jh191002-NWJ

財務ビッグデータの可視化と統計モデリング

地道 正行*, 宮本大輔**, 阪 智香*, 永田 修一*

* 関西学院大学 商学部, **東京大学大学院情報理工学系研究科

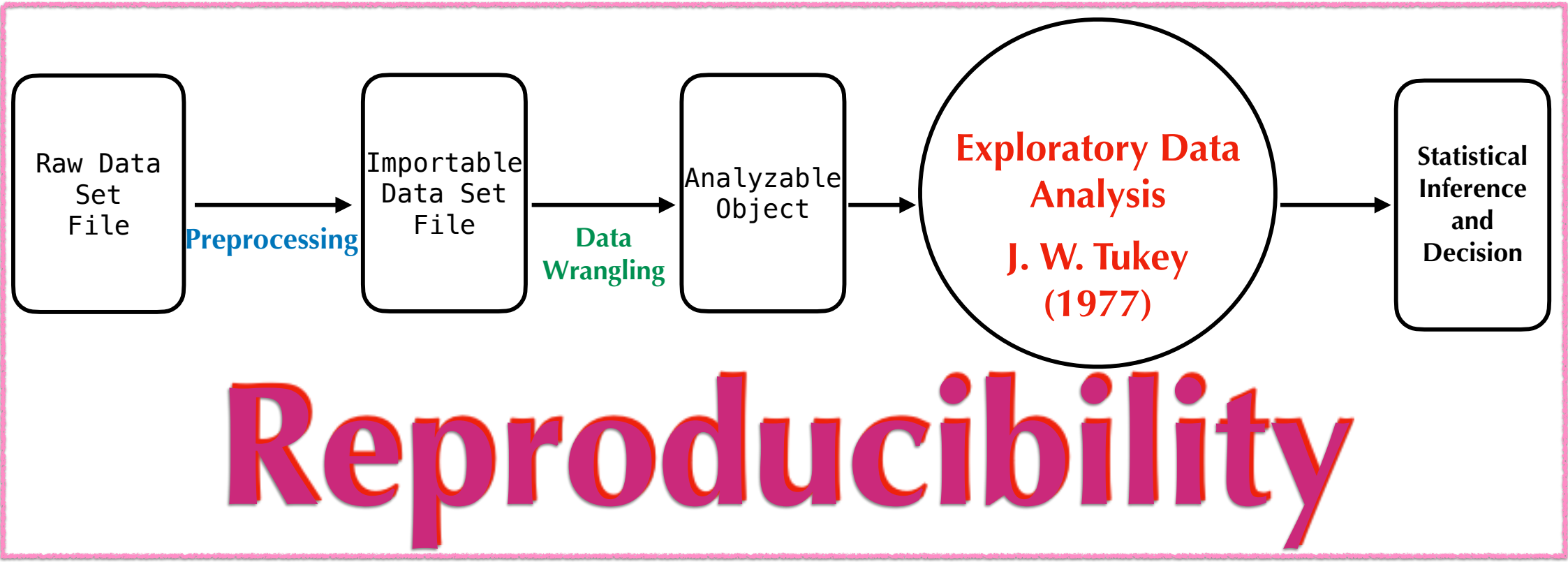
jimichi@kwansei.ac.jp

第12回 JHPCN シンポジウム

2020年7月9日(木)

オンライン開催

Preprocessing, Data Wrangling, Exploratory Data Analysis, and Reproducibility



Computational Environments

Hardware: Local

- macOS
  - Mac Pro
  - iMac
  - MacBook Pro
- Ubuntu
  - Dell Precision T7910

Hardware: FENNEL

(東京大学 専有利用型リアルタイムデータ解析ノード)

node name	OS	node	cores	memory	disk	GPU
GPU node	Ubuntu16.04	2	8	16GB	1TB	8GB
CPU node	Ubuntu16.04	2	8	12GB	1TB	-
total		4	32	56GB	4TB	16GB

Setup / Software

- OS: macOS, Ubuntu
- UNIX shell: `bash, zsh`
- UNIX command: `make, GNU parallel, (g)sed, grep, dos2unix,sftp,`etc
- Apache Spark
- R, RStudio
- R Packages: `tidyverse, SparkR, GGally, rgl, sn, xtable`
- Reproducibility: `Sweave`

2019 Articles and Presentations

学術論文 (2019年度)

(1) 大鹿智基, 阪 智香, 地道正行『企業の租税回避行動をめぐる証拠の可視化 –グローバルデータの探索的解析–』産業経理, 第 79 巻, 第 2 号, pp. 118–128, 産業経理協会, 2019 年 7 月.

(2) 地道正行, 阪 智香『探索的財務ビッグデータ解析 –データ可視化による企業活動の実態解明と統計モデリング–』, 日本経営数学会誌, 2019 年 8 月 13 日, 投稿中.

(3) 地道 正行『変換による財務データの統計解析 –売上高の場合–』, 商学論究, 第 67 巻, 第 1 号, pp. 27 – 46, 関西学院大学商学研究会, 2019 年 10 月.

(4) C. Saka, T. Oshika, and M. Jimichi, Visualization of Tax Avoidance and Tax Rate Convergence: Exploratory Analysis of World-scale Accounting Data, *Meditari Accountancy Research*, Vol. 27 No. 5, 2019, pp. 695–724, Emerald Publishing Limited.

(5) 阪 智香, 國部克彦, 地道正行『探索的データ解析に基づく世界企業の付加価値分配』, 神戸大学ディスカッションペーパー, 2019-28, pp. 1–35, 2020 年 1 月.

(6) 大鹿智基, 阪 智香, 地道正行, 『「社会にとってよい企業」への市場の評価とサステナビリティ』, 企業会計, 第 72 巻, 第 1 号, pp. 74–80, 中央経済社, 2020 年 1 月.

(7) 地道正行『探索的財務ビッグデータ解析 –前処理の並列化–』商学論究, 第 67 巻, 第 3 号, pp. 1–19, 関西学院大学商学研究会, 2020 年 3 月.

国際会議発表 (2019年度)

(1) M. Jimichi*, D. Miyamoto, C. Saka, and S. Nagata, *Exploratory Financial Big Data Analysis and Reproducible Research*, DSSV 2019, Doshisha University, Imadegawa Campus, August 14th, 2019.

(2) C. Saka* and M. Jimichi, *Visualization of Corporate Tax Avoidance and Value Added Distribution: Exploratory Analysis of Financial Big Data*, DSSV 2019, Doshisha University, Imadegawa Campus, August 14th, 2019.

国内会議発表(1) (2019年度)

(1) 地道正行, 宮本大輔, 阪 智香, 永田修一『探索的財務ビッグデータ解析 –前処理, データラングリング, 再現可能性–』, 日本計算機統計学会シンポジウム予稿集, 滋賀大学データサイエンス学部, 2018 年 11 月 11 日 (日).

(2) 地道正行*, 宮本大輔, 阪智香, 永田修一『探索的財務ビッグデータ解析と再現可能研究』, 日本経営数学会第 41 回 (通算 61 回) 研究大会, 拓殖大学茗荷谷キャンパス, 2019 年 6 月 1 日 (土).

(3) 阪 智香*, *Visualization of tax avoidance and tax rate convergence: Exploratory analysis of world-scale accounting data*, 日本会計研究学会特別委員会「税制が企業会計その他の企業行動に及ぼす影響に関する研究」研究会, 慶應義塾大学日吉キャンパス, 2019 年 7 月 6 日 (土).

(4) 地道正行*, 宮本大輔, 阪 智香*, 永田修一『財務ビッグデータの可視化と統計モデリング』, 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 第 11 回 シンポジウム, THE GRAND HALL(品川）, 2019 年 7 月 11 日 (木).

(5) 地道正行*, 宮本大輔, 阪 智香, 永田修一『探索的財務ビッグデータ解析 –前処理の並列化–』, 国際数理科学協会, 2019 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 関西学院大学大阪梅田キャンパス, 2019 年 8 月 24 日 (土).

(6) 齊藤 美桜里 *, 地道正行『R による GIS データの可視化』, 国際数理科学協会 2019 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 関西学院大学大阪梅田キャンパス, 2019 年 8 月 24 日 (土).

国内会議発表(2) (2019年度)

(7) 阪 智香*, 國部克彦, 地道正行『会計と平等 –付加価値分配率の探索的データ解析–』, 日本会計研究学会, 第 78 回大会, 神戸学院大学ポートアイランドキャンパス, 2019 年 9 月 9 日 (月).

(8) M. Jimichi, D. Miyamoto*, C. Saka, and S. Nagata, *Exploratory Financial Big Data Analysis and Reproducible Research*, 2019 年度年会統計関連学会連合大会, 滋賀大学彦根キャンパス, 2019 年 9 月 10 日 (火).

(9) 地道正行*, 宮本大輔, 阪 智香*, 永田修一『探索的財務ビッグデータ解析と再現可能研究』, RIMS 共同研究「マクロ経済動学の非線形数理」, 京都大学数理解析研究所, 2019 年 10 月 17 日 (木).

(10) 地道正行*, 宮本大輔, 阪 智香, 永田修一『探索的財務ビッグデータ解析 –前処理の並列化–』, 日本計算機統計学会 第 33 回シンポジウム, 青山学院大学 青山キャンパス, 2019 年 11 月 30 日 (土).

(11) 阪 智香* 『財務ビッグデータの探索的データ解析 –企業の租税回避と付加価値分配–』, 統計数理研究所・リスク解析戦略研究センター, 第 7 回 金融シンポジウム, フクラシア丸の内オアゾ, 2019 年 12 月 5 日 (木).

(12) 地道正行*, 宮本大輔, 阪 智香, 永田修一『探索的財務ビッグデータ解析 –前処理の並列化–』, 2019 年度日本経営数学会 秋季研究会, 専修大学 神田キャンパス, 2019 年 12 月 7 日 (土).

(13) 地道正行*, 宮本大輔, 阪 智香, 永田修一『探索的財務ビッグデータ解析 –前処理とデータラングリングの並列化–』, 統計数理研究所共同研究集会 2019 年度「データ解析環境 R の整備と利用」, 統計数理研究所, 2019 年 12 月 21 日 (土).

Database, Preprocessing of Osiris2018 Data Set (Consolidate Version)

Database

- Bureau van Dijk (ビューロー・ヴァン・ダイク)社 (以下 BvD と略)
  全世界上場・上場廃止企業データベース Osiris (オシリス)

- 世界の全上場企業90,000社強の情報を 国際比較可能な統一のフォームで収録

  【収録情報】 世界の上場企業・一般事業会社の以下のデータ

  - 財務情報(BS/PL/CF)- 平均15年(最長30年)
  - 株主/関連会社情報 - 出資比率・種別・アーカイブ
  - 株価情報 - 時価総額、β 値、EPS、インデックス
  - 企業概要 - 事業内容、業種分類、創業年、IPO Date 統一のフォームで収録

Data Set Information: Osiris2018

- 世界160カ国の上場企業 (93,836社) の主要財務情報 (売上高, 営業利益, 総資産など84項目) を33年分 (最長30年分) 抽出
  →パネルデータ(経時観測データ)
- 前処理後のデータセットは社名, 社名+BvD ID, 決算年を新たに変数として加えたため, 87項目になっている.

No_	VariableName(data7)	R Variable name	変数説明
1	year(USD)	year_USD	年(通貨単位)
2	BvD ID number	ID	企業コード
3	Address of incorp_ - Country	country	国
4	US SIC, Primary code(s) (M)	SIC_code	業種コード
5	US SIC, primary code(s) description	SIC_name	業種名
6	Main exchange	exchange	主取引所
7	Consolidation code	cons	連結・単独
8	Closing date	date	決算日
9	Number of months	month	月数
10	Audit Status	audit	監査
11	Accounting standard	practice	会計基準
12	Source	source	データの出所
13	Statement unit	units	単位(価格)
14	:	:	:
:	Market price - July	market_price7	市場価格7月末分
78	Market price - August	market_price8	市場価格8月末分
79	Market price - September	market_price9	市場価格9月末分
80	Market price - October	market_price10	市場価格10月末分
81	Market price - November	market_price11	市場価格11月末分
82	Market price - December	market_price12	市場価格12月末分
83	Market Cap	market_cap	時価総額
84	Dividend	dividend	配当金

Data File Informations by Unix Commands

```
$ ls -l dataC.txt
-rwxr-xr-x  1 masa  staff  1548270085  4  9 20:56 dataC.txt

$ wc -l dataC.txt
 3190424 dataC.txt
```

Raw Data File

<U+FEFF>COURTAULDS PLC BvD ID number Address of incorp. - Country US SIC, Primary code(s) (M) US SIC, primary code(s) description Main exchange Consolidation code Closing date Number of months Audit Status Accounting standard Source Statement unit Currency of the statement Exchange Rate from Local Currency Fixed Assets Intangible Fixed Assets Tangible Fixed Assets Other Fixed Assets
Current Assets Stock Debtors Others Cash & Cash Equivalent Total Assets Shareholders Funds Capital Other Non Current Liabilities Long Term Debt Other Non Current Liabilities Provisions Current Liabilities Loans Creditors Other Total Shareh. Funds & Liab.
Working Capital Net Current Assets Enterprise Value Number of Employees Operating Revenue / Turnover Sales Costs of Goods Sold Gross Profit Other Operating Items Operating P/L Financial Revenue Financial P/L Other non Oper./Financial Items P/L before Tax Taxation P/L after Tax Extraord. & Oth. Items P/L for Period Material Costs Costs of Employees Depreciation/Amortization Financial Expenses Research & Development expenses (memo) Cash Flow Added Value EBITDA Income taxes Income Tax Payable Deferred Taxes Def. Inc. Taxes & Invest. Tax Credit Date of current Market capitalisation Current market capitalisation Market price - Year Market price - January Market price - February Market price - March Market price - April Market price - May Market price - June Market price - July Market price - August Market price - September Market price - October Market price - November Market price - December Market Cap. Cash Dividends Paid - Total

1985 (th USD) GB01605EX UNITED KINGDOM 2899 Chemicals and chemical preparations, not elsewhere specified manufacturing Unlisted C1 31/03/1986 12 Qualif n.a. Acc. Std na AR th GBP 1.48531 640,317 n.a. 595,164 45,153 1,429,166
583,281 492,083 353,801 313,400 2,069,483 844,547 146,006 698,541 387,369 276,862 110,507 n.a. 837,566 90,307 508,719 238,541 2,069,483 566,646 591,599 n.a. 68,000 3,227,134 3,227,134 -2,411,104 816,029 -510,353 229,480 n.a. -17,081 n.a. 212,399 -39,064 173,336 -18,715 154,621 n.a. -778,006 -76,196 -17,081 0 230,817 1,064,967 305,677 -39,064 28,369 n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a.

1986 (th USD) GB01605EX UNITED KINGDOM Unlisted C1 31/03/1987 12 Qualif n.a. Acc. Std na AR th GBP 1.60501 842,148 n.a. 795,281 46,866 1,636,947 637,188 558,864 440,896 396,918 2,479,095 1,058,663 160,982 897,681 434,315 301,420 132,895 n.a. 986,117 70,620 611,508 303,988 2,479,095 584,544 650,831 n.a. 65,400 3,630,367 3,630,367 -2,564,963 1,065,404 -628,200 342,830 n.a. -20,063 n.a. 322,767 -74,793 247,974 -33,224 214,750 n.a. -879,063
-94,374 -20,063 0 309,124 1,283,043 437,204 -74,793 59,867 n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a.

1987 (th USD) GB01605EX UNITED KINGDOM Unlisted C1 31/03/1988 12 Qualif n.a. Acc. Std na AR th GBP 1.87981 1,149,877 n.a. 1,089,723 60,154 1,897,287 790,458 713,386 393,443 347,388 3,047,164 1,154,388 190,048 964,340 648,533 515,443 133,090 33,461 1,244,243 113,540 517,510 613,192 3,047,164 986,334 653,044 n.a. n.a.
4,551,384 4,551,384 -3,211,083 1,340,301 -774,292 437,619 40,416 -22,934 n.a. 414,685 -95,682 319,003 -49,627 269,376 n.a. -1,128,823 -128,391 -63,349 0 397,767 1,685,621 566,009 -95,682 91,171 11,091 n.a. n.a.
n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a. n.a.

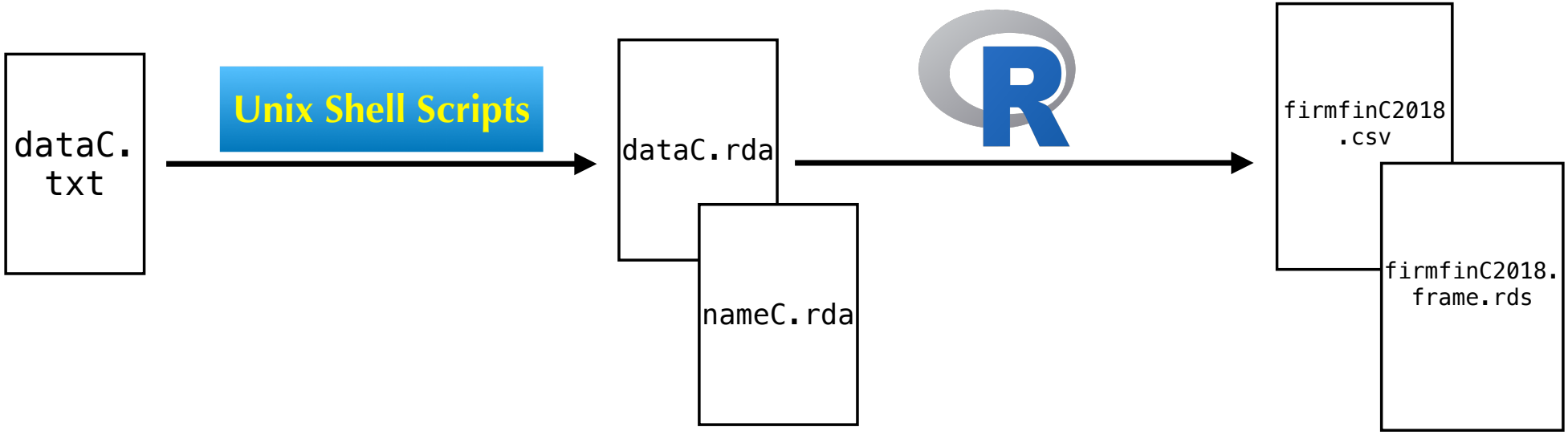
Problems

- 粗データセット(raw dataset)をそのままソフトウェア環境で解析することが難しい.
  - ⋆ Byte Order Mark (BOM) コードの存在
  - ⋆ OS間での行末コードの相違
  - ⋆ 欠測値の存在 (NA記号が存在しない欠測値の存在)
  - ⋆ レイアウトの不統一
  - ⋆ 金額に関するフォーマット(カンマ区切り)
  - ⋆ 特殊記号の存在 (#など)
  - ⋆ ファイルはある程度の規模があるので通常のエディタでは整形が難しい.

Solutions

1. UNIX コマンドやインタプリター (`grep`,`dos2unix`, `|`, `>`, `(g)sed` など) を利用してファイルを整形
2. データ解析環境Rを用いてデータファイルをRに読み込むことができる形式(CSV, RDSファイル)に変換

Preprocessing



Parallelization of Preprocessing of Osiris2018 (Consolidate Version) Data Set

ビッグデータの前処理に関する経験則

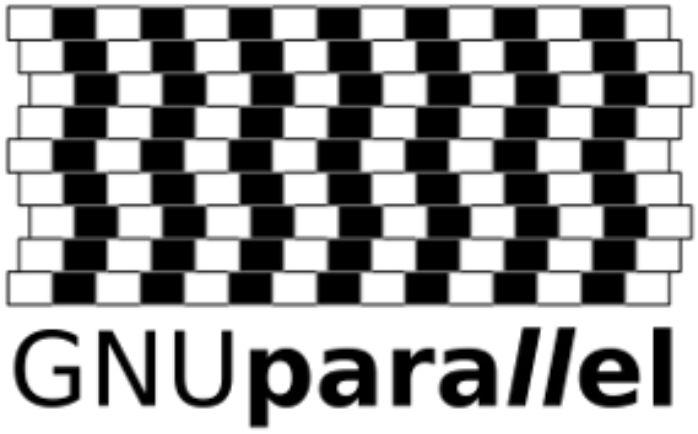
- 「前処理には, 分析・解析を行う全工程の50%〜90%の時間を費やす」

(e.g. Patil(2012))

GNU parallel

https://www.gnu.org/software/parallel/

- GNU parallel is a shell tool for executing jobs in parallel using one or more computers. A job can be a single command or a small script that has to be run for each of the lines in the input. The typical input is a list of files, a list of hosts, a list of users, a list of URLs, or a list of tables. A job can also be a command that reads from a pipe. GNU parallel can then split the input and pipe it into commands in parallel



Makefile

```
all:
  date > start.txt
  /bin/bash ./script.sh
  Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
  date > end.txt
all-p:
  date > start-p.txt
  /bin/bash ./script-p.sh
  Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
  date > end-p.txt
rda:
  date > start-rda.txt
  /bin/bash ./script.sh
  date > end-rda.txt
rda-p:
  date > start-rda-p.txt
  /bin/bash ./script-p.sh
  date > end-rda-p.txt
csv:
  Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
clean-data:
  rm *.rda *.rda-e *.part
clean-csv:
  rm firmfinC2018.csv
```

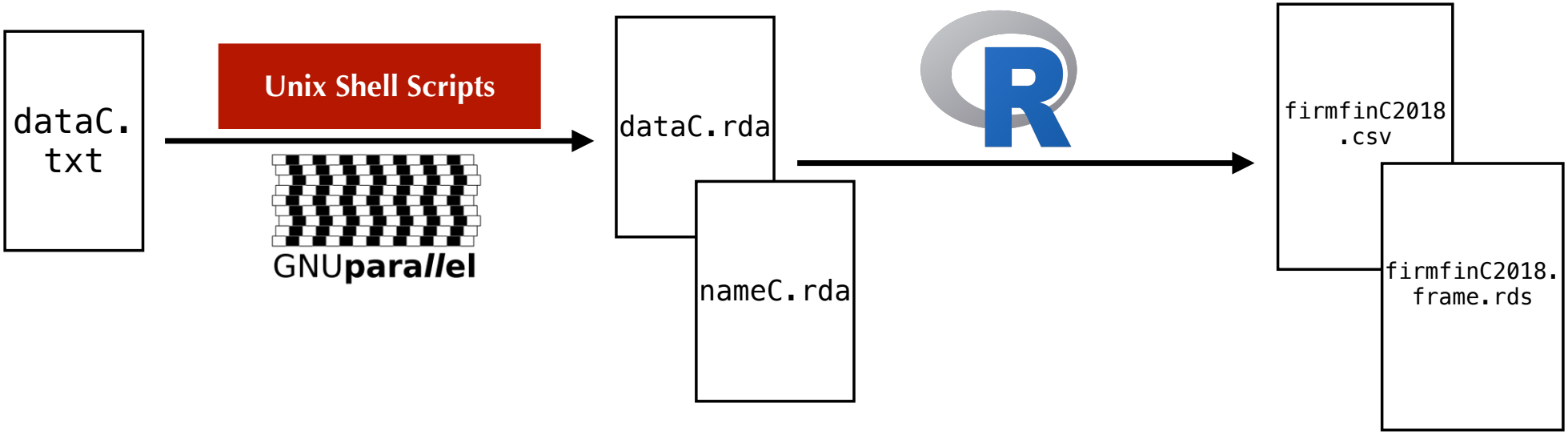
script.sh

```
$ cat script.sh
 #!/bin/bash
 #
 #echo "Remove BOM codes"
 gsed -i -s -e '1s/^\xef\xbb\xbf//' dataC.txt
 #echo "dos2unix"
 dos2unix dataC.txt
 echo "separate data file"
 grep -E "th\sUSD\)" dataC.txt > dataC.part
 grep -v -E "th\sUSD\)" dataC.txt > nameC.part
 echo "replacement special character"
 sed -f sedscr dataC.part > dataC.rda
 sed -i -e s/^$'\t'//g dataC.rda
 sed -e s/#//g nameC.part > nameC.rda
```

script-p.sh

```
$ cat script-p.sh
 #!/bin/bash
 #echo "Remove BOM codes"
 gsed -i -s -e '1s/^\xef\xbb\xbf//' dataC.txt
 echo "dos2unix"
 parallel --pipepart -k --block 100M -a dataC.txt "dos2unix" > tmp
 echo "separate data file"
 echo "separate data file"
 parallel --pipepart -k --block 100M -a tmp 'grep -E "th\sUSD\)"' > dataC.part
 parallel --pipepart -k --block 100M -a tmp 'grep -v -E "th\sUSD\)"' > nameC.part
 echo "replacement special character"
 parallel --pipepart -k --block 100M -a dataC.part "sed -f sedscr" > tmp
 parallel --pipepart -k --block 100M -a tmp "sed s/^$'\t'//g" > dataC.rda
 parallel --pipepart -k --block 100M -a nameC.part "sed s/#//g" > nameC.rda
 rm tmp
```

