An Alternative to the Carnegie Classifications:

Using Structural Equation Models to Identify Similar Doctoral Institutions

Paul Harmon
with Sarah McKnight, Laura Hildreth, Ian Godwin and Mark Greenwood
Montana State University

April 2, 2018

Institutional Classifications

Systems for identifying like institutions, ranking universities, and delineating similar groups of peer-schools are used by students, faculty and administrators alike. They include:

- The Carnegie Classifications for Higher Education
- The US News World Ranking
- Times Higher Ed

However, each one is subjective, and no perfect classifier exists. However, administrators use these to inform policy decisions (Montana State University, Idaho University), so it is useful to know how they work.

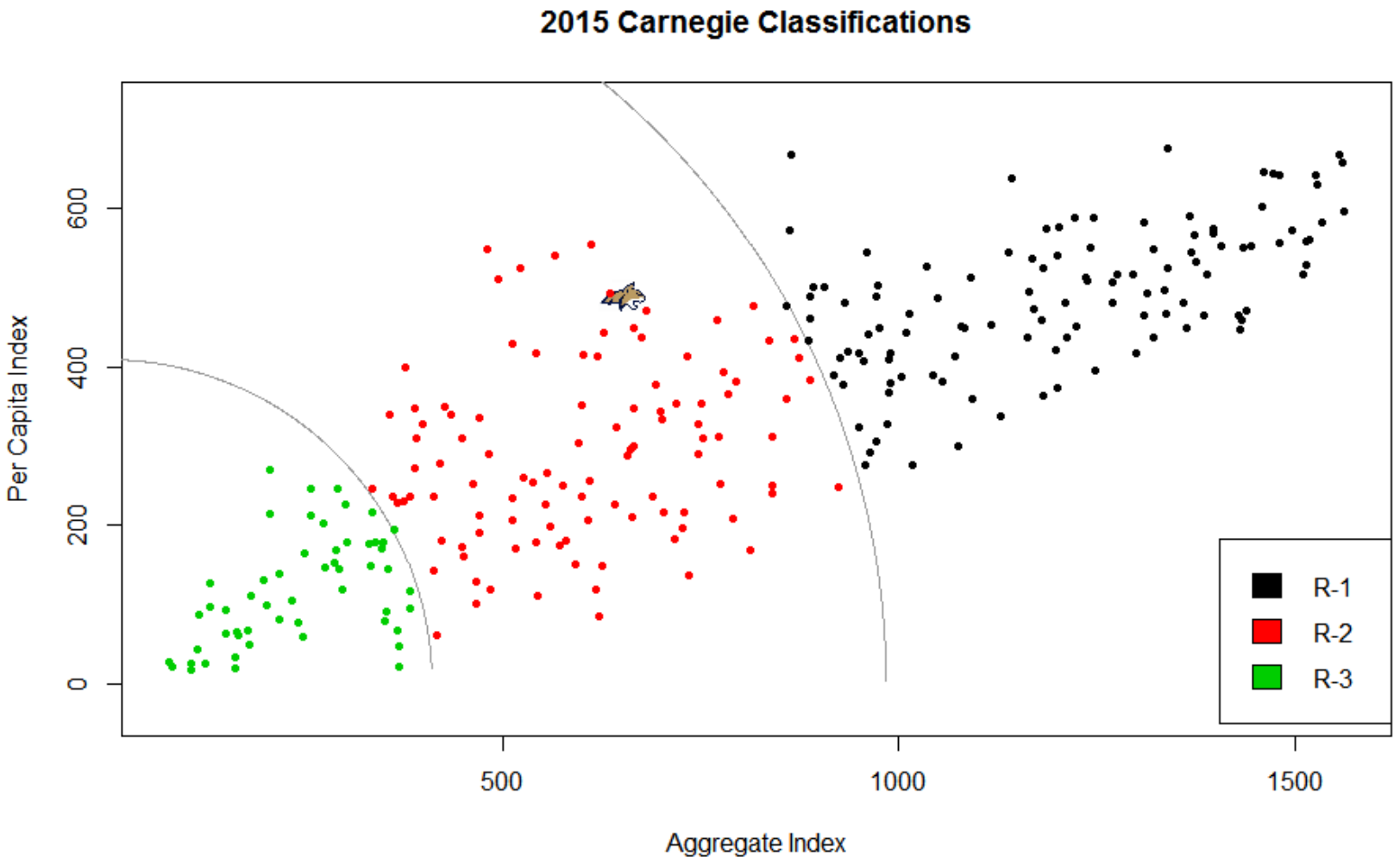
The Carnegie Classifications

The Carnegie Classifications delineate doctoral-granting institutions into three groups:

- R1: Very High Research
- R2: High Research
- R3: Moderate Research

Previously, they had been updated every five years. However, that has changed to a three year cycle in the future. Data are collected in a snapshot from IPEDS and other sources to be used in each update.

