BAN432 Applied Textual Data Analysis for Business and Finance

Geographic Dispersion and Stock Returns

Maximilian Rohrer and Christian Langerfeld

September 28, 2022

Overview

- Collecting textual data
- Preprocessing and cleaning
- Analysis
  - Regex-based application: Geographical dispersion
  - ...

Plan for this lecture

- Garcia, D./Norli, O., 2012, Geographic Dispersion and Stock Returns. Journal of Financial Economics
- Repeating Garc?a & Norli (2012)
- Introduction to corpus we will use in many lectures

- Main take-away
  1. Regular expressions
  2. First replication of a Finance paper using text-mining

Geographic Dispersion and Stock Returns

ABSTRACT:

This paper shows that stocks of truly local firms have returns that exceed the return on stocks of geographically dispersed firms by 70 basis points per month. By extracting state name counts from annual reports filed with the Securities and Exchange Commission (SEC) on Form 10-K, we distinguish firms with business operations in only a few states from firms with operations in multiple states. Our findings are consistent with the view that lower investor recognition for local firms results in higher stock returns to compensate investors for insufficient diversification.

Garcia, D./Norli, O., 2012, Geographic Dispersion and Stock Returns. Journal of Financial Economics 106. 547-565.

Main table

Table 3
Average return on portfolios sorted by geographic dispersion.

The table reports average portfolio returns in percent. Geographic dispersion is measured as the number of U.S. states mentioned in the annual report filed on Form 10-K with the SEC. Five portfolios are formed based on geographic dispersion. A firm that files a 10-K form on or before June of year *t* is eligible for inclusion in a portfolio starting in July of year *t*. The firm carries its state count up to and including June of the next year. A firm gets added to the portfolio of local firms if its state count is below the 20th percentile in the cross-section of state counts. Correspondingly, a firm gets added to the portfolio of dispersed firms if the state count exceeds the 80th percentile. Three more portfolios are formed using the 40th and the 60th percentiles as breakpoints. The sample period is July 1994 through December 2008.

Variable	Local	2	3	4	Dispersed	Local-dispersed
EW returns	1.18	0.97	0.87	0.83	0.62	0.56 (2.73)
VW returns	0.89	0.78	0.74	0.58	0.49	0.40 (2.06)
Average number of firms	1,084	830	784	757	818	

Figure 1: Table 3 from Garcia and Norli (2012)

Summary of the paper

- **Ecnomic mechanism** Merton 1987 characterizes equilibrium stock returns when investors are not aware of all securities. Stocks with lower investor recognition have higher expected returns to compensate investors that hold the stock for insufficient diversification. Stocks that have a very local focus will likely have a smaller investors base.
- **Econometric procedure** The authors download annual reports (10-Ks) from EDGAR, and count how many unique US state names are named within a given report. Subsequently, they regress return on this measure of geographic dispersion.
- **Contribution** Known measures of geographical exposure are not as accurate as the one suggested in this paper.
- **This course** You already know all ingredients to redo this analysis!

Motivating example: Walmart

- In 2005, no sales split data available at all
- US states mentioned in annual report 2005
- Compare to store distribution in 2007

