Statistics and Public Policy

ISSN: (Print) 2330-443X (Online) Journal homepage: https://www.tandfonline.com/loi/uspp20

An Alternative to the Carnegie Classifications: Identifying Similar Doctoral Institutions With Structural Equation Models and Clustering

Paul Harmon, Sarah McKnight, Laura Hildreth, Ian Godwin & Mark Greenwood

To cite this article: Paul Harmon, Sarah McKnight, Laura Hildreth, Ian Godwin & Mark Greenwood (2019) An Alternative to the Carnegie Classifications: Identifying Similar Doctoral Institutions With Structural Equation Models and Clustering, Statistics and Public Policy, 6:1, 87-97, DOI: 10.1080/2330443X.2019.1666761

To link to this article: https://doi.org/10.1080/2330443X.2019.1666761

Accepted author version posted online: 13 Sep 2019.
Published online: 16 Oct 2019.

STATISTICS AND PUBLIC POLICY
2019, VOL. 6, NO. 1, 87–97
https://doi.org/10.1080/2330443X.2019.1666761

Taylor & Francis
Taylor & Francis Group

An Alternative to the Carnegie Classifications: Identifying Similar Doctoral Institutions With Structural Equation Models and Clustering

Paul Harmon[a], Sarah McKnight[a], Laura Hildreth[a*], Ian Godwin[b], and Mark Greenwood[a]

[a]Department of Mathematical Sciences, Montana State University, Bozeman, MT; [b]Office of Planning and Analysis, Montana State University, Bozeman, MT

ABSTRACT

The Carnegie Classification of Institutions of Higher Education is a commonly used framework for institutional classification that classifies doctoral-granting schools into three groups based on research productivity. Despite its wide use, the Carnegie methodology involves several shortcomings, including a lack of thorough documentation, subjectively placed thresholds between institutions, and a methodology that is not completely reproducible. We describe the methodology of the 2015 and 2018 updates to the classification and propose an alternative method of classification using the same data that relies on structural equation modeling (SEM) of latent factors rather than principal component-based indices of productivity. In contrast to the Carnegie methodology, we use SEM to obtain a single factor score for each school based on latent metrics of research productivity. Classifications are then made using a univariate model-based clustering algorithm as opposed to subjective thresholding, as is done in the Carnegie methodology. Finally, we present a Shiny web application that demonstrates sensitivity of both the Carnegie Classification and SEM-based classification of a selected university and generates a table of peer institutions in line with the stated goals of the Carnegie Classification.

ARTICLE HISTORY
Received August 2018
Accepted September 2019

KEYWORDS
Carnegie Classification; Clustering; Institutional research; Multivariate statistics; Structural equation modeling

1. Introduction

Institutional classifications are often important to administrators, faculty, and, to a lesser extent, students at institutions of higher education. The Carnegie Classification of Institutions of Higher Education (CC) has sought to describe "institutional diversity in U.S. higher education for the past four and a half decades" (The Carnegie Classification of Institutions of Higher Education, n.d.). Ideally, the CC is used to develop groups of peer institutions for analysis by separating similar institutions into groups.

Since its publication in 1973, the CC has been updated eight times (1976, 1987, 1994, 2000, 2005, 2010, 2015, and 2018) to account for both changes in the "universe of institutions (the result of openings, closings, and mergers) in the United States and the institutions themselves (the result of changes in offerings and activities)" (McCormick and Zhao 2005). In the 2005 update, instead of a single framework representing similarities and differences among institutions, the classification provided a set of six independent frameworks through which similarities and differences can be viewed (McCormick and Zhao 2005) and also introduced the use of a "multi-measure research index to classify doctoral-granting institutions" (Carnegie Foundation for the Advancement of Teaching 2019). Doctoral-granting institutions typically focus on only one of those frameworks: the Basic Classification. While the 2015 update defined "Doctoral Universities" as including those higher education institutions awarding 20 or more research doctoral degrees in a given year, the 2018 update reshaped those categories to include institutions that confer a given volume of professional practice doctoral degrees (JD, MD, PharmD, DPT, DNP, etc.) (The Carnegie Classification of Institutions of Higher Education, n.d.). While the CC system previously sorted doctoral-granting institutions into three classes: R1—highest research activity, R2—higher research activity, and R3—moderate research activity, the most recent update now identifies R1: Doctoral Universities—Very High Research Activity, and R2: Doctoral Universities—High Research Activity institutions as those where at least 20 research/scholarship doctorates were conferred and a minimum of $5 million dollars of total research expenditures during the reporting year, and the D/PU: Doctoral/Professional Universities class as those with at least 30 professional practice doctoral degree conferrals not meeting the above R1/R2 criteria. Thus, the 2018 version only analyzes and classifies the R1 and R2 categories.

The nominal information (classification) provided by the CC is often either unintentionally or intentionally misrepresented as ordinal information (rankings), akin to such institutional rankings as the US News & World Report Best Colleges Rankings, Shanghai Academic Ranking of World Universities, the Times Higher Education World University Rankings, etc.

CONTACT Paul Harmon paulharmongj@gmail.com Department of Mathematical Sciences, Montana State University, Bozeman, MT.
*Current affiliation: Institute for Defense Analyses, Alexandria, VA.

(Herzog 2016). However, as Brint (2013) stated, "classification is a method for apprehending the structure of a system; ranking is a method for stimulating competition among those at a similar level in the system." The distinction has been reemphasized throughout the classification system's 45-year history. McCormick and Zhao (2005) wrote extensively on the issue, and the current director of the Indiana University Center for Postsecondary Research that produces the CC, Victor Borden, almost perpetually states in interviews that "the label should not be viewed as a ranking or rating" (Herzog 2016).

Clark Kerr, one of the principal architects of the classification, "expressed unhappiness about the amount of organizational striving the Classification had encouraged, as institutions lobbied to move up the levels in the Classification" (McCormick and Zhao 2005). Indeed, even among smaller R1 institutions, maintenance of that "status" can often be a priority. This is because it provides them with the opportunity to market their inclusion in the R1 category as an indicator of quality that makes them potentially more desirable to prospective faculty, funding agencies, graduate students, and (to a much lesser extent) undergraduate students. Leading up to the 2018 update, several institutions in the "higher research activity" (R2) group had explicitly set policy goals directed toward improving their standings both after the 2010 update and 2015 updates to the classifications. Some, like Texas Tech University (Rangel 2015), Temple University (Verghese and Jelesiewicz 2016), and Kansas State University (Tidball 2016), set explicit funding and research related goals to transition from R2 to R1 in the 2015 CC update. Others, including the University of Montana (2017) and the University of Idaho (2016), oriented strategic efforts toward transitioning to the R1 class in a later iteration of the classifications.

Recently, Kosar and Scott (2018) attempted to replicate the 2015 CC and illustrated some of the shortcomings of the CC methodology. The purpose of this article is to further discuss the CC method in 2015 and 2018 and present a possible alternative using mixture modeling on scores obtained from a structural equation model, or SEM (Bollen 1989). Much of the research pertaining to processes for institutional classification deals with variable selection: identifying the data and characteristics that most accurately describe each university (Shin, Toutkoushian, and Teichler 2013). However, we are more concerned with the statistical methodology used to group institutions than with the data themselves. Rather than proposing that new variables be added to the dataset to improve or change the classifications, or even changes in the features derived from the data as Kosar and Scott (2018) advocated, we propose utilization of an alternate statistical methodology that provides an easier model to interpret, a platform to assess diagnostics, and an objective way to cluster that aligns with the stated goals of the CC system.

