counts
- Humanities PhD Counts
- Social Science PhD Counts
- Other PhD Counts
- *Tenured/Tenure Track Faculty Headcount*

Per-Capita: The 3 variables in bold are used in per-capita variables by dividing by Faculty Headcount.

The Carnegie Methodology

The Carnegie Classifications are built on the following methodology:

- **Rank** Institutions on 7 aggregate variables and 3 per-capita variables (*many schools are tied on PhD counts, which is a problem*)
- **Index Creation**: 2 individual PCAs (one aggregate, one per-capita)
- **Plot the Indices**: (per-capita vs aggregate)
- **Create groups**: Groups are delineated via arc-drawing after visual inspection (V. Borden, personal communication, 2017)

Problems with Carnegie System

The process used by the Carnegie Classifications illustrates several problems illustrated by Kosar & Scott (2018):

- **Ranking**: Ranking the data consolidates it, removing the separation that could show group differences. This also leads to institutions with tied ranks.
- **Model Complexity**: Two PCAs are needed, one for Per-Capita and the other for Aggregate production.
- **Group Determination**: The group determination is completely subjective. Arcs are drawn after visual inspection of the plot.
- **Group Interpretation**: Indiana University Center for Postsecondary Research insists they aren't an ordinal "ranking" but they developed classifications that carry that implication.

STEM and Non-STEM Factors: An Alternative

Because of these problems, multiple alternatives have been developed:

- Kosar & Scott (2018) used a single 2D PCA on the entire dataset, rotated to try to have each component match the Carnegie Structure.
- We tried fitting the Aggregate and Per-capita indices as latent traits in a SEM.

A more intuitive (and estimable) method would be to consider two latent factors:

- **STEM** productivity
- **Non-STEM** productivity

Variables loaded onto these factors are not as likely to be correlated.

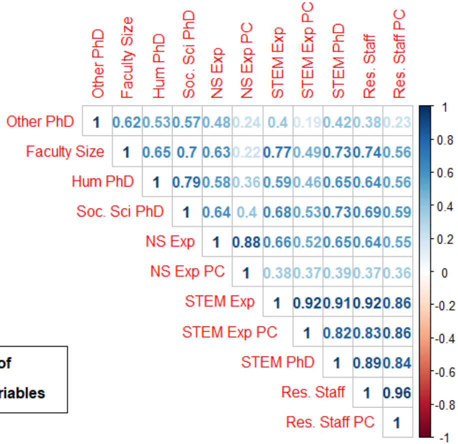
Structural Equation Modeling

Structural Equation Models (SEM) are used to model simultaneous equations, allowing for the use of latent or unobserved variables and variables to be measured with error. See Bollen (1989) for more detail.

- **Latent Variable Model**: Structural model (like a regression) illustrating relationships between latent, unobserved variables.
- **Measurement Model**: Relates the latent variables we are interested in to the manifest that we actually observe.

In R, we use the package **lavaan** (Rosseel, 2012) to fit SEM models.

STEM and Non-STEM Manifest Correlations

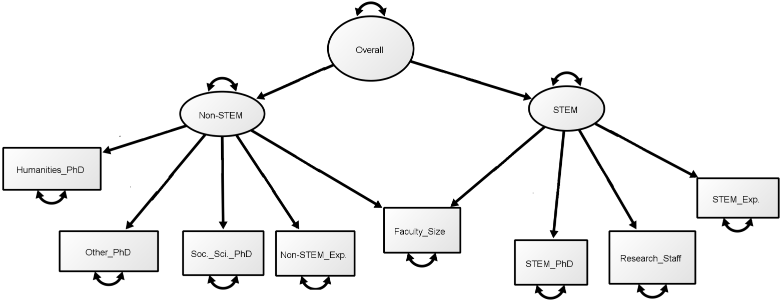


	Other PhD	Faculty Size	Hum PhD	Soc. Sci PhD	NS Exp	NS Exp PC	STEM Exp	STEM Exp PC	STEM PhD	Res. Staff	Res. Staff PC
Other PhD	1	0.62	0.53	0.57	0.48	0.24	0.4	0.19	0.42	0.38	0.23
Faculty Size		1	0.65	0.7	0.63	0.22	0.77	0.49	0.73	0.74	0.56
Hum PhD			1	0.79	0.58	0.36	0.59	0.46	0.65	0.64	0.56
Soc. Sci PhD				1	0.64	0.4	0.68	0.53	0.73	0.69	0.59
NS Exp					1	0.88	0.66	0.52	0.65	0.64	0.55
NS Exp PC						1	0.38	0.37	0.39	0.37	0.36
STEM Exp							1	0.92	0.91	0.92	0.86
STEM Exp PC								1	0.82	0.83	0.86
STEM PhD									1	0.89	0.84
Res. Staff										1	0.96
Res. Staff PC											1

Correlation Plot of

Ranked Observed Variables

Proposed Model: STEM and Non-STEM



Determining Group Membership

Single factor-of-factor score for each university used as inputs to a clustering algorithm to determine cluster membership.