Parallelized Preprocessing



Apache Spark and R Packages


APACHE
Spark
TM

Apache Spark

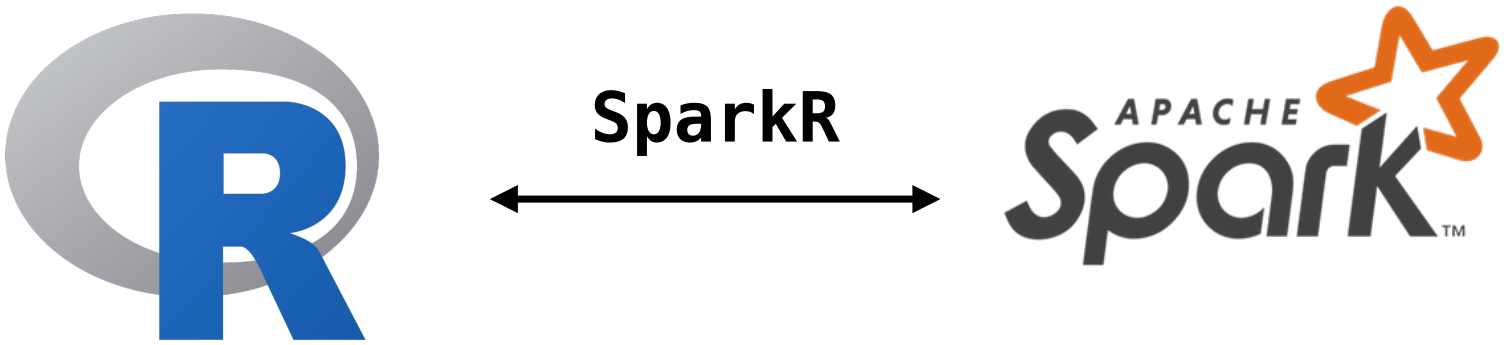
- Apache Spark is a fast and general-purpose cluster computing system.
- It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs.
- It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

Quotation: http://spark.apache.org/docs/latest/index.html

SparkR (R on Spark)

- SparkR is an R package that provides a light-weight frontend to use Apache Spark from R.
- In Spark 2.1.0, SparkR provides a distributed data frame implementation that supports operations like selection, filtering, aggregation etc. (similar to R data frames, dplyr) but on large datasets.
- SparkR also supports distributed machine learning using MLlib.
- Quotation: http://spark.apache.org/docs/latest/sparkr.html#overview

Connect to Spark from R by SparkR on RStudio




```
> Sys.setenv(SPARK_HOME = "/usr/local/Cellar/apache-spark/2.4.1/libexec")

> library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))

> sparkR.session(master = "local[*]", sparkConfig = list(spark.driver.memory =
  "2g"))
```

Data Wrangling and Transformation

Comparison of Data Manipulations by SparkR, dplyr, R

Manipulations	Spark Style	dplyr Style	R Style
Selection of Columns	`select`	`select`	`df[,"colName"]`
Filtering of Rows	`filter`	`filter`	`df[condition,]`
Addition of Columns	`withColumn`	`mutate`	`df$colName<-col`
Permutation of Rows	`orderBy`	`arrange`	X
Grouping	`groupBy`	`group_by`	X
Aggregation	`agg, summarize`	`summarize`	X
Joinning	`join`	`join`	`merge`

Data Wrangling of Financial Data by Spark Environments

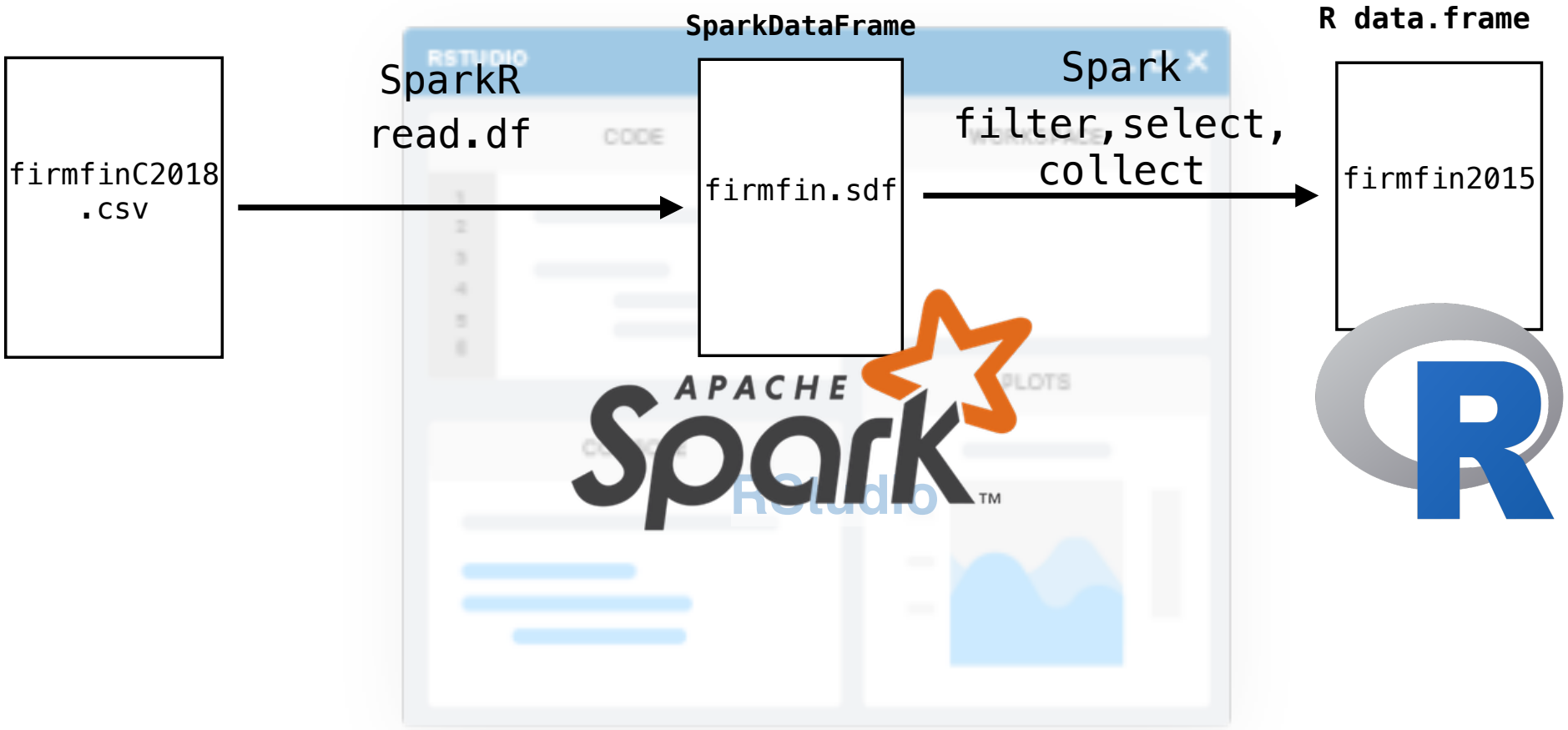
Data Set:
`firmfinC2018.csv`

Data File Information

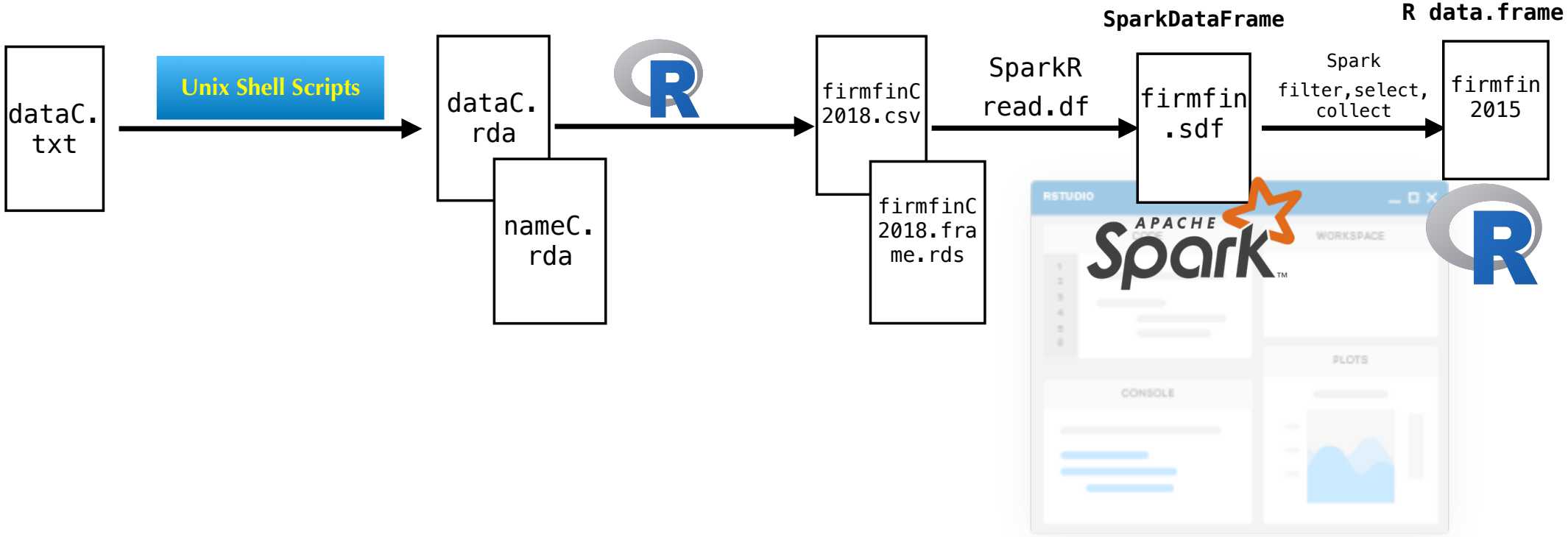
```
$ head -n 2 firmfinC2018.csv
firm,year_USD,ID,country,SIC_code,SIC_name,exchange,cons,date,month,audit,practice,
source,units,currncy,exchange_rate,assets_fix,assets_int,assets_tang,assets_other_f
ix,assets_cur,stock,debtors,assets_other_cur,cash,assets_total,shareholders,capital
,shareholders_other,liabilities_non_cur,debt_long,liabilities_other_non_cur,provisi
ons,liabilities_cur,loans,creditors,liabilities_other_cur,total_s_l,capital_working
,assets_net_cur,enterprise_value,employees,operating_revenue,sales,costs_goods,prof
it_gross,expenses_other,EBIT,revenue_fin,PL_fin,PL_other,PL_before_tax,tax,PL_after
_tax,PL_extr,net_income,costs_material,costs_employees,depr_amor,interest_paid,R_D,
cash_flow,add_value,EBITDA,tax_income,tax_pay_income,tax_deferred,tax_credit,date_c
ur_market_cap,market_cap_cur,year_mp,market_price1,market_price2,market_price3,mark
et_price4,market_price5,market_price6,market_price7,market_price8,market_price9,mar
ket_price10,market_price11,market_price12,market_cap,dividend,firmID,year
COURTAULDS PLC,1985 (th USD),GB01605EX,UNITED KINGDOM,2899,Chemicals and chemical
preparations not elsewhere specified manufacturing,Unlisted,C1,1986-03-31,12,Qualif
NA,Acc. Std
na,AR,th,GBP,1.48531,640317,NA,595164,45153,1429166,583281,492083,353801,313400,206
9483,844547,146006,698541,387369,276862,110507,NA,837566,90307,508719,238541,206948
3,566646,591599,NA,68000,3227134,3227134,2411104,816029,510353,229480,NA,-17081,NA,
212399,39064,173336,-18715,154621,NA,778006,76196,17081,-0,230817,1064967,305677,39
064,28369,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,COURTAULDS PLC
GB01605EX,1985
```

SparkR on RStudio

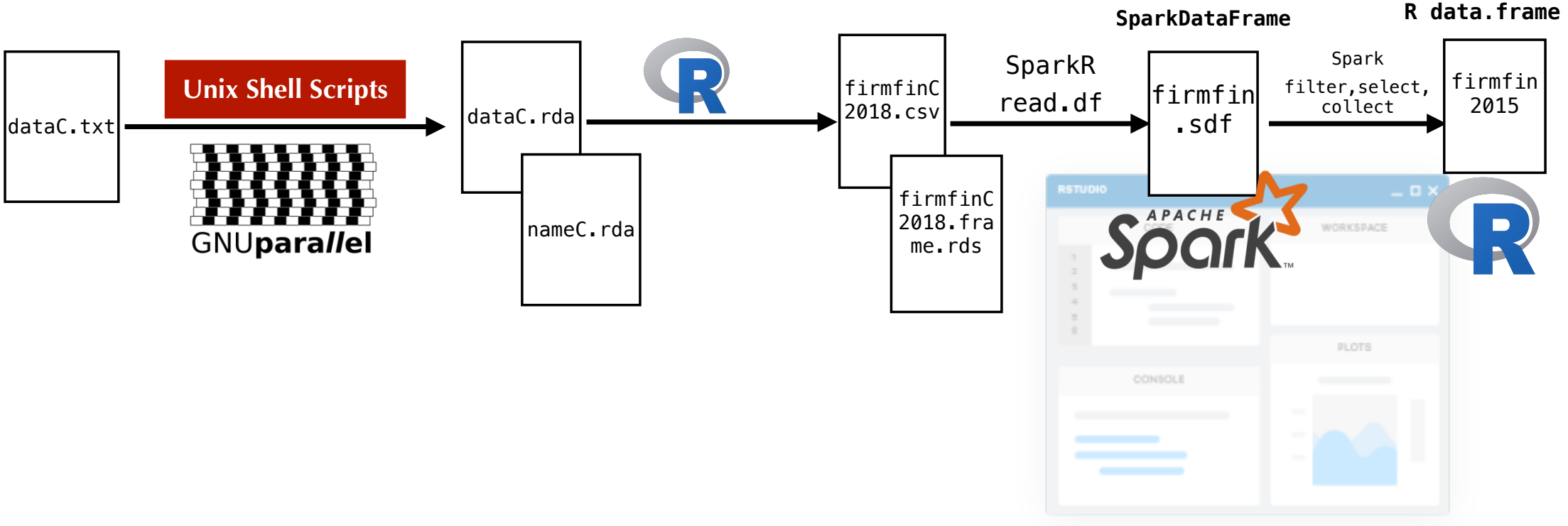
Data Wrangling as SparkDataFrame and Transform SparkDataFrame to R data.frame: SparkR on R and RStudio



Preprocessing and Data Wrangling



Parallelized Preprocessing and Data Wrangling



Statistical Modeling with Osiris2018 Data: Verification of Reproducibility

Modeling of Distribution of `log(sales)`

Skew-Normal Distribution and Related Families

- Azzalini (1985) generalized the normal distribution to have non-zero skewness.
- It is called the *skew-normal distribution*.
- It has been used for modeling and analyzing skewed data.
- Additionally, some related families of distributions included the *skew-t distribution* are also studied in Azzalini and Capitanio (2014).

Skew-Normal Distribution

Notation of Distribution: If the distribution of the random variable (r.v.) X is the skew-normal, then we write as follows:

$$X \sim \mathsf{SN}(\xi, \omega^2, \alpha)$$

where (ξ, ω^2, α) are called *direct parameters* (DP).

Probability Density Function:

$$f_{\mathsf{SN}}(x \mid \xi, \omega, \alpha) := \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{x-\xi}{\omega}\right) \tag{4.1}$$

where $x \in \mathbb{R} := (-\infty, \infty)$, $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+ := (0, \infty)$, $\alpha \in \mathbb{R}$, and

$$\phi(x) := \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right);\ \text{p.d.f. of Standard Normal Distribution,}$$

$$\Phi(x) := \int_{-\infty}^{x}\phi(z)dz;\ \text{c.d.f. of Standard Normal Distribution.}$$

Fitting Skew-Normal Distribution to `log(sales)`

- Maximum Likelihood Estimates:

$$(\widehat{\xi}, \widehat{\omega}, \widehat{\alpha}) = (13.01, 3.29, -1.15)$$

- Statistical Model:

$$f_{\mathsf{SN}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha}) := \frac{2}{\widehat{\omega}} \phi \left(\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}} \right) \Phi \left(\widehat{\alpha} \frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}} \right)$$

- Histogram with Statistical Model and Q-Q Plot:

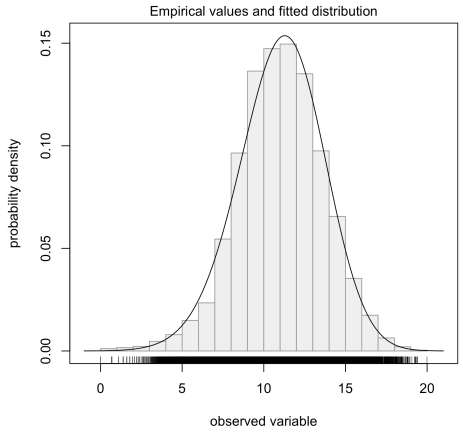


Figure: Histogram of `log(sales)` with Statistical Model $f_{\mathsf{SN}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha})$

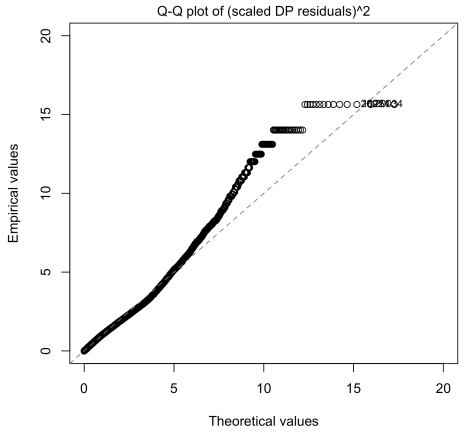


Figure: Q-Q Plot of `log(sales)`

Skew-t Distribution

Notation of Distribution: If the distribution of X is the *skew-t*, then we write as follows:

$$X \sim \mathsf{ST}(\xi, \omega^2, \alpha, \nu)$$

where $(\xi, \omega^2, \alpha, \nu)$ are called the direct parameters.

Probability Density Function:

$$f_{\mathsf{ST}}(x \mid \xi, \omega, \alpha, \nu) = \frac{2}{\omega} f_{\mathsf{t}}\left(\frac{x-\xi_i}{\omega} \middle| \nu \right) F_{\mathsf{t}}\left(\alpha \frac{x-\xi}{\omega} \sqrt{\frac{\nu+1}{\left(\frac{x-\xi}{\omega}\right)^2 + \nu}} \middle| \nu+1 \right) \tag{4.2}$$

where $x \in \mathbb{R} := (-\infty, \infty)$, $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+ := (0, \infty)$, $\alpha \in \mathbb{R}$, $\nu \in \mathbb{R}^+$, and

$$f_{\mathsf{t}}(x \mid \nu) := \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} ; \text{ p.d.f. of t distribution,}$$

$$F_{\mathsf{t}}(x \mid \nu) := \int_{-\infty}^{x} f_{\mathsf{t}}(u \mid \nu) \mathrm{d}u; \text{ c.d.f. of t distribution}$$

Fitting Skew-t Distribution to `log(sales)`

- Maximum Likelihood Estimates:

$$(\widehat{\xi}, \widehat{\omega}, \widehat{\alpha}, \widehat{\nu}) = (12.57, 2.9, -0.83, 22.42)$$

- Statistical Model:

$$f_{\mathsf{ST}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha}, \widehat{\nu}) = \frac{2}{\widehat{\omega}} f_{\mathsf{t}}\left(\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}} \mid \widehat{\nu}\right) F_{\mathsf{t}}\left(\widehat{\alpha}\frac{\log(\texttt{sales}) - \widehat{\xi}}{\widehat{\omega}}\sqrt{\frac{\widehat{\nu}+1}{\left(\frac{\log(\texttt{sales})-\widehat{\xi}}{\widehat{\omega}}\right)^2 + \widehat{\nu}}} \mid \widehat{\nu}+1\right)$$

- Histogram with Statistical Model and Q-Q Plot:

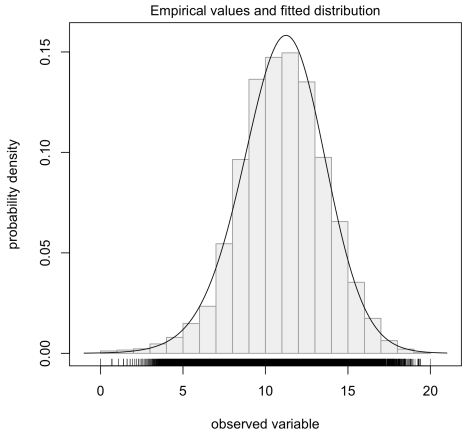


Figure: Histogram of `log(sales)` with Statistical Model $f_{\mathsf{ST}}(\log(\texttt{sales}) \mid \widehat{\xi}, \widehat{\omega}, \widehat{\alpha}, \widehat{\nu})$

