# Spatial Distribution of Forest Research in the Conterminous United States

Hayden Elza[1,2]
Steve Kochaver[1,3]
Dylan Moriarty[1,4]

Geography 578 – Semester Project
May 8th, 2015

---

[1] GIS Certificate Program, Department of Geography, University of Wisconsin - Madison
[2] Forest Landscape Ecology Lab, Department of Forest and Wildlife Ecology, University of Wisconsin - Madison
[3] Townsend Landscape Ecology Lab, Department of Forest and Wildlife Ecology, University of Wisconsin - Madison
[4] Cartography Lab, Department of Geography, University of Wisconsin - Madison

# I. Executive Summary

Modern web-based technologies and attitudes toward data availability are making modern science incredibly accessible. Many institutions are addressing public desires for easy access, bulk information by collecting studies and publications across every scientific discipline, and aggregating them into massive databases. An immediate result of this type of data assembly is the standardization of metadata between previously unconnected fields, and this standardization is opening up new possibilities for comparisons among research. Though still early in their development cycles many of these databases are expansive enough to provide an excellent basis for big-data analytics.

The purpose of this study was to explore possible ways this new type of meta-analysis could be conducted and to provide a proof-of-concept for quantifying bias in scientific focus. Using the United States Geological Survey's open source ScienceBase database this study investigates the relationships between the subfield of forestry research in the United States, and the institutions conducting that research. By spatially characterizing and analyzing these relationships it becomes possible to observe the potential bias in where research is being conducted and may illuminate patterns that can be scaled to explain how science is conducted in a much larger scope.

# II. Introduction

Science, as a framework of knowledge building, is presumed to completely unbiased. It is expected that attempts to explain a phenomena will be broad enough to be able to model and predict new occurrences while having enough detail to still maintain an appropriate focus[1]. Finding the perfect balance in these theoretical forces is the goal of every good, effective scientist, though limited time and limited resources in an imperfect world constrain theories of science to a similarly imperfect level. In particular the field of natural science, which is so dependent on observation of phenomena in that imperfect form, is overtly self-aware of its own constraints. Though it seems obvious that not every aspect of everything in our universe, or even

our world, can be studied and explained[1], it is often taken for granted in what is omitted since predictions easily fill the physical void of a research focus.

As science becomes more public and more accessible these potential gaps are becoming more obvious and, more importantly, they are becoming quantifiable. Digital databases of scientific data, studies, and publications are growing in popularity because of the quick-access services they provide to a scientific community dependent on peer review. As these databases are aggregated, usually by credible public institutions, the data that fill them are standardized, and suddenly comparisons of metadata among and between every discipline can be made. More importantly the gaps and bias of these studies foci can be observed and quantified in ways that haven't existed until this open-source, big-data movement.

One such database hosted by the United States Geological Survey, called the ScienceBase database, has been growing since 2011 and has an impressive structure in place to facilitate a comparison of bias among research[2]. ScienceBase has amassed a considerable number of scientific records and is growing at an incredible rate. This study aims to use the current ScienceBase database to test the feasibility of measuring the type of bias described above, and to provide proof-of-concept on a subset of the scientific community in the hopes that future analysis can observe bias in science as a whole and therefore identify ways to improve this theoretical framework (science) for a more complete understanding of knowledge for knowledge's sake.

The subset chosen here is the field of Forestry because of the completeness of the records currently in the ScienceBase database. More particularly, Forestry research done in the contiguous United States to avoid effects from political boundaries that cannot be as easily measured as spatial ones from the data available. The bias we seek to measure then, by nature of the dependencies of a natural science, is spatial bias. Also, since research is only being done by scientists who do research this spatial bias is measured in the context of locations of institutions with the potential of contributing to Forestry research.

# III. Concepts and Methodology

## *Research Distribution*

The first step in comparing instances of Forestry research was to extract Forestry-specific records within the study area from ScienceBase. The database allows advanced spatial queries (using a within spatial operator and a default polygon of US borders) and a robust tag and id search to retrieve the metadata of every record that would fit within the confines of the study. The study areas for each record was then further extracted from the metadata and converted from points to bounding polygons which were subsequently burned into a raster to show research instance intensities (Figure 1).

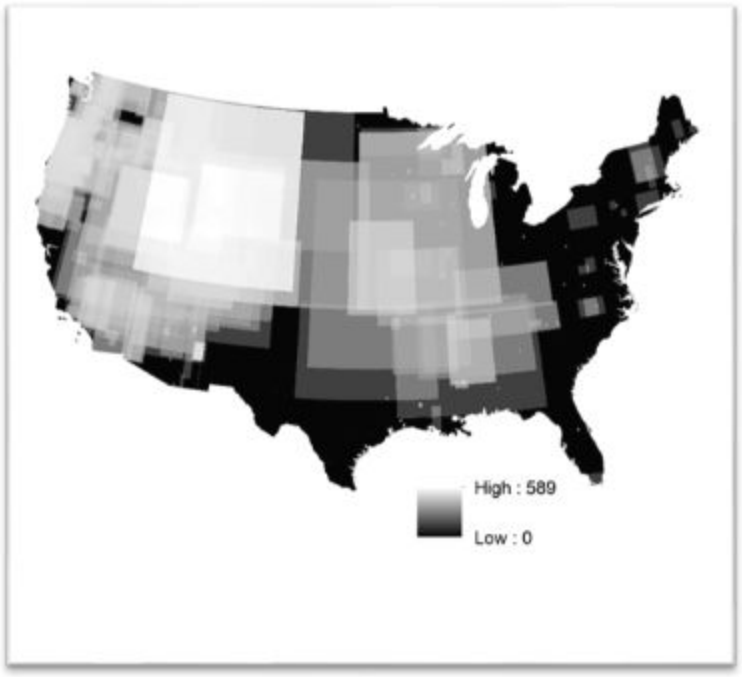



***Figure 1.*** *Raw raster output of research intensity within contiguous US boundaries.*

## *Forest Coverage*

To obtain the Forest coverage for the United States, we took NCLD coverage of the United States and reduced the raster to a binary image of just the forestry coverage. By clipping

the research distribution to the forestry coverage, we had our research frequency by forests (Figure 2).