The rest of the article is organized as follows. In the second section, the methodology of the CC in both 2015 and 2018 is addressed and compared to similar principal component analysis (PCA)-based alternatives presented by Kosar and Scott (2018). Further, we outline a proposed SEM that analyzes the institutional data in a latent-variable framework. In Section 3, we overview a clustering methodology that can be used as an alternative to the Carnegie method to group the factor scores from the SEM. Lastly, in Section 4, we introduce a web-based applet that can be used to assess the sensitivity of both classification systems to changes in the characteristics of a single university, along with introducing the idea of exploring neighboring peer institutions as part of the classification system.

2. Classification Methods

2.1. Data

The data used in the CC are published with each update and are found at: *http://carnegieclassifications.iu.edu/* for both the 2015 and 2018 releases. This analysis focuses specifically on the doctoral-granting schools ($n = 334$ in 2015, $n = 417$ in 2018). The dataset comes from the Integrated Postsecondary Education Data System (IPEDS) and the National Science Foundation (NSF) HERD survey, and data represent a point-in-time snapshot.

In 2015, all three groups of institutions were used; however, in 2018, the methodology was changed so that R3 (D/PU) institutions were filtered out of the data as a first step prior to any statistical analysis. This reduced the clustering problem from a 3-group classification problem in 2015 to a 2-group one in 2018. In both updates, institutions with missing data were removed from the classification to avoid missing data problems.

For the analyses that underpin the 2015 and 2018 CC, the data are separated into two distinct sets of variables, one to describe overall doctoral production (called aggregate productivity) and the other to describe doctoral productivity per tenure-track faculty (called per-capita productivity). The aggregate dataset contains seven variables relating to research production, including counts of awarded doctorates in STEM, humanities, social science, and other research doctorates, as well as STEM and non-STEM expenditures (in thousands of dollars), and the headcount of non-faculty research staff with earned doctorates. The per-capita dataset contains both expenditure types and the research staff headcount from the aggregate dataset, but divides them by the headcount of tenured/tenure-track faculty to obtain a per-capita version of each variable.

2.2. Methodology of the CC

Despite the widespread use of the CC, exactly reproducible documentation did not exist in 2015 and is still opaque in the 2018 update. Kosar and Scott (2018) as well as Harmon (2017) provide some insight into the process used to construct them for the 2015 release. In 2018, some changes were implemented to their methodology and were made available (*http://carnegieclassifications.iu.edu/pdf/CCIHE2018_Research_Activity_Index_Method.pdf*).

Prior to analysis, the variables are placed in ascending order and ranked from smallest to largest. In 2015, schools that are tied on any metric are given the minimum rank of the tied group (Harmon 2017; Kosar and Scott 2018). In 2018, tied institutions are assigned the average rank instead (Carnegie Foundation for the Advancement of Teaching 2019). Both are accepted ways to handle ties but these changes can have important impacts if there are many ties encountered; here ties mostly occur in metrics relating to degree conferral counts. Ranking is done

because the values of the variables for some of the largest universities dwarf the values for the smaller institutions in the dataset, leading to many of the variable distributions being highly right-skewed. In previous work, Scott (2011) considered alternatives to the rank transformation by log-transforming the Carnegie variables instead of ranks; however, they did not substantially change the classifications.

PCA is used on the two ranked aggregate and per-capita datasets to create two indices of research performance. PCA uses eigenvector decompositions of each set of variables to create orthogonal axes that explain most of the variation in the underlying variables (Hastie, Tibshirani, and Friedman 2001). The first component from the PCA of the seven aggregate variables is used to form a single aggregate index. Similarly, a per-capita index is generated using the first principal component from the PCA of the three per-capita variables. This process of dimension reduction involves some loss of information; in previous iterations of the CC, these single indices explained between 68% and 72% of the variation in the underlying data and in 2015 specifically, these were 70% and 72% for the aggregate and per-capita scales, respectively (V. Borden, personal communication, January 8, 2016).

In Figure 1, the per-capita PC scores (y-axis) are plotted against the aggregate PC scores (x-axis) for each school, with the 2015 shown at left and the 2018 update at right. The plot of scores in 2015 was partitioned approximately into thirds by hand by drawing concentric arcs that separate institutions into three groups based on areas of "best separation" in the groups (V. Borden, personal communication, February 2, 2017). In 2015, schools in the bottom left corner of the plot are in the R3 category, and those in the top-right corner are in the R1 category, with the R2 category in the middle. Similarly, in 2018, the schools in the lower left are in the R2 category and the R1 institutions remain on the upper right, with R3 schools not shown. Because both the scores and the boundaries can change in different years, it is possible for schools to move between categories from one update to the next. This can occur based either on changes in where researchers decide to place the arcs or on substantive changes in individual institutional productivity, ranks, or shifts in the PCAs that then alter the relative locations of institutions.

2.3. Shortcomings of the CC

The CC has the potential to be used by administrators on campuses to drive institutional goals and academic development. However, the CC is marked by several statistical concerns. First, it is based on highly variable single-year snapshot data. Second, it uses unsupervised dimension reduction, leading to information loss. Finally, its methodology is not entirely transparent.

Because of this, universities that prioritize moving from one category to another must account for possible changes in the way that future calculations are done. Since the data used to calculate the CC in any given year are based on ranked snapshots at a single time point, the weighting of a single variable can change from year to year. If substantial changes were to occur in the characteristics of a large proportion of schools in the calculation, it is possible that the loading for a given variable could be noticeably different in the next release of the classifications. While unlikely to be extreme, this can affect the impacts of changes in ranks on individual variables because variables may not carry the same amount of weight in each year. For instance, a school may determine based on the weights used in an update of the classifications that it needs to gain a certain number of doctorates in STEM and social science and increase STEM expenditures by a substantial, but attainable, amount. That school might implement those changes, only to find out that because of shifts in the underlying PCA that generates each school's score, those changes were not actually aligned with the most important variables in the updated classifications and thus changes that would have proven effective in the current year to change categories may not be sufficient in the new one. This problem is exacerbated for non-STEM programs because they are broken out into separate components and have separate loadings that can change. Kosar and Scott (2018) present a solution that inputs summed non-STEM PhD counts, mitigating the number of sensitive loadings going into the aggregate index but not completely solving the issue.

Moreover, the PCA-based methodology of the CC reduces data dimensionality in a way that cannot be directly controlled, leading to two problems. First, creating a single index from seven (or even three) variables leads to information loss in the form of unexplained variation in the original variables since a single variable cannot contain all the information in the under-

