FAKECOMPUSTAT

An R package for financial data replication

Dylan Phillips and Candace Jens

PURPOSE

- To produce an R package designed to create fully synthetic Compustat and CRSP-style datasets for use in financial paper replication and methodological testing

PACKAGE GOALS

- Generated data must follow accounting and financial principles
  - Ex. Assets = Liabilities + Equity

PACKAGE GOALS

- Generated data must follow accounting and financial principles
  - Ex. Assets = Liabilities + Equity
- Generated distributions must be realistic on both the firm and market-levels
  - Must simulate "growth," cannot use pure randomness

PACKAGE GOALS

- Generated data must follow accounting and financial principles
  - Ex. Assets = Liabilities + Equity
- Generated distributions must be realistic on both the firm and market-levels
  - Must simulate "growth," cannot use pure randomness
- Generated data cannot be matched to any real company
  - Simple data permutation is not sufficient

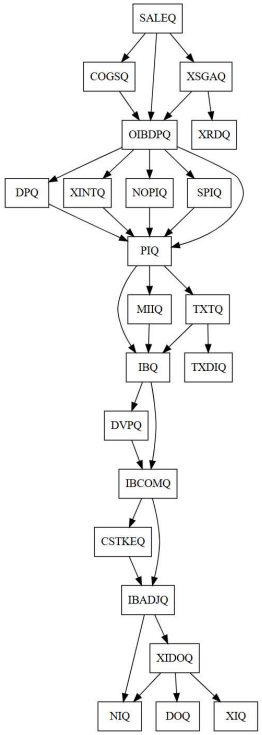
METHODOLOGY

1. Key variables (assets, sales, stock price) will be simulated along a growth path
   - Starting point is randomly drawn from distribution of values at a sector and year level
   - Subsequent values are created based on distribution of growth

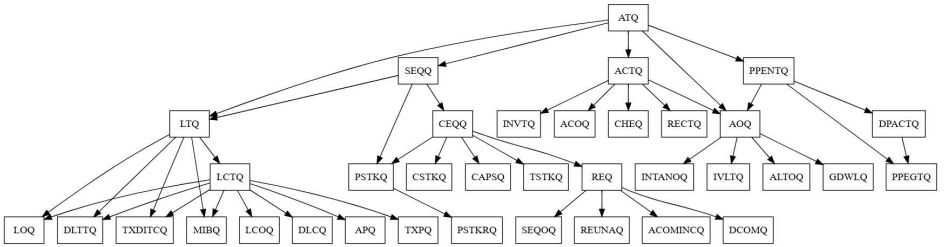
METHODOLOGY

1. Key variables (assets, sales, stock price) will be simulated along a growth path
   - Starting point is randomly drawn from distribution of values at a sector and year level
   - Subsequent values are created based on distribution of growth
2. Remaining variables are simulated using ratios to appropriate key variable

