Demystifying the Carnegie Classifications: A Writing Project

Paul Harmon

April 6, 2017

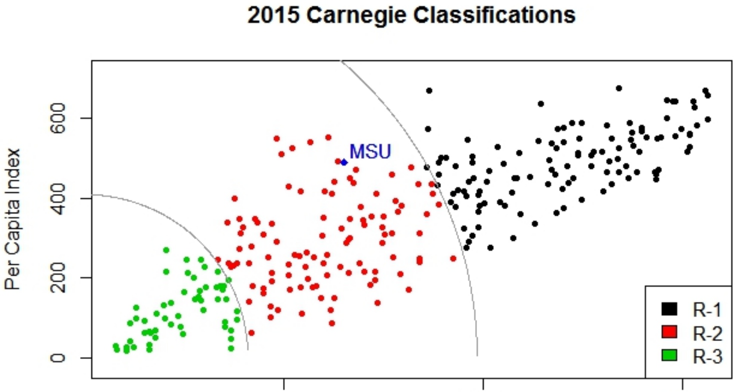
Introduction: What are the Carnegie Classifications?

- The Carnegie Classifications are a metric by which like institutions can be compared.
- Three Classifications of Doctorate-Granting Universities
- Many different classifications of Bachelor's, Associates-only institutions
  In 2015, the classifcations were used on more than 4600 institutions, everything from Stanford University to the Golf Academy of America.

Figure 1: Montana State (R-2), Stanford (R-1), Boise State(R-3):

Montana State University: A History

- Montana State had been classified as R-1: "Very High Research Activity" in 2005 and 2010
- In 2015, Montana State moved to R-2: "High Research Activity"



Montana State's Nearest Neighbors

- The system is designed for **CLASSIFICATION** of institutions, not rankings!
- Institutions near each other can be considered a "peer group" of schools

So what do Montana State's peers look like?

Montana State's Peer Group

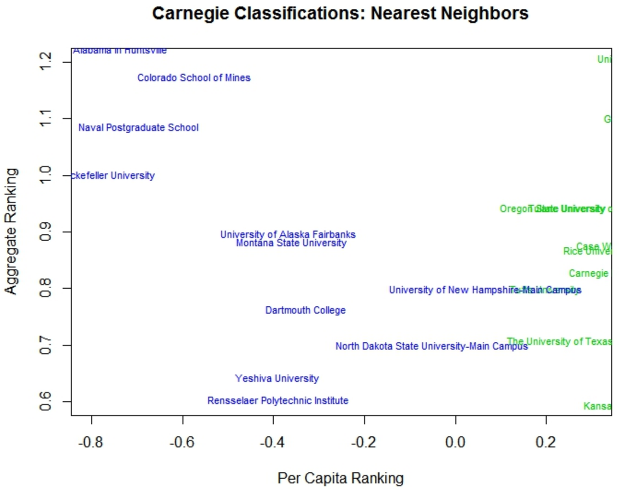


Figure 3: Nearest Institutions to Montana State

How are the Classifications Calculated?

The classifications are calculated based on two indices of institutional output. The first is based on a weighted average of the number of PhDs awarded by the institution; the second is based on a per-capita measurement of research expenditures and research staff.

Aggregate Index:

$$Ag.Index_i = HumPhD_i + StemPhD_i + SocSciPhD_i + OtherPhD_i + StemExp_i + No$$

Per Capita Index:

$$PC.Index_i = \frac{ResearchStaff_i + StemExpenditures_i + NonStemExpenditures_i}{FacultySize_i}$$

Principal Components Analysis

Goal: To reduce p predictors into k components via eigenvalue decomposition. PCA can be done on unscaled raw data or on a scaled covariance matrix.

Given p predictors $x_1, x_2, ...x_p$, we can generate via an eigenvalue decomposition of $\mathbf{X}$ a set of p new variables $y_1, y_2, ..., y_n$. The y's are ordered so that y_1 explains the most variation in the underlying x's, and y_p the least.