Walmart: US state count in 10K

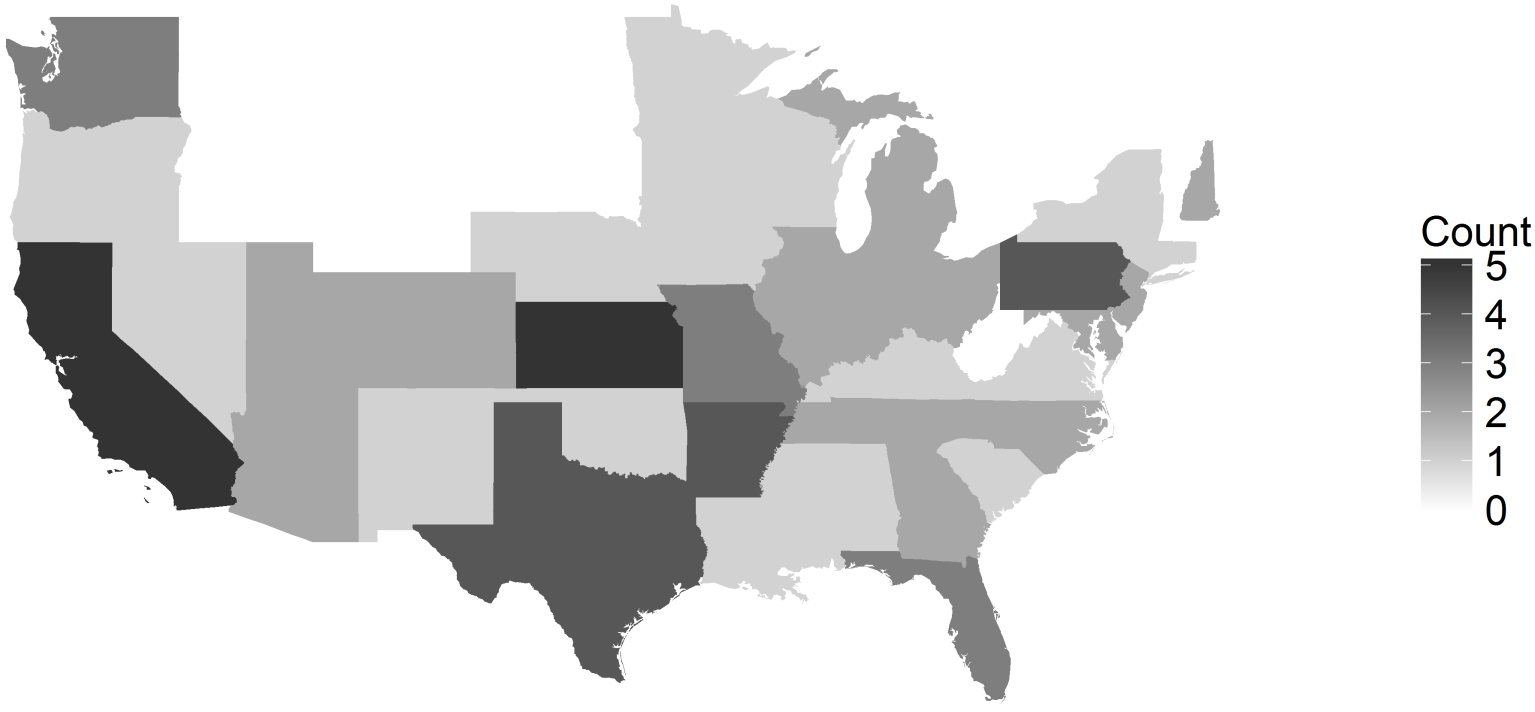


Figure 2: US state count 2005

Walmart: US states store map

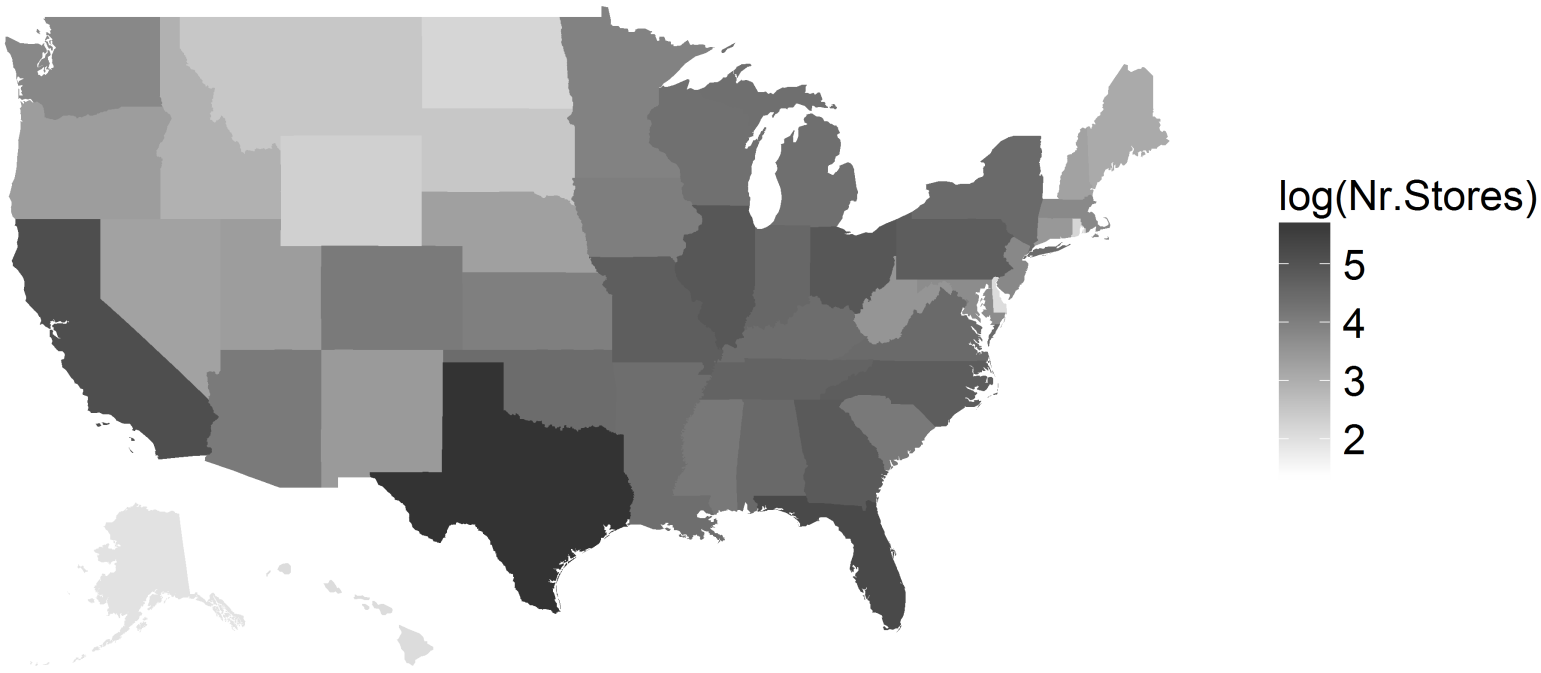


Figure 3: stores per US states

A corpus for the course: `firm_dataset.Rda`

Three files

1. `raw.data`
2. `section.1.business`
3. `section.7.mda`

Tibble covering all 10-K filings for 500 randomly selected firms in calender year 2013

```
require(tibble)
require(dplyr)
load("firm_dataset.Rdata")
```

```
ls()
```

```
## [1] "raw.data"              "section.1.business" "section.7
```