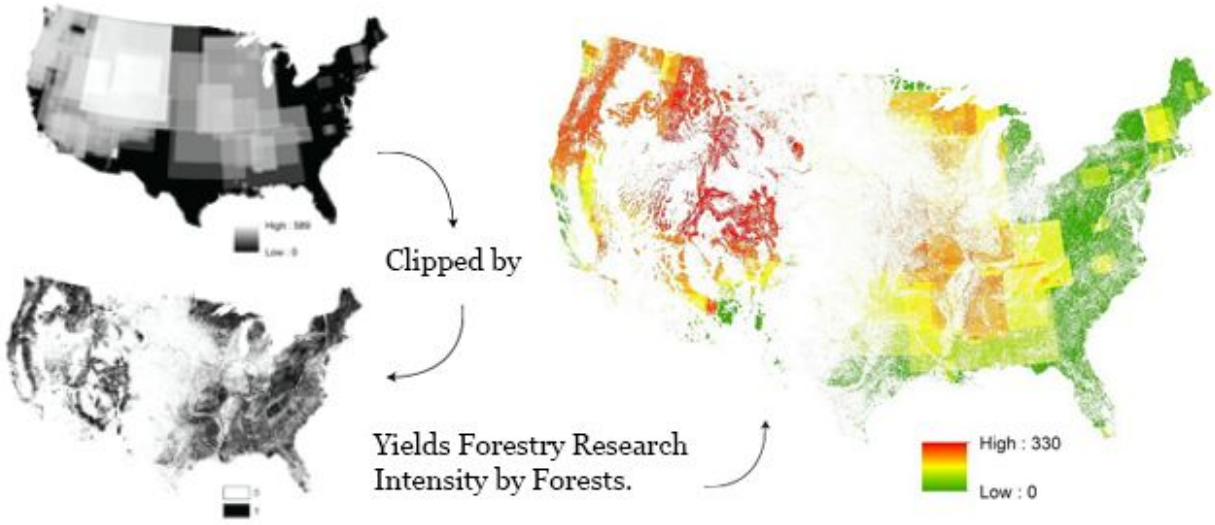



***Figure 2.*** *Process to produce Forestry Research Intensity by Forests.*

## *University Locations*

For the university locations, we also were able to use the ScienceBase-Catalog, which contained a shape file with all Post Secondary Education facilities as defined by the Integrated Post Secondary Education System (IPEDS), National Center for Education Statistics, and US Department of Education. This data set was great, but unnecessarily all-inclusive. We determined that schools that lack qualities that would be conducive for forestry research could be whittled down by factors including: maximum degree length, total enrollment, and physical facilities.

By selecting schools with degree lengths greater than two years, we removed most technical colleges, cosmetology schools, and similar universities which wouldn't be inclined towards research. Eliminating schools with a total enrollment of less than 1,000 students further helped. Finally, to remove online universities which lack the facilities to conduct research, we removed any universities without physical campuses. With those qualities weeded out, we had a data layer containing research-capable US Universities.

## *Statistical Analysis*

In order to properly assess the possible presence of bias in forest research a suitable metric must be chosen that can capture the relationships between the two point patterns (i.e. university locations and instances of forest research). To fit this purpose, Ripley's *K*-function was chosen due its ability to summarize second-order characteristics which describe variation and correlation in point fields. Ripley's *K*-function is often utilized as a self-*K* function, $K_{ii}(t)$, which analyzes second-order characteristics within a single event. Though in the case of this study, the cross-*K* function $K_{ij}(t)$, where $i \neq j$, was utilized in order to assess variation and correlation between the two datasets. Ripley's cross-*K* function is calculated using expanding search radii around a point *i* and counting the number of *j* points within the radius. Calculating the *K*-function over multiple distances (search radii) can show how point pattern distributions can change with scale. Ultimately the *K*-function will describe the data as clustered, dispersed, or randomly distributed throughout the study area.

In this study Ripley's *K*-function was calculated using the statistical software R. This allowed for tighter control of the function rather than trusting the software to make the correct decisions regarding the parameters. Data were imported into the software from shapefile and converted to a format suitable for analysis. The cross-*K* function was run using university locations as the *i* points and using the research locations separated by number of research instances (e.g. areas with only seven research instances) as the *j* points. That is to say, the *K*-function searched for the number of research locations within a distance of a university location. Research locations were separated into subsets containing all research locations with a specific number of research instances to further tease out variation among datasets. This was done under the assumption that areas with a low number of research instances would likely have different distributions than areas with a high number of research instances. The function used also included several border corrections in the output rather forcing the user to choose one. This is advantageous because it allows for the border corrections to be quickly compared often revealing that they show the same trend, but it is important to not always assume this is the case.

An example of the output in graphical form is shown below in Figure 3. On the x-axis is the search radius in meters and on the y-axis is the resulting *K*-function. Three border

corrections, Isotropic, Translate, and Border, are shown in black, red, and green respectively as compared to the perfectly random distribution of the Poisson process in blue. In this case, it can be seen that all three border corrections lie above the Poisson, indicating research points tend to be clustered. This is more prevalent as the search radius increases. Though at the larger search radii, the difference between border corrections is expected to be larger because of the higher likelihood of a large circle lying outside of the study area (i.e. the U.S.).

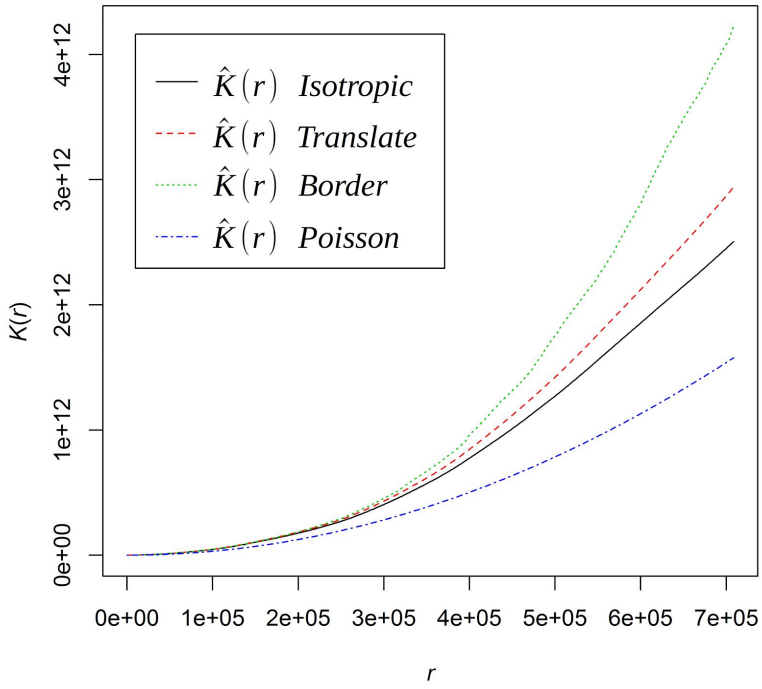



***Figure 3***. *Graph of K-function for the subset of seven research instances. The x-axis represents the search radius and the y-axis represents the resulting K-function.*