The Carnegie Classifications



What Data are used?

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

- **STEM expenditures** (in thousands of dollars)
- **Non-STEM expenditures** (in thousands of dollars)
- **Nontenurable 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
- **Index Creation**: PCA of aggregate and per-capita variables
- **Plot the Indices**: (per-capita vs aggregate)
- **Create groups**: (via line-drawing)

Principal Component Analysis

Principal Componenent Analysis is a method for dimension reduction that takes a set of p predictor variables and decomposes it into k principal components that explain the most variation in the underlying variables.

- Principal Components are weighted averages of the observed variables
- In the Carnegie Classifications, they create the aggregate index by taking the first principal component from a PCA of 7 aggregate variables
- The per-capita index is the first principal component from a PCA of 3 per-capita variables

Problems with Carnegie System

The process used by the Carnegie Classifications illustrates several problems:

- **Based on Snapshot Data**: Data are based on a single snapshot rather than averages during the five-year period.
- **Changing Loadings**: Loadings can change yearly based on variability in the data. This has the potential to change which variables are most important in determining the aggregate and per-capita indices.
- **Dependence on Correlation**:
- **Group Determination**: The group determination is completely subjective. Lines are drawn after visual inspection of the plot.

This causes problems for schools that try to direct policy based on these classifications because the most important variables can be different from release to release based on variability in the data.

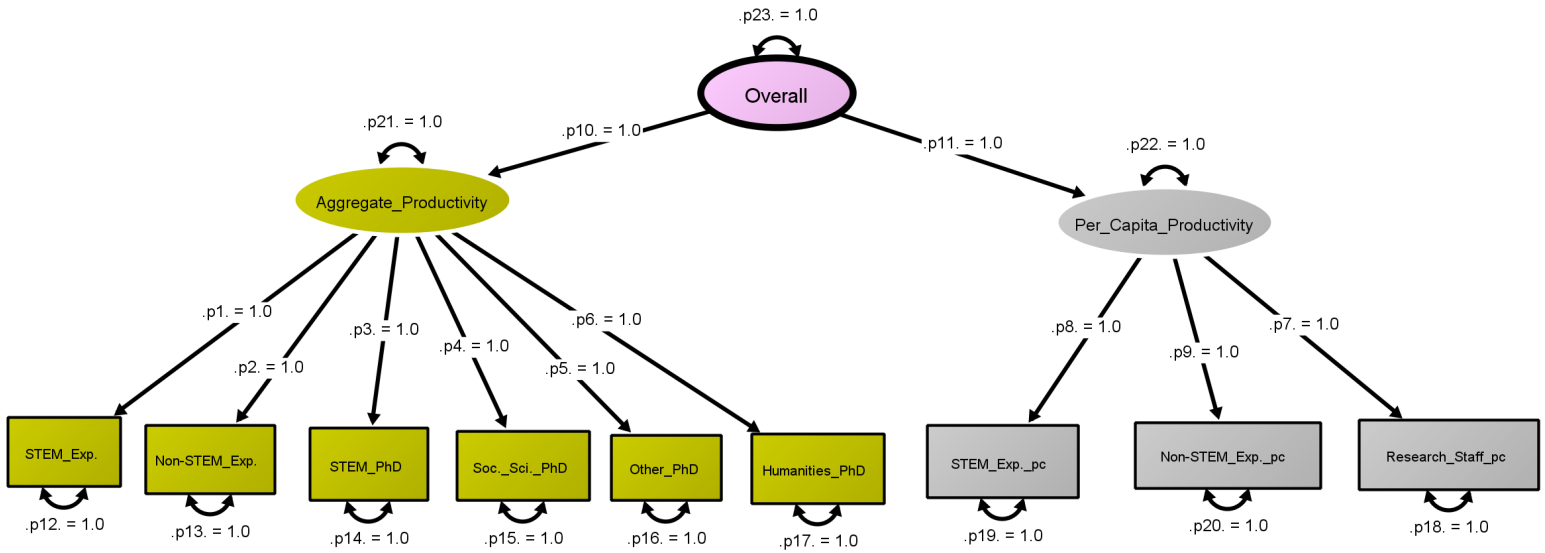
Structural Equation Modeling

Structural Equation Models (SEM) are used to model simultaneous equations, the use of latent or unobserved variables, and variables to be measured with error.