raw.data

- cik = central index key, firm identifier
- permno = firm identifier
- date.filed = date of the corporate filing
- date.filed.y = year of the corporate filing
- exchange.code = code of the exchange, 1=NYSE & 2=NYSE American & 3=NASDAQ
- industry.fama.french.49 = 49 industries as defined by Fama and French
- industry.hoberg.phillips.50 = 50 industries as defined by Hoberg and Philips
- beta.market.monthly.5y = exposure to the market estimated 5 years prior to the filing using monthly returns

raw.data

- Accounting items: `total.assets`, `operating.income`, `total.liabilities`, `revenue`, `book.equity` as released with the 10-K.
- `market.value` = market value end-of-month before the filing date
- `?.lag` & `?.forw` = for the previous two bullet points the value for the previous/following filing is included with the suffix `?.lag` & `?.forw`
- `return.monthly.?` = monthly return during December of the current calender year `?.CY.m12` and all months of the next year `?.NY.m01-12`
- `return.daily.eventday.?` = daily return on days relative to the filing/event day. Starts 10 days before filing `?.lag.10` and ends 30 days after it `.forw.30`
- `filing.size.bytes` = file size of cleaned 10K filing in bytes

section.1.business and section.7

- Vector containing sections from the annual report as string
- Section 1 Business: requires a description of the company's business, including its main products and services, what subsidiaries it owns, and what markets it operates in. (Source: SEC)
- Section 7 Management's Discussion and Analysis of Financial Condition and Results of Operations: gives the company's perspective on the business results of the past financial year. This section, known as the MD&A for short, allows company management to tell its story in its own words. (Source: SEC)

Small reminder on tibbles and `dplyr`

- Improvement on data.frames
- `dplyr`: grammar for data manipulation, see link
  - %>%: pipe to nest functions together
  - `mutate()`: adds new variables that are functions of existing variables
  - `select()`: picks variables based on their names
  - `filter()`: picks cases based on their values
  - `summarise()`: reduces multiple values down to a single summary
  - `arrange()`: changes the ordering of the rows

Our goal

Reproduce Figure 1 and Table 2 from Garcia and Norli (2012)

Construct a measure of dispersion by counting US state names in annual reports

Figure 1

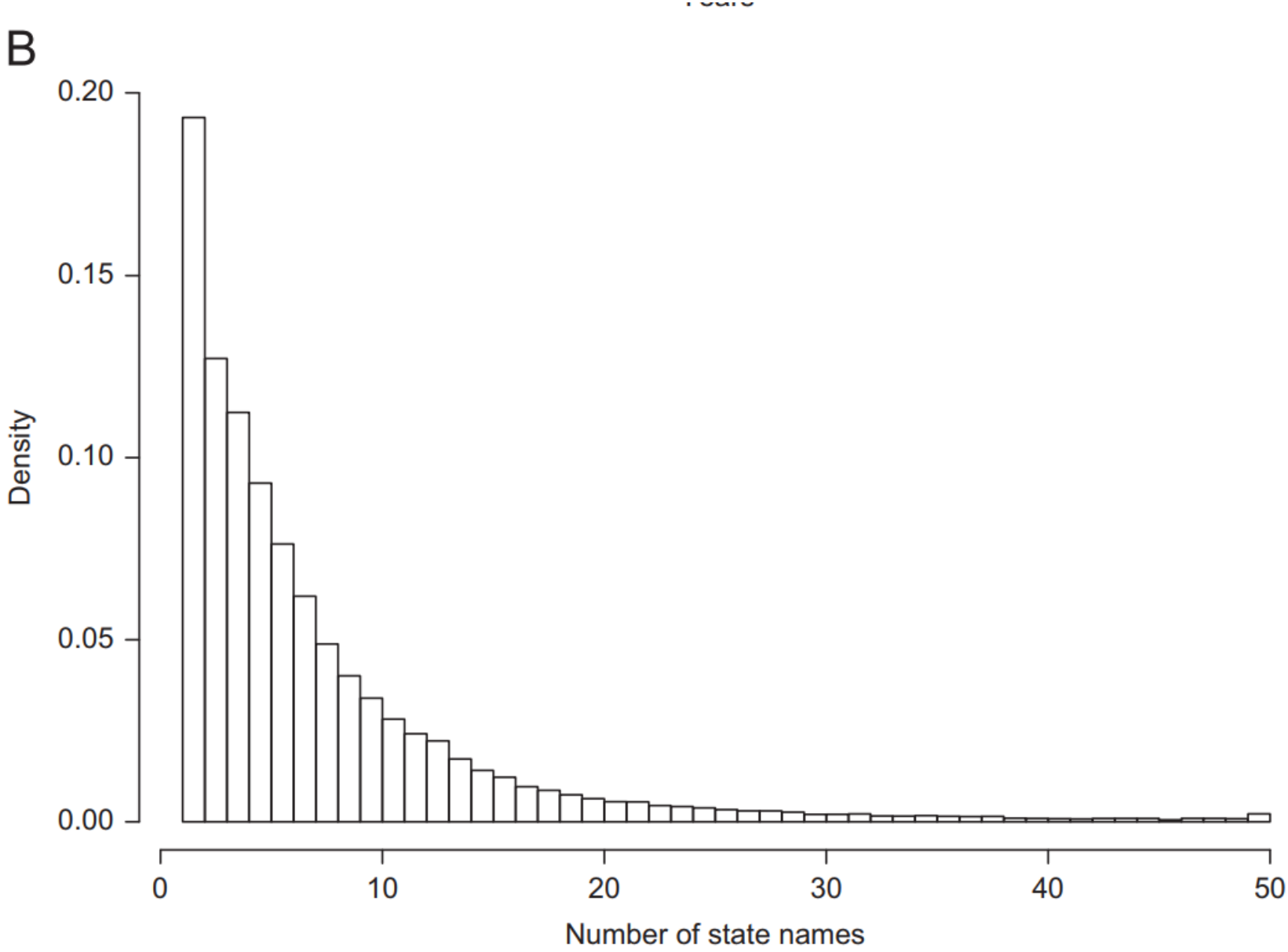

B
0.20
0.15
0.10
0.05
0.00
Density
0
10
20
30
40
50
Number of state names

Table 2

Table 2
Geographic dispersion and other firm characteristics.

