Innovation Landscapes: An Analysis of National Science Foundation Research Grants from 1999-2016

Sahar Alhassan, Kaitlin Lynes, Deepa Mehta, Ilya Perepelitsa

December 21, 2015

Introduction

Innovation is often considered to be considered an engine for development. Governments typically support innovation by funding research and development for universities, corporations, and other institutions. (Etzkowitz and Leydesdorff 2000) Recently, the White House approved a bipartisan budget agreement which will allocate even more funding over the next years in order to maintain the United States' leadership in innovation. (Donavan 2015) One of the few agencies that this bipartisan agreement will support is the National Science Foundation (NSF).

NSF is an independent, federal grantmaking agency that awards nearly one fourth of all federal grants in the United States, nearly 11,000 new grants per year, and has a current annual budget of $7.3 billion (NSF 2015). A majority of these grants support critical research in science and technology. In order to get a sense of how innovation research, as procured by the NSF, is distributed, one can visit the NSF website to locate some summary statistics about grant awards. However, the NSF website provides only basic information about the number of grants awarded and grant allocations by states and institutions. For any serious researcher who is interested in understanding the landscape of federal funding of innovation in the United States, one is met with an aggregate dataset that numbers in the millions of grants awarded since NSF's founding in 1950.

Given the sheer number of grants and the number of variables that denote grant type, this report began with a desire to develop an analytical methodology to understand the distribution of NSF grants and recognize any emergent patterns within these numbers. As such, this report studies NSF grants from the fiscal years 1999 to 2016 and aims to understand the distribution of NSF grants during this period. By developing a subset of the aggregate NSF data, we have been able to study the distribution of grant duration by type of grant (new grants, continuing grants, or revised grants), grants by recipient type (universities, private instititutions or corporations, etc.), by state, and by types of funded research.

The broader aim of this exercise in big data analytics is to be able to apply quantitative methods to larger datasets, acquire subsets, and devise collated dataframes that can then be useful in identifying distributional attributes. The findings in this report and this dataframe may be useful to scholars, grantseeking organizations and individuals, as well as state agencies and private institutions to better understanding the landscape of scientific research funding in the United States.

Methodology

As a starting point, our dataset is comprised of queries obtained from the publicly available https://www.usaspending.gov. Using multiple queries on the followig page https://www.usaspending.gov/Pages/TextView.aspx?data=HomeAwardTypeFunding , the range of fiscal years 1999-2016 was obtained. The variables that were selected for the working dataframe are as follows:

Grant Periods

- fiscal_year - Budget year
- starting_date - Start date of awarded research project

- ending_date - End date of awarded research project
- obligation_action_date - Start date of grant award obligation

Grant Amounts and Grant Type

- fed_funding_amount - Dollar amount of grant This is the total amount awarded for a particular grant.
- action_type - Type of Grant Award
  This variable tells us whether this is a new grant, a continuing grant, or revised grant. A new grant refers to a new award which typically entail a "specific level of support for a specified period of time", while continuing grants provide specified support for an initial period of time, "with a statement of intent to provide additional support of the project for additional periods, provided funds are available", while revised grants refer to existing NSF grants that have been amended with new terms (NSF 2015).

Grant Recipient Information

- recipient_name - Recipient Name
  This variable denotes the name of the recipient, usually a university or organization.
- recipient_type -Type of Recipient
  This variable allows us to identify whether the recipeint organization is a higher education institution, corporation, or non-profit, among other categories.
- cfda_program_title - Field of research
  This refers to the research discipline of a grant.

Grant Recipient Location

- recipient_city_name - Recipient's City
- principal_place_state - Recipient location state
- principal_place_state_code - Recipient location state code These variable allows us to identify the location of the recipient.