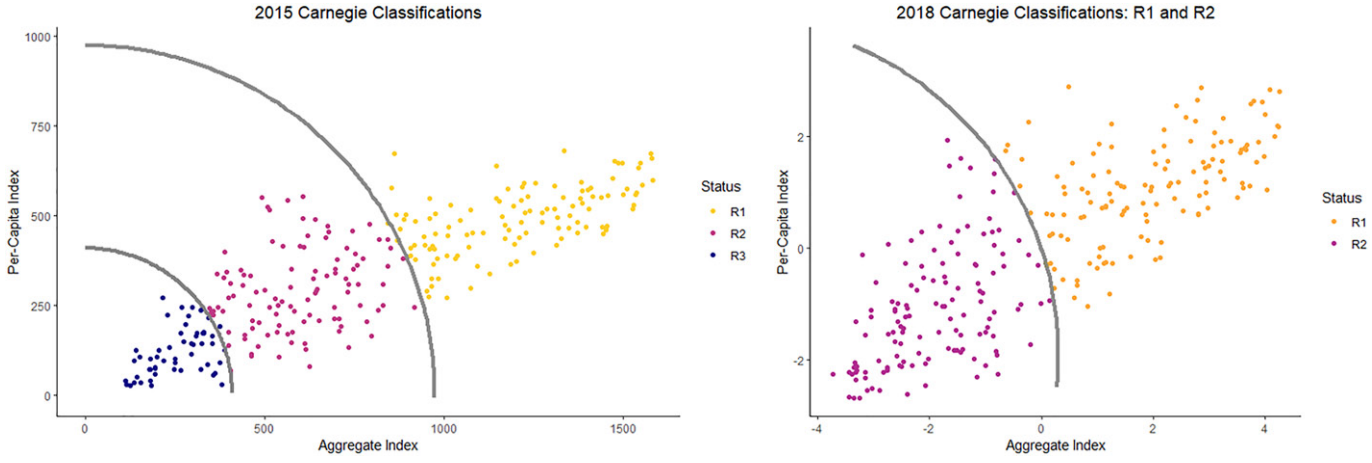

Figure 1. The 2015 and 2018 updates of the CC, with 3 groups in 2015 and 2 groups in 2018 (borders are approximate).

lying data. Second, using just the first principal component is not guaranteed to explain a consistent amount of variation in the underlying variables across years. It is possible that the data could be substantially less or more correlated in future years; in such a case, the amount of variation explained by the single index has the potential to be much lower or much higher, respectively. Interpretation of the PCs could therefore change as well, meaning that each index might have different meanings in each release. Thus, there is not currently a way to directly compare the results from one edition of the CC to another.

Finally, the lack of transparency and statistical objectivity in group membership makes the exact group membership difficult to replicate, even with published work done on the subject (Kosar and Scott 2018). In particular, the arcs used to determine the boundaries can be critiqued both in the choice of their location (length of the diameter of the circles) and the origin of the circle (relative orientation of arcs to data points). If the data are poorly separated and points lie close together in the CC map, a common problem that is exacerbated by the ranking step underlying the scores, determining an optimal place for the lines to be drawn is ambiguous at best. Schools could move back and forth in different classification groups without any (substantive) change in their characteristics or those around them.

2.4. SEM-Based Classifications

Rather than using PCA, we propose using SEM and using estimated factor scores which can then be used to classify institutions (Bollen 1989). In contrast, Kosar and Scott (2018) outline two PCA-based alternatives to the current Carnegie method; both use different formulations of the data matrix as inputs to a more-traditionally applied PCA and then build classifications off of rotations of the first two principal components.

The rotations are designed to provide a single PCA with two components that can match the aggregate and per-capita indices of research productivity (not necessarily the principal components themselves). This has the advantage of creating two indices from a single analysis; however, this method does not allow for easy diagnosis of model fit (other than percent of variation explained) nor will it directly lead to clusters of schools with common characteristics without some additional rotation.

In contrast, SEM with an objective clustering algorithm provides tools for diagnosing model fit and produces clusters that are aligned with research goals. The goal is not to create purely nominal clusters but rather to create clusters that are defensible, reproducible, and objective (or at least more objective than CC). Admittedly, such clusters have the potential to perpetuate the same myth of ordinality as the CC, but only to preserve interpretability of group membership and make for a comparable substitute.

SEM is a statistical methodology that allows for the modeling of simultaneous equations, the use of latent or unobserved variables, and the use of variables measured with error. A typical SEM consists of two parts: the latent variable model which describes the relationships among latent variables and the measurement model which relates the latent variables to their indicators or items. Compared to the PCA-based methods used in the CC and by Kosar and Scott (2018), this method has two notable advantages. First, SEM allows for modeling of correlated latent factors as opposed to orthogonal ones created by PCA (or rotated PCA), allowing a single model to replace the more complex pair of PCAs necessary to calculate aggregate and per-capita indices or Kosar and Scott's rotation with orthogonal components. Even more importantly, SEM includes diagnostic tools that do not exist to assess the efficacy of PCA-based methods as well as methods for estimation for SEMs when missing data are present (discussed later). All SEMs were fit using `lavaan` (Rosseel 2012) in statistical software package R (R Core Team 2018). All path diagrams were created using `OnyxR` (von Oertzen, Brandmaier, and Tsang 2015).

We first considered a SEM where we attempted to replicate the aggregate and per-capita indices as latent factors, as depicted in Figure 2. When fitting a SEM with aggregate and per capita productivity as latent factors, the algorithm does not converge. This is due to the high correlation of the aggregate and per-

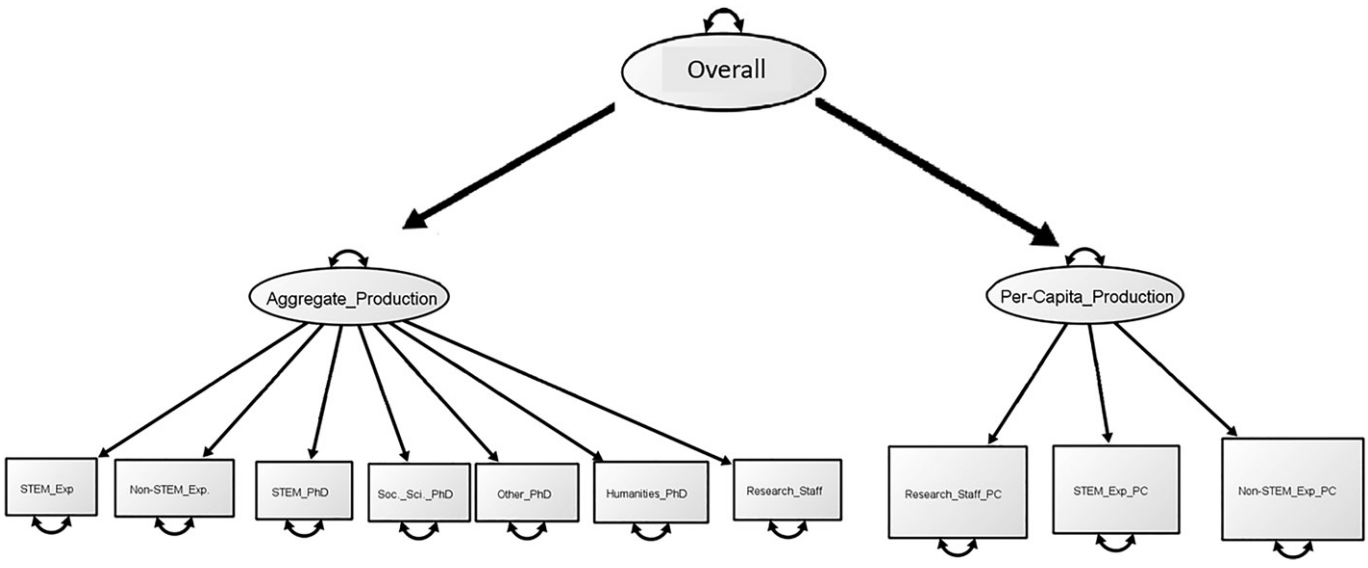

Figure 2. Path diagram of the CC, modeling two latent factors, one for aggregate production and the other for per-capita production.