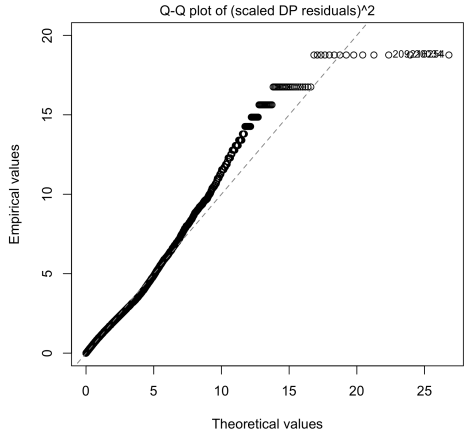
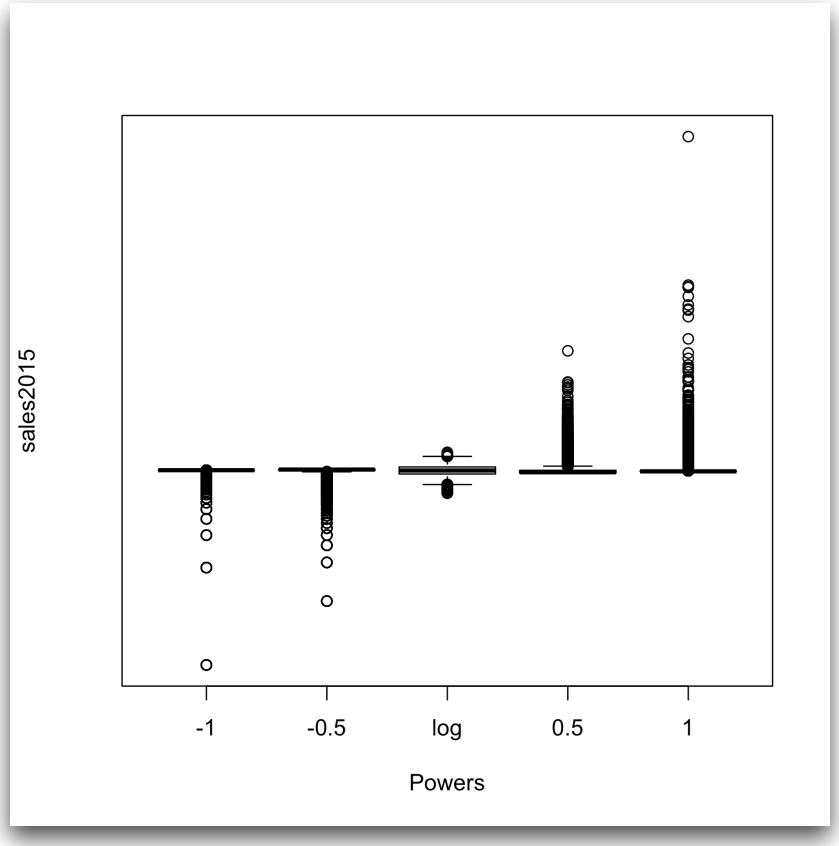
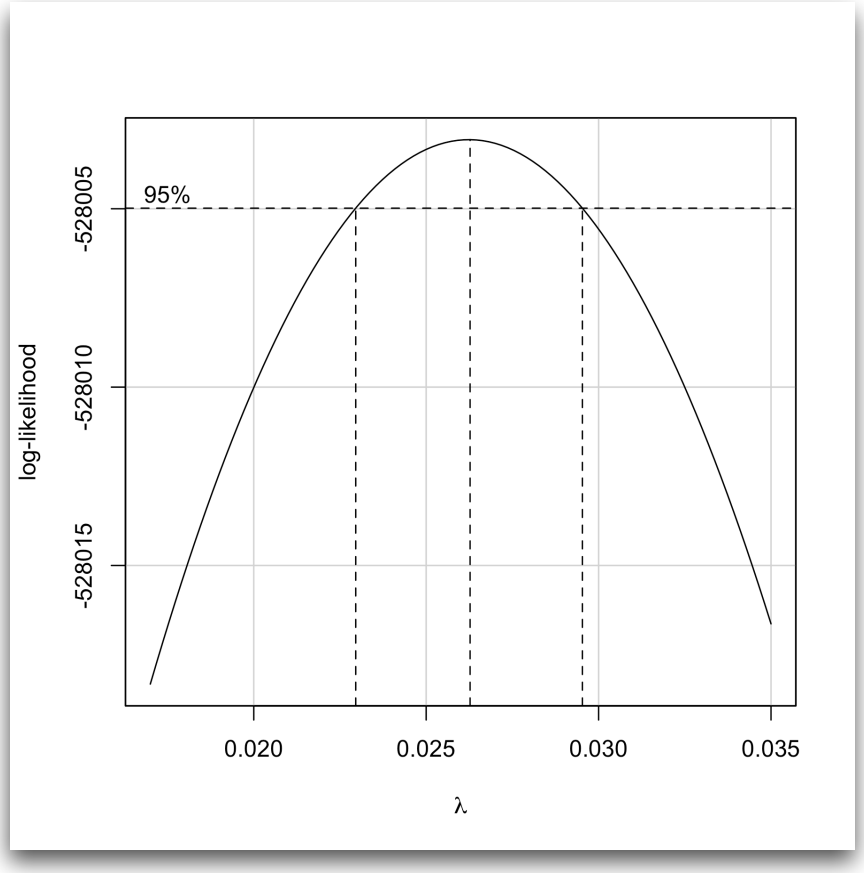


Figure: Q-Q Plot of `log(sales)`

Box-Cox Transformation of `log(sales)` and Maximum Log-likelihood



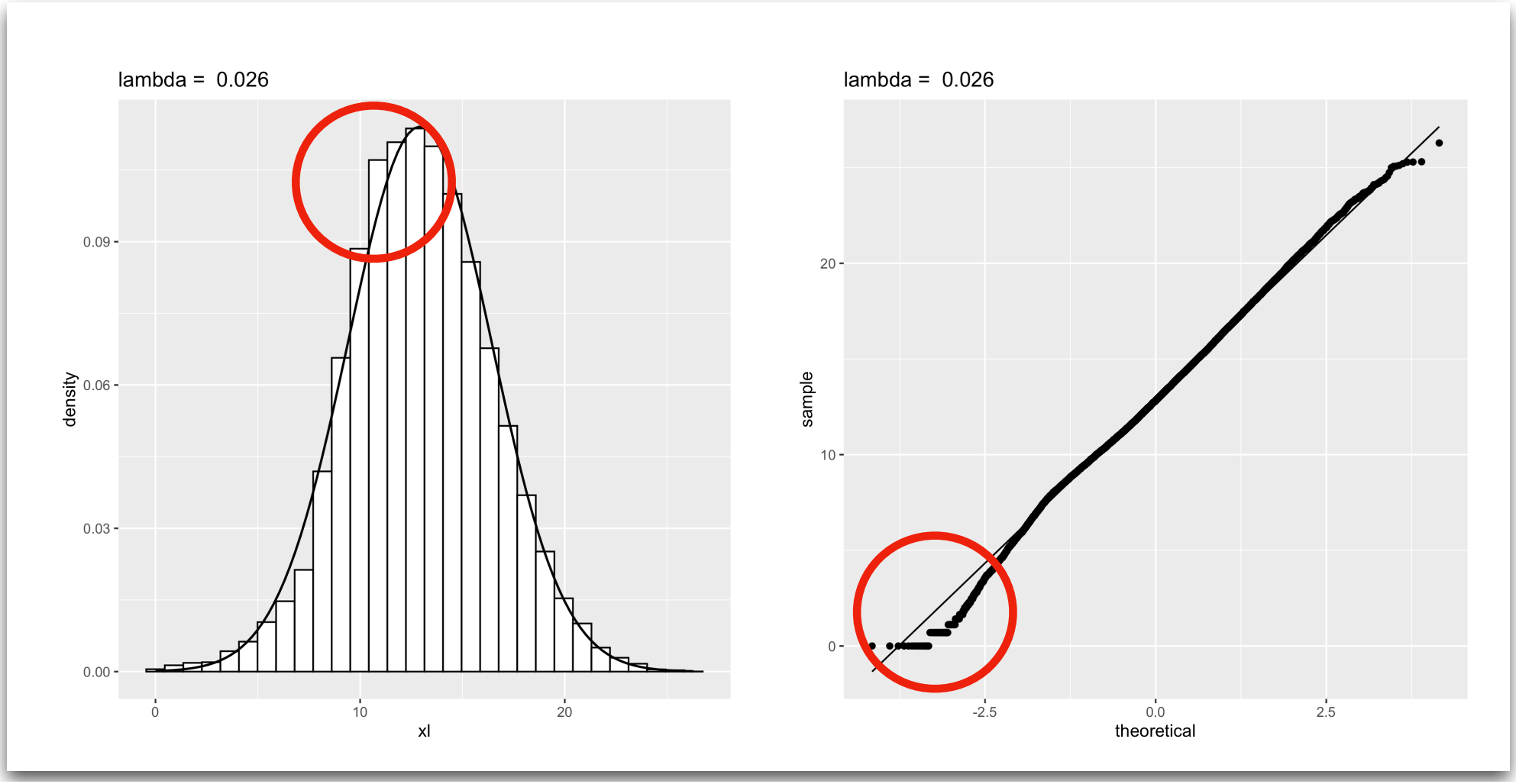


$$X^{(\lambda)} := \begin{cases} \dfrac{X^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(X), & \lambda = 0 \end{cases}$$

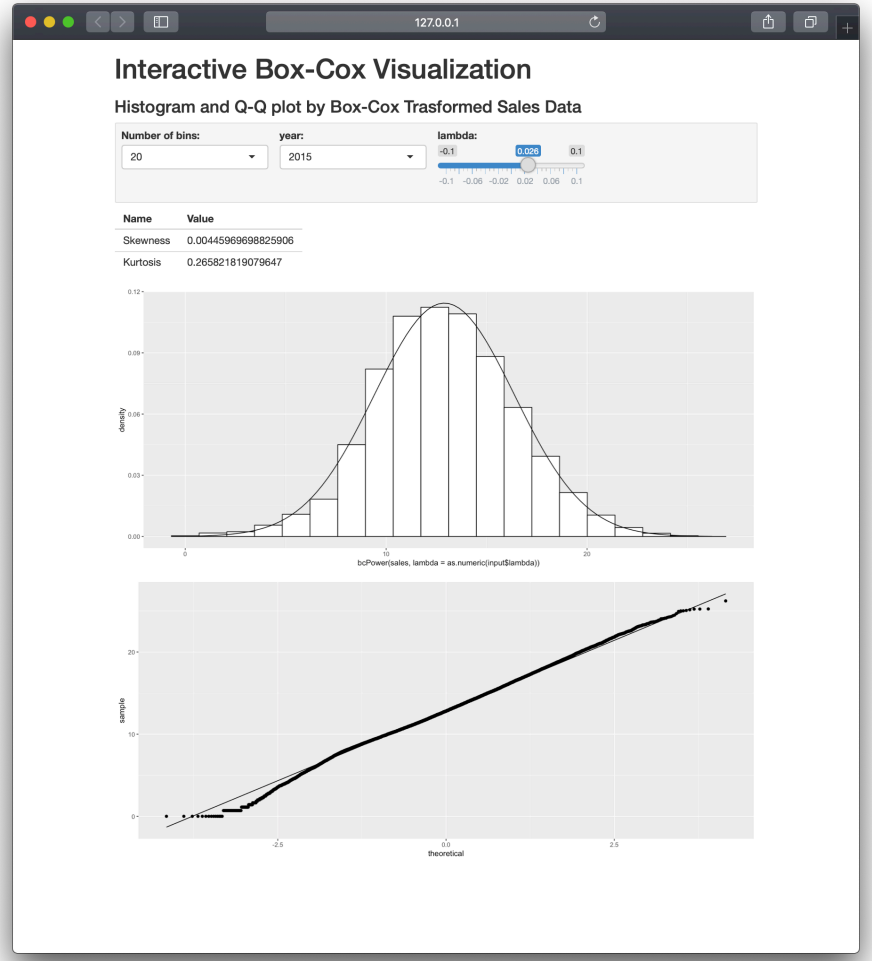
Box and Cox (1964)

$$\ell_{\text{profile}}(\lambda) := -\frac{n}{2}\log(2\pi\widehat{\sigma}^2(\lambda)) - \sum_{i=1}^{n}\frac{(x_i^{(\lambda)} - \widehat{\mu}(\lambda))^2}{2\widehat{\sigma}^2(\lambda)} + (\lambda - 1)\sum_{i=1}^{n}\log x_i$$

Histogram and Normal Q-Q Plot for Box-Cox Transformation of `log(sales)`



Web Application for Box-Cox Transformation



Double-log Modeling for `log(sales)`

Double-log Model

- Cobb-Daglas Type Model:

$$\texttt{sales}_i = \gamma \times \texttt{employees}_i^{\alpha_1} \times \texttt{assets.total}_i^{\alpha_2} \times \varepsilon_i, \quad i = 1, \ldots, n$$

- Log-linear Model:

$$\log(\texttt{sales}_i) = \alpha_0 + \alpha_1 \log(\texttt{employees}_i) + \alpha_2 \log(\texttt{assets.total}_i) + \log(\varepsilon_i)$$

- Assumptions of Error Distributions:

  Normal Case: $\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma^2)$

  Skew-normal Case: $\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{SN}(0, \omega^2, \alpha)$

  Skew-t Case: $\log(\varepsilon_i) \overset{\text{i.i.d.}}{\sim} \mathsf{ST}(0, \omega^2, \alpha, \nu)$

  where $i = 1, \ldots, n$, and the notation "$\overset{\text{i.i.d.}}{\sim}$" denotes *independent and identically distributed*.

Results of Fitting: Normal Case

- t-Values Table:

Table: t-Values Table: Log-normal Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6004	0.0284	21.18	0.0000
log(employees)	0.4660	0.0042	109.87	0.0000
log(assets.total)	0.6552	0.0037	176.30	0.0000

- Estimate of Variance of Errors: $\widehat{\sigma}^2 = 0.974^2$
- Coefficient of Determination: $R^2 = 0.863$
- Adjusted Coefficient of Determination: $\overline{R}^2 = 0.863$

Results of Fitting: Normal Case

- Sample Regression Plane:

$$\begin{aligned}\widehat{\eta}_{\mathsf{LNL}} &= \widehat{\alpha}_0 + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 0.6 + 0.466 \log(\texttt{employees}) + 0.655 \log(\texttt{assets.total})\end{aligned}$$

- Residuals:

$$e_{\mathsf{LNL}i} := \log(\texttt{sales}_i) - \widehat{\eta}_{\mathsf{LNL}i}$$

- Plots of Regression Diagnostics:

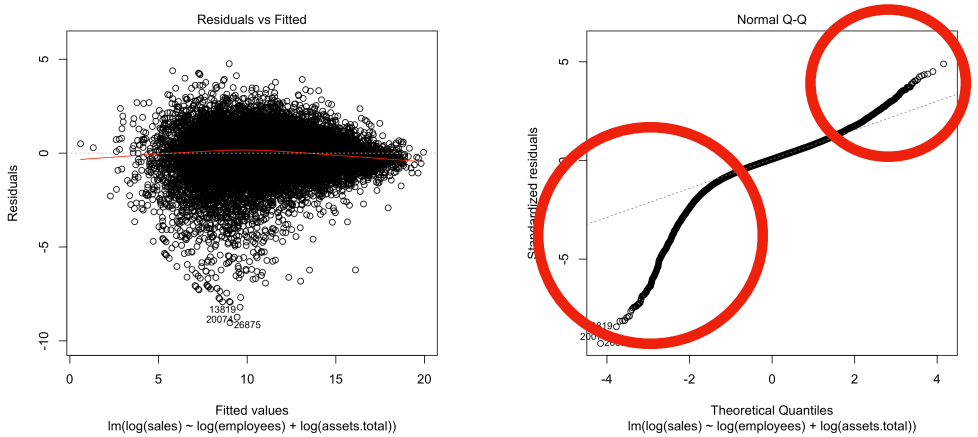


Figure: Plots of Regression Diagnostics: Log-normal Linear Model

Results of Fitting: Skew-normal Case

- t-Values Table:

Table: z-Ratio Table: Log-skew-normal Linear Model

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.62	0.03	59.75	0.00
log(employees)	0.36	0.00	80.89	0.00
log(assets.total)	0.71	0.00	193.20	0.00
omega	1.39	0.01	172.67	0.00
alpha	-2.31	0.04	-63.89	0.00

Results of Fitting: Skew-normal Case

- Adjusted Sample Regression Plane:

$$\begin{aligned}\widetilde{\eta}_{\mathsf{LSNL}} &= \widehat{\eta}_{\mathsf{LSNL}} + \widehat{\omega} b \widehat{\delta} = (\widehat{\alpha}_0 + \widehat{\omega} b \widehat{\delta}) + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 0.6 + 0.358 \log(\texttt{employees}) + 0.709 \log(\texttt{assets.total})\end{aligned}$$

- Centered Parameter (CP) Residuals:

$$e_{\mathsf{LSNL.CP}i} := \log(\texttt{sales}_i) - \widetilde{\eta}_{\mathsf{LSNL}i} = \log(\texttt{sales}_i) - \widehat{\eta}_{\mathsf{LSNL}i} - \widehat{\omega} b \widehat{\delta}$$

- Plots of Regression Diagnostics: CP Case

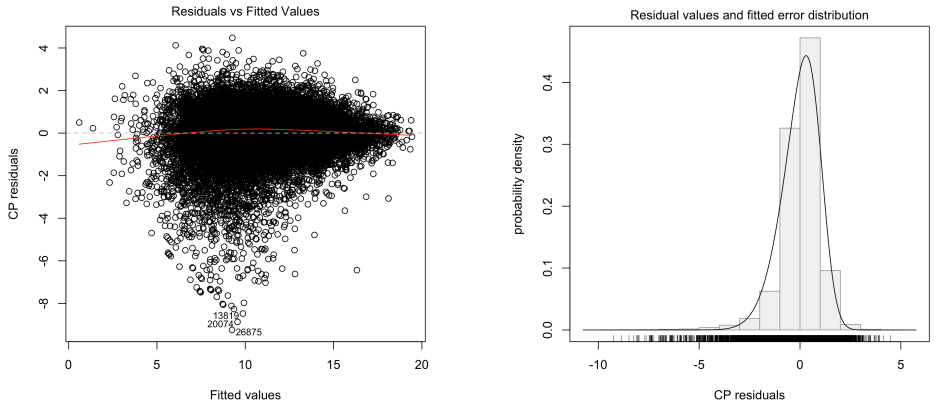


Figure: Plots of Regression Diagnostics: Log-skew-normal Linear Model, CP Case

Results of Fitting: Skew-normal Case

- Scaled DP Residuals:

$$z_{\text{LSNL}i} = \frac{\log(\texttt{sales}_i) - \widehat{\eta}_{\text{LSNL}i}}{\widehat{\omega}}$$

- Note that $z^2_{\text{LSNL}i} \overset{a}{\sim} \chi^2_1$. (See Azzalini and Capitnio (2014), p. 61.)
- Q-Q, P-P Plots of Scaled DP Reisuals:

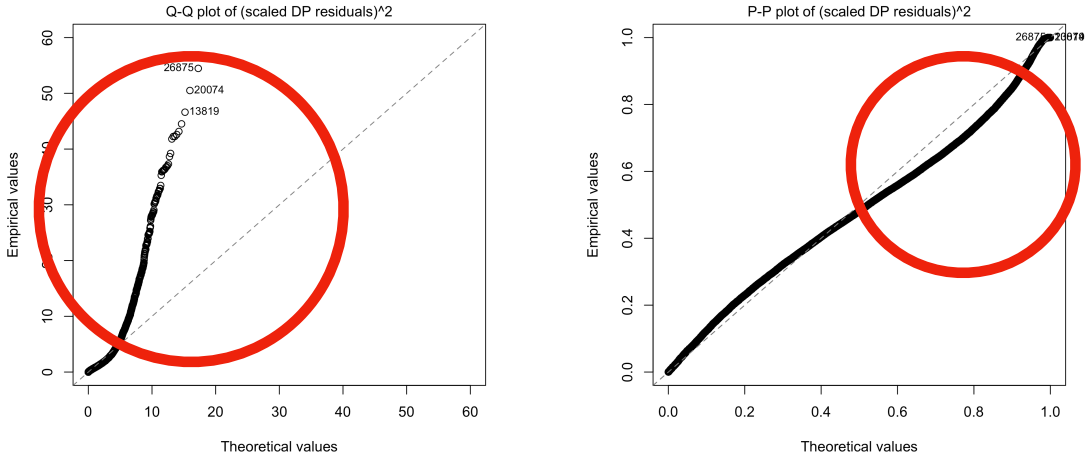


Figure: Q-Q, P-P Plots: Log-skew-normal Linear Model

Results of Fitting: Skew-t Case

- t-Values Table:

Table: z-Ratio Table: Log-skew-t Linear Model

	estimate	std.err	z-ratio	Pr{>|z|}
(Intercept.DP)	1.25	0.03	49.30	0.00
log(employees)	0.35	0.00	86.36	0.00
log(assets.total)	0.71	0.00	213.09	0.00
omega	0.73	0.01	77.67	0.00
alpha	-0.93	0.04	-25.60	0.00
nu	3.40	0.07	47.04	0.00

Results of Fitting: Skew-t Case

- Adjusted Sample Regression Plane:

$$\begin{aligned}\widetilde{\eta}_{\text{LSTL}} &= \widehat{\eta}_{\text{LSTL}} + \widehat{\omega} b_{\widehat{\nu}+1} \widehat{\delta} = (\widehat{\alpha}_0 + \widehat{\omega} b_{\widehat{\nu}+1} \widehat{\delta}) + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total}) \\ &= 0.721 + 0.351 \log(\texttt{employees}) + 0.706 \log(\texttt{assets.total})\end{aligned}$$

- Pseudo-Centered Parameter (CP) Residuals:

$$e_{\text{LSTL.PCP}i} := \log(\texttt{sales}_i) - \widetilde{\eta}_{\text{LSTL}i} = \log(\texttt{sales}_i) - \widehat{\eta}_{\text{LSTL}i} - \widehat{\omega} b_{\widehat{\nu}+1} \widehat{\delta}$$

- Plots of Regression Diagnostics: Pseudo-CP Case

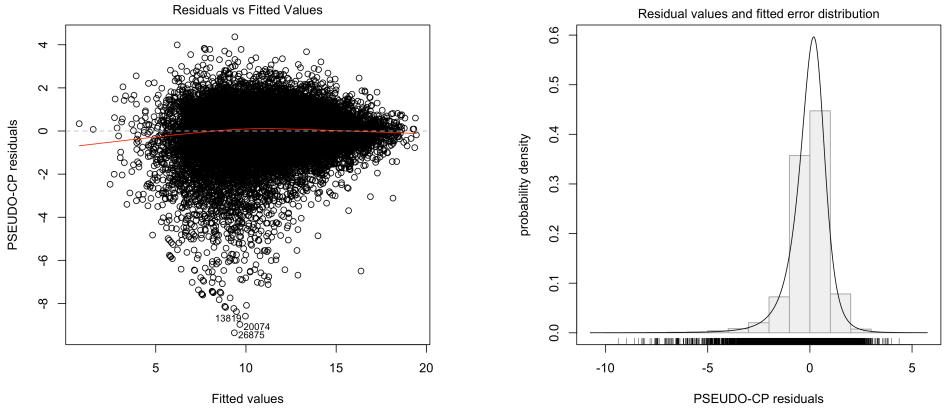


Figure: Plots of Regression Diagnostics: Log-skew-t Linear Model, Pseudo-CP Case

Results of Fitting: Skew-t Case

- Scaled DP Residuals:

$$z_{\mathrm{LSTL}i} = \frac{\log(\texttt{sales}_i) - \widehat{\eta}_{\mathrm{LSTL}i}}{\widehat{\omega}}$$

- Note that $z^2_{\mathrm{LSTL}i} \overset{a}{\sim} \mathrm{F}^1_\nu$ (:F distribution with degree of freedom $(1, \nu)$). (See Azzalini and Capitnio (2014), p. 102.)
- Q-Q, P-P Plots of Scaled DP Reisuals:

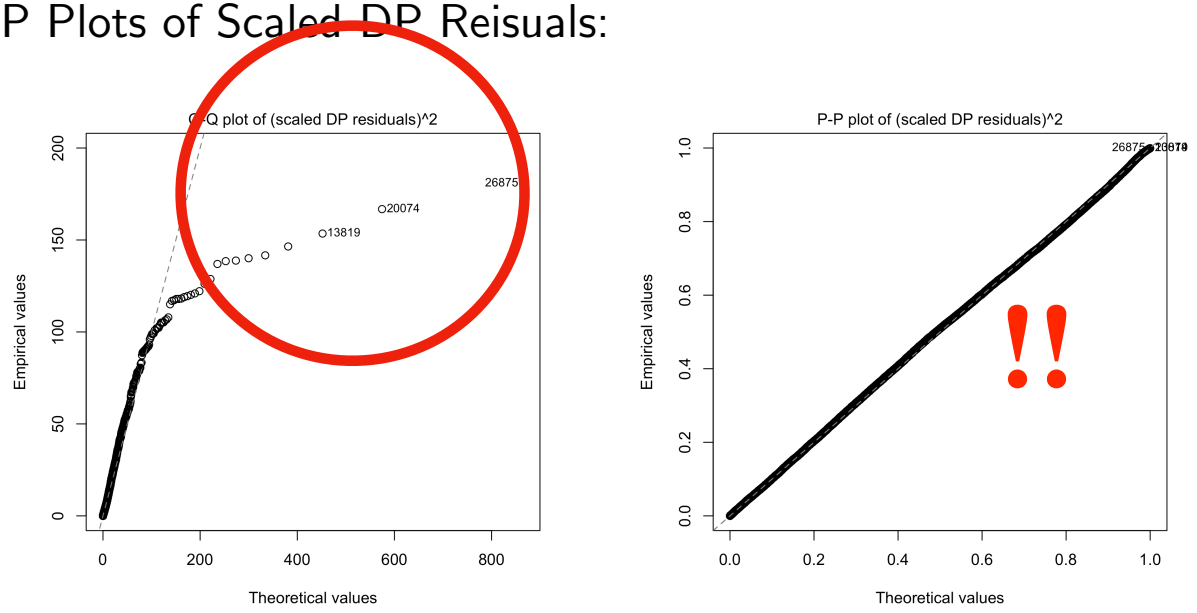
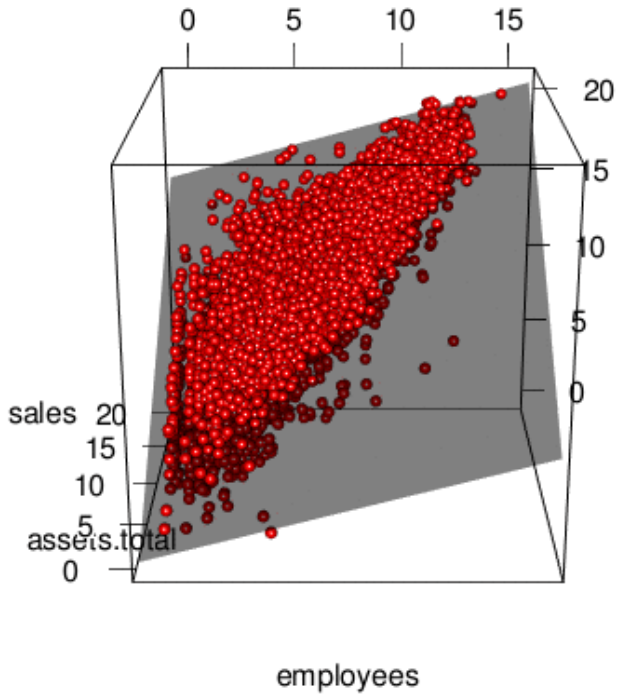


Figure: Q-Q, P-P Plots: Log-skew-t Linear Model

Adjusted Sample Regression Plane: Double-log Model with Skew-t Error



$$\widetilde{\eta}_{\mathsf{LSTL}} = \widehat{\eta}_{\mathsf{LSTL}} + \widehat{\omega} b_{\widehat{\nu}+1} \widehat{\delta} = (\widehat{\alpha}_0 + \widehat{\omega} b_{\widehat{\nu}+1} \widehat{\delta}) + \widehat{\alpha}_1 \log(\texttt{employees}) + \widehat{\alpha}_2 \log(\texttt{assets.total})$$
$$= (1.247 + 0.732 \times 0.976 \times (-0.683)) + 0.351 \log(\texttt{employees}) + 0.706 \log(\texttt{assets.total})$$
$$= 0.759 + 0.351 \log(\texttt{employees}) + 0.706 \log(\texttt{assets.total})$$

Model Selection

Akaike Information Criterion

- Log-likelihood:

$$\ell(\boldsymbol{\theta}) := \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta})$$

  where f is a probability density function and $\boldsymbol{\theta}$ is a parameter vector.

- Maximum Likelihood Estimate (MLE):

$$\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta})$$

  where Θ is a parameter space.

- Definition of Akaike Information Criterion (AIC):

$$\begin{aligned} \text{AIC} &:= -2\ell(\widehat{\boldsymbol{\theta}}) + 2\dim(\boldsymbol{\theta}) \\ &\quad (:= -2 \times \text{(Maximum Log-likelihood)} \\ &\quad + 2 \times \text{(Dimension of Parameter Vector)}) \end{aligned}$$

  See Akaike(1973), Konishi and Kitagawa(2008).

Model Selection: Distribution of `log(sales)`

- AIC Table:

Table: AIC Table: Models for Log of Sales

	df	AIC
lm.log.sales2015	2.00	146767.84
selm.log.sales2015	3.00	146532.92
selm.ST.log.sales2015	4.00	146461.11

- The best distribution for $\log(\texttt{sales})$ in the above them is the skew-t one (selm.ST.log.sales2015).

Model Selection: Log-linear Models for `log(sales)`

- AIC Table:

Table: AIC Table: Log-linear Models for Sales

	df	AIC
lm.log.firmfin2015	4.00	85692.05
selm.log.firmfin2015	5.00	82161.55
selm.ST.log.firmfin2015	6.00	77304.87

- The best model for $\log(\texttt{sales})$ in the above them is the log-skew-t linear one (selm.ST.log.firmfin2015) .

Model Evaluation

K-fold Cross-Validation: MSEP

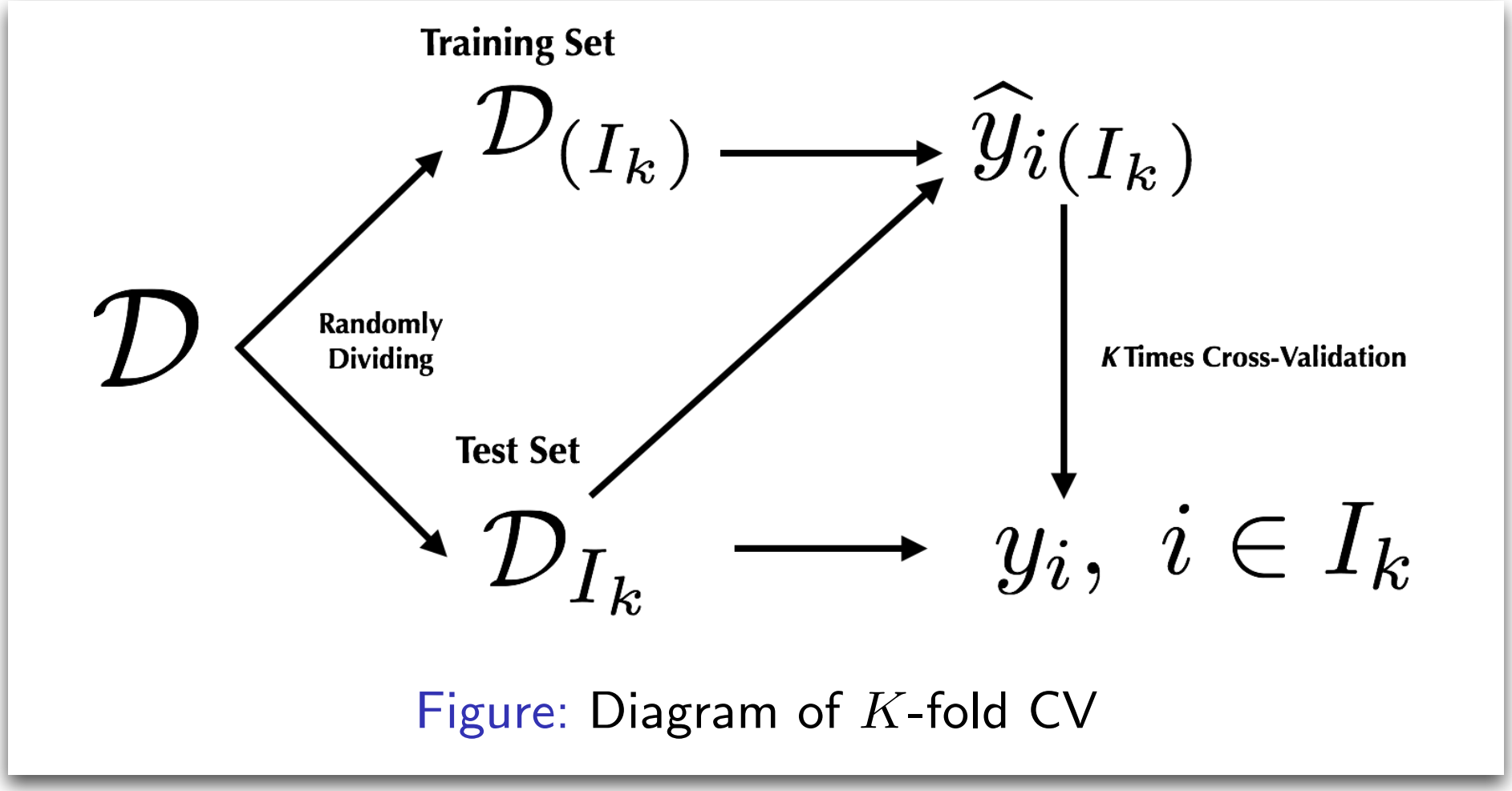


Figure: Diagram of K-fold CV

Procedure of K-fold Cross-Validation: MSEP

(CV1) Randomly divide the data set $\mathcal{D} := \left\{(\boldsymbol{x}^{i\prime}, y_i);\ i = 1, \ldots, n\right\}$ into K sets of approximately the same size,

$$\mathcal{D}_{I_k} := \left\{(\boldsymbol{x}^{i\prime}, y_i);\ i \in I_k\right\} (:\ \text{Training Set})$$

where I_k, $k = 1, \ldots, K$, are set of indices of the data set $\mathcal{D}_{I_k}$ and $n_k := \sharp I_k$. Note that $I := \{1, \ldots, n\} = I_1 \cup \cdots \cup I_k \cup \cdots \cup I_K$.

(CV2) Obtain the linear predictor $\widehat{\eta}_{i(I_k)}$ and its adjusted form $\widetilde{\eta}_{i(I_k)}$ from the data set by

$$\mathcal{D}_{(I_k)} := \mathcal{D} \setminus \mathcal{D}_{I_k} = \left\{(\boldsymbol{x}^{i\prime}, y_i);\ i \in I \setminus I_k\right\} (:\ \text{Test Set})$$

where the notation "$\setminus$" denotes the set difference.

(CV3) Calculate the following criterion (MSEP):

$$\mathrm{CV}(K) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} D(y_i, \widehat{y}_{i(I_k)}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in I_k} \left(y_i - \widehat{y}_{i(I_k)}\right)^2 \tag{17}$$

where the discrepancy function is given by

$$D(y_i, \widehat{y}_{i(I_k)}) := \left(y_i - \widehat{y}_{i(I_k)}\right)^2 (: \text{Squared Error of Prediction}) \tag{18}$$

and

$$\widehat{y}_{i(I_k)} := \widehat{\eta}_{i(I_k)} \text{ or } \widetilde{\eta}_{i(I_k)}$$

K-fold Cross-Validation: AIC

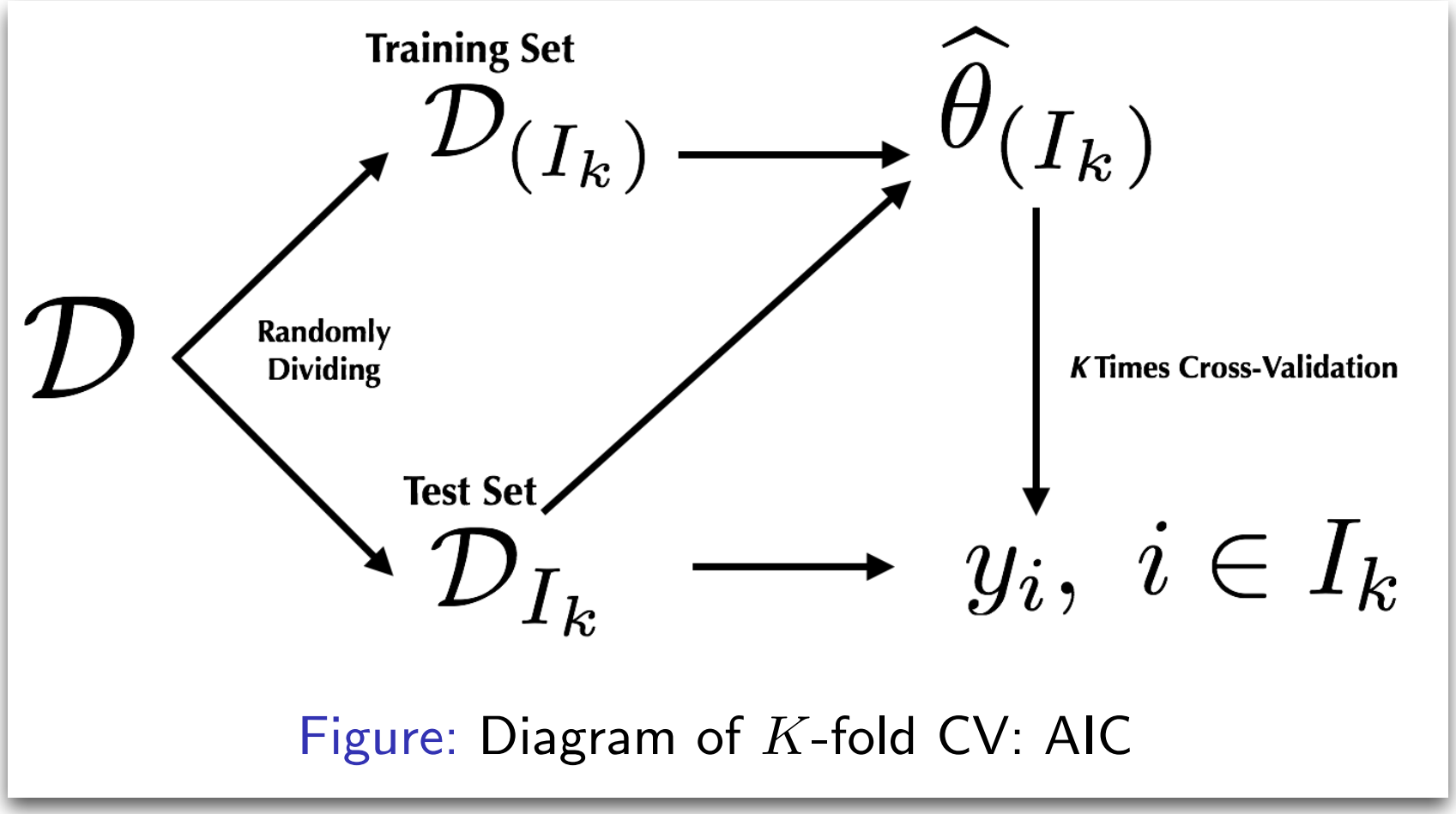


Figure: Diagram of K-fold CV: AIC

Procedure of K-fold Cross-Validation: AIC

(CV-AIC1) Randomly divide the data set $\mathcal{D}$ into K sets of approximately same-size $\mathcal{D}_{I_k}$.

(CV-AIC2) Estimate the MLE $\widehat{\boldsymbol{\theta}}_{(I_k)}$ from the data set $\mathcal{D}_{(I_k)}$.

(CV-AIC3) Calculate the following criterion:

$$\begin{aligned}\mathrm{CV}_{\mathrm{AIC}}(K) &:= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in I_k} D_{\mathrm{AIC}}(y_i, \widehat{\boldsymbol{\theta}}_{(I_k)}) \\ &= -\frac{2}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in I_k}\log f(y_i \mid \widehat{\boldsymbol{\theta}}_{(I_k)}) + \frac{2\dim(\boldsymbol{\theta})}{n}\end{aligned}$$

where the discrepancy function

$$\begin{aligned}D_{\mathrm{AIC}}(y_i, \widehat{\boldsymbol{\theta}}_{(I_k)}) &:= -2\ell_i(\widehat{\boldsymbol{\theta}}_{(I_k)} \mid y_i) + \frac{2\dim(\boldsymbol{\theta})}{n} \\ &= -2\log f(y_i \mid \widehat{\boldsymbol{\theta}}_{(I_k)}) + \frac{2\dim(\boldsymbol{\theta})}{n} \qquad (19)\end{aligned}$$

See Efron and Hastie (2016), p. 226.

Results of Cross-Validation

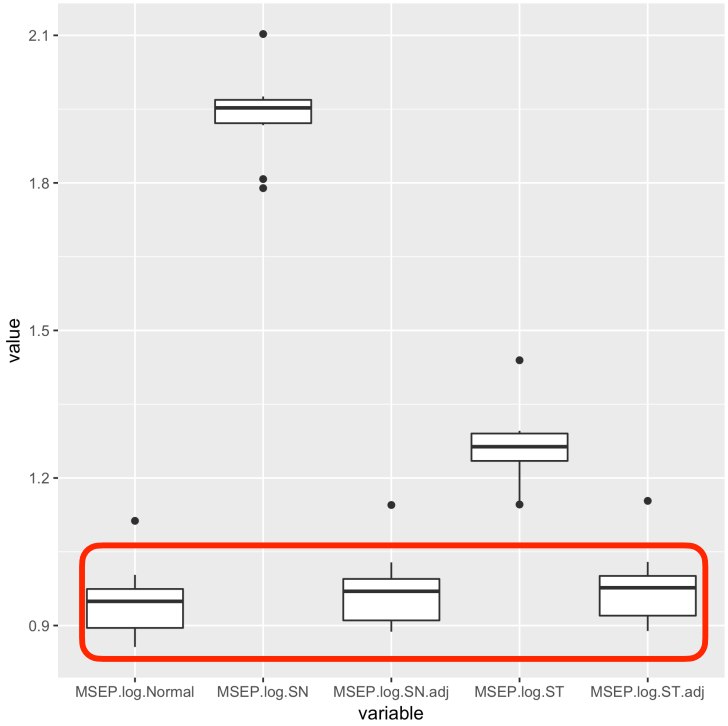


Figure: K-fold CV: MSEP, $K = 10$

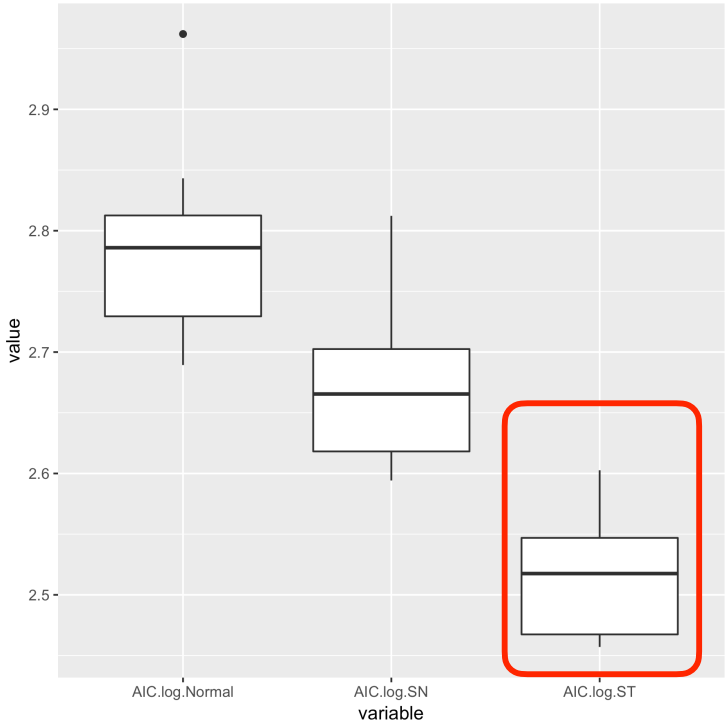


Figure: K-fold CV: AIC, $K = 10$

Dynamic Documents and Reproducible Research

Peng's Reproducibility Spectrum
Peng, R. D. (2011) Reproducible research in computational science, *Science*, Vol. 334, pp. 1226–1227.

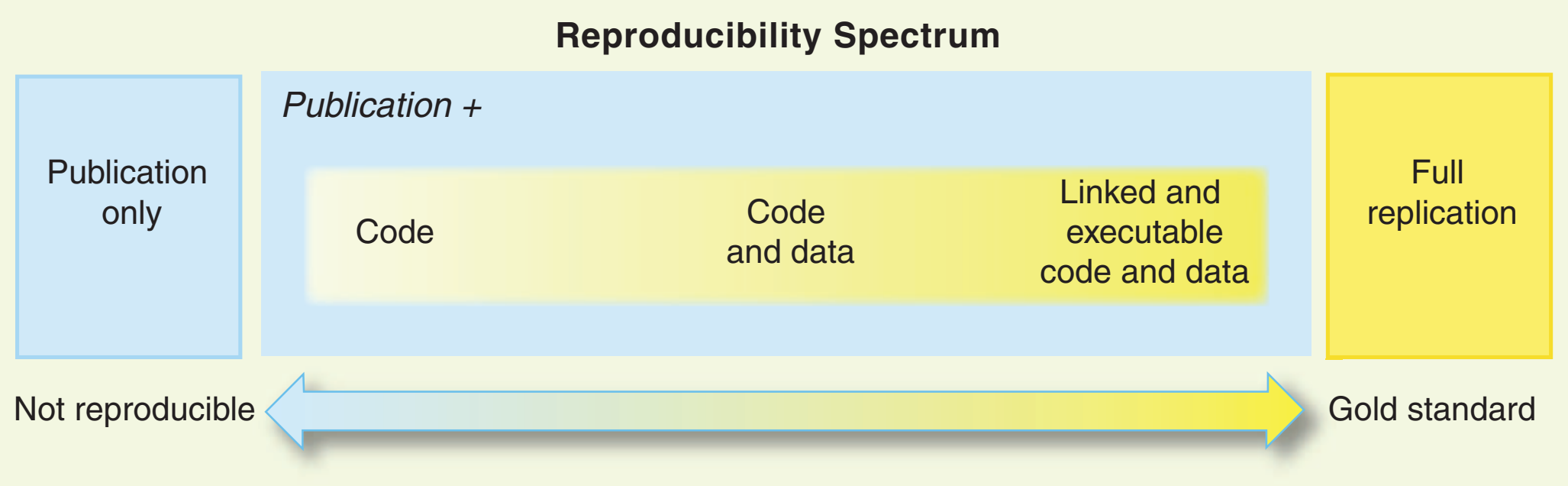
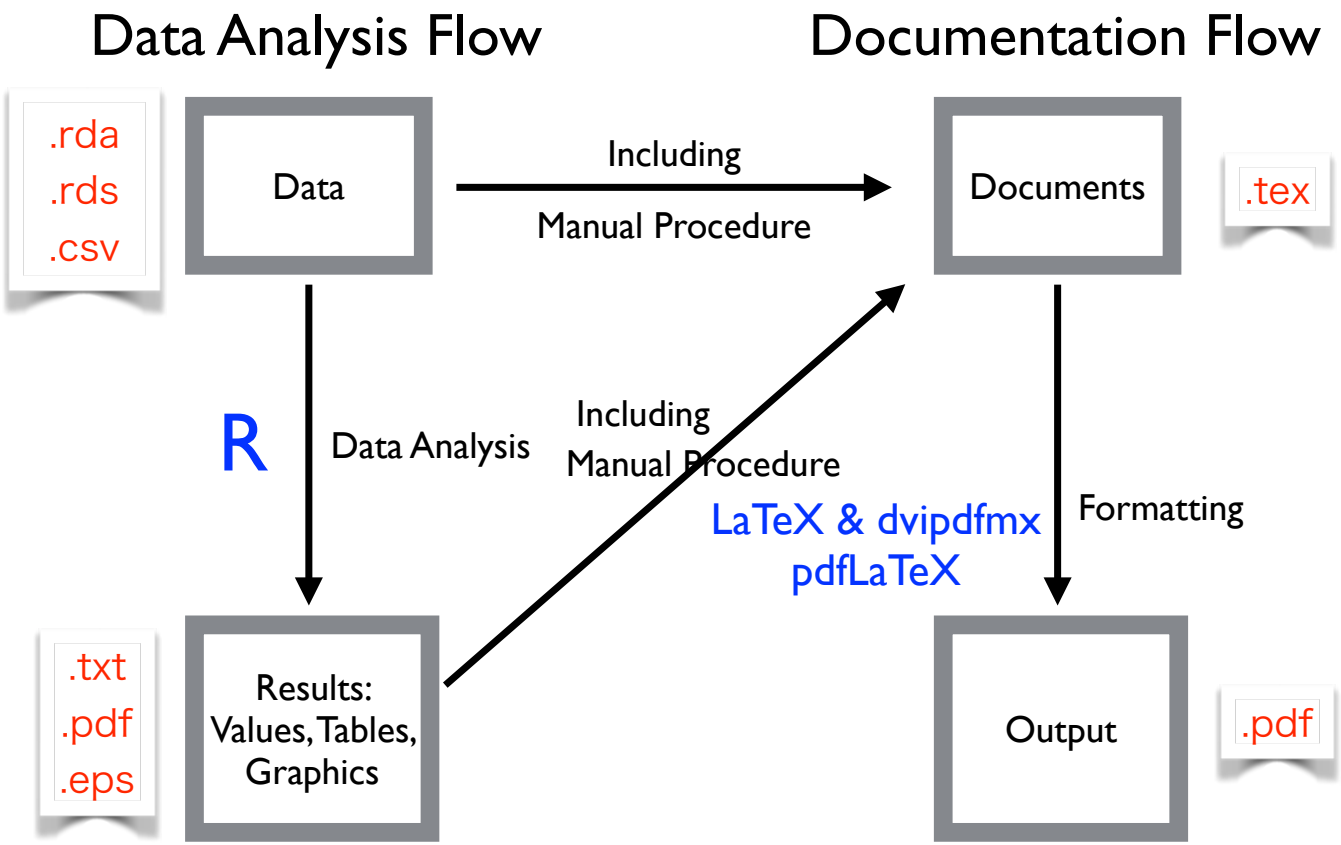


Fig. 1. The spectrum of reproducibility.