Geographic dispersion is measured as the number of U.S. states mentioned in the annual report filed on Form 10-K with the SEC. Geographic dispersion for year t is the number of U.S. states mentioned in the last annual report filed prior to July of year t. Size (the market value of common equity) and the Book-to-market ratio are computed as described in Fama and French (1993). Amihud illiquidity is the price impact liquidity measure of Amihud (2002). Bid–ask spread is the proportional quoted spread measured as: $100(P_A-P_B)/(0.5P_A+0.5P_B)$, where P_A is the ask price and P_B is the bid price. Volatility is computed as the standard deviation of the error term from a regression using the three-factor model of Fama and French (1993) on one month worth of daily data. Momentum is the buy and hold return for months $t-12$ through $t-2$. In Panel A, all variables are measured as of July each year and the panel reports time series averages of cross-sectional averages. Panel B reports results from a pooled time series cross-sectional regression. All variables in the regression are measured in natural logs. For the momentum variable, the natural log is computed from 1+Momentum. YEARS indicates the presence of dummy variables for each year. DIVS indicates the presence of dummy variables for each of nine U.S. Census divisions. INDS indicates the presence of dummy variables for each of 12 industries from Ken French's Web site. The column labeled R^2 reports adjusted R-squared. The column labeled N reports the number of firm-months used in the regression. Parentheses contain t-statistics computed from the heteroskedasticity-consistent standard errors of White (1980). The sample period is July 1994 through December 2008.

Panel A: Averages by geographic dispersion quintiles

Variable	Local	2	3	4	Dispersed
Geographic dispersion	1.9	3.8	5.7	8.8	19.9
Size (ME)	1,732	1,640	1,862	2,645	3,963
Book-to-market ratio (BEME)	0.689	0.688	0.707	0.760	0.767
Amihud illiquidity (AMI)	0.028	0.021	0.012	0.014	0.006
Bid–ask spread (SPR)	0.030	0.028	0.026	0.023	0.018
Volatility (VOL)	0.031	0.032	0.031	0.028	0.025
Momentum (MOM)	0.150	0.119	0.108	0.107	0.110

Figure 5: Table 2 from Garcia and Norli (2012)

Ingredients

- Data:
  - Business description in annual reports: `section.1.business`
  - Firm data: `raw.data`
- Coding:

Ingredients

- Data:
  - Business description in annual reports: `section.1.business`
  - Firm data: `raw.data`
- Coding:
  1. ...
  2. ...
  3. ...

US States

R offers a vector with the 50 states of the US

```
us.states <- datasets::state.name

head(us.states)
```

```
## [1] "Alabama"    "Alaska"     "Arizona"    "Arkansas"
## [6] "Colorado"
```

Are those regex? Some adjustments?

```
us.states <- gsub("\\s", ".", us.states)
us.states <- tolower( us.states)
```