- **Latent Variable Model**: Measures unobserved variables.
- **Measurement Model**: Relates the latent variables we are interested in to the manifest that we actually observe.

The Carnegie Method using SEMs

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



Problems with the Carnegie Method for SEM:

The variables used to measure the aggregate and per-capita latent variables are the same (with a slight per-capita transformation). The SEM cannot handle this level of correlation in the two latent factors; it **does not converge** and thus paramaters cannot be estimated.

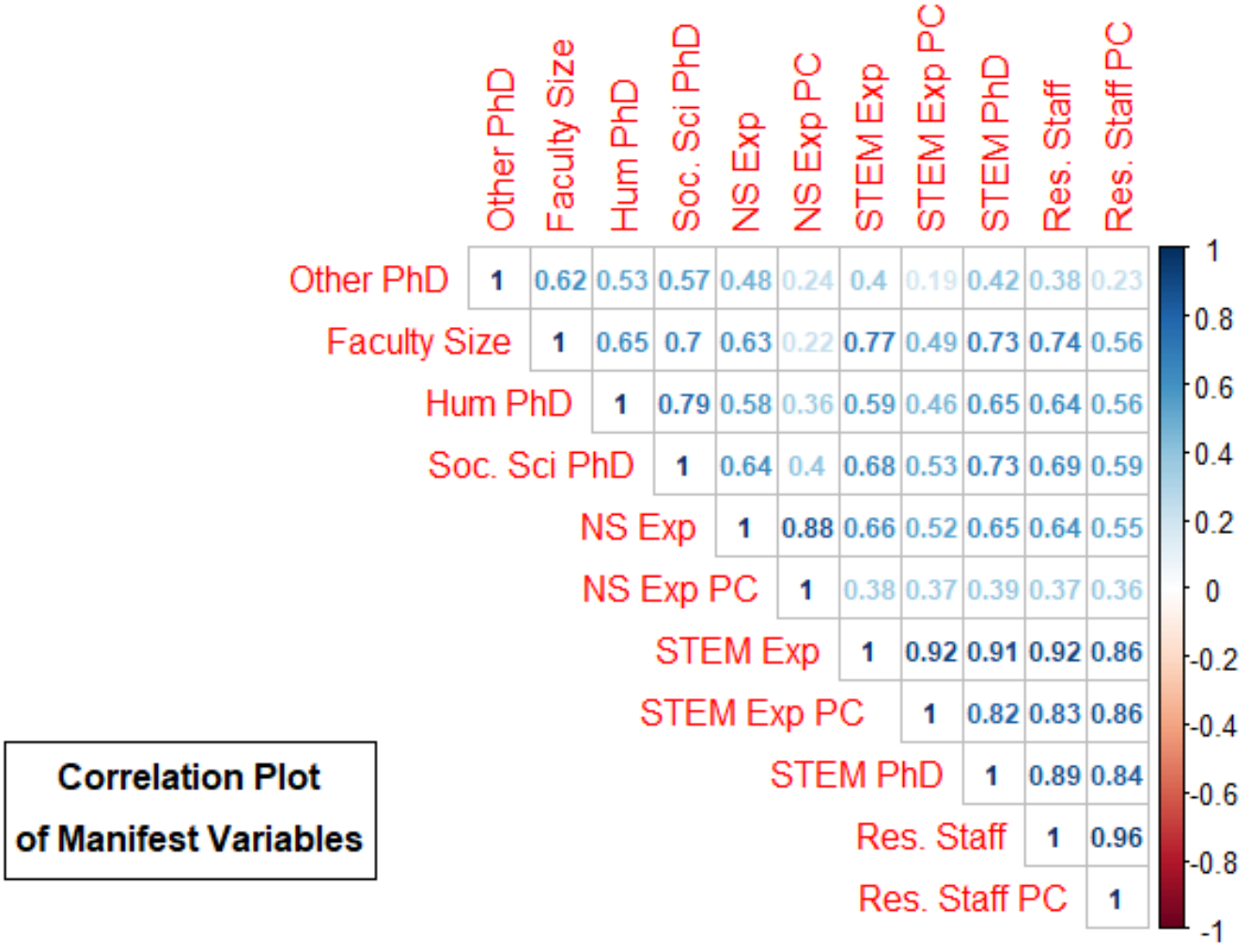
STEM and Non-STEM Factors: An Alternative