Potential Problems:

- Several clustering methods that could be used (hierarchical clustering, mixture-model based methods, kmeans, etc.).
- Optimal Number of Clusters may be too large/small: We can fix this to a reasonable number if necessary (results here are optimal at 3 clusters).
- We used model-based clustering based on mixtures of normal distributions with equal variances.

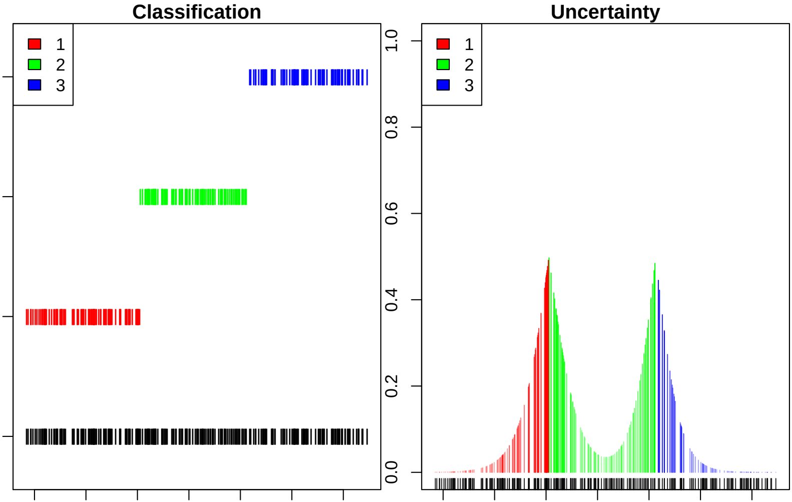
In R, we used **mclust** (Fraley and Raftery, 2002) for clustering.

Uncertainty in Classifications

In model-based clustering, uncertainty of group membership is defined as $1 - P(C_i)$ where $P(C_i)$ is the estimated probability of being in a specific cluster, conditional on the cluster solution.

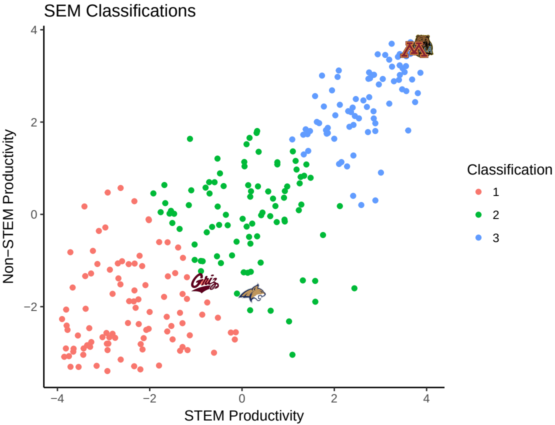
Our method can be used to visualize clusters and associated uncertainty in cluster assignments.

Uncertainty Plots



Final SEM Classifications:

Using a mixture model, we can objectively define clusters based on **one** score and can illustrate uncertainty in classification for schools near the boundary.



Teaching Administrators: Shiny Applications

Administrators, faculty, and students want to know two things:

- Why did their institution move up/down?
- What can they do to improve?

These relationships are complex due to ties in the ranked data and non-linear relationships between observed variables and final classifications.

Shiny Applications

We developed R Shiny (Chang et al., 2017) applications to assess sensitivity of both the Carnegie Classifications and the proposed SEM-Classifications for any institution of your choosing.