Construct an aggregate state count per state

Figure 1

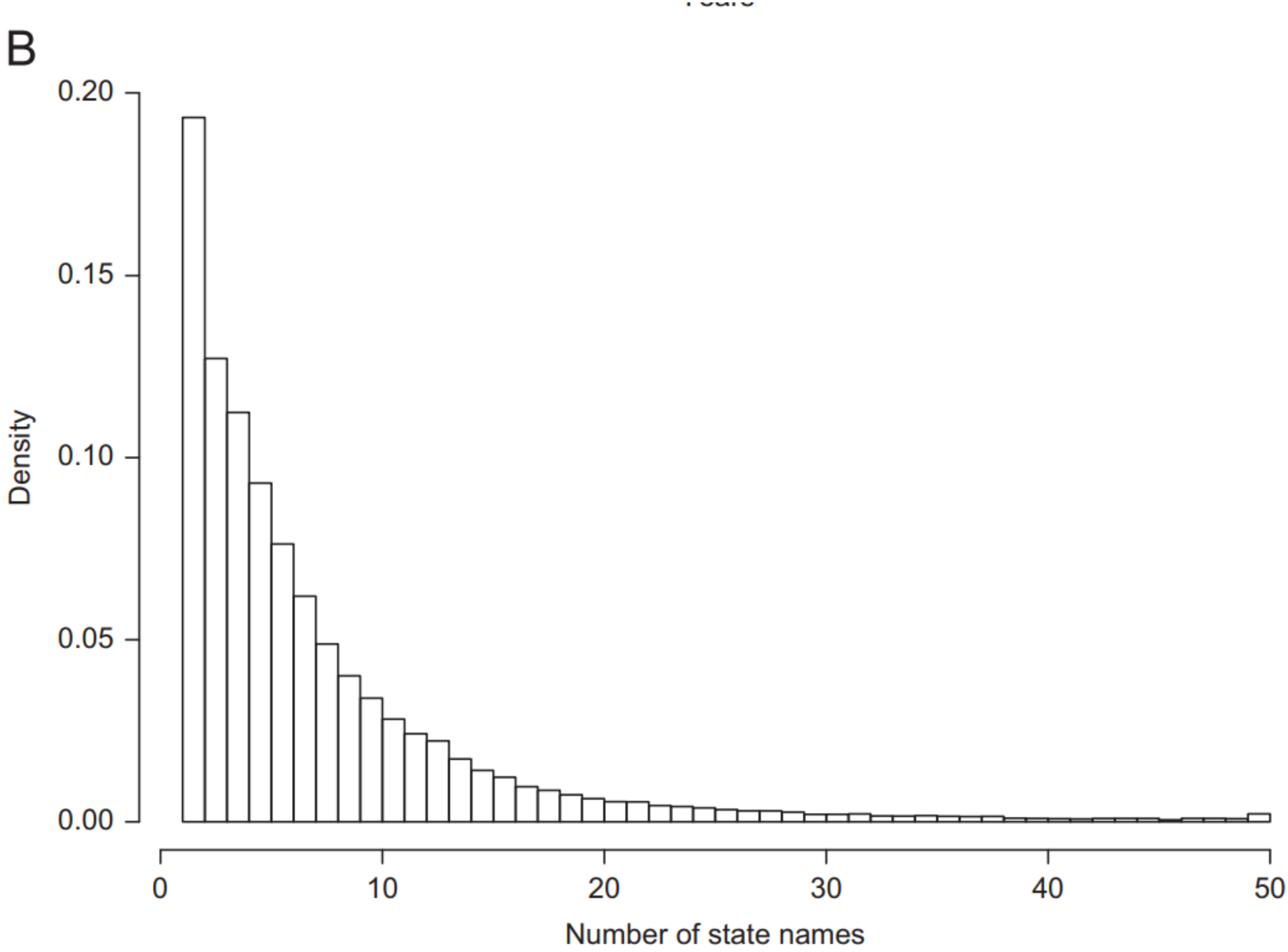
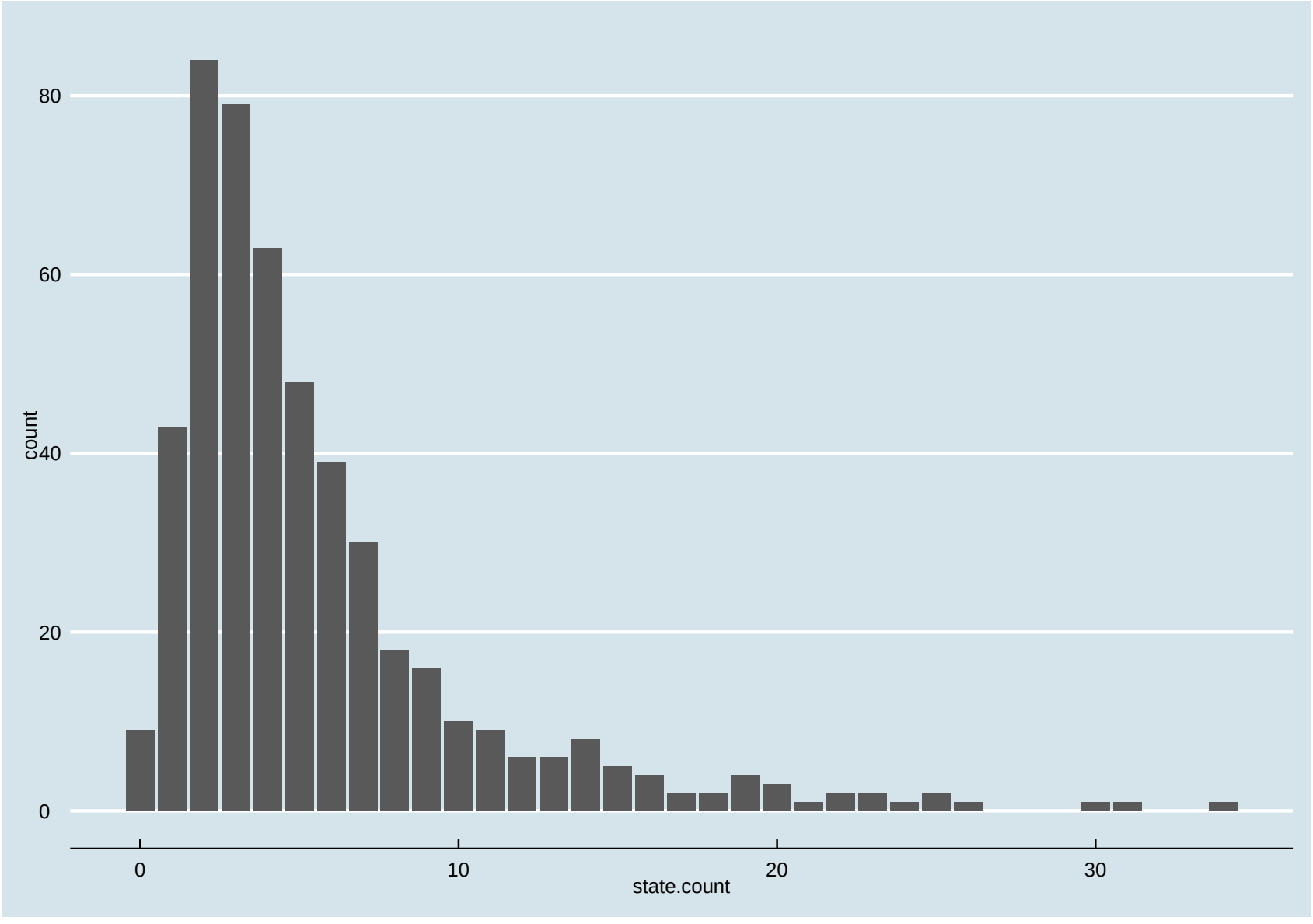