Scores: The new set of covariates. These are functions (weighted averages) of the old covariates.
Loadings: The loadings give the formula used to calculate the scores from the original covariates.

But how do we do dimension reduction? Since we know how much variation in x is explained by each y-score, we can make a new factor matrix of some subset of the scores. **The Carnegie Classifications use only the first score from each PCA run.**

PCA: Up for Debate

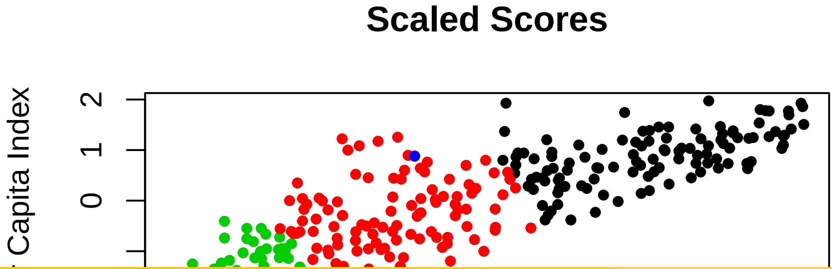
Some of the decisions made by the researchers creating the Carnegie Classifications are worth discussion:

- Using only the **first** PC to create the index. The Aggregate index explains 70% of the variation in the original aggregate variables, the Per-Capita index explains 71% (Borden 2016).
- However, the second PC score explains an additional 25% more variation for the per capita index and 12 percent more for the aggregate index. Should they have included those?
- Finally, is PCA the best way to work with these data? Only 3 variables are used to create the per-capita index and only 7 for the aggregate - dimension reduction is usually more of interest with **big** sets of covariates.

Replicating the Classifications

The scores were calcluated using the following methods:

- Rank each instution on the covariates
- Calculate Principal Component Scores for each index
- Plot the first PC score from the aggregate index vs the first PC score from the per-capita index

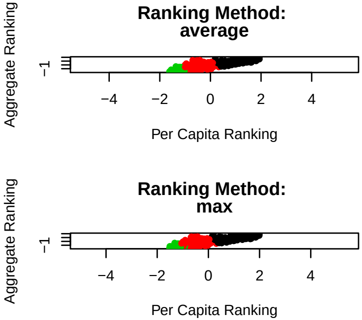


Methods for Ties

Because the scores are based on ranks, **not** the original data. Since most of the variables used are counts, many schools will be tied, especially for the aggregate index. R calculates ties in several ways:

- **Average** (default): If the first 3 schools are tied, each would get rank 1.5. Tie=breaking leads to gain by about half.
- **Minimum**: If the first 3 schools are tied, each would get rank 1. Tie-breaking leads to gain by the number of institutions that are tied.
- **Maximum**: If the first 3 schools are tied, each would get rank 3. Tie-breaking leads to 1-rank unit gain.

Methods for Ties:



Where were lines drawn?

The Carnegie Classifications are based on three groups. However, the data do not form 3 neat clusters.

- In 2010, lines were hand drawn. (Borden 2017)
- In 2015, scores were un-standardized and circles with radii 984.007 and 409.461.
- The choices of radii were somewhat arbitrary.