The following modifications were conducted to the existing variables: 1. conversion of non-numeric variable into categorical variables
2. dates were transformed to be treated as dates by R
3. dublicates of variable values were merged to be treated as identcal values
4. cfda_programm_title values were replaced by more concise Continue, New and Revision

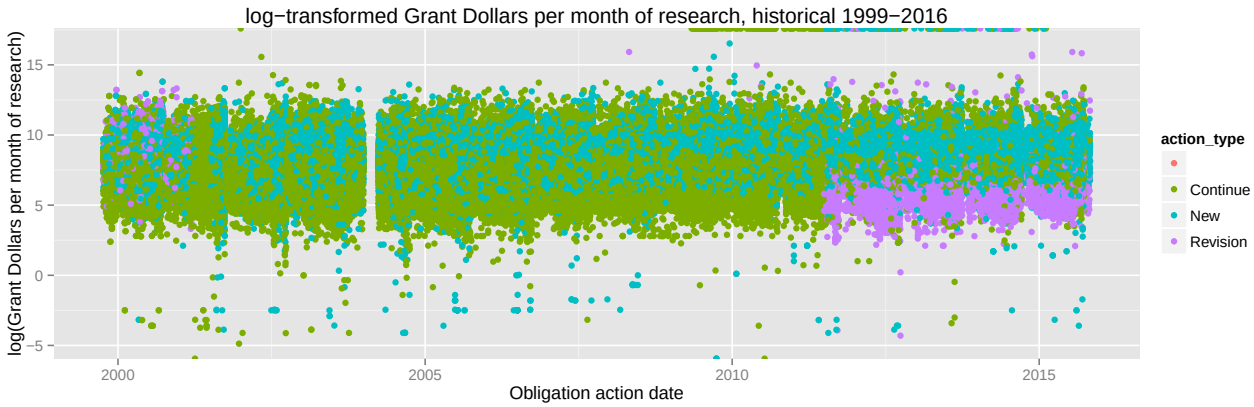
The following variables were added as new columns * res_duration - This is the duration of research obtained by calculating the difference between the ending and starting dates of research. Such values were divided by 30 to obtain duration in months.

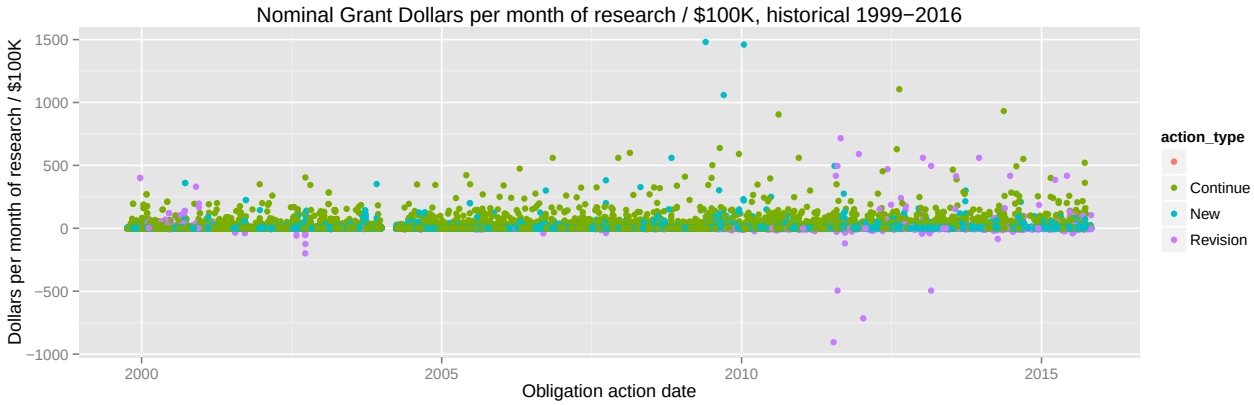
- costmon - The variable fed_funding_amount was divided by res_duration to establish the common base for comparing amounts alocated to different recipients.
- monstar, mondec - Month of research start and obligation action date respectively.
- logcost - Log-transormed costmon calculated for plotting these data.

The resulting dataset consists of 410,292 observations of 17 variables. Various subsets were created to calculate cumulative amounts granted to states, programs and by months. Subsets of top recipient states was also created for similar purposes.