B
0.20
0.15
0.10
0.05
0.00
Density
0
10
20
30
40
50
Number of state names

Figure 1 using our results



Reproducing Figure 1

- alternative plots

```
hist(raw.data$state.count)
barplot(table(raw.data$state.count))
```

- Why differences?
  - Selection of 10-K section
  - Selection of sample year

Figure 1

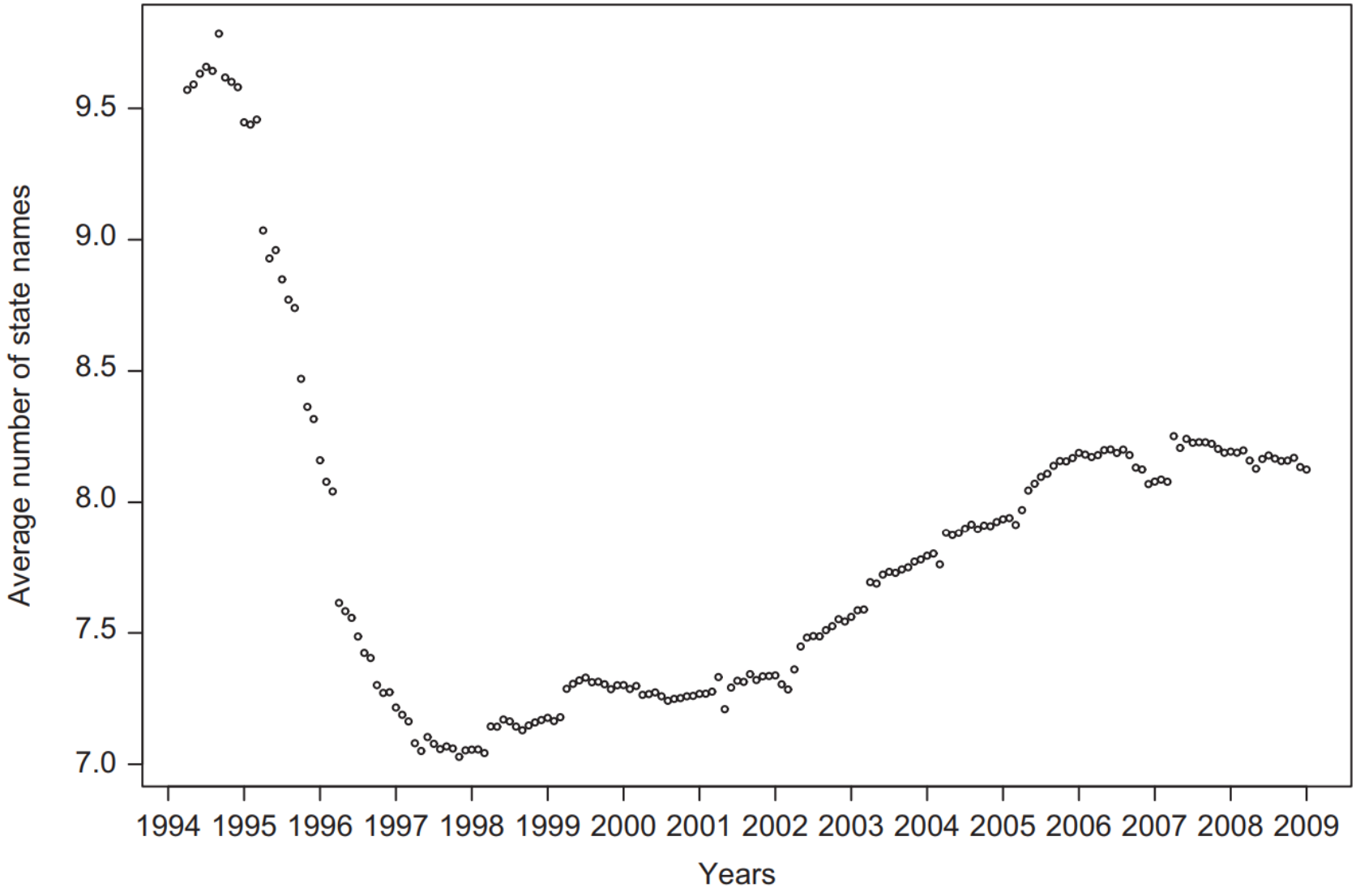


Figure 7: Figure 1 from Garcia and Norli (2012)

Table 2

Table 2
Geographic dispersion and other firm characteristics.