After plotting the *K*-function for each unique number of research instances, the data were combined to assess trends over all research intensities. In this way, variation between less researched areas and highly researched areas could be quickly compared. An example of this is shown below in Figure 4 As before, the x-axis is the search radius in meters and the y-axis is the *K*-function. New on this graph is the z-axis which shows the different subsets by number of

research instances. This graph shows the distribution for the Border type border correction. Most notable is the large peak at a low number of research instances and a second much smaller peak in the higher numbers of research instances. Comparing Figure 4 with a graph of the Poisson process in figure 5, one can see that the large peak is much higher than the Poisson indicating clustering, while the smaller peak sits just below indicating slight dispersion. The trough in the middle range of research instances can be explained by the lack of data in those ranges.

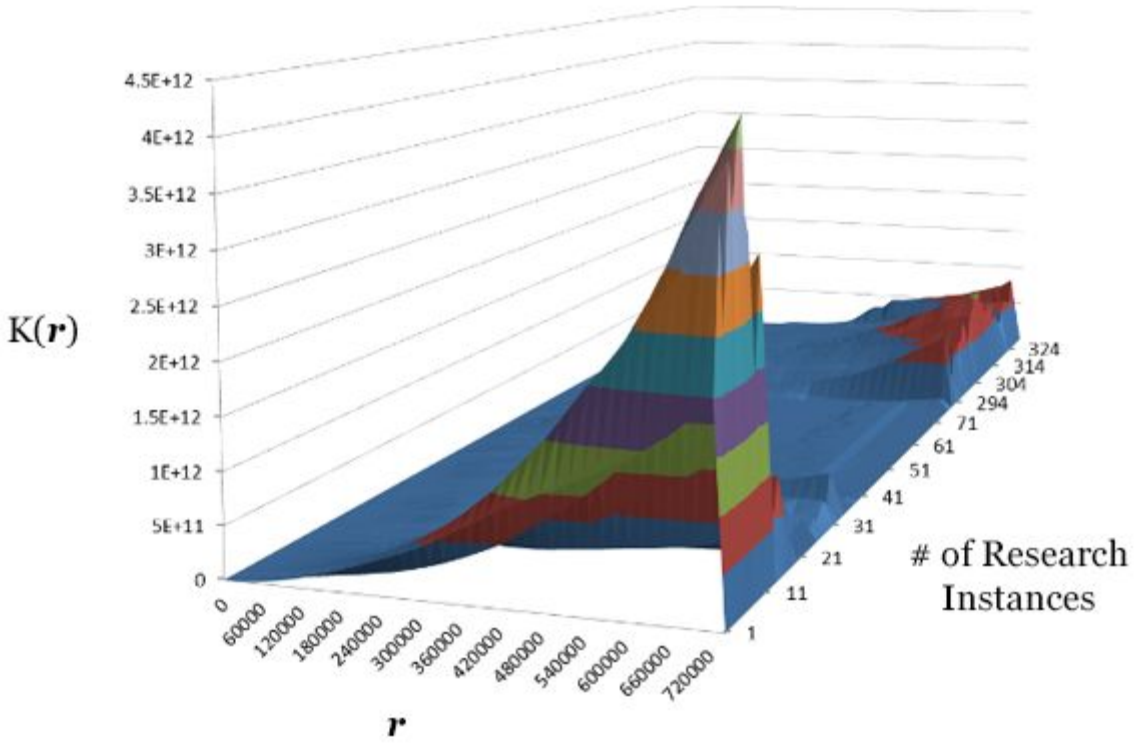



***Figure 4.*** *Plot of K-function over all subsets of different numbers of research instances using the Border method of border correction. The x-axis is the search radius in meters, the y-axis in the resulting K-function, and the z-axis is the number of research instances.*