Standard Data Analysis and Documentation with R and TeX

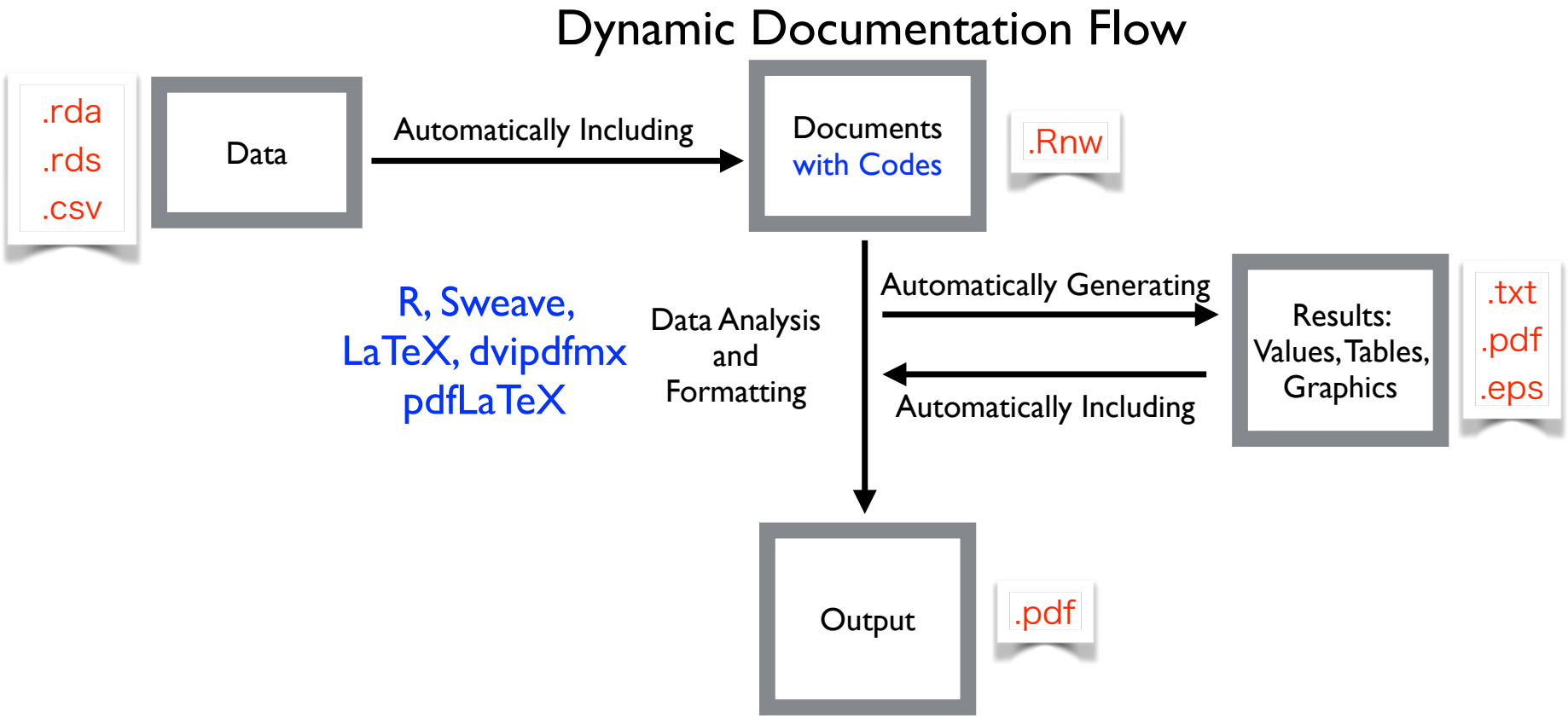


Current Dynamic Documents Tools

- Sweave (Friedrich Leisch and R-core) and Rnw (R Noweb (Sweave)) file based on noweb (Norman Ramsey)
  `https://www.statistik.lmu.de/~leisch/Sweave/`

- knitr (Yihui Xie) and Rmd (R Markdown) file based on Markdown (John Gruber)
  `http://yihui.name/knitr/`

Dynamic Documents with R and TeX

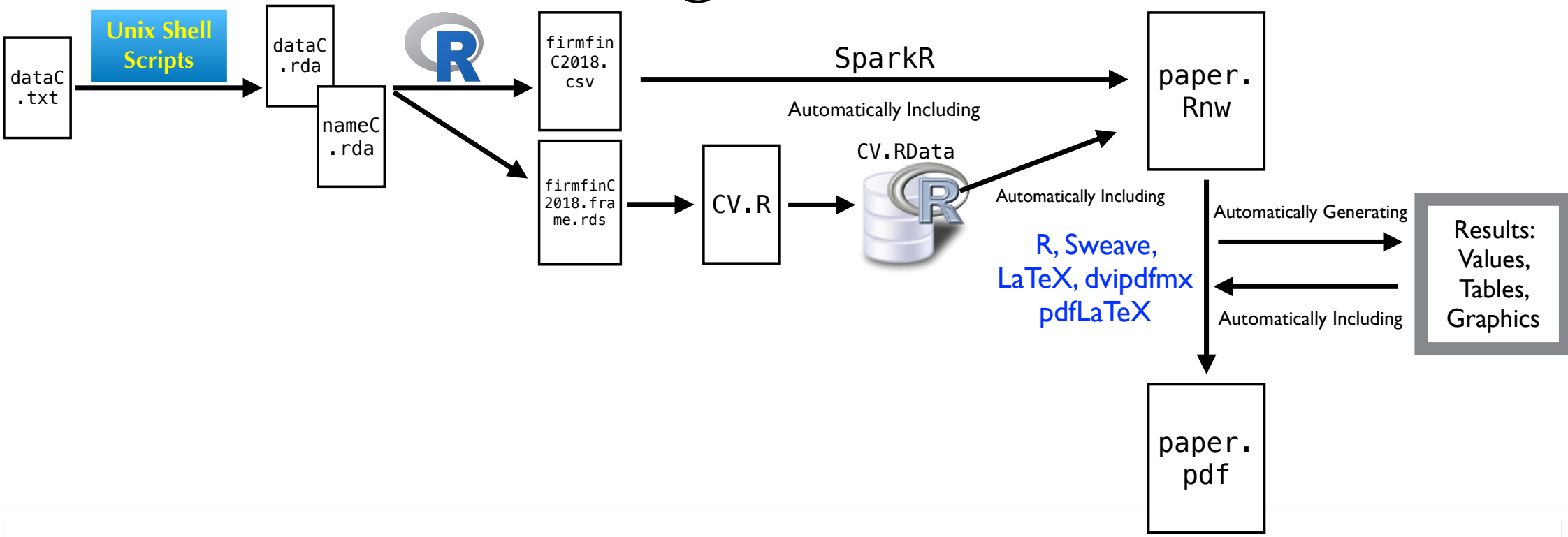


Automation for All Process by `make` Command

Makefile

```
all:
	date > start-all.txt
	/bin/bash ./script.sh
	Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
	Rscript CV.R
	~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
	date > end-all.txt
all-p:
	date > start-all-p.txt
	/bin/bash ./script-p.sh
	Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
	Rscript CV.R
	~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
	date > end-all-p.txt
csv:
	date > start-csv.txt
	/bin/bash ./script.sh
	Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
	date > end-csv.txt
csv-p:
	date > start-csv-p.txt
	/bin/bash ./script-p.sh
	Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
	date > end-csv-p.txt
CV:
	date > start-CV.txt
	Rscript CV.R
	date > end-CV.txt
paper:
	~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
paper-without-CV:
	/bin/bash ./script.sh
	Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
	~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
```

Target: `all`




```
all:
   /bin/bash ./script.sh
   Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
   Rscript CV.R
   ~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
```

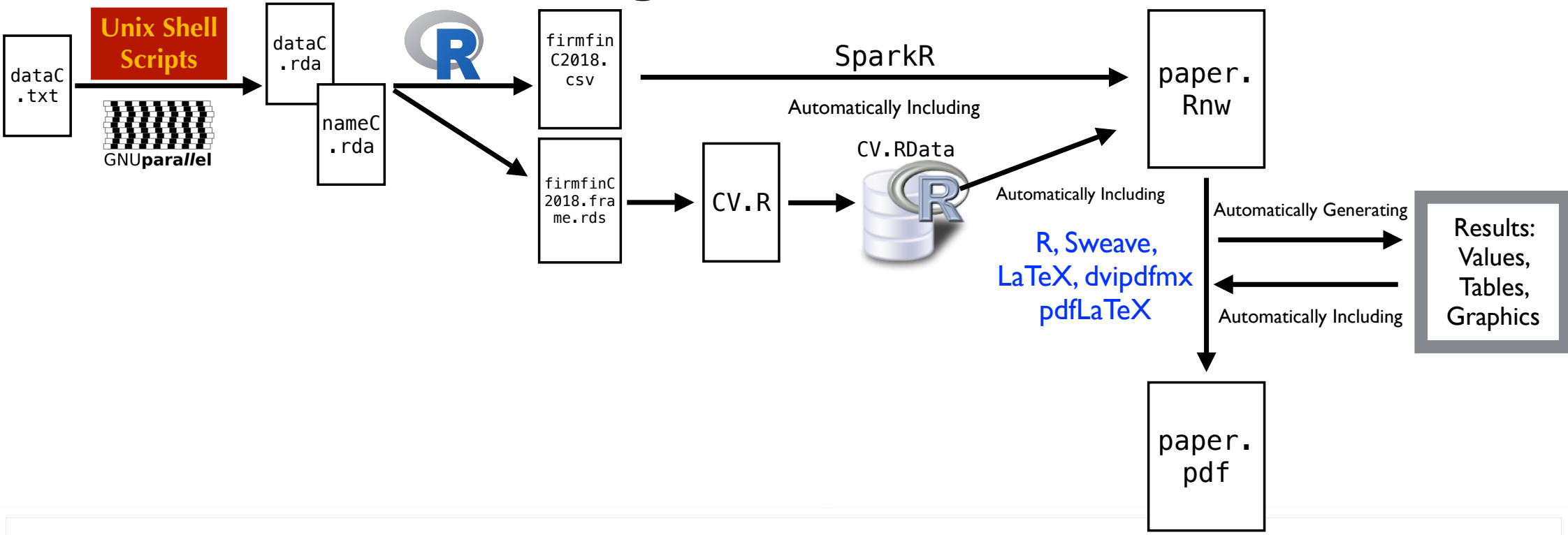
Time: `make all` (with CV Simulation)

(University of Tokyo, FENNEL)

```
$ cat start-all.txt
2019年  9月 16日 月曜日 13:01:10 JST

$ cat end-all.txt
2019年  9月 16日 月曜日 13:16:47 JST
```

処理時間15分37秒

Target: all-p




```
all-p:
   /bin/bash ./script-p.sh
   Rscript datadump.R "dataC.rda" "nameC.rda" "firmfinC2018.csv" "firmfinC2018.frame.rds"
   Rscript CV.R
   ~/Library/TeXShop/Engines/Sweave-utf8.engine paper.Rnw
```

Time: `make all-p` (with CV Simulation)

(University of Tokyo, FENNEL)

```
$  cat start-all-p.txt
 2019年  9月 16日 月曜日 15:48:06 JST

$ cat end-all-p.txt
$ 2019年  9月 16日 月曜日 16:01:23 JST
```

処理時間 13分17秒

(並列処理をしない場合(15分37秒)から2分20秒の短縮)

Peng's Reproducibility Spectrum
Peng, R. D. (2011) Reproducible research in computational science, *Science*, Vol. 334, pp. 1226–1227.

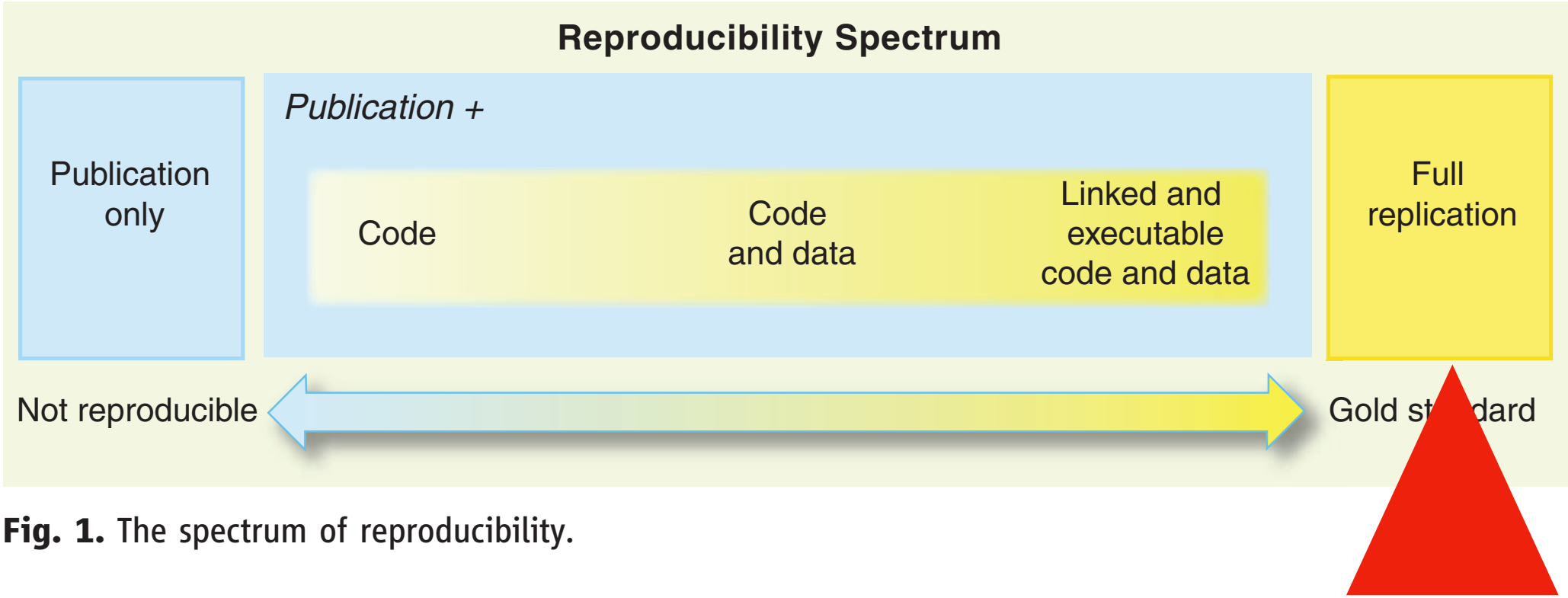
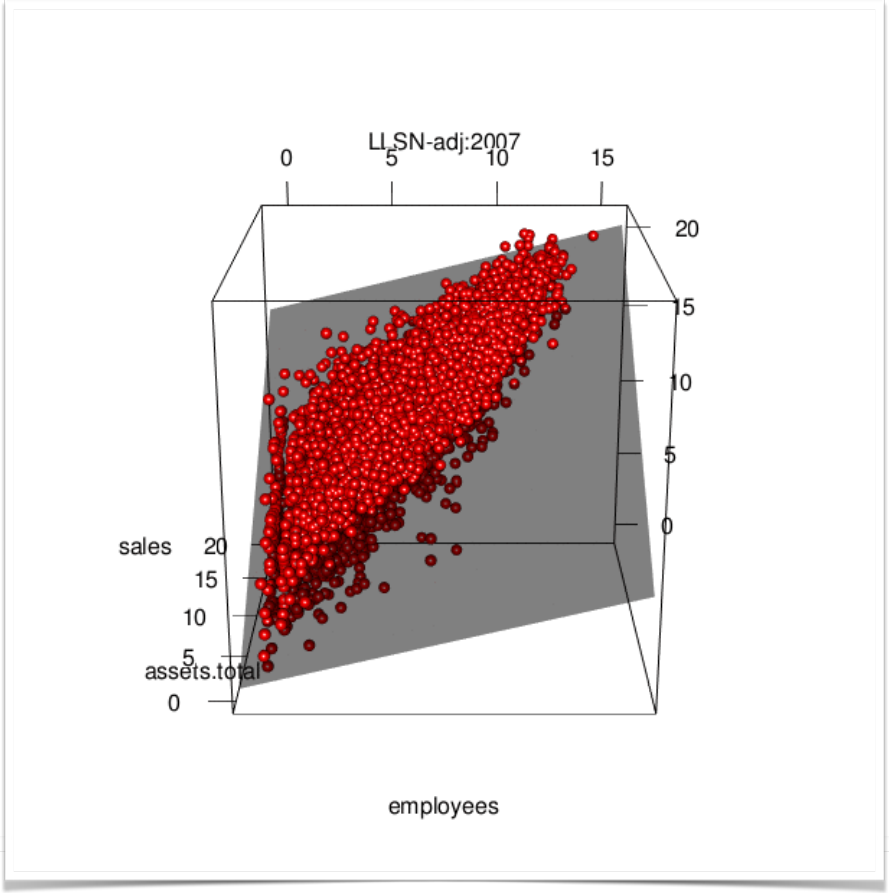


Fig. 1. The spectrum of reproducibility.

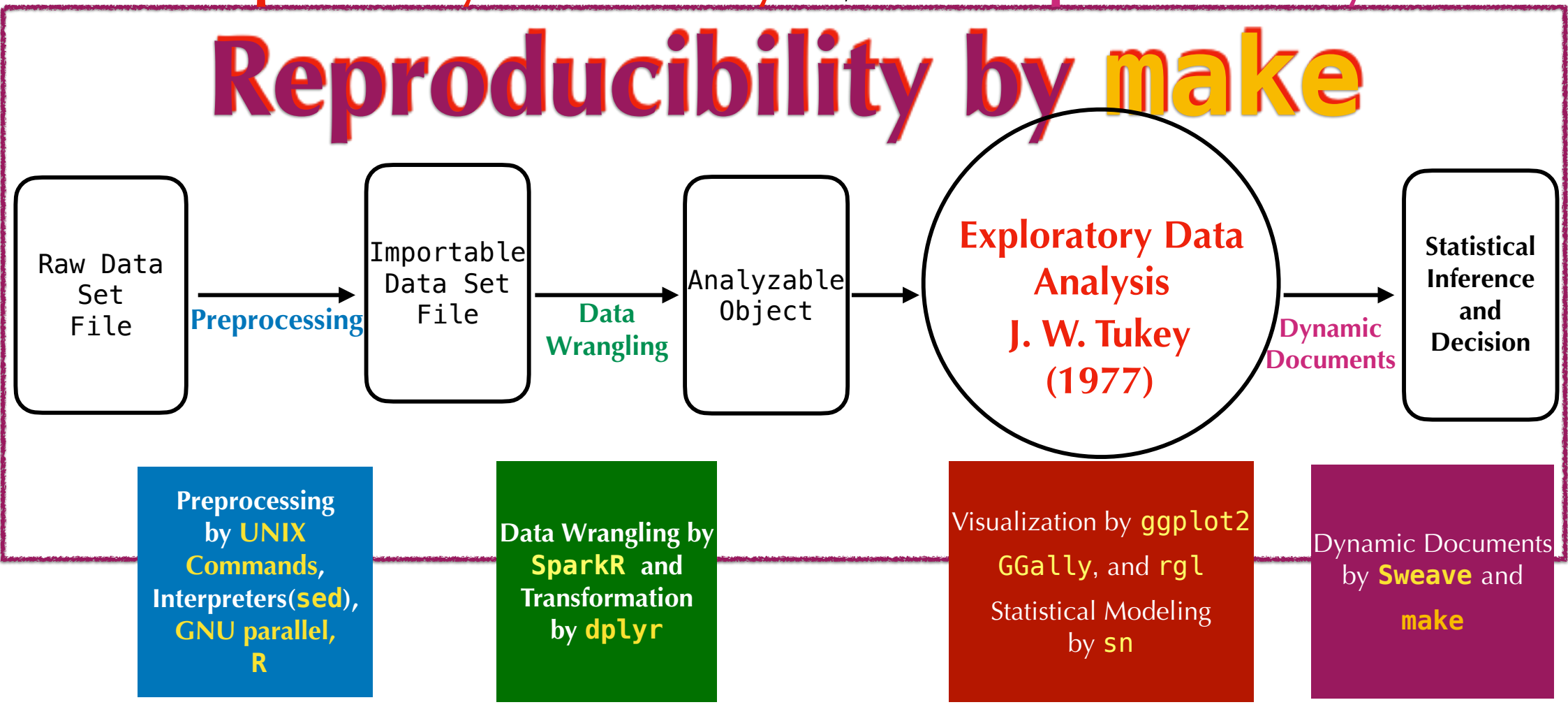
Automation for Making Documents for One Decade(2007–2016) by `make` Command