INCOME STATEMENT



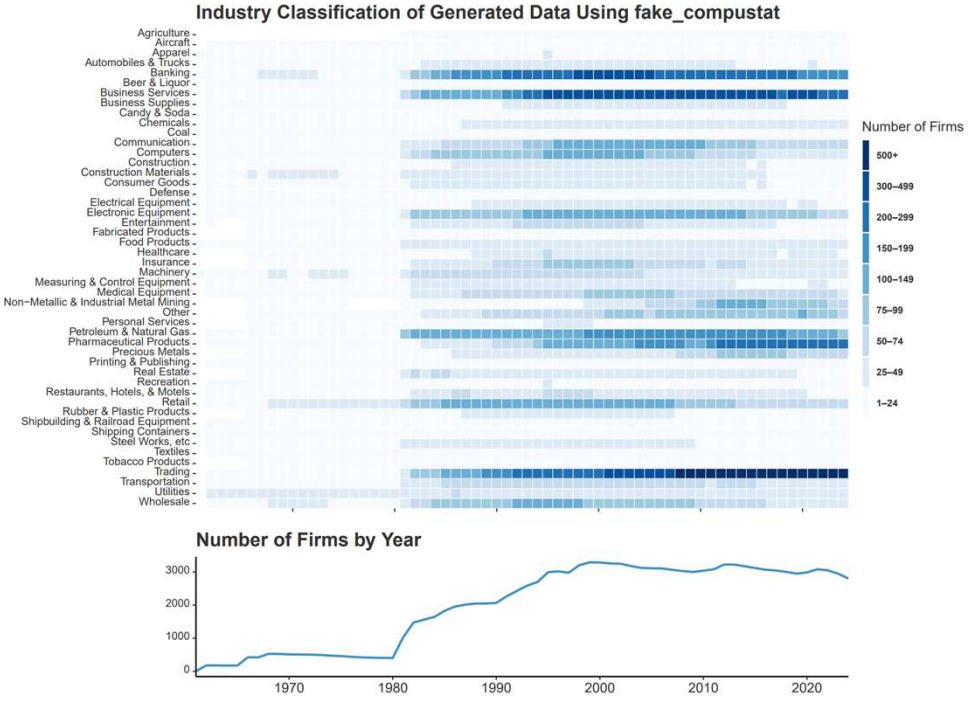
BALANCE SHEET



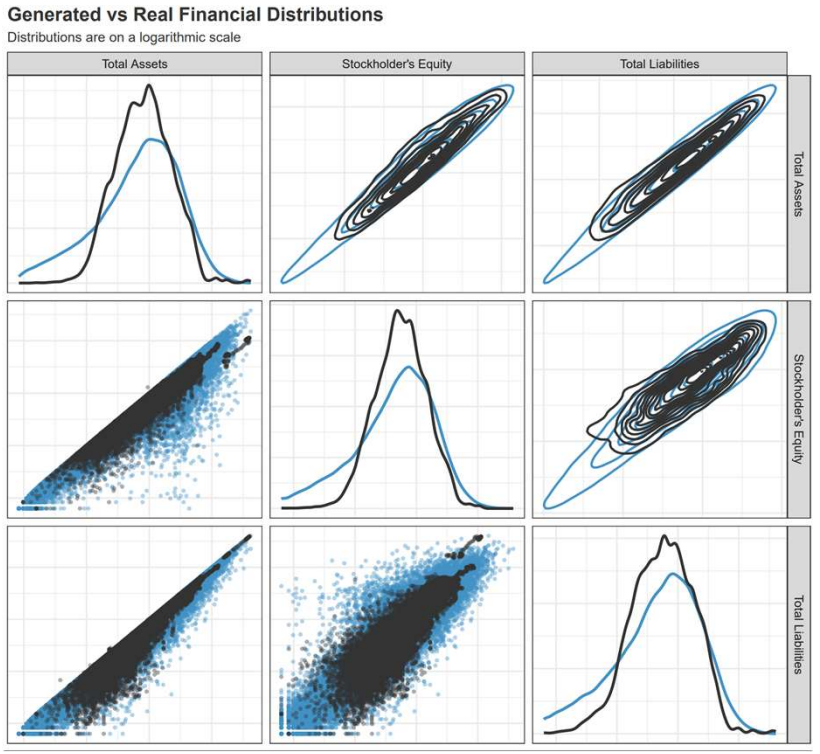
METHODOLOGY

1. Key variables (assets, sales, stock price) will be simulated along a growth path
   - Starting point is randomly drawn from distribution of values at a sector and year level
   - Subsequent values are created based on distribution of growth
2. Remaining variables are simulated using ratios to appropriate key variable
3. Data is mutated to fulfill principle and distribution goals