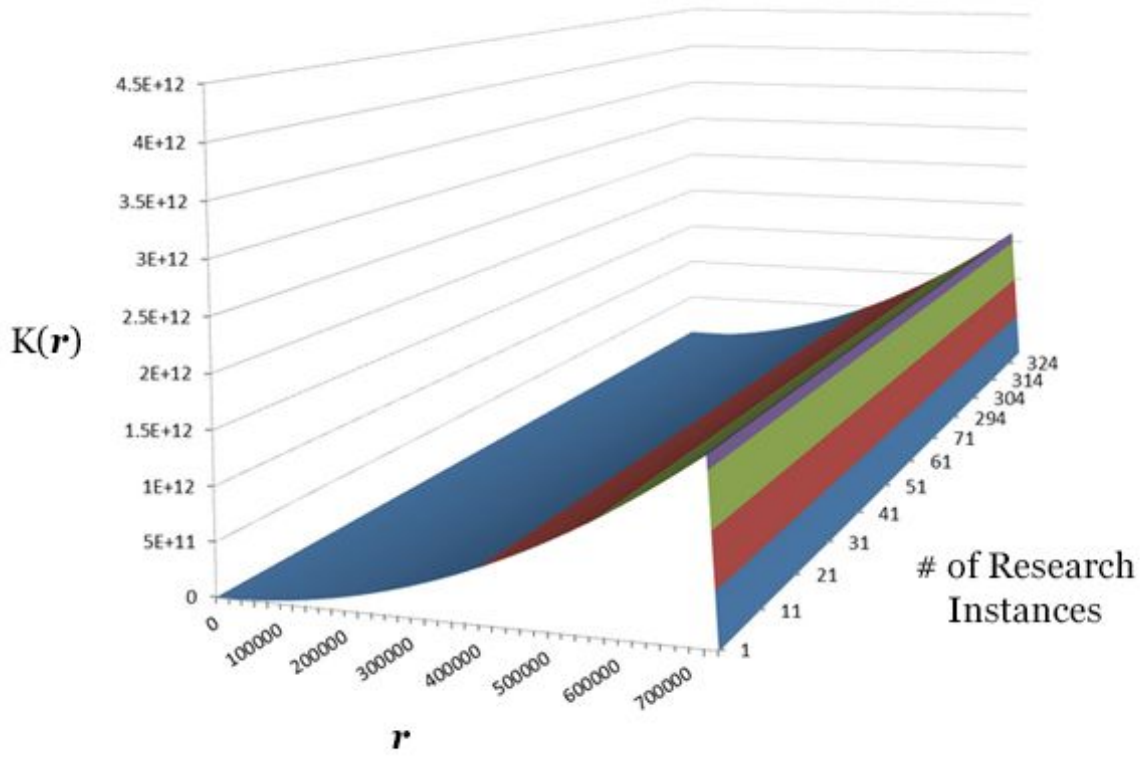



***Figure 5.*** *Plot of K-function over all subsets of different numbers of research instances for the Poisson process (i.e. randomly distributed). The x-axis is the search radius in meters, the y-axis in the resulting K-function, and the z-axis is the number of research instances.*

A second way to assess the *K*-function over different numbers of research instances is by utilizing a weighted *K*-function. This was done in the exact same manner as the previous K cross functions save that each research instance was treated as series of stacked points of research intensity. One point for the value of the research intensity found there; this was to provide an unseparated, generalized K function.

# IV. Results and Discussion

## *Meta Statistics*

There was not an equal distribution of research intensity subsets and this can clearly be seen in Figure 12 below. The first thing to note about this graph is that there are missing values along the x-axis. This is because there are no research areas with those number of research instances. This causes a more or less bimodal distribution with a high frequency of areas with low research intensity and relatively smaller but still high frequency of areas with high research intensity. There was a total of 117 subsets with a mean of 131.8, a median of 59, and a standard deviation of 130.8.

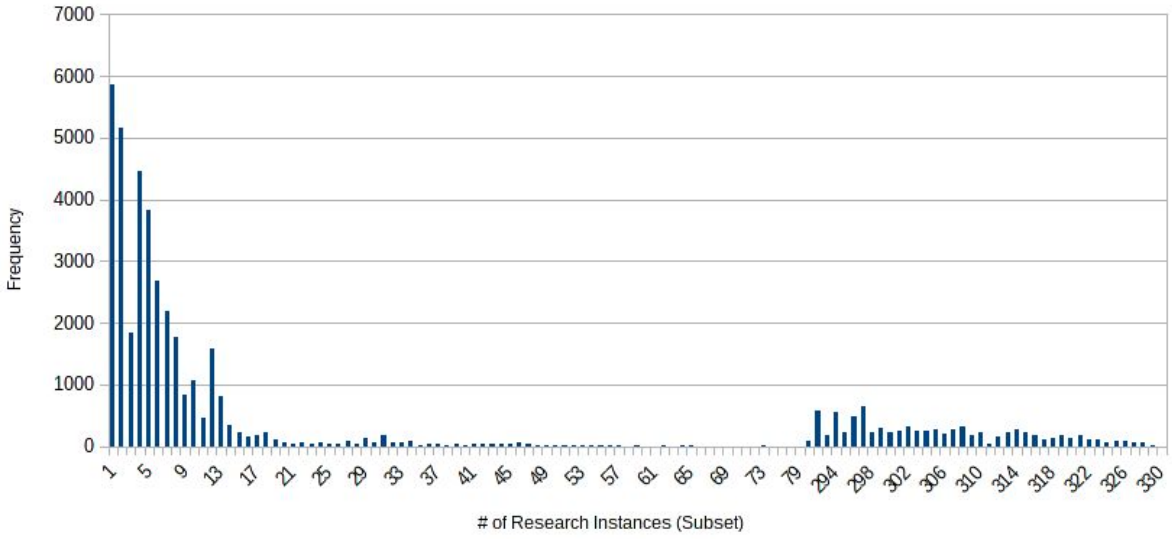



***Figure 12.*** *This graph shows the frequency of each research intensity.*

## *Example Cases*

Although there are too many subsets to show the *K*-function for each research intensity in this report, examples of low, medium, and high research intensity are shown below. Figure shows areas of exactly seven research instances. The map shows university locations in black

and research areas in blue. On this map one can see numerous research areas located in the midwest, as well as a group along west coast possibly following the Rocky mountains. Looking at the graph, one may notice that all three border corrections lie above the Poisson process, indicating clustering. This is especially true for larger radii. Note the high variation between border corrections at large radii is a product of the differences in how the border corrections are calculated, because as the search radius increases so does the likelihood that a circle will lie outside the study area.