3-d Scatter Plot Animation: Log-Linear Model with Skew-t, Adjusted Version

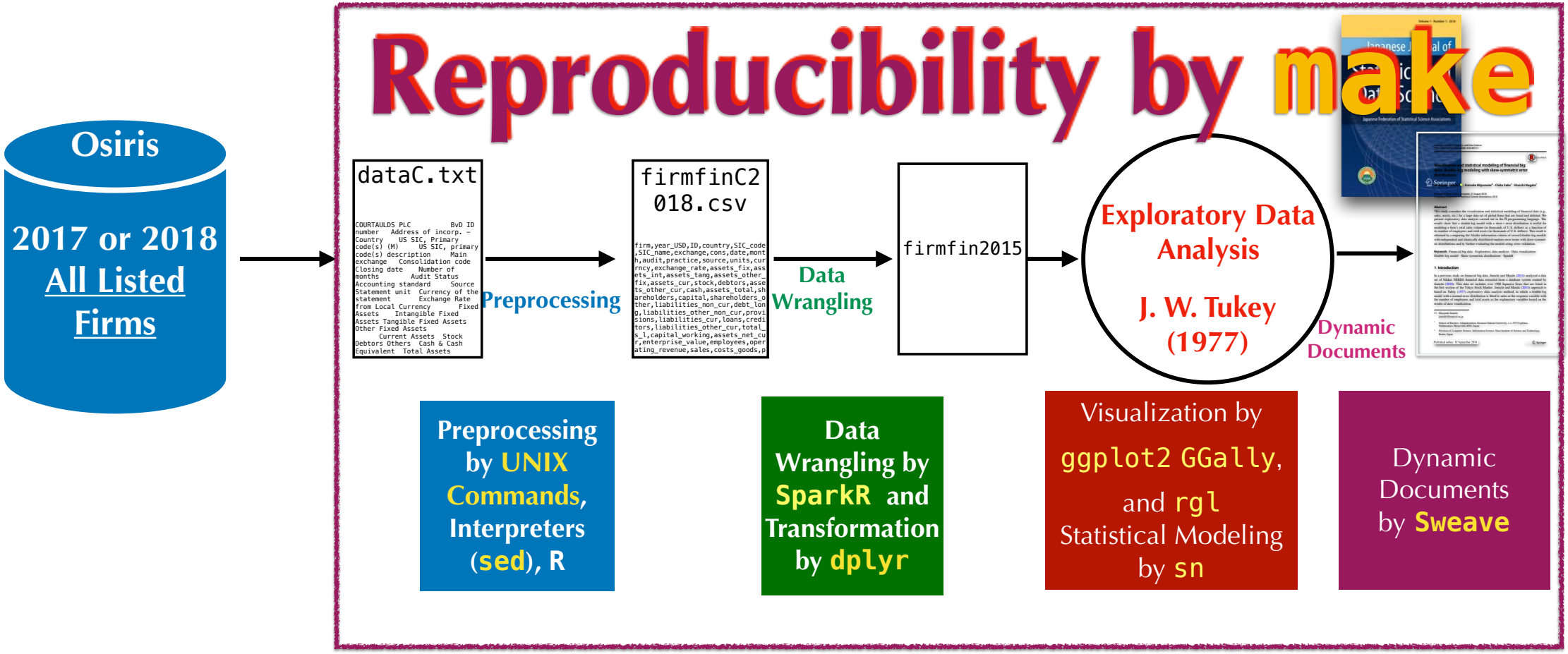



```
animation:
 convert -layers optimize -loop 0 -delay 40 LogLinearModelingDecade-scatter3dlstlm-adj*.png animationscatterplot-LLMST-adj.gif
```

Preprocessing, Data Wrangling, Exploratory Data Analysis, and Reproducibility

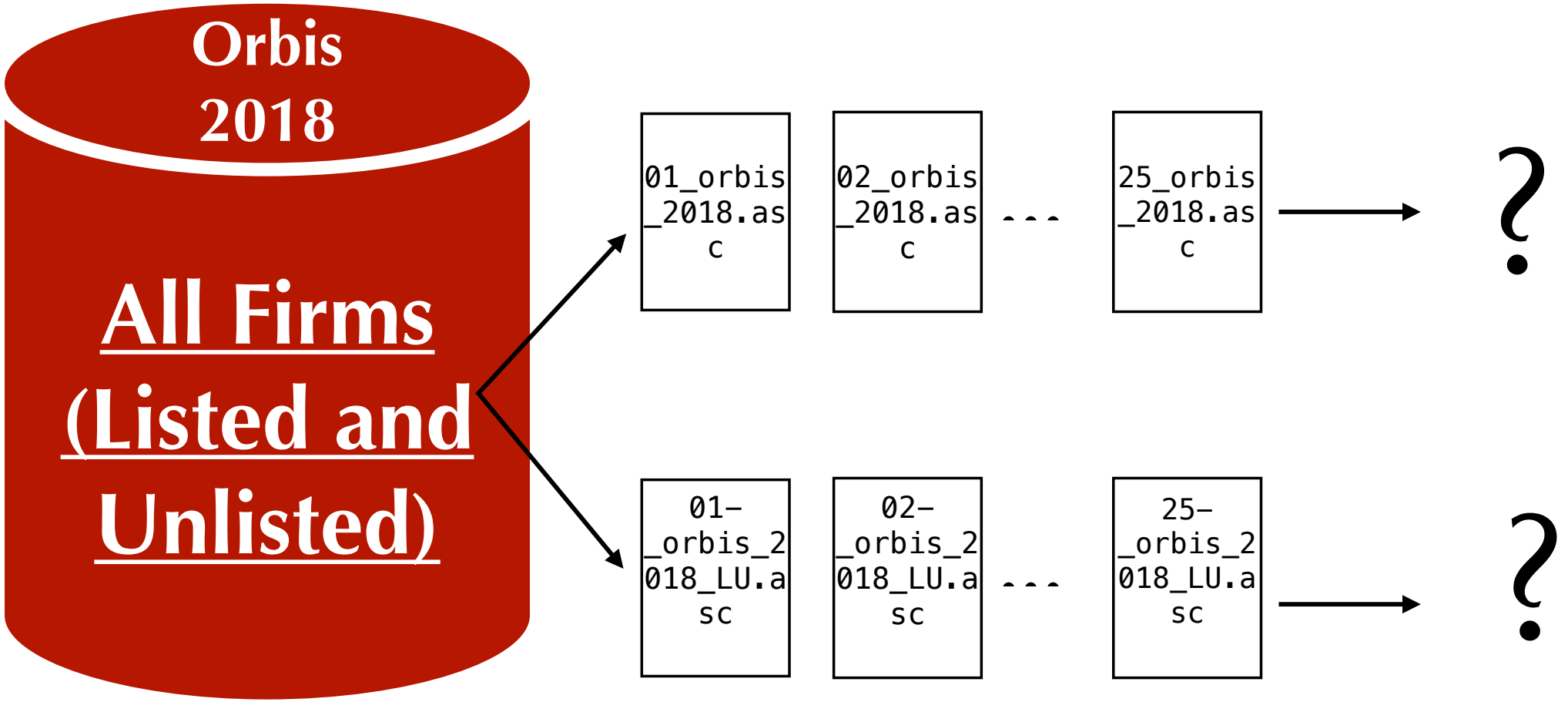


Osiris Database and Its Data Set



Next Challenges

Orbis Database and Its Data Sets



Database, Data Sets and Preprocessing

Database and Data Information

- Bureau van Dijk (ビューロー・ヴァン・ダイク)社 (以下 BvD と略)
  世界の全企業 (上場・非上場) のデータベース Orbis (オービス)
- 世界の企業約2.4億社以上の情報を 国際比較可能な統一のフォームで収録
- 「連結主体」(Consolidate)と 「非連結主体」(Un-consolidate) の二種類で抽出
- 世界の全企業 (連結主体抽出対象 24,014,352社, 非連結主体抽出対象 24,012,807社) の主要財務情報 (売上高, 営業利益, 総資産など) を最長10年分抽出
  →パネルデータ(経時観測データ)
- 「項目(フィールド, カラム)間はタブ(\t)区切りで抽出 (TSV ファイル)

Data File Information by Unix Commands: Consolidate Data Sets

```
$ du -hc *.asc
5.7G  01_orbis_2018.asc
5.5G  02_orbis_2018.asc
5.4G  03_orbis_2018.asc
5.4G  04_orbis_2018.asc
5.3G  05_orbis_2018.asc
5.3G  06_orbis_2018.asc
5.3G  07_orbis_2018.asc
5.3G  08_orbis_2018.asc
5.3G  09_orbis_2018.asc
5.2G  10_orbis_2018.asc
5.2G  11_orbis_2018.asc
5.2G  12_orbis_2018.asc
5.2G  13_orbis_2018.asc
5.3G  14_orbis_2018.asc
5.3G  15_orbis_2018.asc
5.2G  16_orbis_2018.asc
5.3G  17_orbis_2018.asc
5.4G  18_orbis_2018.asc
5.3G  19_orbis_2018.asc
5.2G  20_orbis_2018.asc
5.2G  21_orbis_2018.asc
5.3G  22_orbis_2018.asc
5.3G  23_orbis_2018.asc
5.4G  24_orbis_2018.asc
79M 25_orbis_2018.asc
127G  合計
```

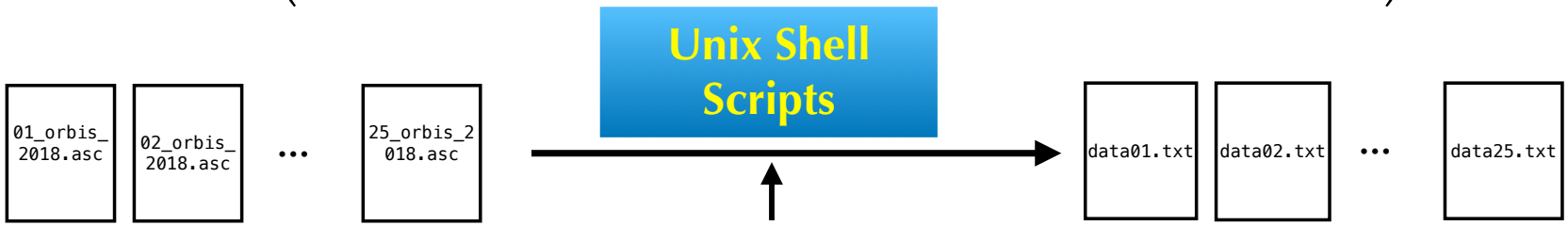
```
$ wc -l *.asc
 11000000 01_orbis_2018.asc
 11000000 02_orbis_2018.asc
 11000000 03_orbis_2018.asc
 11000000 04_orbis_2018.asc
 11000000 05_orbis_2018.asc
 11000000 06_orbis_2018.asc
 11000000 07_orbis_2018.asc
 11000000 08_orbis_2018.asc
 11000000 09_orbis_2018.asc
 11000000 10_orbis_2018.asc
 11000000 11_orbis_2018.asc
 11000000 12_orbis_2018.asc
 11000000 13_orbis_2018.asc
 11000000 14_orbis_2018.asc
 11000000 15_orbis_2018.asc
 11000000 16_orbis_2018.asc
 11000000 17_orbis_2018.asc
 11000000 18_orbis_2018.asc
 11000000 19_orbis_2018.asc
 11000000 20_orbis_2018.asc
 11000000 21_orbis_2018.asc
 11000000 22_orbis_2018.asc
 11000000 23_orbis_2018.asc
 11000000 24_orbis_2018.asc
  157872 25_orbis_2018.asc
 264157872 total
```

Data File Information by Unix Commands: Un-consolidate Data Sets

```
$ du -hc *.asc
5.6G  01_orbis_2018_LU.asc
5.4G  02_orbis_2018_LU.asc
5.3G  03_orbis_2018_LU.asc
5.3G  04_orbis_2018_LU.asc
5.2G  05_orbis_2018_LU.asc
5.2G  06_orbis_2018_LU.asc
5.2G  07_orbis_2018_LU.asc
5.2G  08_orbis_2018_LU.asc
5.2G  09_orbis_2018_LU.asc
5.2G  10_orbis_2018_LU.asc
5.2G  11_orbis_2018_LU.asc
5.2G  12_orbis_2018_LU.asc
5.2G  13_orbis_2018_LU.asc
5.2G  14_orbis_2018_LU.asc
5.2G  15_orbis_2018_LU.asc
5.1G  16_orbis_2018_LU.asc
5.2G  17_orbis_2018_LU.asc
5.4G  18_orbis_2018_LU.asc
5.3G  19_orbis_2018_LU.asc
5.2G  20_orbis_2018_LU.asc
5.1G  21_orbis_2018_LU.asc
5.2G  22_orbis_2018_LU.asc
5.2G  23_orbis_2018_LU.asc
5.3G  24_orbis_2018_LU.asc
69M 25_orbis_2018_LU.asc
125G  合計
```

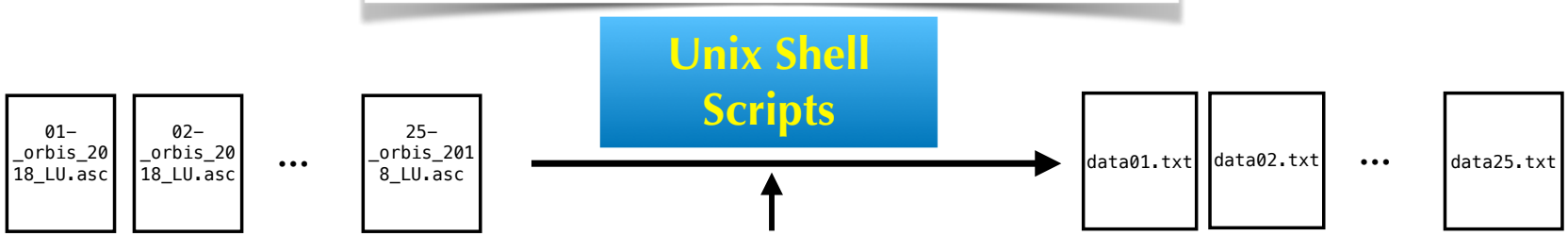
```
$ wc -l *.asc
11000000 01_orbis_2018_LU.asc
11000000 02_orbis_2018_LU.asc
11000000 03_orbis_2018_LU.asc
11000000 04_orbis_2018_LU.asc
11000000 05_orbis_2018_LU.asc
11000000 06_orbis_2018_LU.asc
11000000 07_orbis_2018_LU.asc
11000000 08_orbis_2018_LU.asc
11000000 09_orbis_2018_LU.asc
11000000 10_orbis_2018_LU.asc
11000000 11_orbis_2018_LU.asc
11000000 12_orbis_2018_LU.asc
11000000 13_orbis_2018_LU.asc
11000000 14_orbis_2018_LU.asc
11000000 15_orbis_2018_LU.asc
11000000 16_orbis_2018_LU.asc
11000000 17_orbis_2018_LU.asc
11000000 18_orbis_2018_LU.asc
11000000 19_orbis_2018_LU.asc
11000000 20_orbis_2018_LU.asc
11000000 21_orbis_2018_LU.asc
11000000 22_orbis_2018_LU.asc
11000000 23_orbis_2018_LU.asc
11000000 24_orbis_2018_LU.asc
  140888 25_orbis_2018_LU.asc
264140888 合計
```

Copy (Renumbering)
(Consolidate and Un-consolidate)



renumbering.sh

```
#!/bin/bash
for i in $(seq -w 25); do
cp "./rawdata/"$i"_orbis_2018.asc" "./data"$i".txt";
done
```



renumbering.sh

```
#!/bin/bash
for i in $(seq -w 25); do
cp "./rawdata/"$i"_orbis_2018_LU.asc" "./data"$i".txt";
done
```

Raw Data File: Consolidate Version

<U+FEFF>WALMART INC. Country City Postcode Telephone number BvD ID number National ID number National ID label National ID type IP identification number IP identification label ISIN number Ticker symbol Main exchange Listed/Delisted/Unlisted Consolidation code Closing date Number of months Audit status Accounting practice Source (for publicly quoted companies) Original units Original currency Exchange rate from original currency Fixed assets Intangible fixed assets Tangible fixed assets Other fixed assets Current assets Stock Debtors Other current assets Cash & cash equivalent Total assets Shareholders funds Capital Other shareholders funds Non-current liabilities Long term debt Other non-current liabilities Provisions Current liabilities Loans Creditors Other current liabilities Total shareh. funds & liab. Working capital Net current assets Enterprise value Number of employees Operating revenue (Turnover) Sales Costs of goods sold Gross profit Other operating expenses Operating P/L [=EBIT] Financial revenue Financial expenses Financial P/L P/L before tax Taxation P/L after tax Extr. and other revenue Extr. and other expenses Extr. and other P/L P/L for period [=Net income] Export revenue Material costs Costs of employees Depreciation & Amortization Other operating items Interest paid Research & Development expenses Cash flow Added value EBITDA US SIC, Primary code(s) Ibid, text description US SIC, Secondary code(s) Ibid, text description BvD major sector Information provider

2008 (th USD) United States of America BENTONVILLE 72716 +1 479 273 4000 US710415188 71-0415188 EIN

VAT/Tax number 9556N Reuters number US9311421039 WMT New York Stock Exchange (NYSE) Listed C1 31/01/2009 12 Unqualified Local GAAP 10-K thousands USD 1 114,480,000 15,260,000 95,653,000 3,567,000 48,949,000 34,511,000 3,905,000 10,533,000 7,275,000 163,429,000 65,285,000 393,000 64,892,000 42,754,000 34,549,000 8,205,000 n.a. 55,390,000 6,163,000 28,849,000 20,378,000

163,429,000 9,567,000 -6,441,000 219,773,662 2,100,000 404,254,000 404,254,000 297,202,000

107,052,000 84,285,000 22,767,000 284,000 2,184,000 -1,900,000 20,867,000 7,133,000 13,734,000 n.a. n.a. -353,000 13,381,000 n.a. n.a. n.a. 6,739,000 77,546,000 2,184,000 n.a.

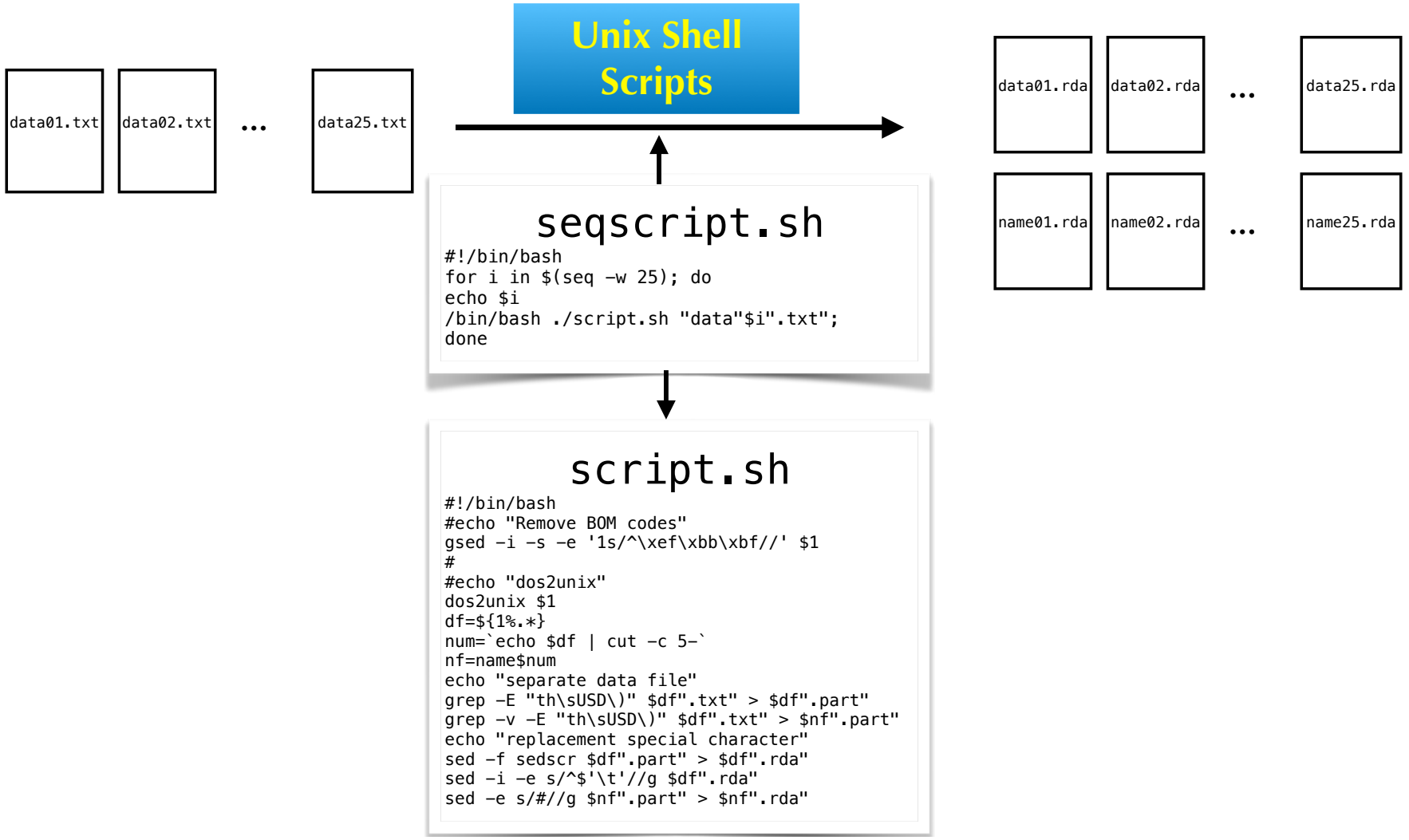
20,120,000 n.a. 29,506,000 5331 Variety stores 5411 Grocery stores Wholesale & retail trade Reuters

Raw Data File: Un-consolidate Version

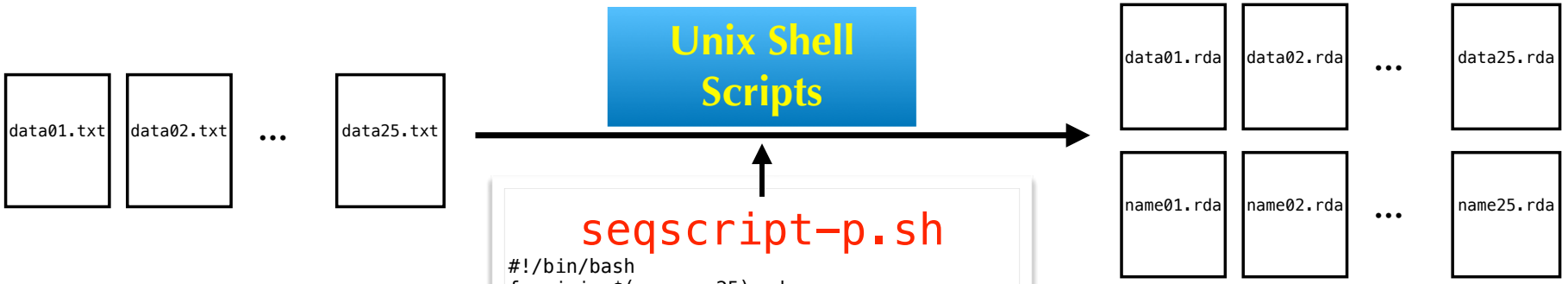
<U+FEFF>WALMART INC. Country City Postcode Telephone number BvD ID number National ID number National ID label National ID type IP identification number IP identification label ISIN number Ticker symbol Main exchange Listed/Delisted/Unlisted Consolidation code Closing date Number of months Audit status Accounting practice Source (for publicly quoted companies) Original units Original currency Exchange rate from original currency Fixed assets Intangible fixed assets Tangible fixed assets Other fixed assets Current assets Stock Debtors Other current assets Cash & cash equivalent Total assets Shareholders funds Capital Other shareholders funds Non-current liabilities Long term debt Other non-current liabilities Provisions Current liabilities Loans Creditors Other current liabilities Total shareh. funds & liab. Working capital Net current assets Enterprise value Number of employees
Operating revenue (Turnover) Sales Costs of goods sold Gross profit Other operating expenses Operating P/L [=EBIT] Financial revenue Financial expenses Financial P/L P/L before tax Taxation P/L after tax Extr. and other revenue Extr. and other expenses Extr. and other P/L P/L for period [=Net income] Export revenue Material costs Costs of employees Depreciation & Amortization Other operating items
Interest paid Research & Development expenses Cash flow Added value EBITDA US SIC, Primary code(s) Ibid, text description US SIC, Secondary code(s) Ibid, text description BvD major sector Information provider
2008 (th USD) United States of America BENTONVILLE 72716 +1 479 273 4000 US710415188 71-0415188 EIN VAT/Tax number 9556N Reuters number US9311421039 WMT New York Stock Exchange (NYSE) Listed C1 31/01/2009 12 Unqualified Local GAAP 10-K thousands USD 1 114,480,000
15,260,000 95,653,000 3,567,000 48,949,000 34,511,000 3,905,000 10,533,000 7,275,000 163,429,000 65,285,000 393,000 64,892,000 42,754,000 34,549,000 8,205,000 n.a. 55,390,000 6,163,000 28,849,000 20,378,000 163,429,000 9,567,000 -6,441,000 219,773,662 2,100,000 404,254,000 404,254,000 297,202,000 107,052,000 84,285,000 22,767,000 284,000 2,184,000 -1,900,000 20,867,000 7,133,000 13,734,000 n.a. n.a. -353,000 13,381,000 n.a. n.a. n.a. 6,739,000 77,546,000 2,184,000 n.a. 20,120,000
n.a. 29,506,000 5331 Variety stores 5411 Grocery stores Wholesale & retail trade Reuters