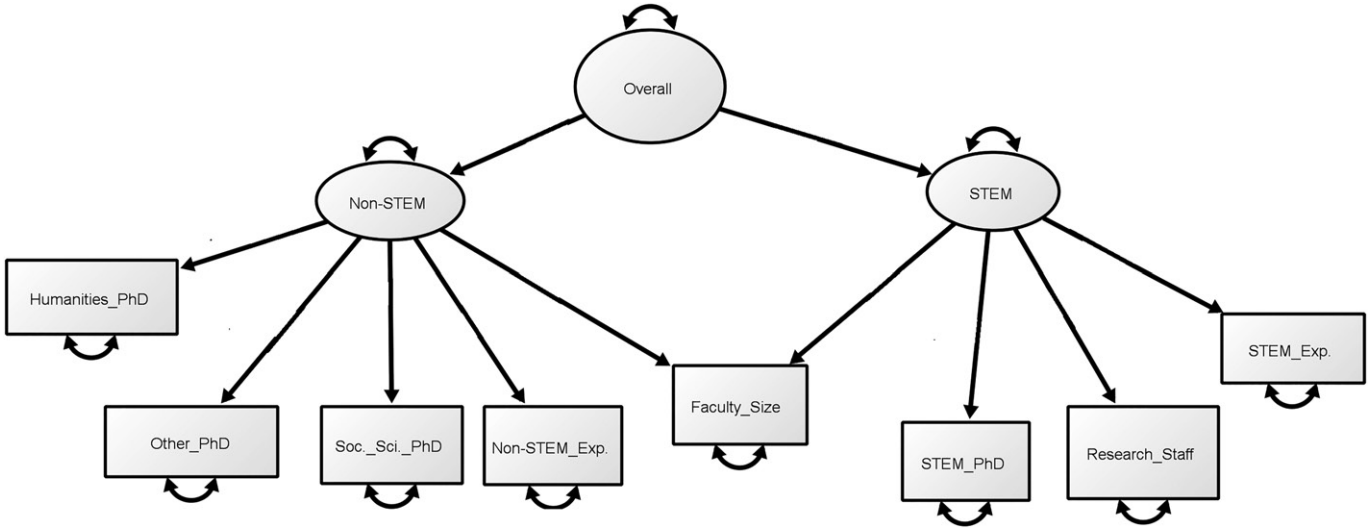

Figure 3. Our SEM specification, which models STEM and non-STEM production as two separate latent traits.

capita manifest variables (0.92, 0.88, and 0.96 for STEM, non-STEM, and research staff, respectively). These factors share nearly the same set of indicators, resulting in a model specification problem not unlike extreme multicollinearity in a linear regression setting.

Because the Carnegie indices cannot be reliably reproduced in a latent-variable framework, we instead formulate an alternate latent variable structure. As depicted in Figure 3, we use a second-order latent factor model. The first-order latent factors are STEM productivity and non-STEM productivity. These latent factors are assumed to be measures of the second-order factor for overall productivity. We chose to use STEM and non-STEM productivity as latent factors as opposed to two factors for aggregate productivity and per capita productivity for several reasons. First, it provides a single estimated factor that will be useful for comparing the institutions on a univariate scale. Second, this model provides a more intuitive set of factors to use to develop the classifications.

The use of a second-order factor allows for the latent STEM and non-STEM productivity factors to be distinct but related concepts that can be accounted for by one underlying factor (Chen, Sousa, and West 2005) of overall institutional productivity. This allows for easier interpretation of this model and allows us to obtain a single score for overall productivity as opposed to two scores used by the CC.

The second part of a SEM is the measurement model that relates the STEM and non-STEM latent factors to their items. As shown in Figure 3, the items for STEM productivity are STEM PhDs produced, STEM expenditures, and research staff size while the items for non-STEM productivity are humanities PhDs produced, social science PhDs produced, other PhDs produced, and non-STEM research expenditures. We opted to use research staff size as a measure of STEM productivity but not non-STEM productivity as research staff typically are employed in STEM fields and tend to be associated with labs and other physical infrastructure. Although we initially tried using per-capita features in line with the CC, we found that they did not add much additional information to the model after incorporating the aggregate variables. Instead, the number of tenured and tenure-track faculty is cross-loaded as an item for both STEM and non-STEM productivity as the total number of tenure/tenure-track faculty is clearly related to productivity in STEM and non-STEM fields (and some faculty may produce in both fields). Note this also mitigates the potential for perverse incentives relating to reducing faculty size to raise per-capita scores without increasing production in the CC. While our method penalizes smaller specialized schools more than the CC, for most institutions, those differences were not compelling enough to include the per capita versions of variables.

We chose to use a specific variable for an item of a given latent factor as the choice intuitively makes sense. Our choice of items is further confirmed when examining the correlation matrix of the items in Figure 4 built using the R package `corrplot` (Wei and Simko 2017) as the items of the latent factor for STEM productivity are highly correlated and the items for the latent factor for non-STEM productivity are also highly correlated while

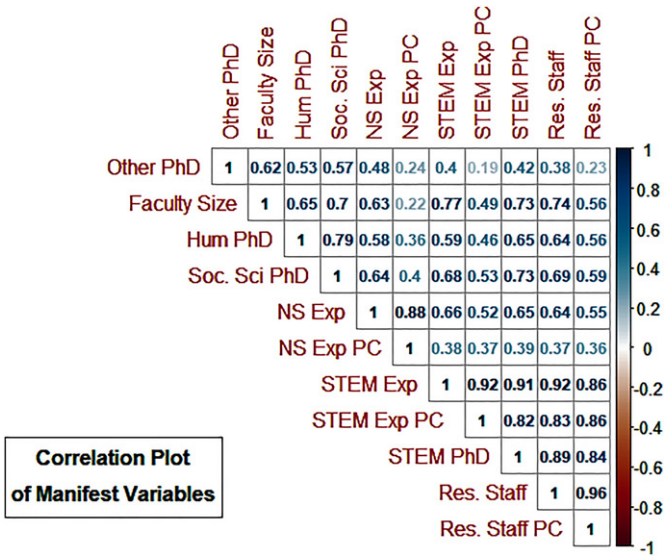

Figure 4. Pearson correlations of the observed ranked variables for the 2015 data. Many of the observed variables are highly correlated, especially with their per-capita counterparts.

Table 1. Standardized loading estimates, test statistics, and *p*-values for the SEM on 2015 data.

	Variable	Std. estimate	*Z*-value	*p*-value
STEM				
	STEM PHD	0.939	3.596	<0.001
	Research Staff	0.953	3.883	<0.001
	STEM Exp	0.967	3.676	<0.001
	Faculty Size	0.400	2.279	0.001
Non-STEM				
	Humanities PhD	0.847	3.703	0.006
	Other PhD	0.639	3.477	0.010
	Social science PhD	0.906	3.740	0.006
	Non-STEM Exp.	0.729	3.712	0.004
	Faculty Size	0.482	2.584	0.007
Overall				
	STEM	0.889	3.007	0.001
	Non-STEM	0.895	2.905	0.03

NOTE: Latent variables are shown in italics.

items of different factors are at most moderately correlated. The number of tenure and tenure-track faculty is moderately to highly correlated with items for both latent factors, with an average Pearson correlation of the ranked data of 0.77 for STEM factors and 0.65 for non-STEM factors. Note that none of the indicators were negatively correlated with each other, so variables are all oriented in the same direction, meaning larger scores indicate higher productivity.