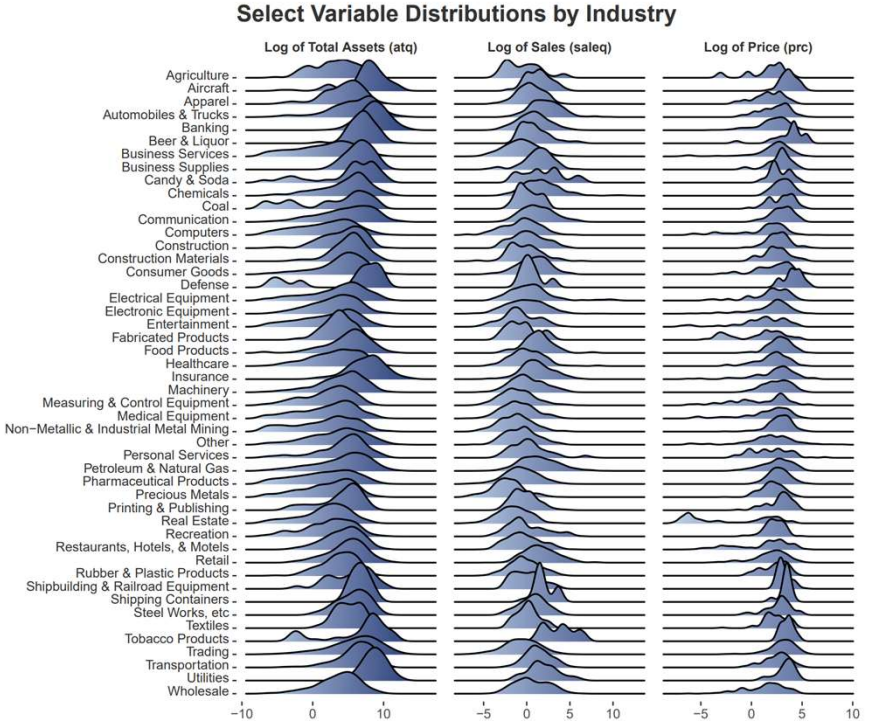
RESULTS


Industry Classification of Generated Data Using fake_compustat
Agriculture
Aircraft
Apparel
Automobiles & Trucks
Banking
Beer & Liquor
Business Services
Business Supplies
Candy & Soda
Chemicals
Coal
Communication
Computers
Construction
Construction Materials
Consumer Goods
Defense
Electrical Equipment
Electronic Equipment
Entertainment
Fabricated Products
Food Products
Healthcare
Insurance
Machinery
Measuring & Control Equipment
Medical Equipment
Non−Metallic & Industrial Metal Mining
Other
Personal Services
Petroleum & Natural Gas
Pharmaceutical Products
Precious Metals
Printing & Publishing
Real Estate
Recreation
Restaurants, Hotels, & Motels
Retail
Rubber & Plastic Products
Shipbuilding & Railroad Equipment
Shipping Containers
Steel Works, etc
Textiles
Tobacco Products
Trading
Transportation
Utilities
Wholesale
Number of Firms
500+
300–499
200–299
150–199
100–149
75–99
50–74
25–49
1–24
Number of Firms by Year
3000
2000
1000
0
1970
1980
1990
2000
2010
2020

RESULTS


Generated vs Real Financial Distributions
Distributions are on a logarithmic scale
Total Assets
Stockholder's Equity
Total Liabilities
Total Assets
Stockholder's Equity
Total Liabilities

RESULTS


Select Variable Distributions by Industry
Log of Total Assets (atq)
Log of Sales (saleq)
Log of Price (prc)
Agriculture
Aircraft
Apparel
Automobiles & Trucks
Banking
Beer & Liquor
Business Services
Business Supplies
Candy & Soda
Chemicals
Coal
Communication
Computers
Construction
Construction Materials
Consumer Goods
Defense
Electrical Equipment
Electronic Equipment
Entertainment
Fabricated Products
Food Products
Healthcare
Insurance
Machinery
Measuring & Control Equipment
Medical Equipment
Non−Metallic & Industrial Metal Mining
Other
Personal Services
Petroleum & Natural Gas
Pharmaceutical Products
Precious Metals
Printing & Publishing
Real Estate
Recreation
Restaurants, Hotels, & Motels
Retail
Rubber & Plastic Products
Shipbuilding & Railroad Equipment
Shipping Containers
Steel Works, etc
Textiles
Tobacco Products
Trading
Transportation
Utilities
Wholesale
−10
0
10
−5
0
5
10
−5
0
5
10

THANK YOU