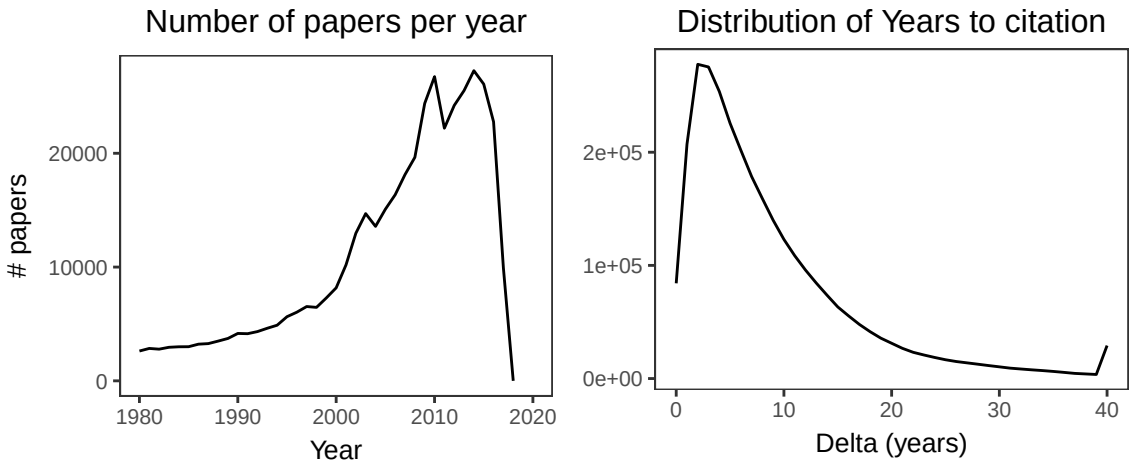
MAG - 'Climatology'

Jan Overgoor

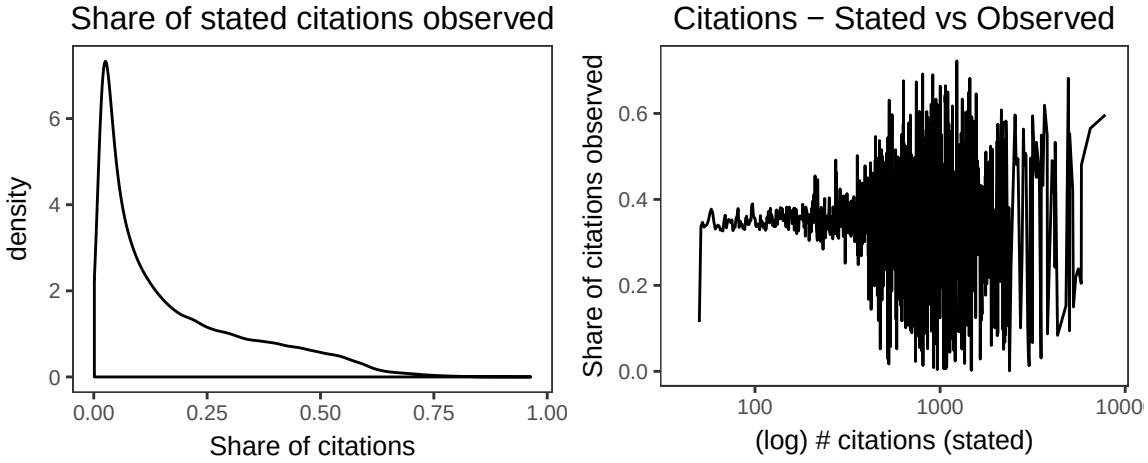
```
to build from command-line, run with:
R -e "rmarkdown::render('data_mag_cli.Rmd', output_file='data_mag_cli.pdf')"
```

- total number of papers (=nodes): 459309
- total number of references: 6464173
- total number of references to known nodes (=edges): 3039103
- share of references to known nodes: 0.4701457