The model in Figure 3 was fit to the 2015 CC data. Standardized parameter estimates for the hypothesized model in 2015 are displayed in Table 1. These results indicate that the hypothesized model fits the data moderately well ($\chi^2 = 110.024$ with 17 df, RMSEA = 0.141, CFI = 0.958). The standardized factor loadings are all above 0.7, with the exception of number of tenure/tenure-track faculty as it cross-loads on both latent factors (factor loadings are 0.482 and 0.400 for STEM and non-STEM productivity, respectively) and number of other PhDs produced (0.639), which indicates that most of the variability in each of the items is explained by its associated latent factor. The largest standardized factor loadings were for the number of Humanities PhDs awarded on the non-STEM factor. For the STEM productivity variables, the standardized loadings were 0.939, 0.953, and 0.967 for the number of STEM PhDs awarded, research staff, and STEM expenditures, respectively. The path coefficients relating overall productivity to STEM and non-STEM productivity are 0.900 and 0.883, respectively. This indicates that the variability of STEM and non-STEM productivity is largely explained by overall productivity. These results are consistent with what is expected.

To compare institutions, ideally we would compare the values of the latent factor for overall productivity. Because latent factors are unobserved by definition, these values must be estimated. This is done by creating factor scores which can then be used in subsequent analyses (DiStefano, Zhu, and Mindrila 2009). Factor scores are computed using a weighted average of the items with a number of options available for weighting. The most common method used to calculate factor scores is Bartlett's method (Bartlett 1937) as it leads to unbiased estimators of the true factor scores. In subsequent analyses, we use the factor scores created using Bartlett's method.

3. Classification Using SEM and Clustering

As noted previously, the CC is based on subjective decisions rather than an objective process. By using two PCAs from similar information, the two indices must be combined in some fashion to create groups. Kosar and Scott (2018) advocate for the creation of two independent indices from a single PCA rather than the correlated indices used by the CC. In either case, the two indices still have to be combined to create groups. The groups should split based on shared overall results from some or all of the original multivariate data; however, this is difficult to do objectively because the resulting clusters may not necessarily create intuitive groupings. Our SEM-based method is easily and reliably reproduced using the methodology outlined in this article. Classification relies on a univariate clustering algorithm that is better aligned with research goals, and changes to the data that substantively change the model can be assessed and even tested for across different iterations of the estimated model.

We suggest a clustering-based approach to identifying groups of institutions that are similar within group and different between using a method such as k-means or model-based clustering. A problem with clustering in two dimensions (as would be required to use the two indices from the CC or the methods of Kosar and Scott (2018)) is that the clusters may not be organized along a common direction of "higher" quality based on the original ranking values. In fact, some of our attempts to cluster the two-dimensional scores used in the CC led to clusters that were oriented from upper left (low aggregate, high per-capita) to lower right (high aggregate, low per-capita). In such a case, the groups were not tied directly to overall productivity. Unsupervised classification methods tend to identify groups in data but those groupings cannot be controlled. They may not correspond to desired groupings nor match the constrained splits used to divide schools in the CC based on the implicit belief that more aggregate and per-capita production is better.

From a clustering perspective, the SEM model described above has a notable advantage over developing two-dimensional scores because it creates a single score. This means that the schools are ordered from lowest to highest; they are given a numerical value that can be used to identify groups of schools while maintaining a clear ordering despite using unsupervised classification methods.

3.1. Model-Based Clustering

In this situation, identifying groups of observations in a single-dimension is particularly well-suited to considering the distribution of scores to be made up of a finite mixture of a suite of univariate normal distributions with different centers and possibly different variances. Moreover, this constitutes the key difference between the solution proposed by Kosar and Scott (2018) and our SEM-based classifications. The proposed PCA-based solutions require clustering algorithms to operate on two (or more) dimensions as opposed to a single variable in our approach. There are dozens of algorithms for clustering data (Everitt and Hothorn 2011). We employ mixture modeling (also called latent variable modeling) to estimate an overall distribution based on the mixtures of these score distributions as well as

to identify the group memberships and, importantly, quantify uncertainty in those group memberships (Banfield and Raftery 1993).

The basic idea for univariate (single variable) mixture modeling is to define the overall density of the n observations in the vector y as

$$f(y, \psi) = \sum_{i=1}^{g} \pi_i f(y_i),$$

where π_i are the mixing proportions across the g clusters and sum to 1, $f_i(y)$ is the density of the ith cluster, and ψ is a vector of parameters used to define the mixture. The densities are assumed to be normally distributed with different means and variances that are either the same or allowed to differ across the clusters as defined in ψ. The densities are estimated based on the observations assigned to each cluster. The challenge in this problem is to define the cluster memberships of the n observations; the EM algorithm is used to iteratively search for optimal allocations and, given a current allocation, estimate the mean, variance, and mixing proportions for the different clusters. This process results in optimal choices of cluster assignments and densities of clusters for a given choice of g and whether variances are assumed to be the same or differ by cluster. More detailed discussion can be found in Banfield and Raftery (1993) and Fraley and Raftery (2002).

We used mixture modeling as defined in the `mclust` R package (Fraley et al. 2012) which uses the Bayesian information criterion (BIC) to select both the optimal number of clusters and whether the clusters are best described with the same variance or different variances (in this case, bigger BICs are better). Results of this exploration are in Figure 5(a), where the mixture model with three groups and equal variances was the optimal cluster solution based on BICs. This is fortuitous as it aligns with the number of groups used in the 2015 CC but it is possible to consider an optimal solution for a predefined number of clusters. The clusters themselves are displayed in Figure 5(b), showing the non-overlapping low, medium, and high splits for the three clusters. Figure 5(c) shows that at the boundaries of the clusters, the cluster assignments become less certain (the uncertainty is defined as 1 minus the estimated probability of the assignment of the observation to the identified cluster, so taller bars indicate less certainty about group membership). It seems reasonable to assume that the classifications for the standard Carnegie results are less certain near the boundaries. Unfortunately, arbitrarily defined boundaries do not lend themselves to quantification of such uncertainty. Figure 5(d) shows the estimated distribution of scores based on the selected mixture model.

4. Demonstration and Shiny App

It is of interest to determine how sensitive results are to changes in the underlying data, especially since institutions routinely implement policies intended to change locations in the CC (among other ranking/classification systems). We independently developed a Shiny (Chang et al. 2017) application in R designed to allow the user to select a school and assess the sensitivity of that school's classification to changes in the underlying variables for both the Carnegie and SEM-based methods and have modified it to be similar in style to the one developed by Kosar and Scott (2018). In much of the same way that their application functions, the user can select a school and then use a slide bar to either increase or decrease the number of PhDs awarded in each category, research staff size, or research