- Sliders allow user to change counts of PhDs, Staff/Faculty size, Expenditures
- User can select any Doctoral-granting institution in the dataset
- The SEM application uses a fixed number of three clusters

These applications can be found at:
https://ccsemclassifications.shinyapps.io/SEM_App2/

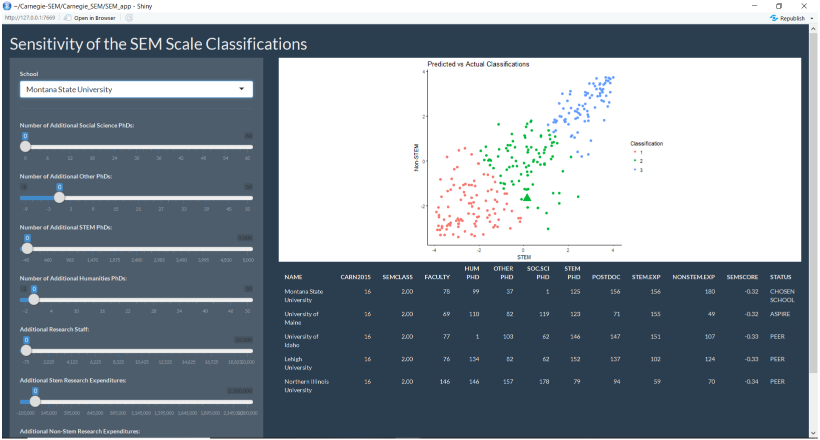
Identifying a Cohort of Institutions

How can we communicate the stated objective of the system: "identification of peer institutions"?

Our solution: Tables of 5-10 schools that are closest to the chosen institution. Administrators can move the sliders and see which schools they would be similar to, given a policy action.

These applications can be found at:
https://ccsemclassifications.shinyapps.io/sem_app/

Shiny Applications:



Sensitivity of the SEM Scale Classifications

School: Montana State University

Number of Additional Social Science PhDs: 0 (0 – 60)

Number of Additional Other PhDs: 0 (-9 – 50)

Number of Additional STEM PhDs: 0 (-45 – 5,000)

Number of Additional Humanities PhDs: 0 (-2 – 50)

Additional Research Staff: 0 (-75 – 20,000)

Additional Stem Research Expenditures: 0 (-105,000 – 2,300,000)

Additional Non-Stem Research Expenditures:

NAME	CARN2015	SEMCLASS	FACULTY	HUM PHD	OTHER PHD	SOC.SCI PHD	STEM PHD	POSTDOC	STEM.EXP	NONSTEM.EXP	SEMSCORE	STATUS
Montana State University	16	2.00	78	99	37	1	125	156	156	180	-0.32	CHOSEN SCHOOL
University of Maine	16	2.00	69	110	82	119	123	71	155	49	-0.32	ASPIRE
University of Idaho	16	2.00	77	1	103	62	146	147	151	107	-0.33	PEER
Lehigh University	16	2.00	76	134	82	62	152	137	102	124	-0.33	PEER
Northern Illinois University	16	2.00	146	146	157	178	79	94	59	70	-0.34	PEER

Questions:

Thank you!

Bibliography

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington

Chris Fraley and Adrian E. Raftery (2002) Model-based Clustering, Discriminant Analysis and Density Estimation Journal of the American Statistical Association 97:611-631

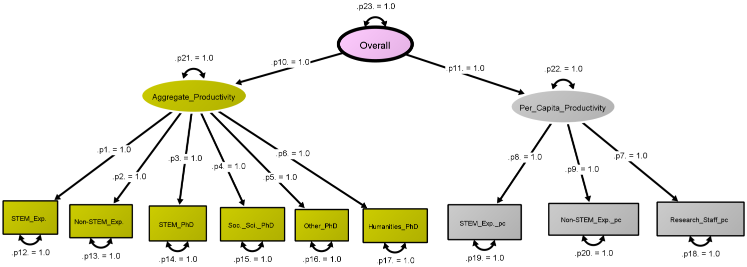
Robert Kosar & David W. Scott (2018). Examining the Carnegie Classification Methodology for Research Universities, Statistics and Public Policy, 5:1, 1-12, DOI: 10.1080/2330443X.2018.1442271

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny

Supplemental: The Carnegie Method using SEMs

We can think of the Carnegie Classifications in a latent modeling framework using this path diagram:



Conclusions

The SEM-based model allows for several benefits compared to the Carnegie Classifications:

- Factor of factors used to combine non-orthogonal components
- Latent variable modeling allows for control of dimension reduction structure and ability to assess it
- Single-factor scores allow for comparison on a single dimension rather than using two scores, making classification much easier
- SEM has diagnostics that can be explored to compare classifications in different years
- The SEM model and Shiny app can be used to better identify similar cohorts of schools rather than applying a broad classification that fails to recognize uncertainty at either side of the boundaries.