Geographic dispersion is measured as the number of U.S. states mentioned in the annual report filed on Form 10-K with the SEC. Geographic dispersion for year t is the number of U.S. states mentioned in the last annual report filed prior to July of year t. Size (the market value of common equity) and the Book-to-market ratio are computed as described in Fama and French (1993). Amihud illiquidity is the price impact liquidity measure of Amihud (2002). Bid–ask spread is the proportional quoted spread measured as: $100(P_A-P_B)/(0.5P_A+0.5P_B)$, where P_A is the ask price and P_B is the bid price. Volatility is computed as the standard deviation of the error term from a regression using the three-factor model of Fama and French (1993) on one month worth of daily data. Momentum is the buy and hold return for months $t-12$ through $t-2$. In Panel A, all variables are measured as of July each year and the panel reports time series averages of cross-sectional averages. Panel B reports results from a pooled time series cross-sectional regression. All variables in the regression are measured in natural logs. For the momentum variable, the natural log is computed from 1+Momentum. YEARS indicates the presence of dummy variables for each year. DIVS indicates the presence of dummy variables for each of nine U.S. Census divisions. INDS indicates the presence of dummy variables for each of 12 industries from Ken French's Web site. The column labeled R^2 reports adjusted R-squared. The column labeled N reports the number of firm-months used in the regression. Parentheses contain t-statistics computed from the heteroskedasticity-consistent standard errors of White (1980). The sample period is July 1994 through December 2008.

Panel A: Averages by geographic dispersion quintiles

Variable	Local	2	3	4	Dispersed
Geographic dispersion	1.9	3.8	5.7	8.8	19.9
Size (ME)	1,732	1,640	1,862	2,645	3,963
Book-to-market ratio (BEME)	0.689	0.688	0.707	0.760	0.767
Amihud illiquidity (AMI)	0.028	0.021	0.012	0.014	0.006
Bid–ask spread (SPR)	0.030	0.028	0.026	0.023	0.018
Volatility (VOL)	0.031	0.032	0.031	0.028	0.025
Momentum (MOM)	0.150	0.119	0.108	0.107	0.110

Figure 8: Table 2 from Garcia and Norli (2012)

Reproducing Table 2

- Constructing bins

```
cut(x,
    breaks,
    include.lowest = FALSE)
```

- Computing a measure for each bucket

```
aggregate(x,
          by = list(),
          FUN)
```

Reproducing Table 2

```
# we use quintiles
q <- quantile(raw.data$state.count,
              c(0, .2, .4, .6, .8, 1),
              na.rm=T)
raw.data$state.count.quintile <- cut(raw.data$state.count,
                                     breaks = q,
                                     include.lowest = T)
table(raw.data$state.count.quintile)
```

```
##
##   [0,2]   (2,3]   (3,5]   (5,8] (8,34]
##     136      79     111      87      87
```

Reproducing Table 2: validation

```
##   Group.1 state.count
## 1   [0,2]    1.551471
## 2   (2,3]    3.000000
## 3   (3,5]    4.432432
## 4   (5,8]    6.758621
## 5  (8,34]   14.436782
```

Reproducing Table 2: Mean

```
##   Group.1 market.value total.assets book.equity
## 1   [0,2]     6103.511     5303.851    2112.814
## 2   (2,3]     3356.935    16199.091    2278.111
## 3   (3,5]     3112.740     4661.894    1229.156
## 4   (5,8]     3347.774     4224.829    1824.588
## 5  (8,34]     5862.868    10079.270    3481.914
```

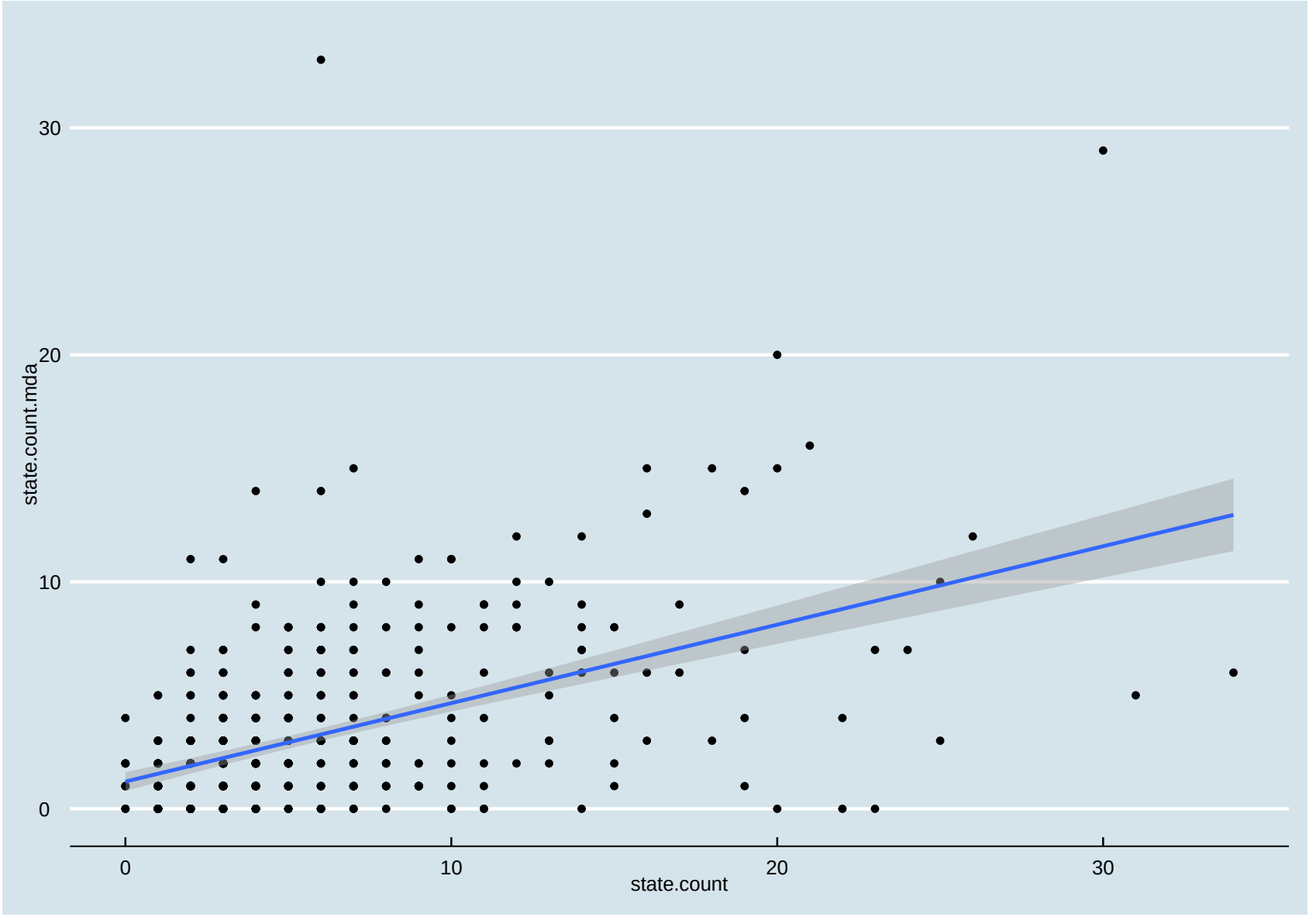
Reproducing Table 2: Median

```
##   Group.1 market.value total.assets book.equity
## 1   [0,2]     617.7509      908.289    295.2925
## 2   (2,3]     811.0892     1043.183    407.5620
## 3   (3,5]     680.7192      919.853    308.8330
## 4   (5,8]     899.5433     1040.798    460.4290
## 5  (8,34]    1551.1078     1955.972    694.2100
```