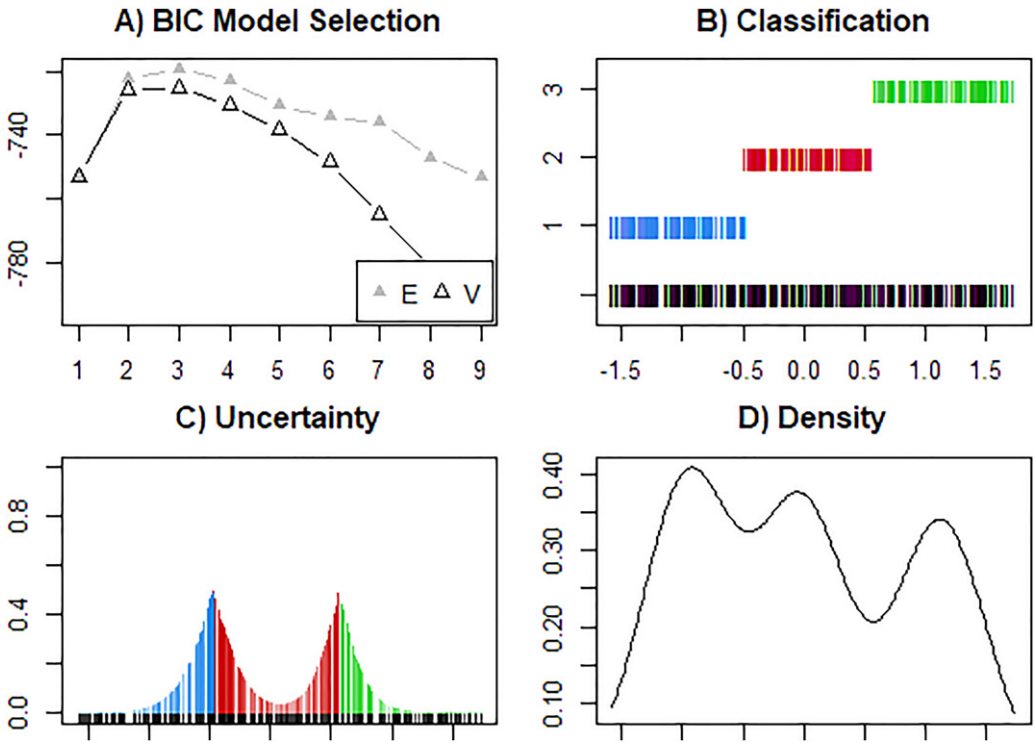

Figure 5. (a) Plot of BIC values associated with normal mixtures. (b) Plot of cluster membership. (c) Uncertainty plot associated with each cluster. (d) Density plot based on clusters.

expenditures. Sliders can be reset to the true value by pressing the reset button. Motivated by the stated goals of the CC, the application also allows for identification of a five-school cohort of objectively selected peer institutions based on the SEM factor scores.

Figure 6 demonstrates how this application (found at: *https://ccsemclassifications.shinyapps.io/SEM_App2/*) can be used by administrators and institutional researchers to assess the efficacy of a proposed policy change. As an example, we highlight Oregon State University (OSU) in Figure 6, which is classified in the SEM-medium group in the SEM-based classification and R1 in the CC.

Changes can be made to either a single variable, all of them, or a selected few. The application takes the user input and recalculates the PCA-based indices or SEM-based results on the new dataset, depending on which tab in the app is being used. It then shows where the university would be relative to other schools in that update. In Figure 7, we show the effect of substantive changes to the counts of PhDs awarded at the selected institution. In 2015, if OSU had awarded an additional 15 PhDs

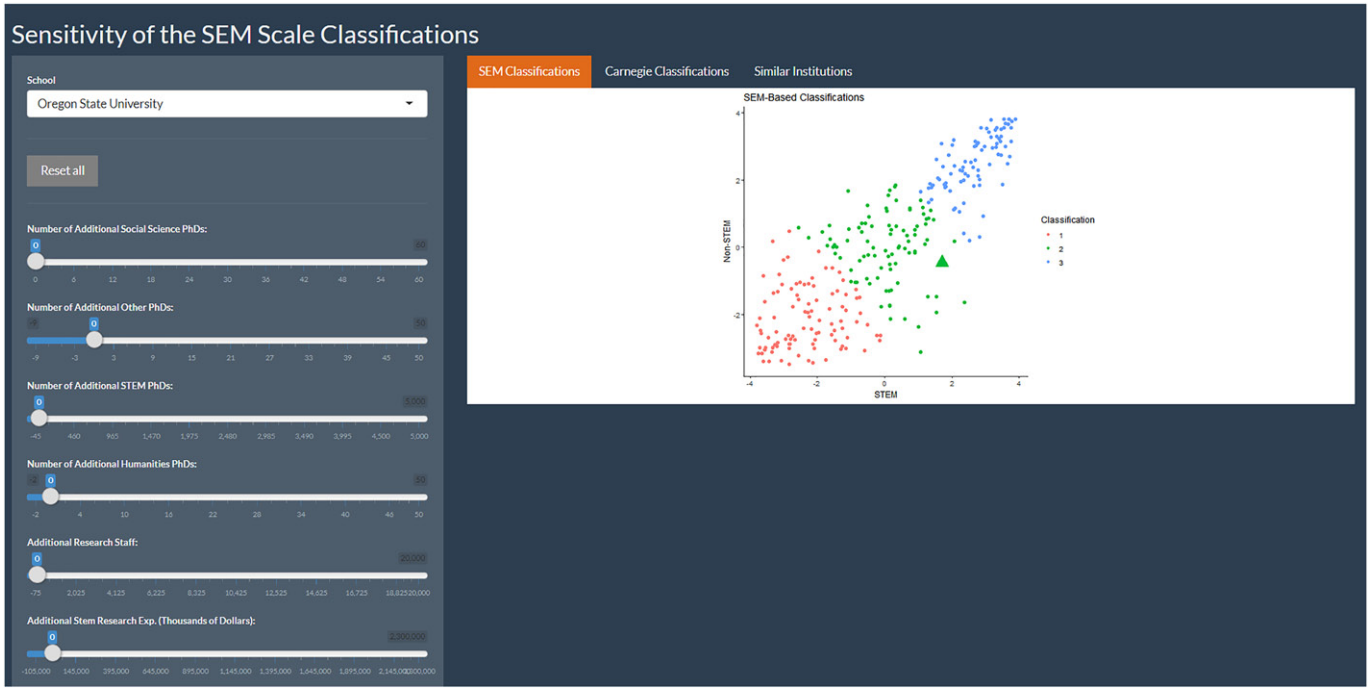

Figure 6. Screenshot of the SEM-classifications application assessing Oregon State University in 2015 data.