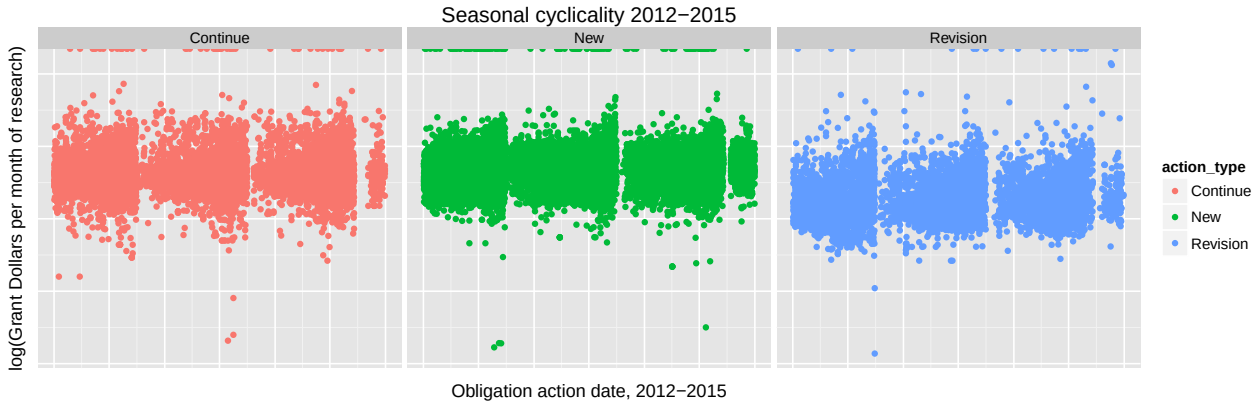
General observations

Initially, we mapped the total number of awarded grants (410,292) from 1999 to 2016 based on the *Obligation Action Date* and *Awarded Grant Dollars per month* divided by *Action Type*. Here, we observed a concentration of continuing grants using a log transformation to identify any relationships between the type of grant over this 17 year period. We also observed a steady number of all grants to be somewhere between $0 to $500,000, with some continuing awards reaching $1,000,000 and some new awards reaching $1,500,000. We also observed revised grants in the negatives. We attribute these negative numbers of nominal grant dollars (mostly in revision grants) to the fact that these grants often seek to make up for cost overruns.

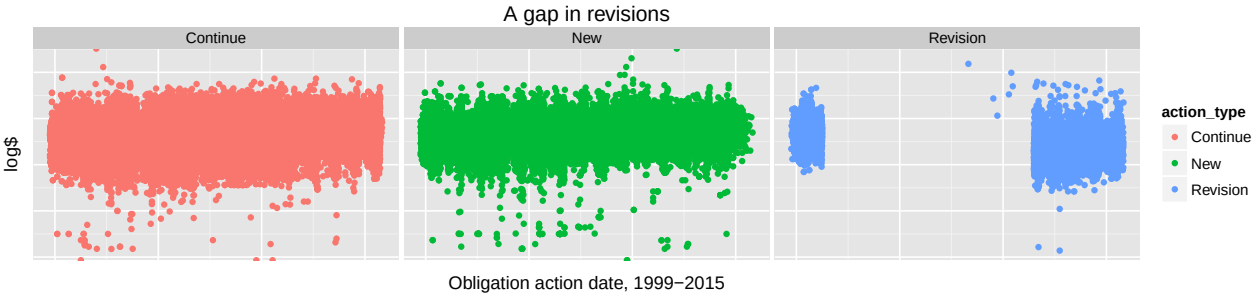




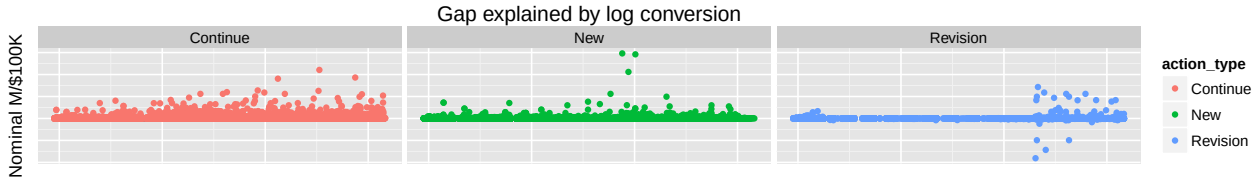
We then categorically analyzed the difference between continuing, new, and revised grants over the 17 year period using a log transformation of their allocated amounts. Here, we continued to see the steady flow of continuing grants, with some outliers of low grant awards. We also observed a more narrow, but steady concentration, of new grants, with even more outliers of lower grant awards. With this log transformation, we observed a large gap in the amount of revised grants.