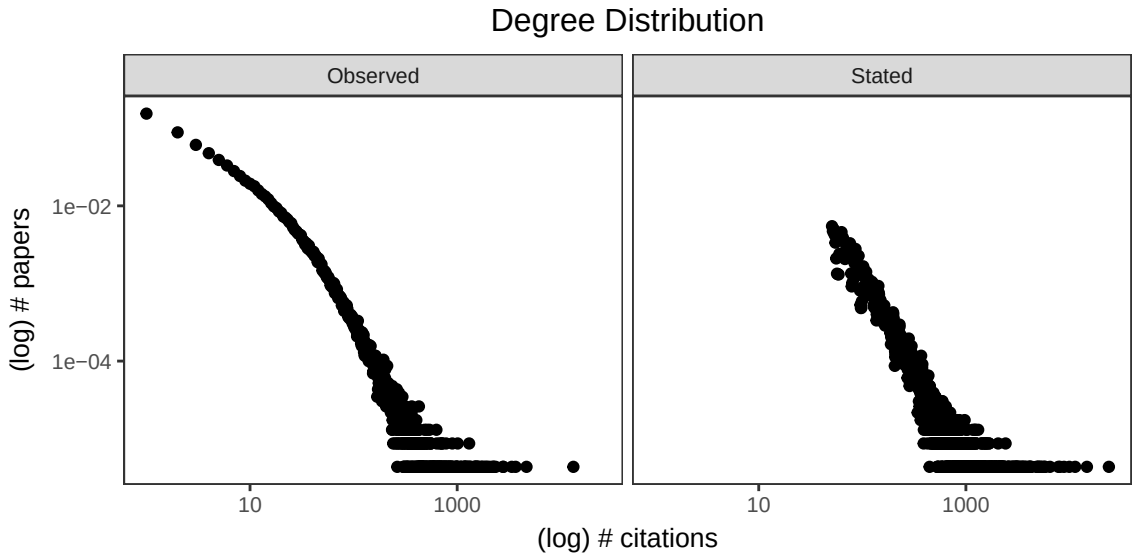
- Left: number of papers per year, linear increase since 2000, drop for recent years
- Right: distribution of years between publication and getting cited. Most citations happen within 2-3 years of publication.

Next, we compare the stated number of citations to the amount we can actually find in the data.



- Left: distribution of "share of stated observations observed", for mostly papers this is <25%. Since I look at the graph filtered by field of study, citing papers might not be included. The whole graph is hard to work with, so this is what we got.
- Right: per "stated number of citations", what is the average "share of citations observed"? Very stable by x, with a much higher variance for the highly cited papers (as there are fewer of them).

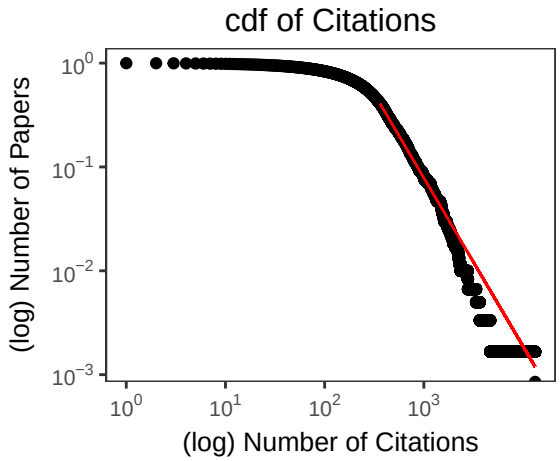
Are the degree distributions similar for the stated and observed citation counts?

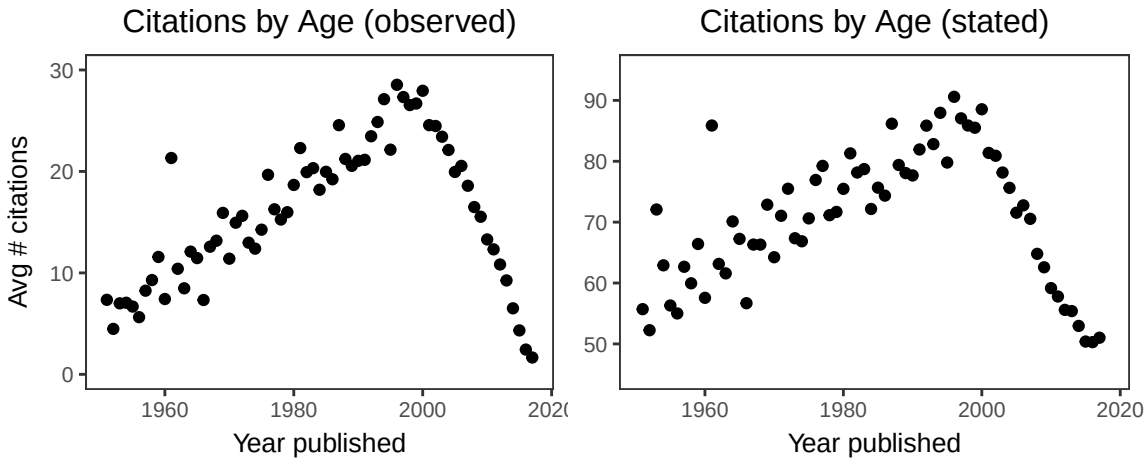


Yes, and this shows the censoring at 50 very clearly.