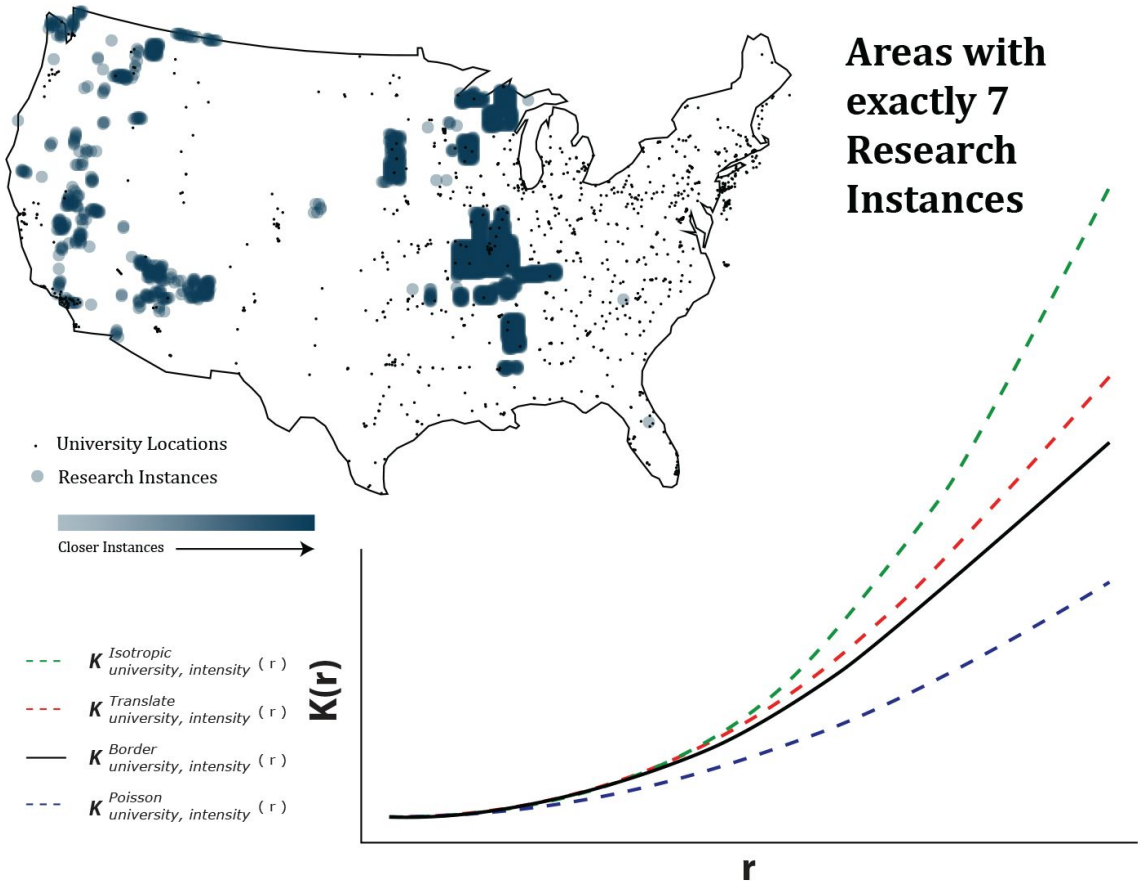



***Figure 6.*** *In the top left portion of the figure is a map of the U.S. showing the location of universities in black and the location of areas with exactly seven research instances in blue. In the lower portion of the figure is a graph showing the distributions of the Isotropic, Translate, and Border border corrections in green, red, and black respectively as well as the Poisson process distribution in black.*

In Figure 7 below is a map similar that of Figure 6, though this time the map is showing the locations of areas with exactly 46 research instances. One will likely notice there far less research areas on the map in Figure 6, as compared to the map in Figure 7; looking back at Figure 12 confirms this. Interestingly the points seem to be grouped in a single area in the northwest portion of the U.S., just south of Portland, OR. Looking at the graph, a distribution much different from the one in figure **v**. can be seen. This time the smaller radii show a very slight clustering, but perhaps too small to be significant, while the larger radii show significant dispersion. Interesting as well is that the isotropic border correction is much lower than the other two border corrections. This is likely due to the fact the majority of the research points lie very close to the study area border causing the results of the border corrections to vary greatly.

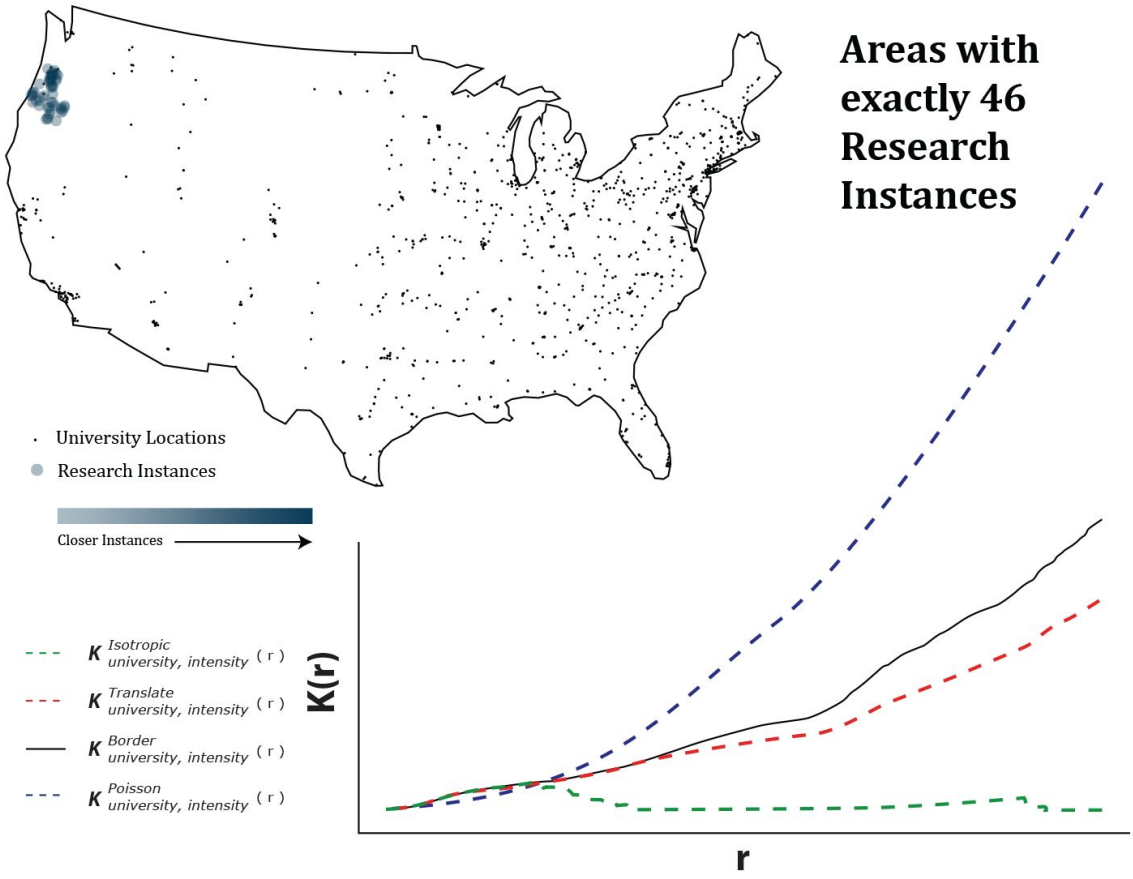



***Figure 7.*** *In the top left portion of the figure is a map of the U.S. showing the location of universities in black and the location of areas with exactly forty-six research instances in blue. In the lower portion of*