Unstandardized Scores

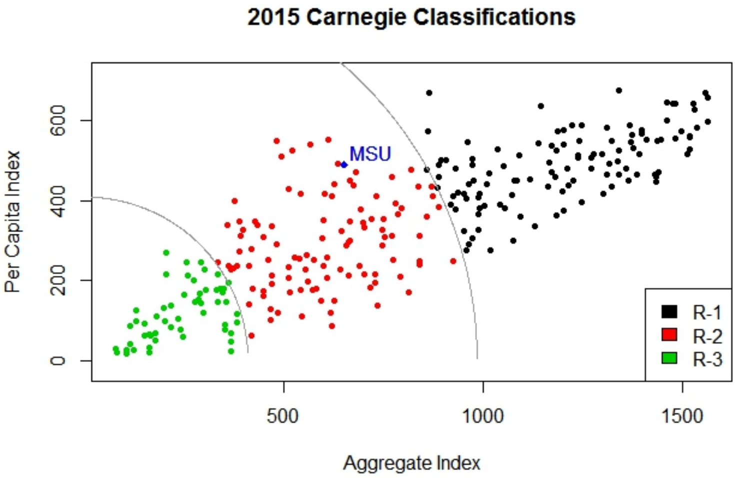


Figure 4: Indexes for each ranking type

Single Metric Changes: Aggregate Index

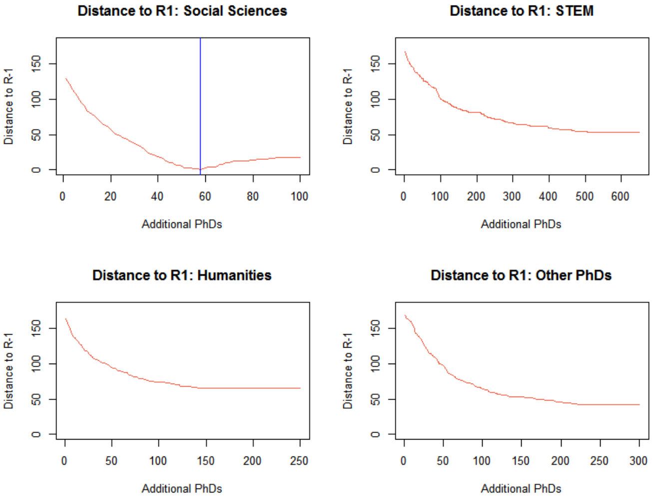


Figure 5: Indexes for each ranking type

Single Metric Changes: Per Capita Index

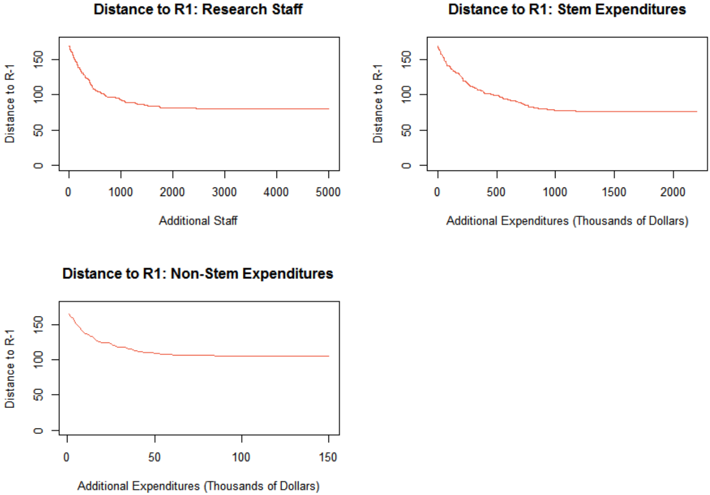


Figure 6: Indexes for each ranking type

Multi-Dimensional Movement

It is clear that to move up, Montana State needs to focus on increasing on **all** dimensions simultaneously. Focusing on the more sensitive metrics (Social Science PhDs) is important, but gains on all metrics together move things more quickly.

- Aggregate variables only move us to the right or left
- Per-Capita Variables move us up or down
- The way that we report/classify research staff could impact classifications

Conclusions

- Maintaining high per-capita metrics is helpful
- While we are closer to R-1 than R-3, getting over the border would involve spending more money and producing more PhDs
- To get to R-1, we need to focus on growing **all** PhD programs, but adding Social Science PhDs would have a big effect: Just a single Social Sciences PhD would help MSU move from rank 1 to rank 62!

Next Steps:

There may be better, less subjective ways to do this. Some changes to the methodology I'm trying are:

- Rather than clustering on the first PC score for two indices, why not include the first two PC scores?
- Model-Based Clustering - lets see how many groups are formed if we let the data speak for themselves

Acknowledgements:

- Dr. Mark Greenwood
- Dr. Christina Fastnow, Dr. Ian Godwin, Rebecca Belou - Office of Planning and Analysis
- Dr. Vic Borden, Indiana University

Questions:

Thanks for coming!