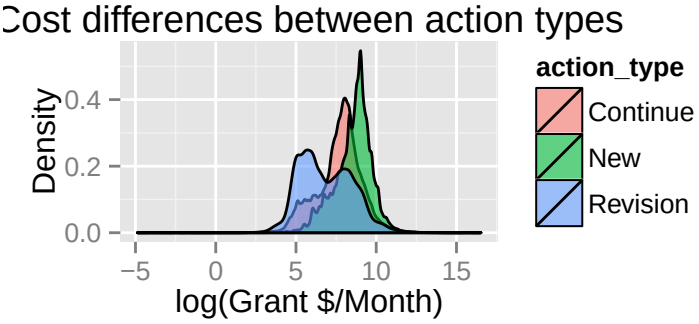
About this plot

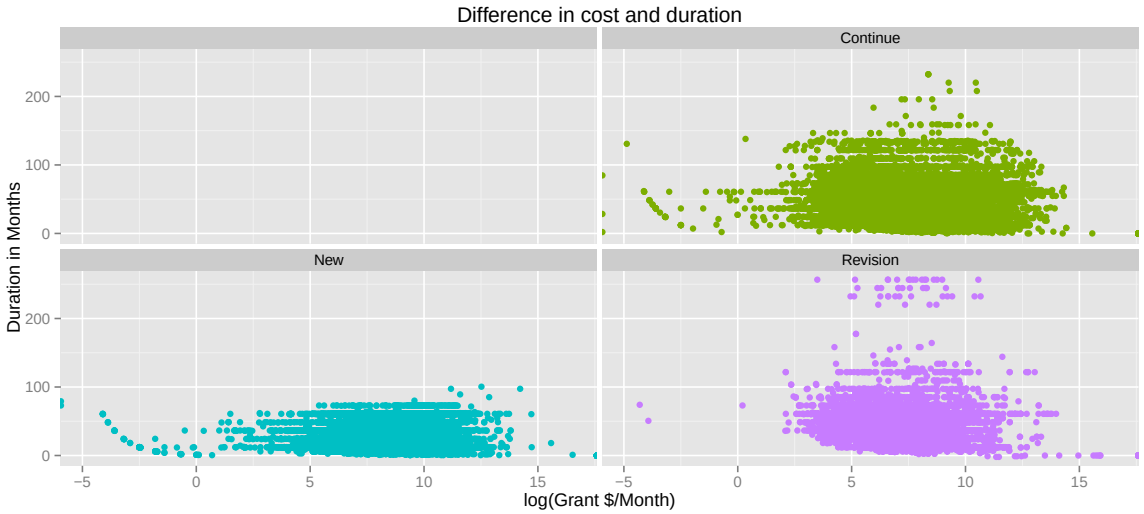


About this plot

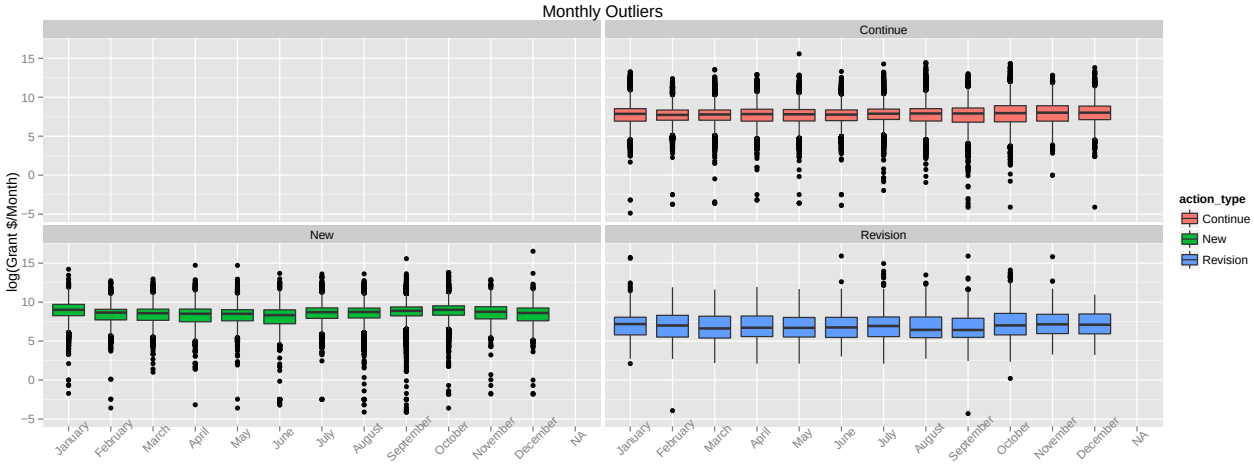


About this plot

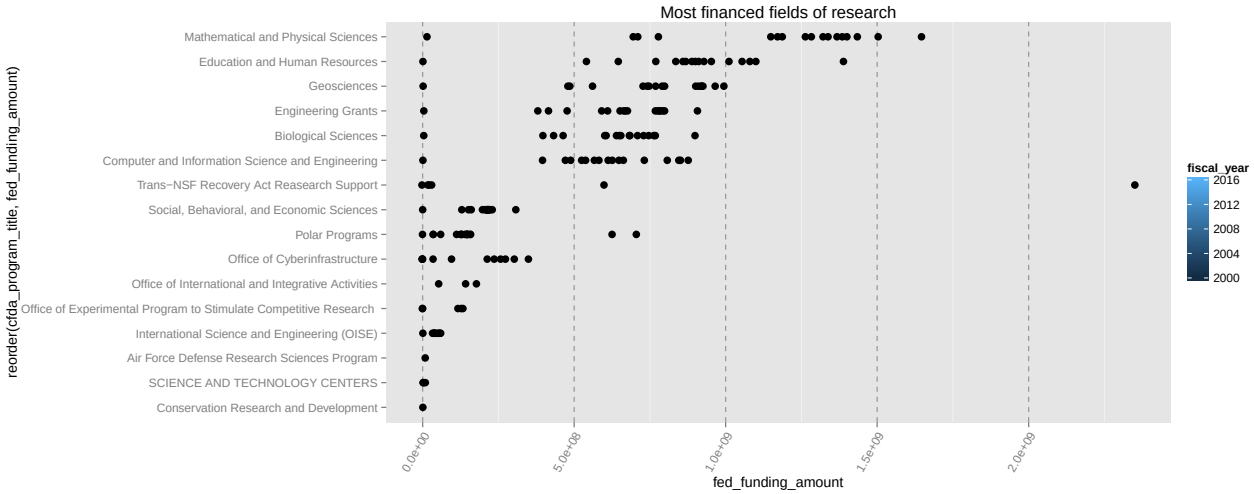




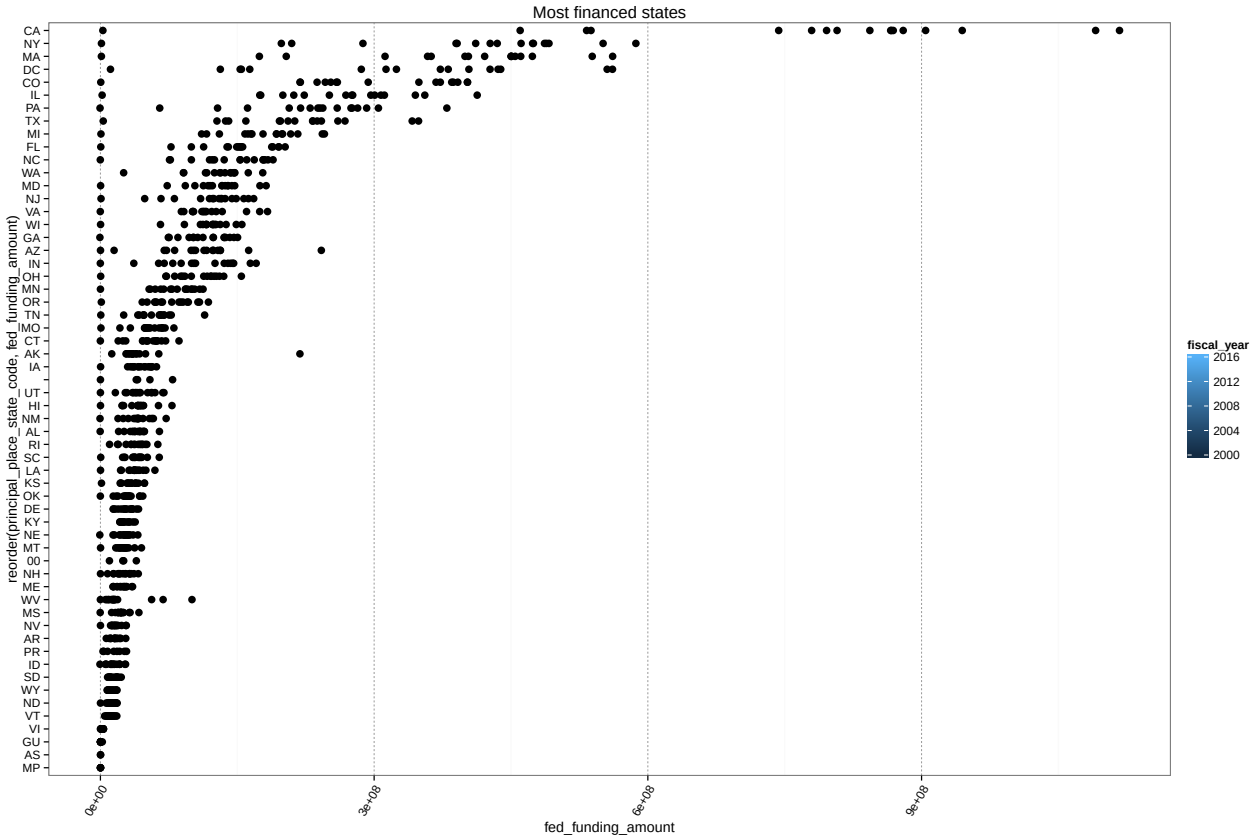
About this plot