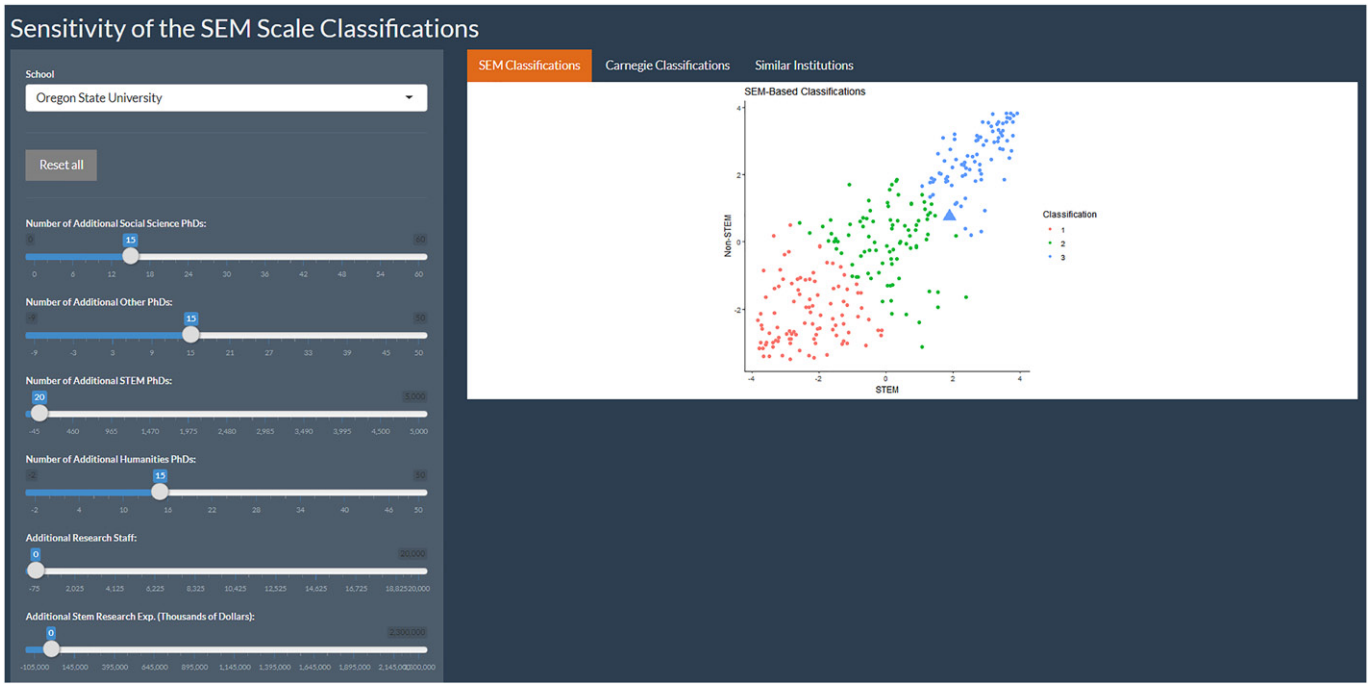

Figure 7. After adding PhDs to Oregon State, it is now classified in the first group of the SEM-based classifications in 2015 data.

in social science and other categories, as well as 20 STEM PhDs and 15 Humanities PhDs, their score on the univariate factor of factors would place them in the SEM-large group.

The app also gives a table of the selected institution's nearest neighbors (on the overall factor) for a comparison peer set. Simply click on the "Similar Institutions" tab to see a dynamic list of nearest neighbors for a given school. Schools with slightly higher factor scores are considered "aspirational" peers and neighbors with slightly lower scores are considered "peer" institutions.

For administrators at institutions, an interactive application presents several advantages. First, users can explore and assess sensitive spots in both classification systems. Administrators can test hypothetical policy actions to compare the efficacy of such actions across both classification systems. Moreover, the applications can be used to inform realistic, feasible policy goals at a given institution. Finally, the availability of all code and data underlying the Shiny app makes the results transparent and reproducible and would make updating the results to a new dataset relatively trivial.

5. Summary and Conclusions

5.1. Comparing Carnegie Versus SEM Classifications

It is useful to illustrate differences in classification between the CC and our SEM analogue. While they function in similar ways, the SEM classifications do not deal with institutional size in the same way that the CC does; rather, the SEM classifications more harshly penalize smaller institutions based on doctoral productivity and faculty size. This drives separation in institutions (a goal of the analysis) but does tend to put smaller schools farther away from large ones.

Table 2 provides the comparison between the CC and SEM-based classifications. None of the institutions that were categorized as R1 in the 2015 CC were categorized by the SEM-method in the SEM-small group; however, 31 of them were moved into the SEM-medium classification. Within the 2015 CC R2 schools, 46 institutions were moved into the SEM-small class. All of the CC R3 institutions were classified in the SEM-small class. The remaining 58 institutions were omitted from the analysis because they do not contain information about STEM and non-STEM expenditures. These are classified as R3 institutions in the CC and are similarly classified in the SEM-small group.

The SEM classifications tended to separate large institutions from small ones more than the CC. No institutions in the CC R2 or R3 groups were present in the SEM Large category in 2015. Some notable R1 institutions that would be reclassified into the SEM-medium class based on the SEM-based classification include: Rice University, Carnegie Mellon University, University of Oregon, and Tufts University. While it may be surprising to see Rice and Carnegie Mellon, for instance, in the SEM-middle category, this reflects the more severe penalty the model places on overall institutional productivity relative to the CC. Indeed, Rice and Carnegie Mellon are both on the border of R1 and R2 in the Carnegie space. The SEM-based classifications contain nearly 47% of the institutions in the SEM-small group whereas the CC categorizes roughly 33% of the schools into the R3 group.

Table 2. SEM classifications versus CC groups indicate some disagreement about the middle group of institutions in 2015.

	R1	R2	R3
SEM large	84	0	0
SEM medium	31	61	0
SEM small	0	46	54

Table 3. SEM classifications versus CC groups in 2018 indicate better agreement than in 2015.

	R1	R2
SEM large	110	0
SEM medium	20	131

In Table 3, we provide a direct comparison to the R1 and R2 institutions, with the R3 institutions held out as in the 2018 CC. Using the same framework, we applied the SEM to the 2018 CC data and clustered schools. There were more R1 and R2 institutions in 2018 than in 2015, so the cluster sizes are higher in 2018. The CC and SEM-based classifications largely agree, with only 20 institutions from the R1 category being classified in the SEM-medium group. This improved agreement may be due to the removal of smaller institutions in the 2018 data.

5.2. Advantages of SEM-Based Classifications

The rotated solution of Kosar and Scott (2018) is a notable improvement over using two separate PCAs to generate the scores for the CC system that retain similar definitions while directly assessing the overall percentage of variation explained. However, this approach does not necessarily lead to easily defined groupings any more than the original Carnegie methodology. We employed SEM methods to first assess an equivalent approach using the same structure their loadings' target and the two separate PCAs of the CC target—but found that there is too much shared information to estimate that structure using SEM. Instead, we propose to separate STEM and non-STEM components where possible and share information where it is not, and then create a factor of factors to map these two correlated but somewhat independent factors into a single scale that can be easily and explicitly used for classification.

Finally, the SEM-framework provides a robust set of tools to deal with the missing data problem that leads to the removal of some R3 institutions in 2015 and all R3 schools in 2018 (Allison 2003). SEMs can be estimated using Full Information Maximum Likelihood estimation under an assumption that data are missing at random (Cham et al. 2017). Multiple imputation could be used to estimated scores for all institutions in the data, even those with a few missing observations (van Buuren and Groothuis-Oudshoorn 2011).

SEM-based classifications do not solve every shortcoming of the CC. They rely on the same data and are therefore subject to the same measurement problems, possible biases, and other data-related issues. In the same way that the PCA loadings can change from one release to another, the SEM's weights relating items to latent factors can also change. However, unlike PCA, the SEM is able to test for different structures of loadings in two datasets, such as between 2015 and 2018 (Beaujean 2014).

The SEM-based classification method allows for determining classifications based on a single factor-of-factor score rather than on hand-drawn delineations. While any automated method for determining both the number of clusters and cluster membership has the possibility of selecting either an overly complicated solution or too few groups, the mixture model resulted in a reasonable three-group solution for the 2015 data. It is also possible to fix the number of clusters if necessary; however, we did not need to do so. The ability to measure uncertainty in an institution's classification is also beneficial. For administrators at schools on either side of a partition, it is useful to know if their institution's classification is highly uncertain or if it is relatively well-classified in a certain group. Further, while one could apply bivariate clustering algorithms to the results from the CC (e.g., Kosar and Scott 2018), many of the two-variable clustering methods give results that are not well-behaved relative to the research goals.

Finally, this method of institutional classification is well documented and reproducible. It can be applied to new datasets and consistently compared. The Shiny applications created to assess the sensitivity of each classification method can be used to identify a cohort of similar institutions, determine key differences between institutions, and predict the outcome of certain policy actions.