Standard and Parallel Preprocessing

Standard Preprocessing
(Common Format between Consolidate and Un-consolidate)



Parallel Preprocessing
(Common Format between Consolidate and Un-consolidate)



seqscript-p.sh

```
#!/bin/bash
for i in $(seq -w 25); do
echo $i
/bin/bash ./script-p.sh "data"$i".txt";
done
```

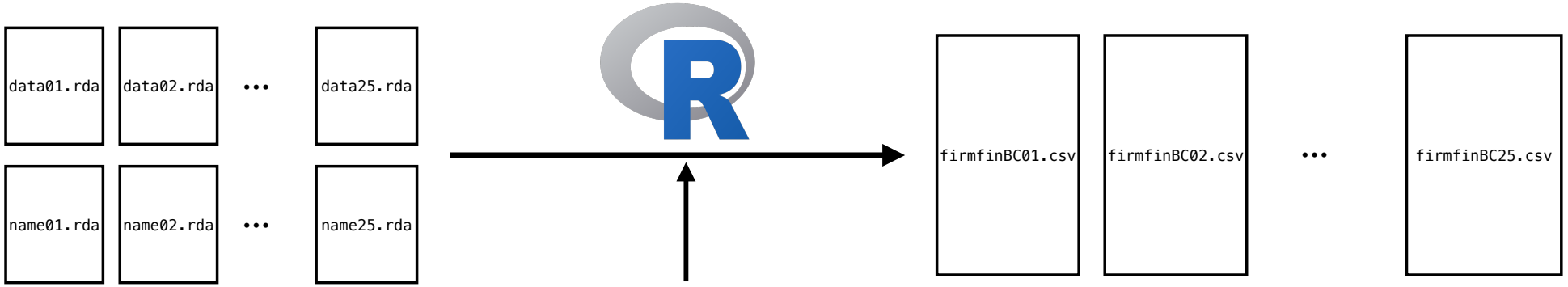
script-p.sh

```
#!/bin/bash
echo "Remove BOM codes"
sed -i -s -e '1s/^\xef\xbb\xbf//' $1
#
echo "dos2unix"
parallel --pipepart -k --block 100M -a $1 "dos2unix -f" > tmp
mv tmp $1
df=${1%.*}
num=`echo $df | cut -c 5-`
nf=name$num
echo "separate data file"
parallel --pipepart -k --block 100M -a $df".txt" 'grep -E "th\sUSD\)"' > $df".part"
parallel --pipepart -k --block 100M -a $df".txt" 'grep -v -E "th\sUSD\)"' > $nf".part"
echo "replacement special character"
parallel --pipepart -k --block 100M -a $df".part" "sed -f sedscr" > tmp
parallel --pipepart -k --block 100M -a tmp "sed s/^$'\t'//g" > $df".rda"
parallel --pipepart -k --block 100M -a $nf".part" "sed -e s/#//g"  > $nf".rda"
rm tmp
```

Preprocessing by R and Merge by Unix Shell Scripts

Consolidate Version

Sequential Dump to CSV Files: Consolidate Version

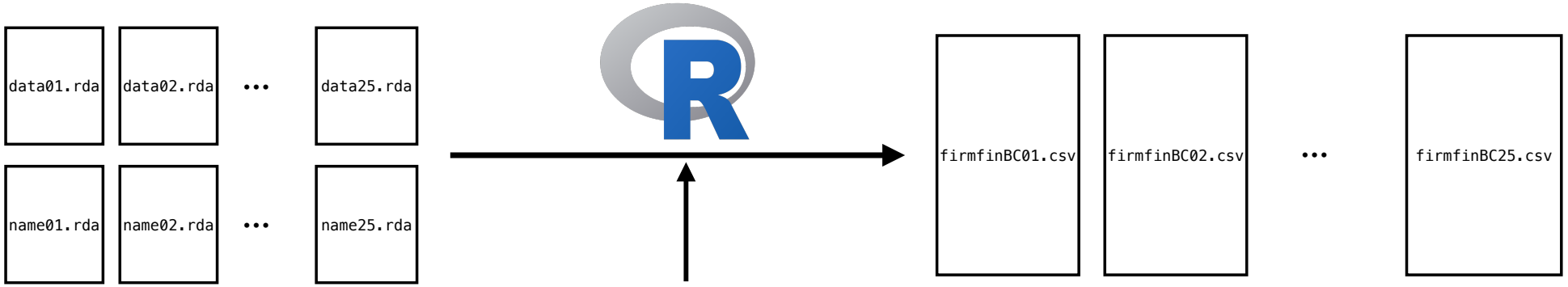


seqdatadump.sh

```
#!/bin/sh
for i in {1..25} ; do
    echo $i
    Rscript datadump.R "data$i.rda" "name$i.rda" "firmfinBC$i.csv"
done
```

Datadump.R

Parallelized Sequential Dump to CSV Files: Consolidate Version



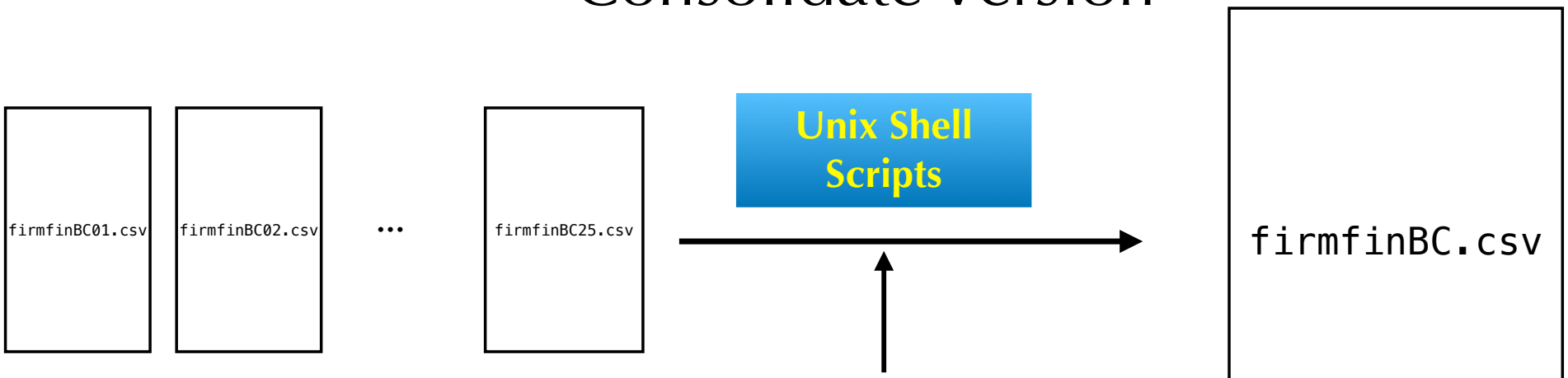
seqdatadump-p.sh

```
#!/bin/bash
echo "Start datadump parallel"
seq -w 10 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBC"{}".csv"
seq -w 11 20 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBC"{}".csv"
seq -w 21 25 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBC"{}".csv"
echo "End datadump parallel"
```

↓

datadump.R

Merge CSV Files: Consolidate Version

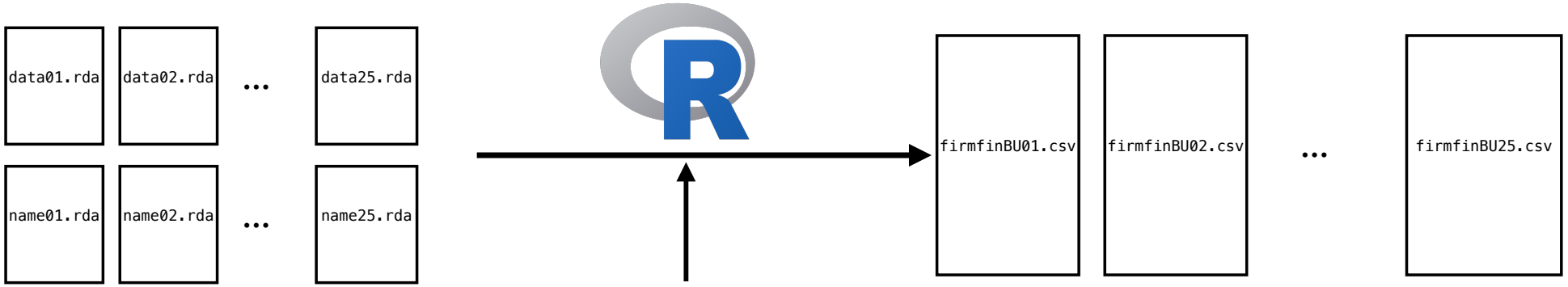


mergedata.sh

```
#!/bin/bash
ofn=firmfinBC
cat ${ofn}01.csv > $ofn.csv
for i in $(seq -w 02 25) ; do
    echo $i
    tail -n +2 $ofn$i.csv >> $ofn.csv
done
```

Un-consolidate Version

Sequential Dump to CSV Files: Un-consolidate Version

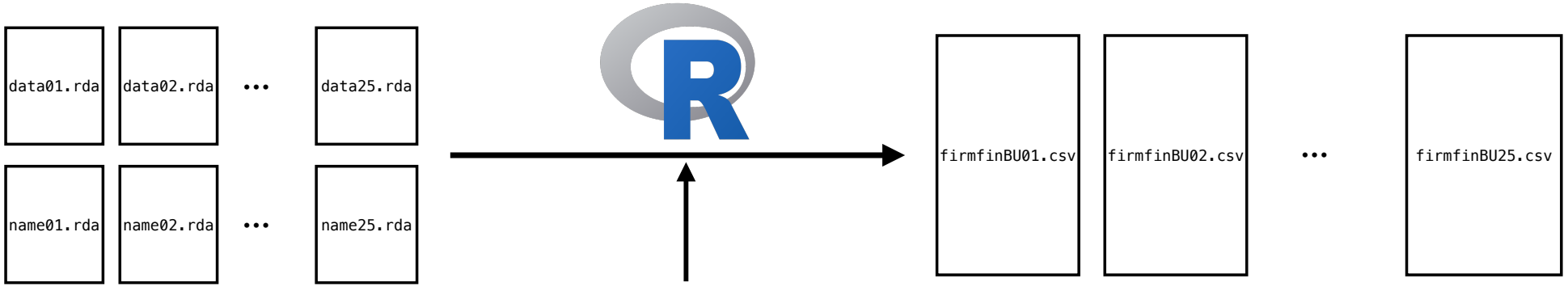


seqdatadump.sh

```
#!/bin/sh
for i in {1..25} ; do
    echo $i
    Rscript datadump.R "data$i.rda" "name$i.rda" "firmfinBU$i.csv"
done
```

datadump.R

Parallelized Sequential Dump to CSV Files: Un-consolidate Version



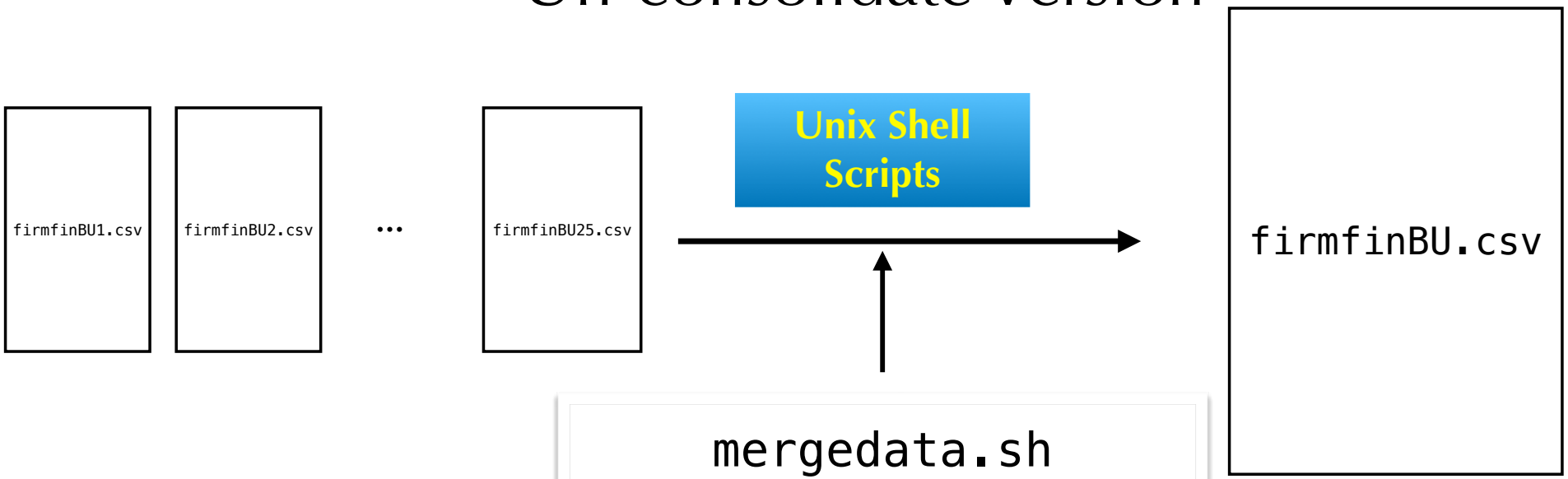
seqdatadump-p.sh

```
#!/bin/bash
echo "Start datadump parallel"
seq -w 10 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBU"{}".csv"
seq -w 11 20 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBU"{}".csv"
seq -w 21 25 | parallel --jobs 100% Rscript datadump.R "data"{}".rda" "name"{}".rda" "firmfinBU"{}".csv"
echo "End datadump parallel"
```

↓

datadump.R

Merge CSV Files: Un-consolidate Version



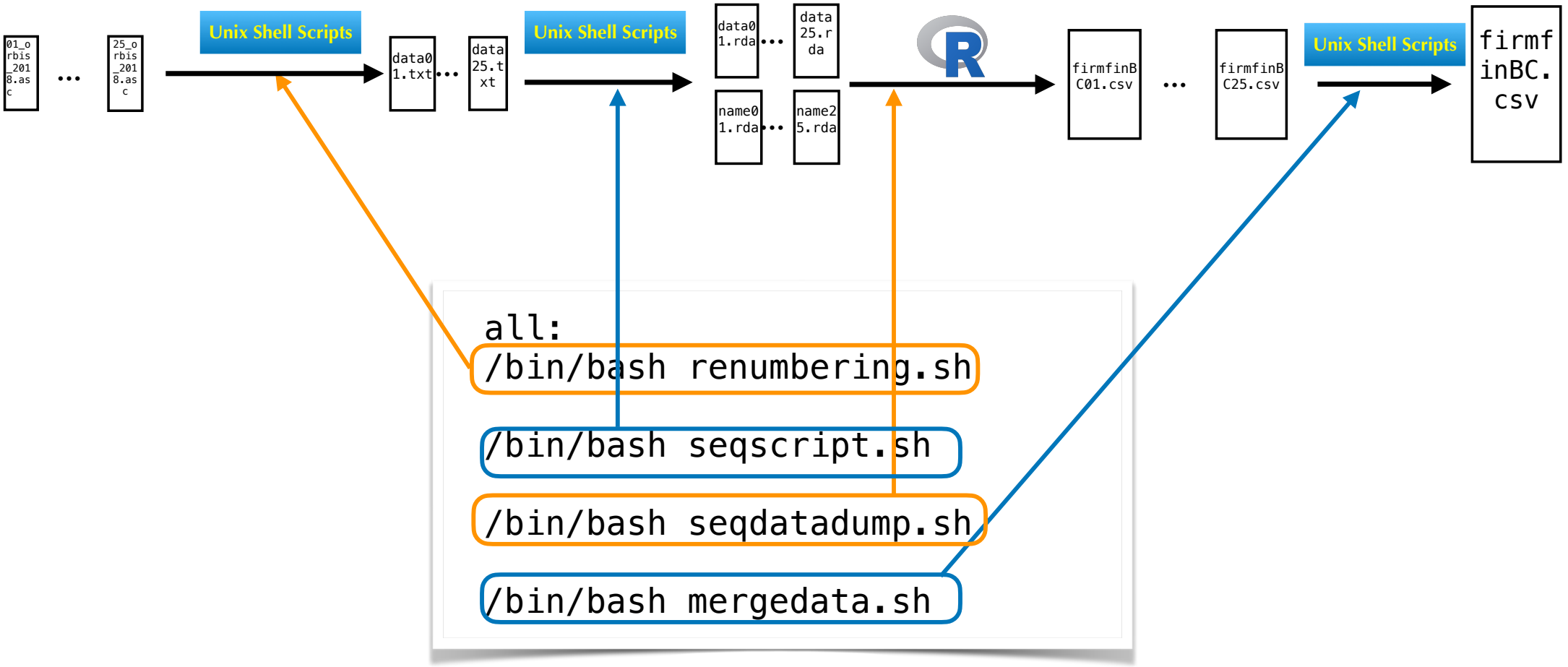
mergedata.sh

```
#!/bin/bash
ofn=firmfinBU
cat ${ofn}01.csv > $ofn.csv
for i in $(seq -w 02 25) ; do
    echo $i
    tail -n +2 $ofn$i.csv >> $ofn.csv
done
```

Automation for All Process of Making CSV Files by `make` Command

Consolidate Version

Preprocessing (All Process for Making CSV File): Consolidate Version (`firmfinBC*.csv`,`firmfinBC.csv`)



Preprocessing Time on iMac Pro and Dell Precision T7910: Consolidate Version

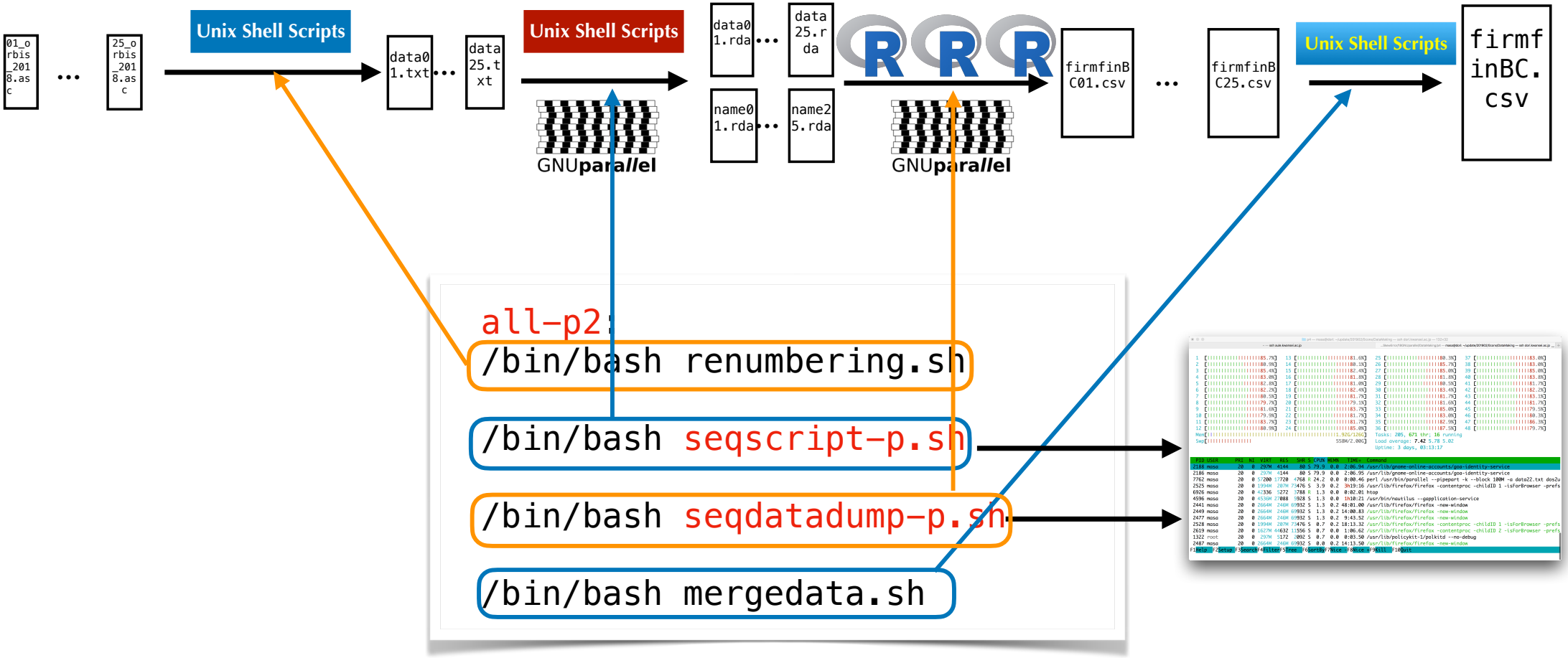
```
# iMac Pro
$ time make
real    806m2.943s
user    768m6.310s
sys     31m43.936s
```

処理時間: 13時間26分

```
# Dell
$ cat starttime.txt
2019年  2月 22日 金曜日 12:30:22 JST
$ cat endtime.txt
2019年  2月 22日 金曜日 18:48:13 JST
```

処理時間: 6時間18分

Double Parallelized Preprocessing (All Process for Making CSV File): Consolidate Version (`firmfinBC*.csv`, `firmfinBC.csv`)



Comparison of Double Parallelized Preprocessing Times on iMac Pro, Dell Precision T7910 Consolidate Version

```
# iMac Pro
aule$ cat starttime-p2.txt

aule$ cat endtime-p2.txt
```

処理時間: 3時間57分

（従来: 13時間26分→**9時間30分程度の短縮**）

```
# Dell
dori$ cat starttime-p.txt
2019年  3月 21日 木曜日 13:06:57 JST
dori$ cat endtime-p.txt
2019年  3月 21日 木曜日 14:26:51 JST
```

処理時間: 1時間20分

Best!