A more intuitive method would be to consider two latent factors:

- STEM productivity
- Non-STEM productivity

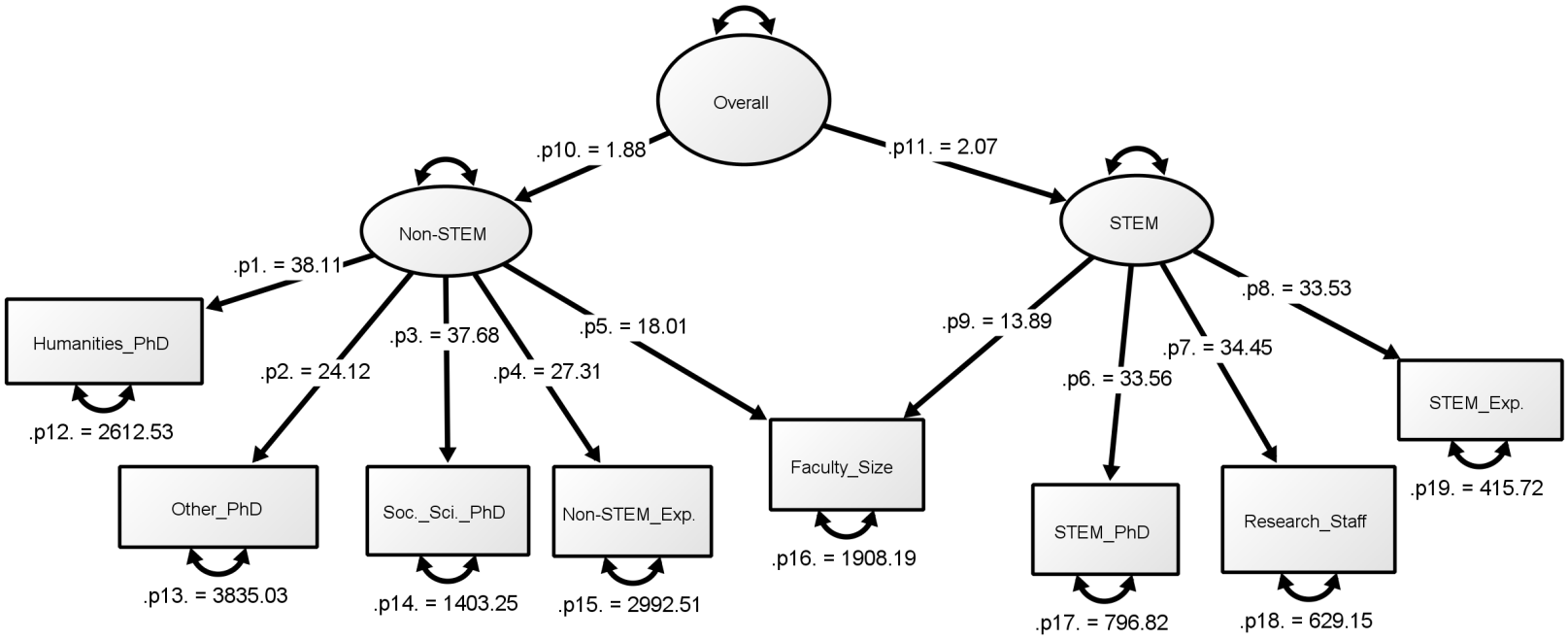
Variables loaded onto these factors are not as likely to be correlated.