*the figure is a graph showing the distributions of the Isotropic, Translate, and Border border corrections in green, red, and black respectively as well as the Poisson process distribution in black.*

Figure 8 again shows a map and graph similar to Figure **7** and Figure 6, but this time for areas with exactly 299 research instances. Looking at the map we see the majority of the research points located in the western portion of the U.S. with a large portion seemingly residing over Yellowstone National Park. This is likely related to the large amount of data generated during the Yellowstone fires of 1988 attracting many fire ecology researchers. The graph is very interesting in that in this case all three border corrections reside below the Poisson process for the entire graph indicating dispersion at all distances. Also interesting is that all three border corrections are very close together. This is likely due to all the research points being located well within the study area and therefore not experiencing much border effect.

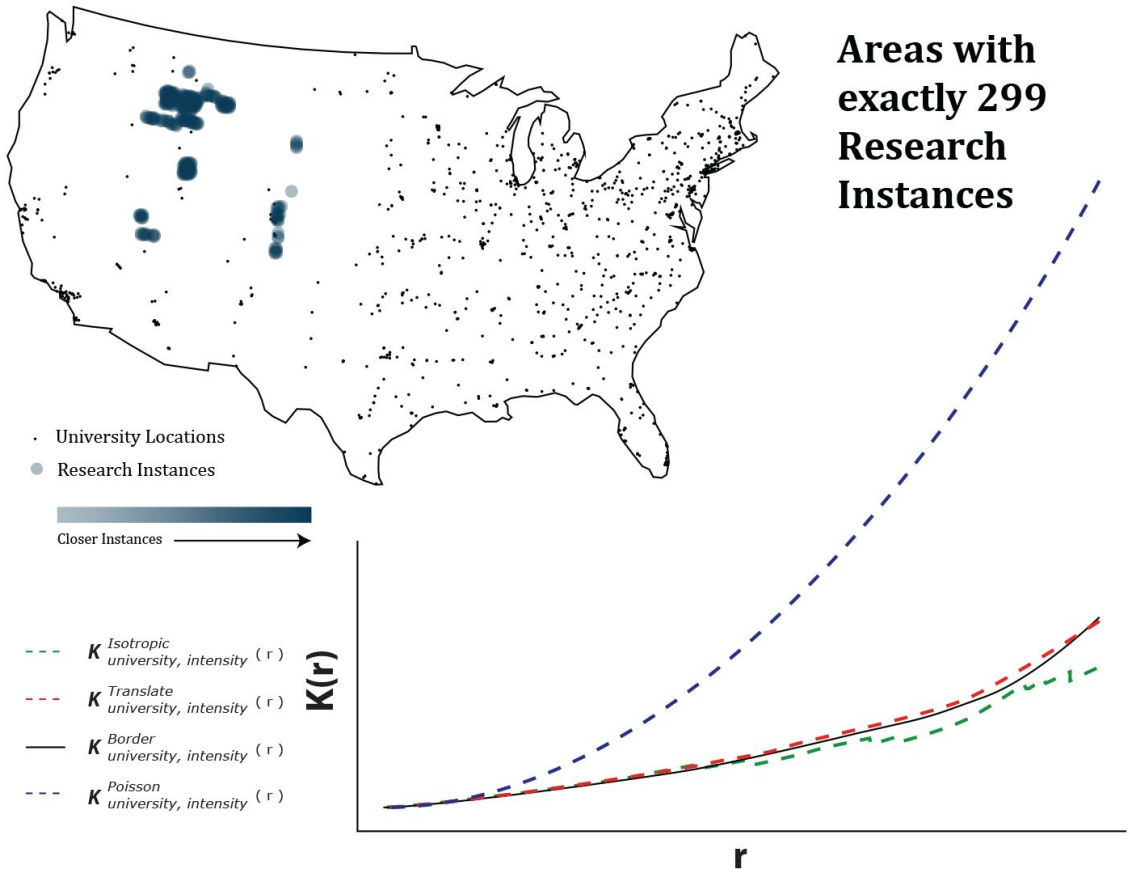



***Figure 8.*** *In the top left portion of the figure is a map of the U.S. showing the location of universities in black and the location of areas with exactly two hundred and ninety-nine research instances in blue. In the lower portion of the figure is a graph showing the distributions of the Isotropic, Translate, and Border border corrections in green, red, and black respectively as well as the Poisson process distribution in black.*

## *Combined Results*

Adding a third axis to the *K*-function plot allows for quick assessment of trends over all research intensities. Combined plots for isotropic, translate, and border type border corrections are shown below in Figures 4, 9, 10 respectively. The perfectly random distribution, i.e. the Poisson process, is shown below in figure **5**. Both the isotropic and translate border corrections look fairly similar in shape. They both have peaks in the low range of research intensity with the translate border correction peaking slightly higher than that of the isotropic border correction. The low intensity peaks rise far above the Poisson distribution indicating a high level of

clustering. The rest of each border corrections graph below the Poisson distribution, indicating dispersion. The border type border correction shows a fairly different distribution than the other two border corrections. It has a large peak at a low number of research instances and a second much smaller peak in the higher numbers of research instances. The large peak is much higher than the Poisson indicating clustering, while the smaller peak sits just below indicating slight dispersion. The trough in the middle range of research instances indicates significant dispersion.