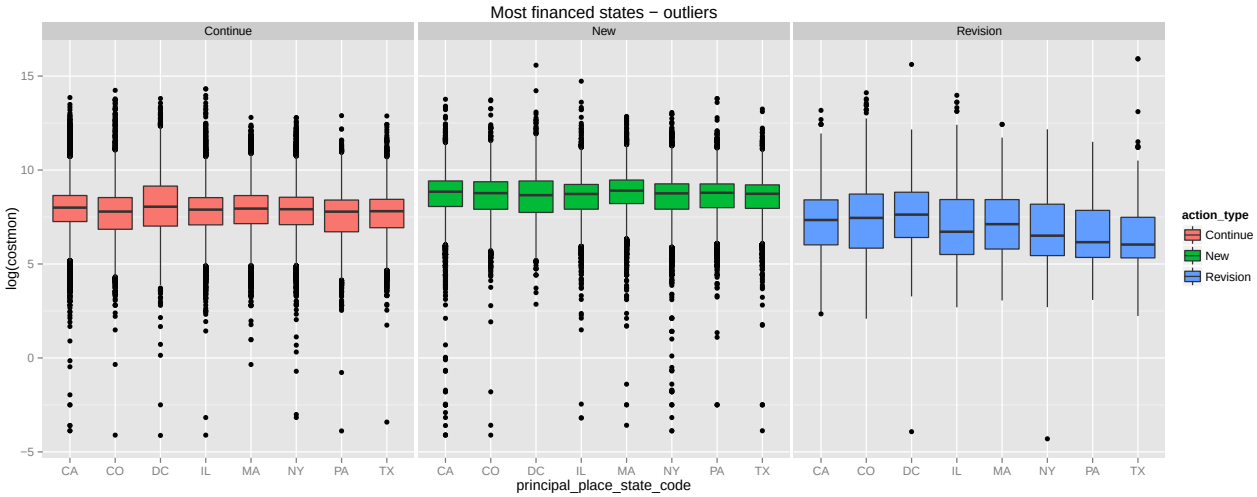
Fields of research



States

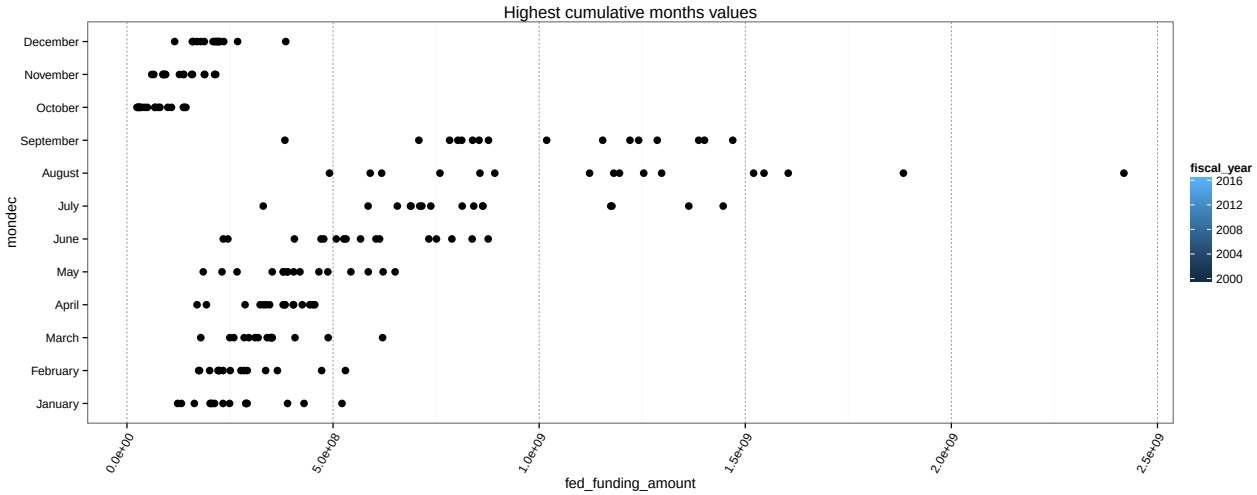


California, New York, Massachussetts and DC are the states with the highest annual cimilative allocations. Annual variance is high for these four but being a recipient state with