Extension you can evaluate

- Does it make a difference to use Section 7 MD&A?
- What about returns? You can reproduce Table 3 (ambiguous results)
- Which states do firms mention most?

Extension 1



Extension 3

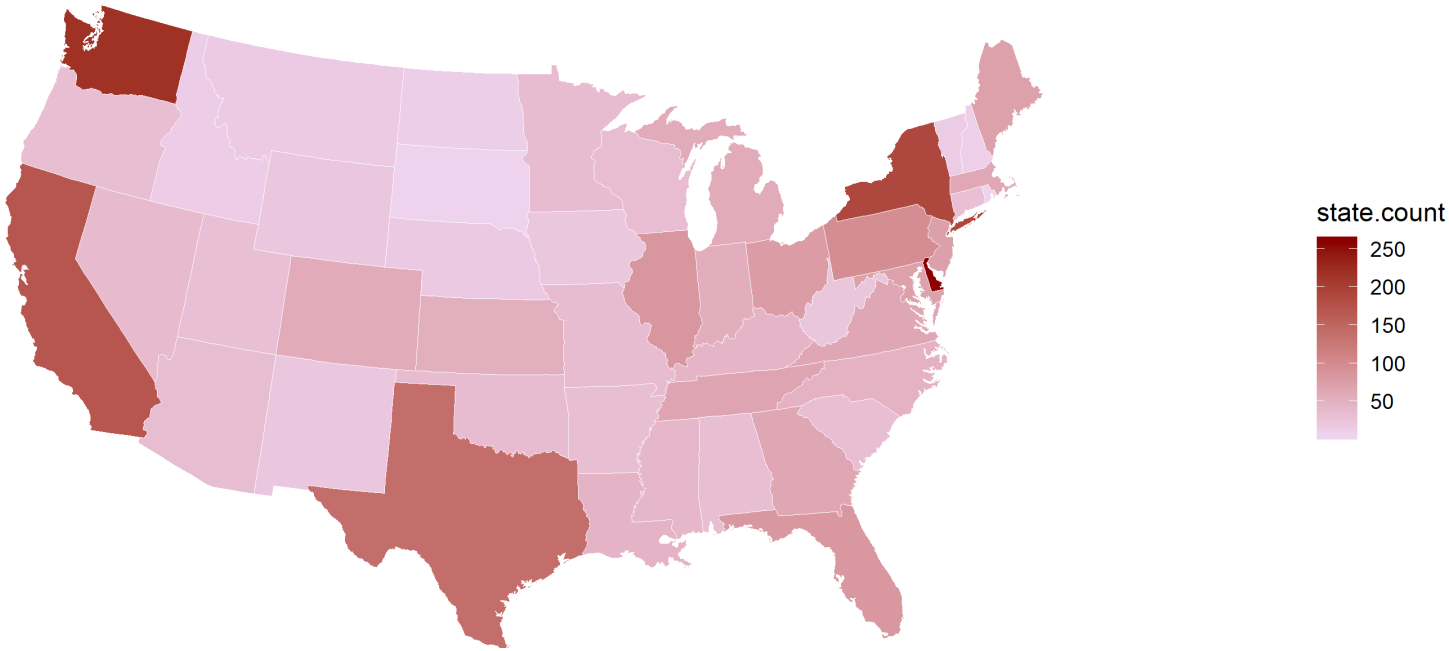


Figure 9: State count

Summary

- Using textual analysis one can extract a good measure of geographic dispersion of activities from annual reports
- Albeit only using regular expressions, we already reproduced a very well published and cited finance paper

Appendix: A tidy way to count states

This would give the number of times any state is mentioned

```
require(stringr)

sapply(us.states, function(x) {
  str_count(string = tolower(section.1.business),
            pattern = x)}) %>%
  rowSums() -> raw.data$state.count
```