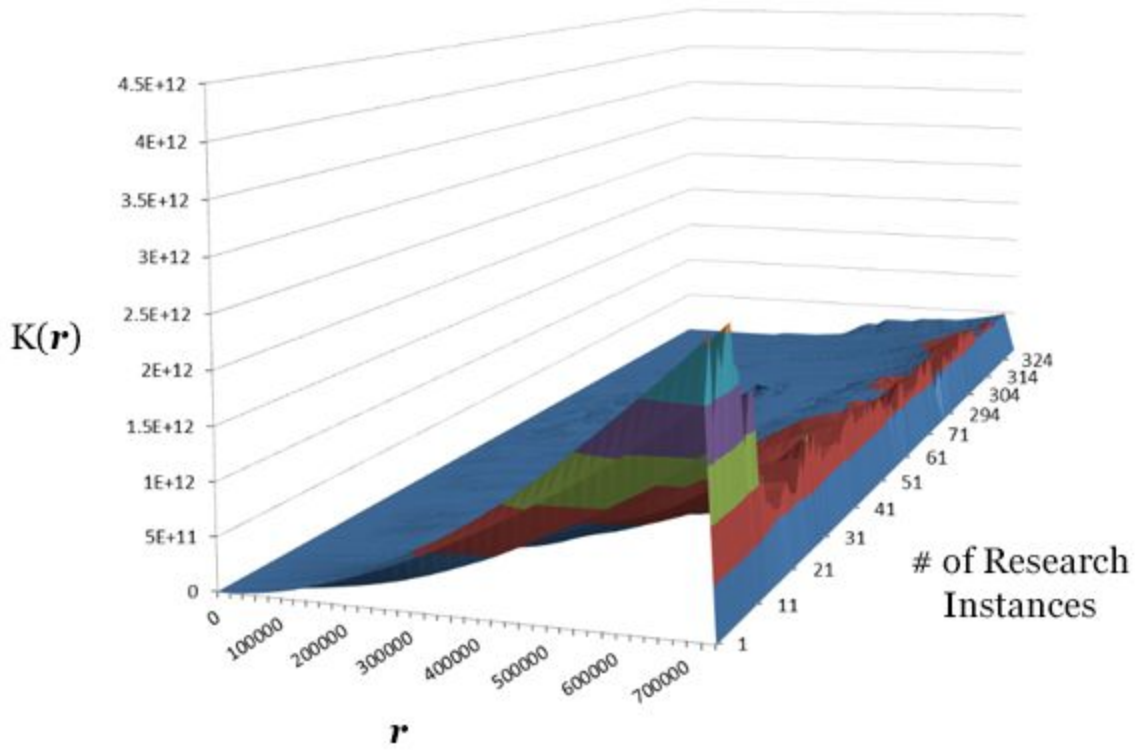



***Figure 9.*** *Plot of K-function over all subsets of different numbers of research instances using the isotropic method of border correction. The x-axis is the search radius in meters, the y-axis in the resulting K-function, and the z-axis is the number of research instances.*

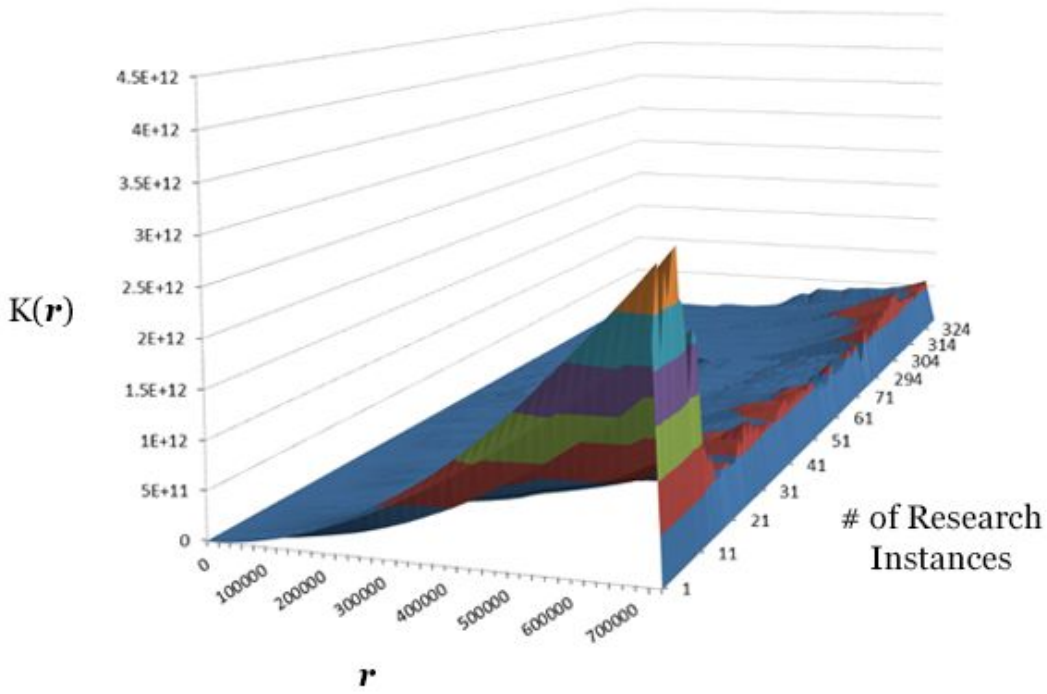



***Figure 10.*** *Plot of K-function over all subsets of different numbers of research instances using the translate method of border correction. The x-axis is the search radius in meters, the y-axis in the resulting K-function, and the z-axis is the number of research instances.*

## *Weighted K-function*

Rather than run the cross-*K* function for each level of research intensity, a special *K*-function can be performed that weights each research area point with a value equal to the number of research instances at that location. This weighted *K*-function, seen below in Figure 11, summarizes the data even further than the three axis plots and supplies an overall assessment of bias. Figure 11. shows all three border corrections, isotropic, translate, and border in black, red, and green respectively, with very similar distributions. All three border corrections light below the perfectly random Poisson process. This indicates dispersion, with the degree of dispersion increasing as the search radius increases.