the most allocations is consistent across 1999-2016. The range is significant - every tick is 300 million dollars so for example California range is 500 m - 1 billion which is the difference of 500 million dollars between the highest and lowest years for this state.

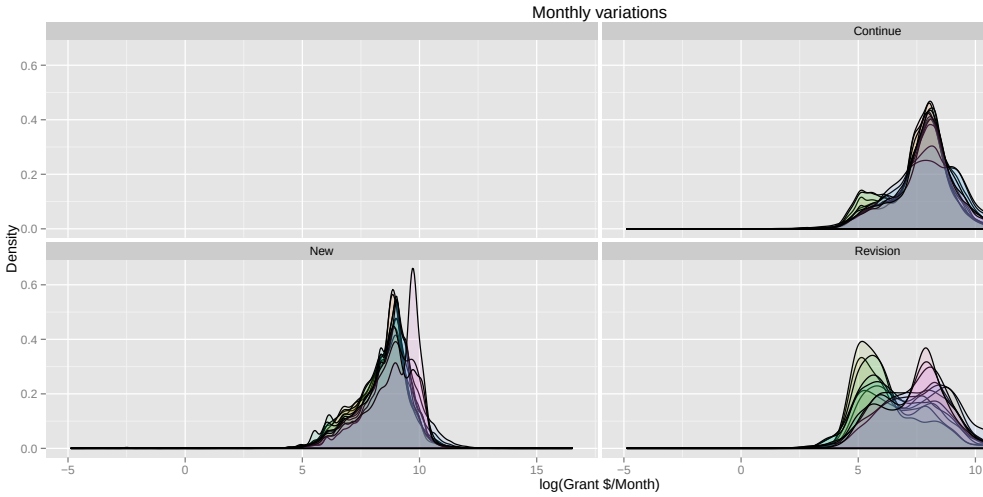


California, New York and DC consistently show more negative otliers (small grants) than other states given that medians with minor variations are roughly similar for most of the larfest states and the upper bounds for larger grants are roughly the same as well.