5.3. Limitations

The SEM-based approach provides a robust set of tools for posing the CC problem in a setting that can handle missing data, model correlated indices of production, and provide a more intuitive univariate clustering solution. However, like any method, it is not without drawbacks and limitations. For one, there is no intuitive way to control for per-capita production in the model. Because per-capita features are so highly correlated with their aggregate counterparts for most of the institutions in the data, they cannot be modeled as an additional latent factor and they do not substantively change the resulting clusters when mapped to the existing latent factors. Thus, the SEM-based classifications focus more heavily on raw aggregate production than the CC does.

A critique of this approach (and that of the CC) is that we are proposing a two-step method where we first define scores on this new factor of factors and then perform clustering of these scores as opposed to directly allocating schools to different clusters from the original measurements or rankings. While a more direct approach has advantages in terms of using the original information for creating groups, the unsupervised nature of the group allocation, especially in higher dimensional spaces, makes it hard to get cluster solutions that always match with expectations and desires for a school-to-group allocation model. By taking a first step to create a reasonable single dimensional scale, the mixture model is able to allocate schools into groups along this single gradient. Results that allow exploration of both the optimal number of clusters and the uncertainty of the classifications at the boundary provide tangible improvements over the arbitrariness of both choices in the CC method. Further, it allows for group determination that is clearly better aligned with the research goals than other approaches.

5.4. Further Research

This research addresses a statistical question, not a qualitative one. The CC (especially for doctoral institutions) has changed over time both in their explicit efforts to de-emphasize hierarchical interpretations by institutions as well as their interest in capturing the more nuanced differences between institutions and the classes that they fall within, and it is likely that they will change in the future. Indeed, the 2018 update reshapes the "membership of the Doctoral Universities and Master's Colleges and Universities categories... to accommodate 'Doctor's degree: professional practice' within their methodology" (Carnegie Foundation for the Advancement of Teaching 2019), among other changes. The efficacy of including different variables into the SEM-based classification system is not something that we assessed; however, the SEM method would allow for addition of new items or more latent factors. This would allow for the integration of professional doctorates into the model, analogous to Carnegie's inclusion of them in 2018.

Other methods could also be applied to the classification stage. Mixture modeling is not the only method for determining clusters with univariate data, even though it does provide a reasonable solution here. A one-step solution that combines building the latent factors and binning the factor-of-factor scores into three groups, if it were possible, would be ideal. Further work could focus on comparing different univariate clustering on the SEM scores or even focus on bivariate clustering algorithms that are constrained to provide more reasonable groups with the original Carnegie indices.

Supplementary Materials

Shiny Application: A Shiny application designed to assess the sensitivity of the SEM classifications, as well as to define a cohort of similar institutions given a proposed policy change. (Availeble at: *https://ccsemclassifications.shinyapps.io/SEM_App2/*)

Carnegie Dataset: Dataset used in the development of the CC, the SEM-classifications, and related Shiny app. (Availeble at: *http://carnegieclassifications.iu.edu/*)

ORCID

Paul Harmon http://orcid.org/0000-0003-0724-6938
Mark Greenwood http://orcid.org/0000-0001-6933-1201

References

Allison, P. (2003), "Missing Data Techniques for Structural Equation Modeling," *Journal of Abnormal Psychology*, 112, 545–557. [95]

Banfield, J., and Raftery, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [93]

Bartlett, M. (1937), "The Statistical Conception of Mental Factors," *British Journal of Psychology*, 28, 97–104. [92]

Beaujean, A. (2014), *Latent Variable Modeling Using F: A Step-by-Step Guide*, New York, NY: Routledge/Taylor & Francis Group. [95]

Bollen, K. (1989), *Structural Equations With Latent Variables*, New York: Wiley. [88,90]

Brint, S. (2013), "A Priori and Empirical Approaches to the Classification of Higher Education Institutions: The United States Case," *Pensamiento Educativo*, 50, 96–114. [88]

Carnegie Foundation for the Advancement of Teaching (2019), *The Carnegie Classification of Institutions of Higher Education*, Bloomington, IN: Indiana University Center for Postsecondary Research. Avail-

able at http://carnegieclassifications.iu.edu/pdf/CCIHE2018_Research_Activity_Index_Method.pdf. [87,88,96]

Cham, H., Reshetnyak, E., Rosenfeld, B., and Breitbart, W. (2017), "Full Information Maximum Likelihood Estimation for Latent Variable Interactions With Incomplete Indicators," *Multivariate Behavioral Research*, 52, 12–30. [95]

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017), "Shiny: Web Application Framework for R," R Package Version 1.0.5. [93]

Chen, F., Sousa, K., and West, S. (2005), "Testing Measurement Invariance of Second-Order Factor Models," *Structural Equation Modeling*, 12, 471–492. [91]

DiStefano, C., Zhu, M., and Mindrila, D. (2009), "Understanding and Using Factor Scores: Considerations for the Applied Researcher," *Practical Assessment, Research and Evaluation*, 14, 1–11. [92]

Everitt, B., and Hothorn, T. (2011), *An Introduction to Applied Multivariate Analysis With R*, New York: Springer. [92]

Fraley, C., and Raftery, A. (2002), "Model-Based Clustering, Discriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [93]

Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. (2012), "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation," *The R Journal*, 8(1), 205–233. [93]

Harmon, P. (2017), "Demystifying the Carnegie Classifications," Montana State University. MS Statistics Writing Project, available at *http://www.math.montana.edu/graduate/writing-projects/2017/Harmon17.pdf*. [88]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, New York: Springer New York Inc. [89]

Herzog, K. (2016), "UW-Milwaukee Elevated to Elite Status as Research University," *Milwaukee Journal Sentinel*, Milwaukee, WI. [88]

Kosar, R., and Scott, D. (2018), "Examining the Carnegie Classification Methodology for Research Universities," *Statistics and Public Policy*, 5, 1–12. [88,89,90,92,93,95,96]

McCormick, A., and Zhao, C. (2005), "Rethinking and Reframing the Carnegie Classification," *Change: The Magazine of Higher Learning*, 37, 51–57. [87,88]

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [90]

Rangel, E. (2015), "Tier One: Texas Tech Making Progress Towards Prestigious Designation," *The Amarillo Globe-News*, Amarillo, TX. [88]

Rosseel, Y. (2012), "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, 48, 1–36. [90]

Scott, D. (2011), "Can the Carnegie Research University Classification Methodology be Improved?," in *Joint Statistical Meetings*. [89]

Shin, J., Toutkoushian, R., and Teichler, U. (2013), *University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education*, New York: Springer. [88]

The Carnegie Classification of Institutions of Higher Education (n.d.). *About Carnegie Classification*. Available at: *http://carnegieclassifications.iu.edu/*. [87]

Tidball, J (2016), *K-State Upgraded to "Highest Research Activity" in Carnegie Classification*, Manhattan, KS: Kansas State University. [88]

University of Idaho (2016), "University of Idaho Strategic Plan 2016–2025," Technical Report. [88]

University of Montana (2017), "University of Montana Strategic Vision," Technical Report. [88]

van Buuren, S., and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67. [95]

Verghese, A. P., and Jelesiewicz, E. (2016), *Temple University Reaches Height of Carnegie Research Classification*, Temple University Press. Available at https://news.temple.edu/news/2016-02-01/temple-university-reaches-height-carnegie-research-classification. [88]

von Oertzen, T., Brandmaier, A.M., and Tsang, S. (2015). Structural Equation Modeling with Ωnyx. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 148–161. doi:10.1080/10705511.2014.935842. [90]

Wei, T., and Simko, V. (2017), *R Package "corrplot": Visualization of a Correlation Matrix*. Available at *https://github.com/taiyun/corrplot*. [91]