STEM and Non-STEM Manifest Correlations



Correlation Plot of Manifest Variables

Proposed Model: STEM and Non-STEM



Determining Group Membership

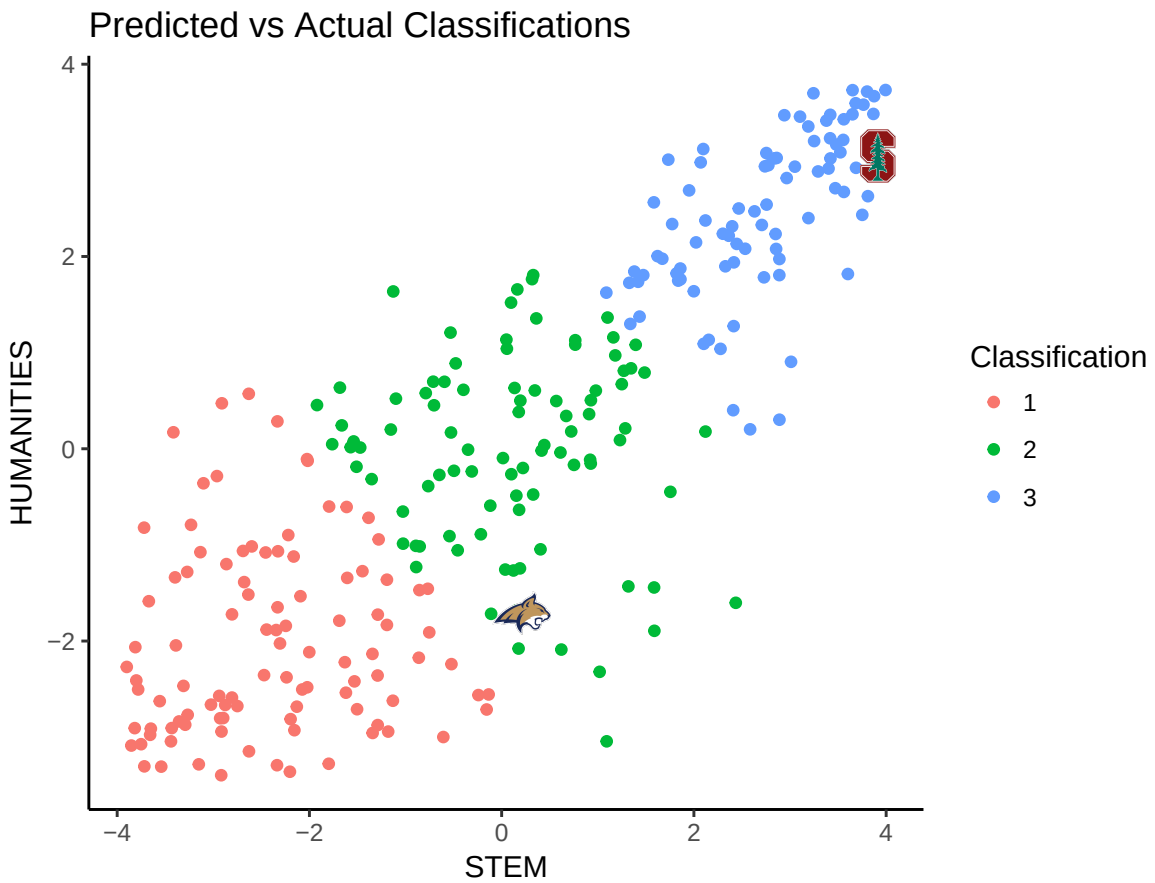
The SEM model returns a single factor-of-factor score for each university. These can be used as inputs to a clustering algorithm to determine both the optimal number of clusters and the optimal cluster membership.

Potential Problems:

- Optimal Number of Clusters may be too large/small: We can fix this to a reasonable number if necessary (results here are fixed at 3).
- There are many different clustering methods to use (hierarchical clustering, mixture-model based methods, etc.).

Group Membership with Mixture Model:

Using a mixture model, we can objectively define clusters. Additionally, we can illustrate uncertainty in classification for schools near the boundary.