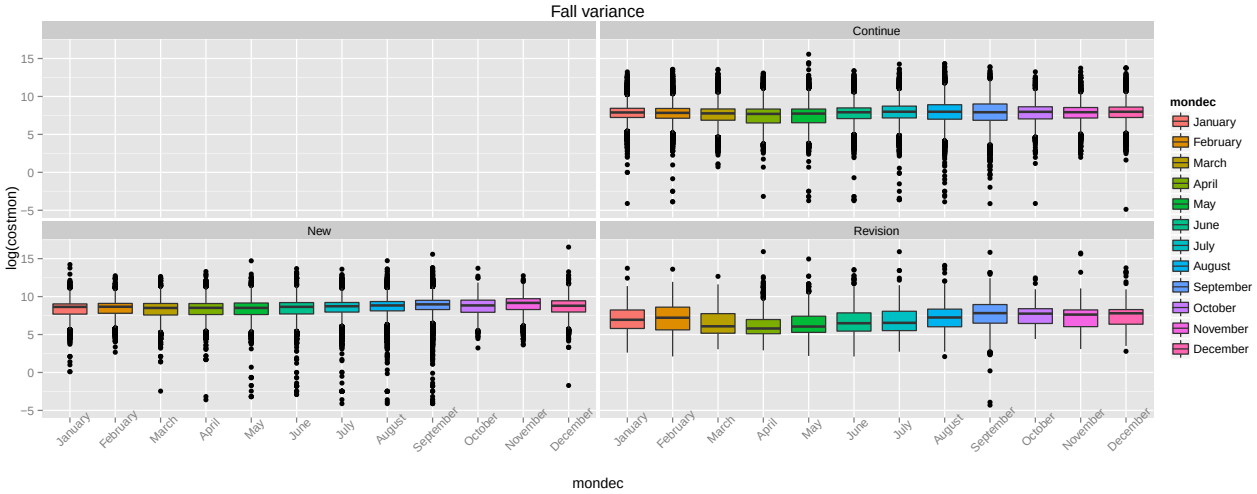
Months



Largest amounts consistently distributed in Summmer months and September - activity significantly slows down in the fall and winter Activity peaks in August and is slightly lower in September. We can conclude that

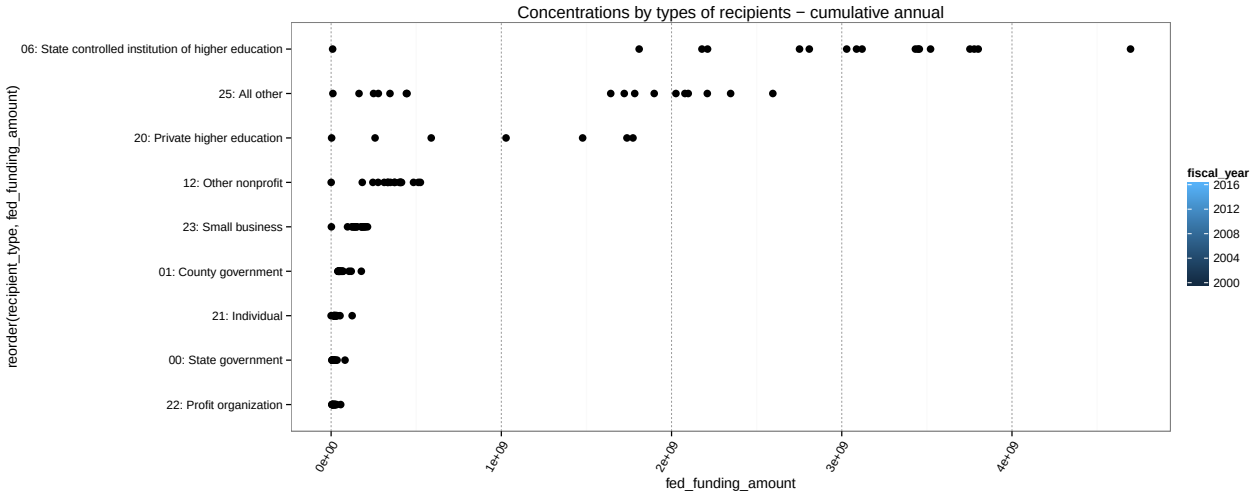


annual NSF cycle is October - September
Frequency changes month - to month, most significantly in Revised grants, creating a bimodal distribution
Continue and New - some shifts in allocations plus frequency varies making most frequent values more or less distinct depending on the month



There is consistent increase in variance during the summer, it slows down in September (activity in general)
A strong positive skew - a lot of negative outliers

Recipient



Highest cumulative amounts among state and private higher education institutions and "All other" but also the dispersion is quite high - it seems like there is no fixed proportion of total grant dollars that higher education institutions are recieving each year.

Discussion

Bibliography