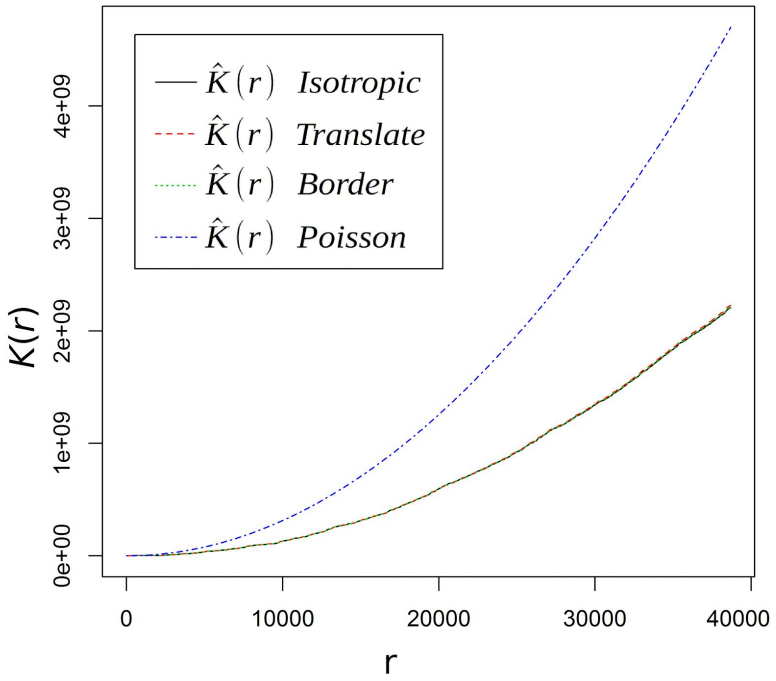
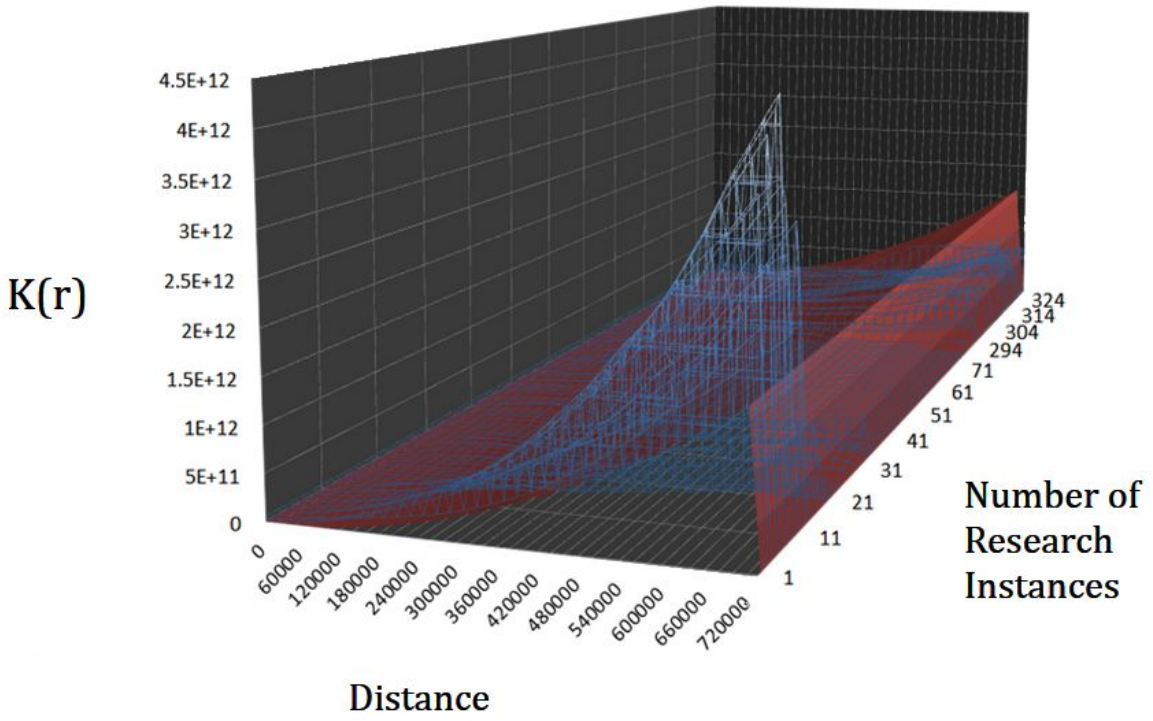



***Figure 11.*** *A graph of the weighted K-function. The x-axis represents the search radius in meters while the y-axis represents the resulting K(r) value.*

## V. Conclusions and Future Efforts




Our incremental K functions showed near random distribution of research sites at low distances from universities and extremely high clustering at high distances, which we can assume to mean that the forests of interest are close together but are far away from most universities (Figure 7). Meanwhile we see near perfect dispersion away from universities at all distances at high research intensities. In the context of Forestry this is evidence of the huge number of studies that were done in the Yellowstone area, which is made extremely attractive to researchers because of its amount of recorded data following the massive 1988 fire season. Since then the area has been a major concern to all ecological researchers as fire is the main driving disturbance in forest ecosystems and this area provides a massive time-series that attracts almost all disturbance ecologists in the US.

In the larger context of scientific bias this shows a clear and explainable spatial bias that was driven by an event generating a great deal of raw data. Scientists followed this data and used it to create models and form explanations that may be currently attributed to forests where it is

assumed function is the same, but there is no basis other than intelligent assumption. While scientists are constrained by that data they have, it can be shown through this analysis that the results of reusing existing data for studies inherently creates a bias and prevents similar phenomena from being studies at other locales.

This is just one manner of bias in scientific coverage. If this, or a similar analysis were performed on other subfields and on larger scales it might soon be shown that massive gaps exist in surprising places within and across all disciplines. As the ScienceBase and other databases grow it will be possible to revisit these analysis with a larger goal in mind and clarification of bias can mean prevention of similar pitfalls in the future, and a correction and refocus on areas previously neglected. Being so self-aware in a meta-approach is essential in maintaining integrity in the core of all scientific endeavors and ensuring that the coverage of knowledge is as near complete as pure logic allows.

## VI. References


1. Martin, Brian. *The bias of science*. Canberra, Australia: Society for Social Responsibility in Science, 1979.
2. https://my.usgs.gov/confluence/display/sciencebase/ScienceBase+Information+Model
3. http://github.com/skochaver/sciencebase_analysis. 2015.
4. Turner, M. G., Hargrove, W. W., Gardner, R. H., & Romme, W. H. (1994). Effects of fire on landscape heterogeneity in Yellowstone National Park, Wyoming. *Journal of Vegetation Science*, *5*(5), 731-742.