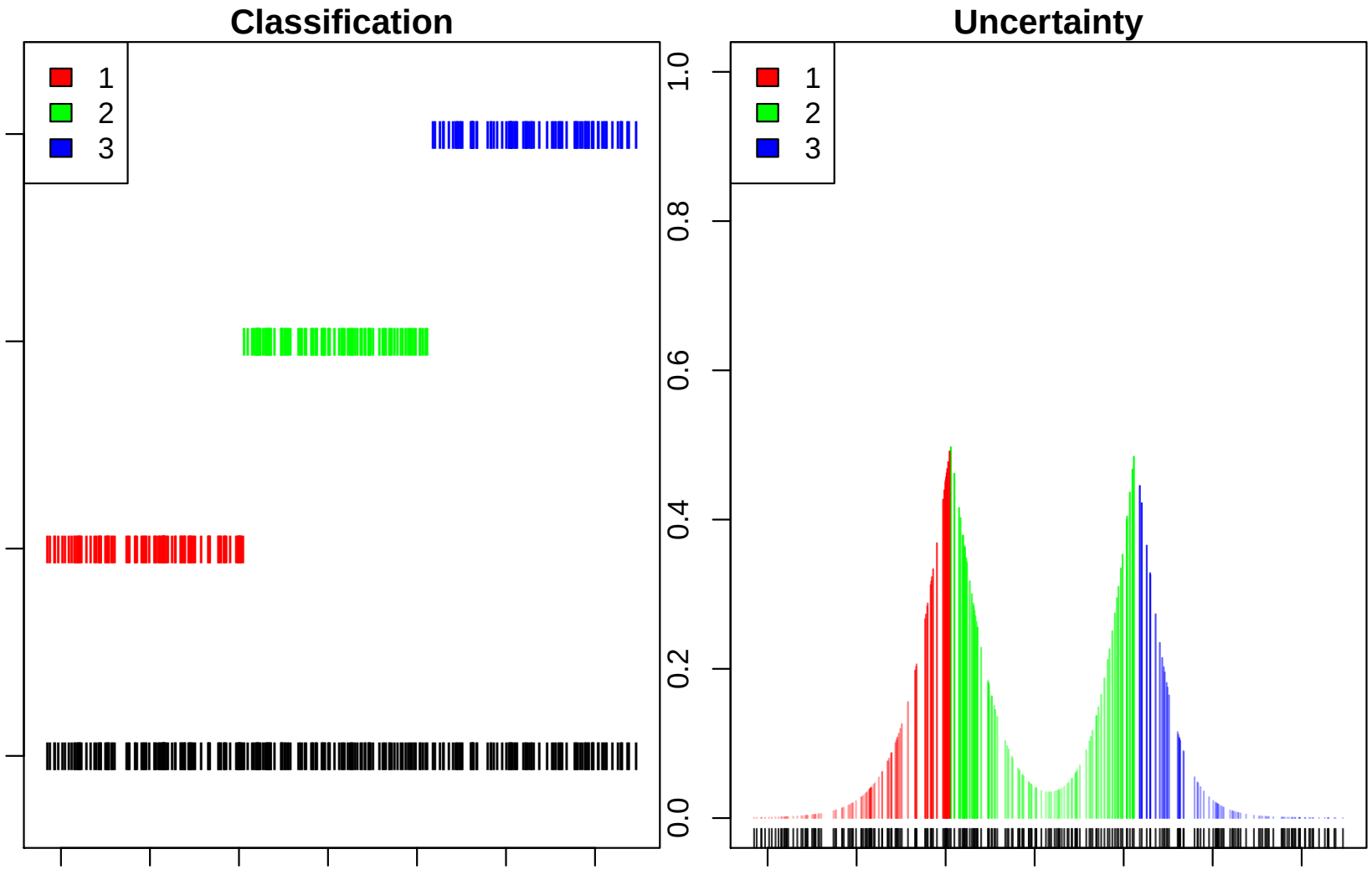
Uncertainty in Classifications

The Carnegie Classifications treat each classification as a fixed value, without accounting for the uncertainty inherent in the process.

- Variability in underlying data
- Uncertainty in PCA
- Uncertainty in Mixture Modeling

Our method can be used to visualize differences in the data.