Here is the cdf and Clauset-Shalizi-Newman powerlaw fit:

```
## [1] "plfit:  alpha=2.621  xmin=361"
```

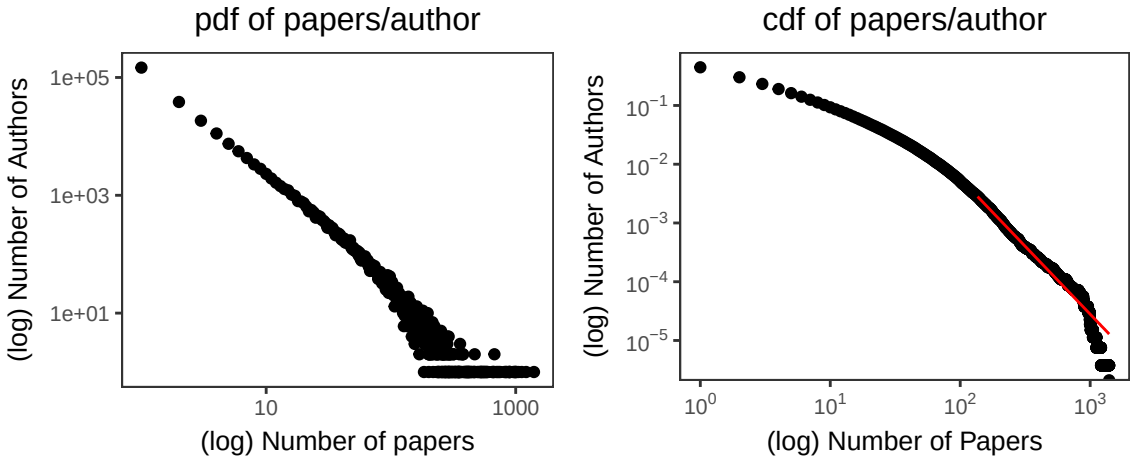




- Left: average number of citations by year of publishing (as observed). linear increase until 2000 (newer papers more cited), but then drops off
- Right: same, but as stated. The trend is the same, but the numbers are inflated by about 50.

Here is the distribution of papers/author:

```
## [1] "plfit:  alpha=3.311  xmin=136"
```



Very heavy-tailed as well.

What are the top keywords?

keywords	n
climate change	38570
seasonality	33465
pacific ocean	15323
sea surface temperature	14418
climate	14368
north america	13163
north atlantic	12576
atmospheric circulation	12270
climate variability	11833
temperature	11154

Model

Data construction process:

- sample 5000 citations from after 2011
- for each actual citation, sample 24 non-cited papers (from before publication date)
- for each of the (paper,option) pairs, compute features (n citations, years since, has same author)

```
##
## ================================================================================
##                                            y
##                           (1)        (2)        (3)        (4)        (5)
## --------------------------------------------------------------------------------
## log Citations          0.955***   1.281***   1.062***   1.311***   1.231***
##                        (0.011)    (0.015)    (0.013)    (0.017)    (0.018)
##
## log Age                          -1.142***             -1.078***  -0.864***
##                                   (0.024)              (0.026)    (0.031)
##
## Has same author                                5.985***   5.277***   5.061***
##                                                (0.131)    (0.132)    (0.137)
##
## # same keywords                                                     1.613***
##                                                                    (0.047)
##
## log(max_n_papers + 1)                                              0.213***
##                                                                    (0.017)
##
## --------------------------------------------------------------------------------
## Observations            4,831      4,831      4,831      4,831      4,831
## Log Likelihood        -9,785.178 -8,373.577 -7,936.322 -6,911.177 -6,377.661
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01

## [1] "Train accuracy:"

## [1] 0.3507205 0.4171994 0.4703967 0.5238489 0.5678299 0.5657214

## [1] "Test accuracy:"

## [1] 0.3631890 0.4378109 0.4852071 0.5359281 0.5912263 0.5900000
```