Uncertainty Plots



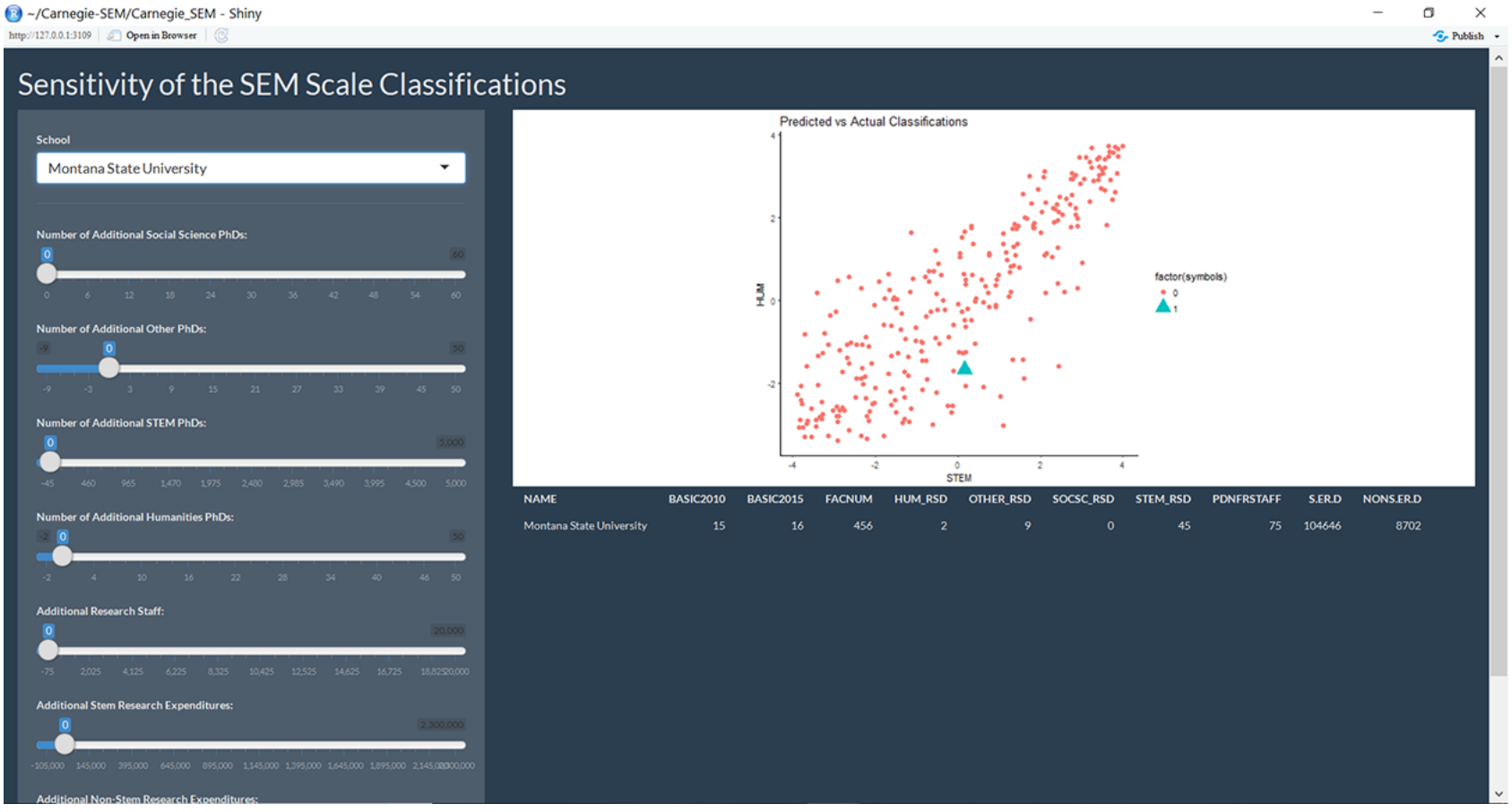
Shiny Applications:

We developed shiny applications to assess sensitivity of both the Carnegie Classifications and the proposed SEM-Classifications for any institution of your choosing.

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

These applications can be found at: paulharmon.shiny.io/SEM-class

Shiny Applications:



Conclusions

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

- Latent variables do not have to be orthogonal like PCA-based indices
- Latent variable modeling makes more sense than dimension-reduction (especially since we only have 7 variables)
- Single-factor scores allow for comparison on a single dimension rather than using two scores.

Future Work

Future work will focus on determining an optimal clustering strategy. The Carnegie Classifications themselves are changing, so if they change the variables that are used, we will investigate development of either a new latent trait or sensitivity of additional manifest variables to measure the STEM and non-STEM latent traits.

Questions:

Thank you!