（従来: 6時間18分→**5時間の短縮**）

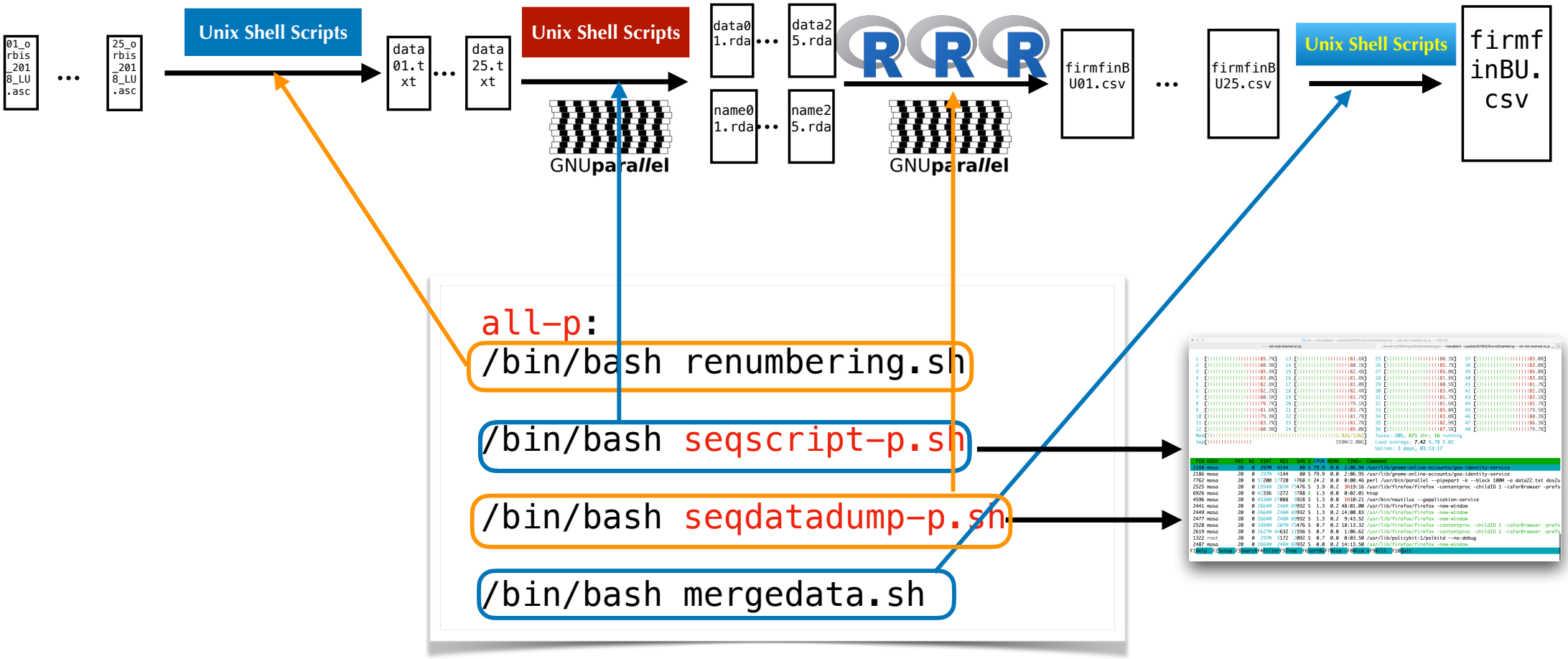
CSV File Informations by Unix Commands: Consolidate Data Sets

```
$ ls -la *.csv
-rw-rw-r-- 1 masa masa 124553951464  3月 21 14:26 firmfinBC.csv
-rw-rw-r-- 1 masa masa   5852932119  3月 21 14:03 firmfinBC01.csv
-rw-rw-r-- 1 masa masa   5588445830  3月 21 14:03 firmfinBC02.csv
-rw-rw-r-- 1 masa masa   5454191450  3月 21 14:02 firmfinBC03.csv
-rw-rw-r-- 1 masa masa   5357048340  3月 21 14:03 firmfinBC04.csv
-rw-rw-r-- 1 masa masa   5275804443  3月 21 14:03 firmfinBC05.csv
-rw-rw-r-- 1 masa masa   5216140330  3月 21 14:03 firmfinBC06.csv
-rw-rw-r-- 1 masa masa   5173893813  3月 21 14:03 firmfinBC07.csv
-rw-rw-r-- 1 masa masa   5140801824  3月 21 14:03 firmfinBC08.csv
-rw-rw-r-- 1 masa masa   5122375105  3月 21 14:02 firmfinBC09.csv
-rw-rw-r-- 1 masa masa   5119743638  3月 21 14:03 firmfinBC10.csv
-rw-rw-r-- 1 masa masa   5130426953  3月 21 14:11 firmfinBC11.csv
-rw-rw-r-- 1 masa masa   5127787733  3月 21 14:11 firmfinBC12.csv
-rw-rw-r-- 1 masa masa   5117004646  3月 21 14:11 firmfinBC13.csv
-rw-rw-r-- 1 masa masa   5111994733  3月 21 14:11 firmfinBC14.csv
-rw-rw-r-- 1 masa masa   5151261919  3月 21 14:11 firmfinBC15.csv
-rw-rw-r-- 1 masa masa   5041672611  3月 21 14:11 firmfinBC16.csv
-rw-rw-r-- 1 masa masa   5082681293  3月 21 14:11 firmfinBC17.csv
-rw-rw-r-- 1 masa masa   5005063589  3月 21 14:10 firmfinBC18.csv
-rw-rw-r-- 1 masa masa   5552883760  3月 21 14:11 firmfinBC19.csv
-rw-rw-r-- 1 masa masa   5117784891  3月 21 14:11 firmfinBC20.csv
-rw-rw-r-- 1 masa masa   4915310319  3月 21 14:19 firmfinBC21.csv
-rw-rw-r-- 1 masa masa   4754012008  3月 21 14:18 firmfinBC22.csv
-rw-rw-r-- 1 masa masa   5041641202  3月 21 14:19 firmfinBC23.csv
-rw-rw-r-- 1 masa masa   5036169356  3月 21 14:19 firmfinBC24.csv
-rw-rw-r-- 1 masa masa     66900703  3月 21 14:12 firmfinBC25.csv
```

```
$wc -l *.csv
 240143521 firmfinBC.csv
 10000001 firmfinBC01.csv
 10000001 firmfinBC02.csv
 10000001 firmfinBC03.csv
 10000001 firmfinBC04.csv
 10000001 firmfinBC05.csv
 10000001 firmfinBC06.csv
 10000001 firmfinBC07.csv
 10000001 firmfinBC08.csv
 10000001 firmfinBC09.csv
 10000001 firmfinBC10.csv
 10000001 firmfinBC11.csv
 10000001 firmfinBC12.csv
 10000001 firmfinBC13.csv
 10000001 firmfinBC14.csv
 10000001 firmfinBC15.csv
 10000001 firmfinBC16.csv
 10000001 firmfinBC17.csv
 10000001 firmfinBC18.csv
 10000001 firmfinBC19.csv
 10000001 firmfinBC20.csv
 10000001 firmfinBC21.csv
 10000001 firmfinBC22.csv
 10000001 firmfinBC23.csv
 10000001 firmfinBC24.csv
  143521 firmfinBC25.csv
 480287066 total
```

Un-consolidate Version

Double Parallelized Preprocessing (All Process for Making CSV File): Un-consolidate Version (`firmfinBU*.csv,firmfinBU.csv`)



Comparison of Double Parallelized Preprocessing Times on iMac Pro, Dell Precision T7910 Un-consolidate Version

```
# iMac Pro
aule$ cat starttime-p2.txt
aule$ cat endtime-p2.txt
処理時間: 4時間5分
```

```
# Dell
dori$ cat starttime-p2.txt
2019年  3月 21日 木曜日 20:55:47 JST
dori$ cat endtime-p2.txt
2019年  3月 21日 木曜日 22:16:04 JST
処理時間: 1時21間分
```

CSV File Informations by Unix Commands: Un-consolidate Data Sets

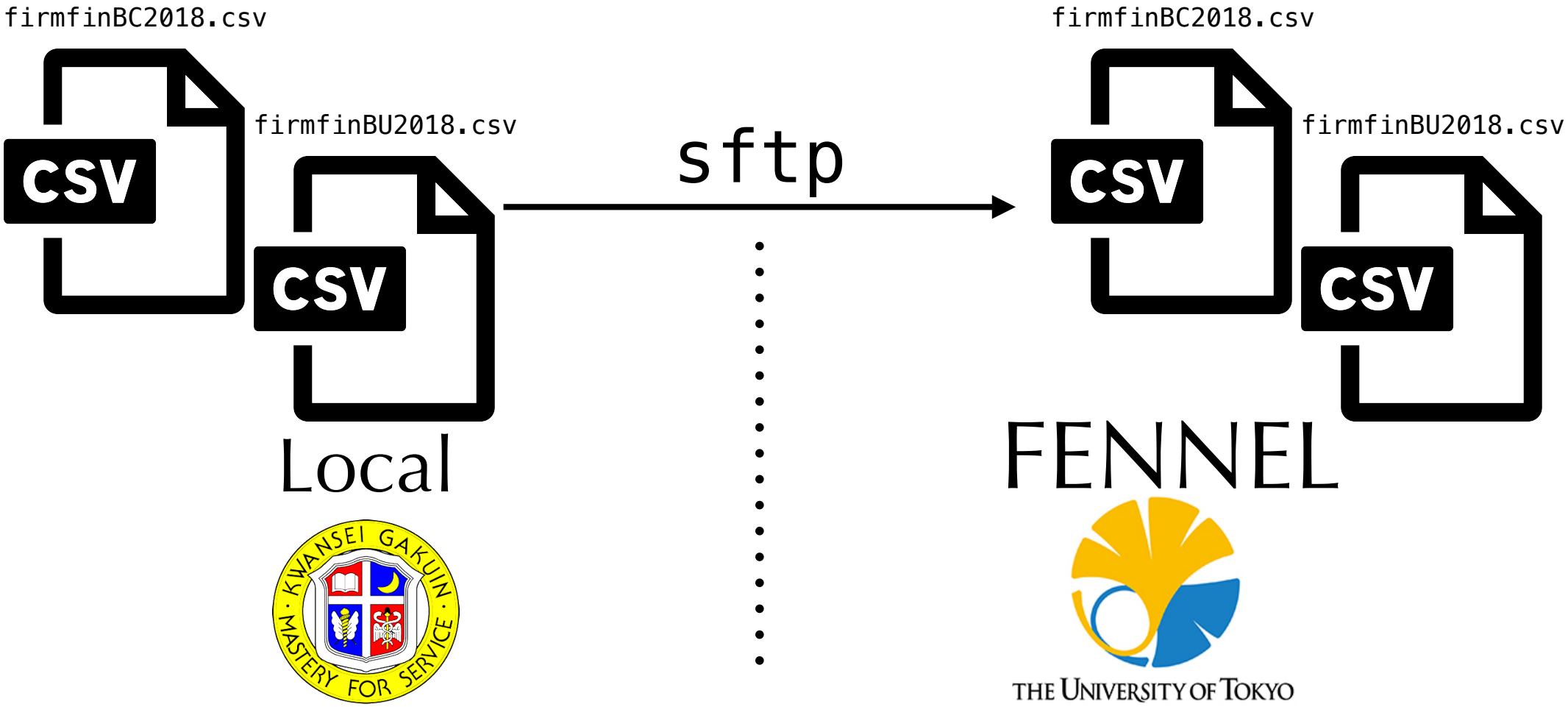
```
$ ls -l *.csv
-rw-rw-r-- 1 masa masa 124803858467  3月 21 22:16 firmfinBU.csv
-rw-rw-r-- 1 masa masa   5847448950  3月 21 21:51 firmfinBU01.csv
-rw-rw-r-- 1 masa masa   5587303337  3月 21 21:52 firmfinBU02.csv
-rw-rw-r-- 1 masa masa   5459850808  3月 21 21:52 firmfinBU03.csv
-rw-rw-r-- 1 masa masa   5365275527  3月 21 21:51 firmfinBU04.csv
-rw-rw-r-- 1 masa masa   5285725615  3月 21 21:51 firmfinBU05.csv
-rw-rw-r-- 1 masa masa   5227485687  3月 21 21:51 firmfinBU06.csv
-rw-rw-r-- 1 masa masa   5186860391  3月 21 21:51 firmfinBU07.csv
-rw-rw-r-- 1 masa masa   5155128183  3月 21 21:51 firmfinBU08.csv
-rw-rw-r-- 1 masa masa   5137839899  3月 21 21:51 firmfinBU09.csv
-rw-rw-r-- 1 masa masa   5136451434  3月 21 21:51 firmfinBU10.csv
-rw-rw-r-- 1 masa masa   5148892364  3月 21 22:00 firmfinBU11.csv
-rw-rw-r-- 1 masa masa   5144579167  3月 21 22:00 firmfinBU12.csv
-rw-rw-r-- 1 masa masa   5133760550  3月 21 22:00 firmfinBU13.csv
-rw-rw-r-- 1 masa masa   5128683345  3月 21 22:00 firmfinBU14.csv
-rw-rw-r-- 1 masa masa   5155868613  3月 21 22:00 firmfinBU15.csv
-rw-rw-r-- 1 masa masa   5077238035  3月 21 22:00 firmfinBU16.csv
-rw-rw-r-- 1 masa masa   5141892758  3月 21 22:00 firmfinBU17.csv
-rw-rw-r-- 1 masa masa   5012961567  3月 21 22:00 firmfinBU18.csv
-rw-rw-r-- 1 masa masa   5560456100  3月 21 22:00 firmfinBU19.csv
-rw-rw-r-- 1 masa masa   5108010761  3月 21 22:00 firmfinBU20.csv
-rw-rw-r-- 1 masa masa   4913515208  3月 21 22:07 firmfinBU21.csv
-rw-rw-r-- 1 masa masa   4753376288  3月 21 22:07 firmfinBU22.csv
-rw-rw-r-- 1 masa masa   5040806936  3月 21 22:07 firmfinBU23.csv
-rw-rw-r-- 1 masa masa   5034119546  3月 21 22:07 firmfinBU24.csv
-rw-rw-r-- 1 masa masa     60348542  3月 21 22:00 firmfinBU25.csv
```

```
$ wc -l *.csv
    240128071 firmfinBU.csv
    10000001 firmfinBU01.csv
    10000001 firmfinBU02.csv
    10000001 firmfinBU03.csv
    10000001 firmfinBU04.csv
    10000001 firmfinBU05.csv
    10000001 firmfinBU06.csv
    10000001 firmfinBU07.csv
    10000001 firmfinBU08.csv
    10000001 firmfinBU09.csv
    10000001 firmfinBU10.csv
    10000001 firmfinBU11.csv
    10000001 firmfinBU12.csv
    10000001 firmfinBU13.csv
    10000001 firmfinBU14.csv
    10000001 firmfinBU15.csv
    10000001 firmfinBU16.csv
    10000001 firmfinBU17.csv
    10000001 firmfinBU18.csv
    10000001 firmfinBU19.csv
    10000001 firmfinBU20.csv
    10000001 firmfinBU21.csv
    10000001 firmfinBU22.csv
    10000001 firmfinBU23.csv
    10000001 firmfinBU24.csv
      128071 firmfinBU25.csv
   480256166 合計
```

Rename Data Set Files

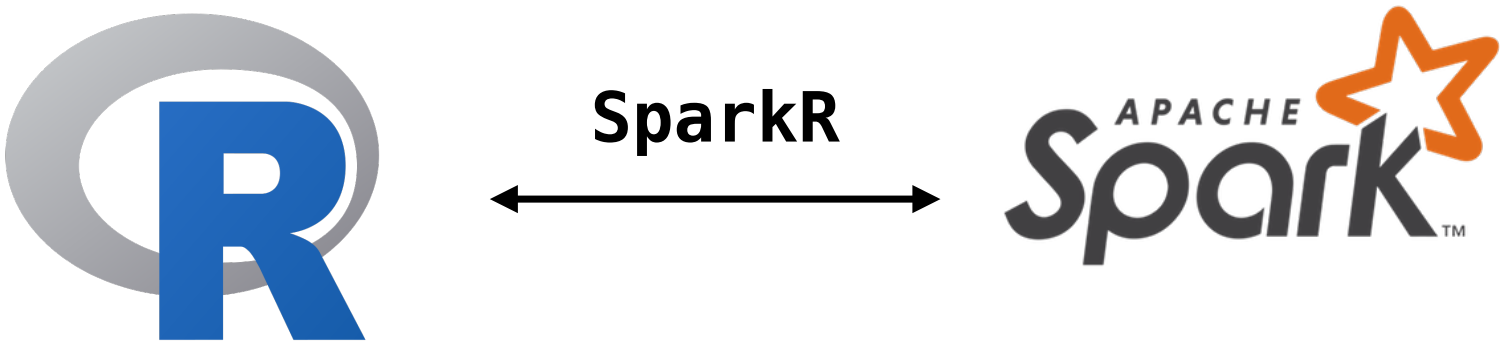
```
$ mv firmfinBC.csv firmfinBC2018.csv
$ mv firmfinBC.csv firmfinBU2018.csv
```

Transfer CSV File from Local to FENNEL



R + Spark + SparkR on **FENNEL**

Connect to Spark from R by SparkR on FENNEL




```
> Sys.setenv(SPARK_HOME = "/home/masa/spark/spark-2.2.0-bin-hadoop2.7")

> Sys.setenv(JAVA_HOME = "/usr/lib/jvm/java-8-oracle")

> library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib","")))

> sparkR.session(master = "local[*]", sparkConfig = list(spark.driver.memory = "12g"),
  spark.debug.maxToStringFields = "200")
```

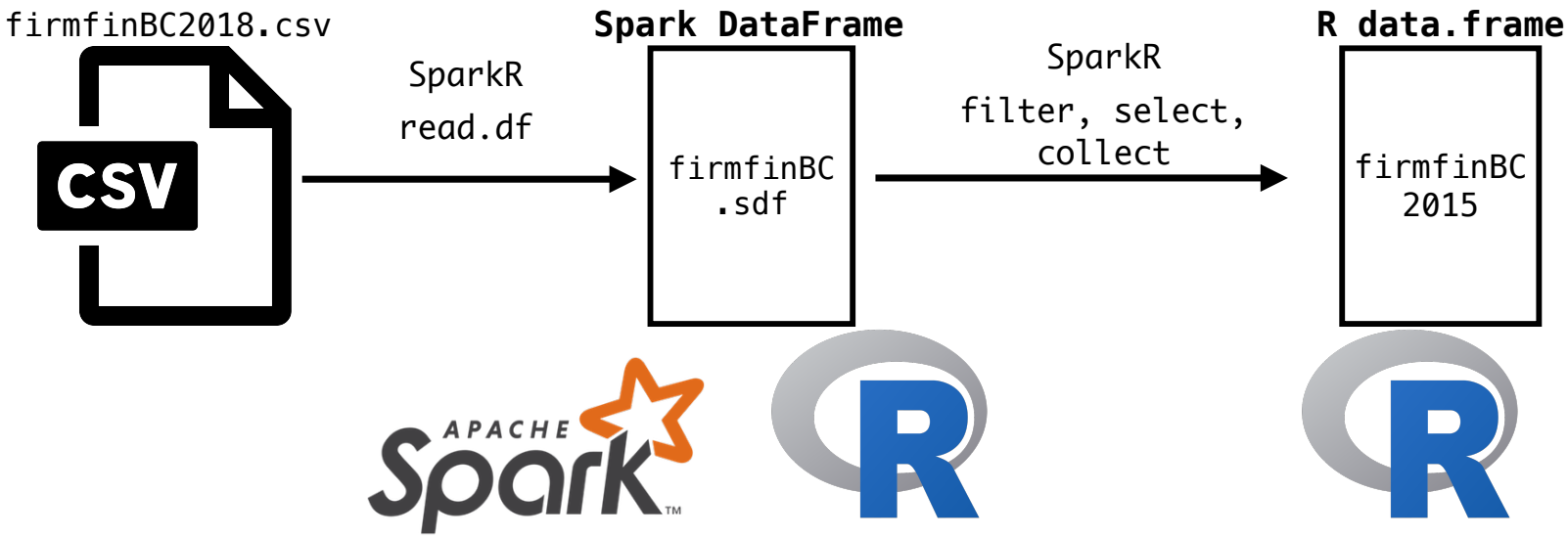
Data Wrangling with SparkR (`read.df`)

```
> firmfinBC.sdf <- read.df("../CSV/firmfinBC2018.csv",sourc="csv", header=TRUE,
  inferSchema = "true", na.strings = "NA")

> library(magrittr)

> firmfinBC2015 <- firmfinBC.sdf %>%
    select(firmfinBC.sdf$firmID, firmfinBC.sdf$country,
           firmfinBC.sdf$cons, firmfinBC.sdf$listed,
           firmfinBC.sdf$exchange, firmfinBC.sdf$InfoProv,
           firmfinBC.sdf$sales, firmfinBC.sdf$employees, firmfinBC.sdf$assets_total) %>%
    filter(firmfinBC.sdf$year == "2015" &
             firmfinBC.sdf$sales > 0 &
             firmfinBC.sdf$employees > 0 &
             firmfinBC.sdf$assets_total > 0 &
             firmfinBC.sdf$month == 12) %>% collect()

> colnames(firmfinBC2015)<-
  c("firmID","country","cons","listed","exchange","infoprov","sales","employees","assets.
  total")
```

Time for Data Wrangling with SparkR

```
> library(tictoc)

> tic()

> firmfinBC.sdf <- read.df("../CSV/firmfinBC2018.csv",sourc="csv", header=TRUE, inferSchema = "true", na.strings = "NA")

> library(magrittr)

> firmfinBC2015 <- firmfinBC.sdf %>%
   filter(firmfinBC.sdf$year == "2015" &
            firmfinBC.sdf$sales > 0 &
            firmfinBC.sdf$employees > 0 &
            firmfinBC.sdf$assets_total > 0 &
            firmfinBC.sdf$month == 12) %>%
   select(firmfinBC.sdf$firmID, firmfinBC.sdf$country,
          firmfinBC.sdf$cons, firmfinBC.sdf$listed,
          firmfinBC.sdf$exchange, firmfinBC.sdf$InfoProv,
          firmfinBC.sdf$sales, firmfinBC.sdf$employees, firmfinBC.sdf$assets_total) %>% collect()

> colnames(firmfinBC2015) <- c("firmID","country","cons","listed","exchange","infoprov","sales","employees","assets.total")

> toc()
 1765.32 sec elapsed (about 30 minutes)
```

Data Wrangling:
All Process for Loading CSV File to Spark and Transforming to R data frame on FENNEL



FENNEL

Next Stage

+α Setup (共同研究者持ち込み環境) / Hardware / Software

- **Hardware: Dell PowerEdge R740**
  - CPU: **Intel® Xeon® Bronze 3104**
  - Main Memory: **128 GB RDIMM**
  - Storage: **HDD 600GB**
  - Network Card: **QLogic FastLinQ 41164**
  - GPU: **NVIDIA Tesla V100 32G Passive GPU**
- **Software:**
  - **OS:** CentOS Linux release 7.7.1908 (Core)
  - **RDBMS:** psql (PostgreSQL) 10.10 + **PG-Strom**
  - **R Package: RPostgreSQL**

PG-Strom

https://heterodb.github.io/pg-strom/ja/

- PostgreSQL (RDBMS) の拡張モジュール
- GPL(GNU Public License) v2 に基づいて公開・配布されている**オープンソースソフトウェア**
- データベースを操作する標準命令 "SQL" の命令から, GPUプログラムを生成し、GPU上で非同期かつ並列に実行する

データベース

ID	NAME	POINT
1	鈴木	90
2	田中	70
3	山田	50

SQL命令

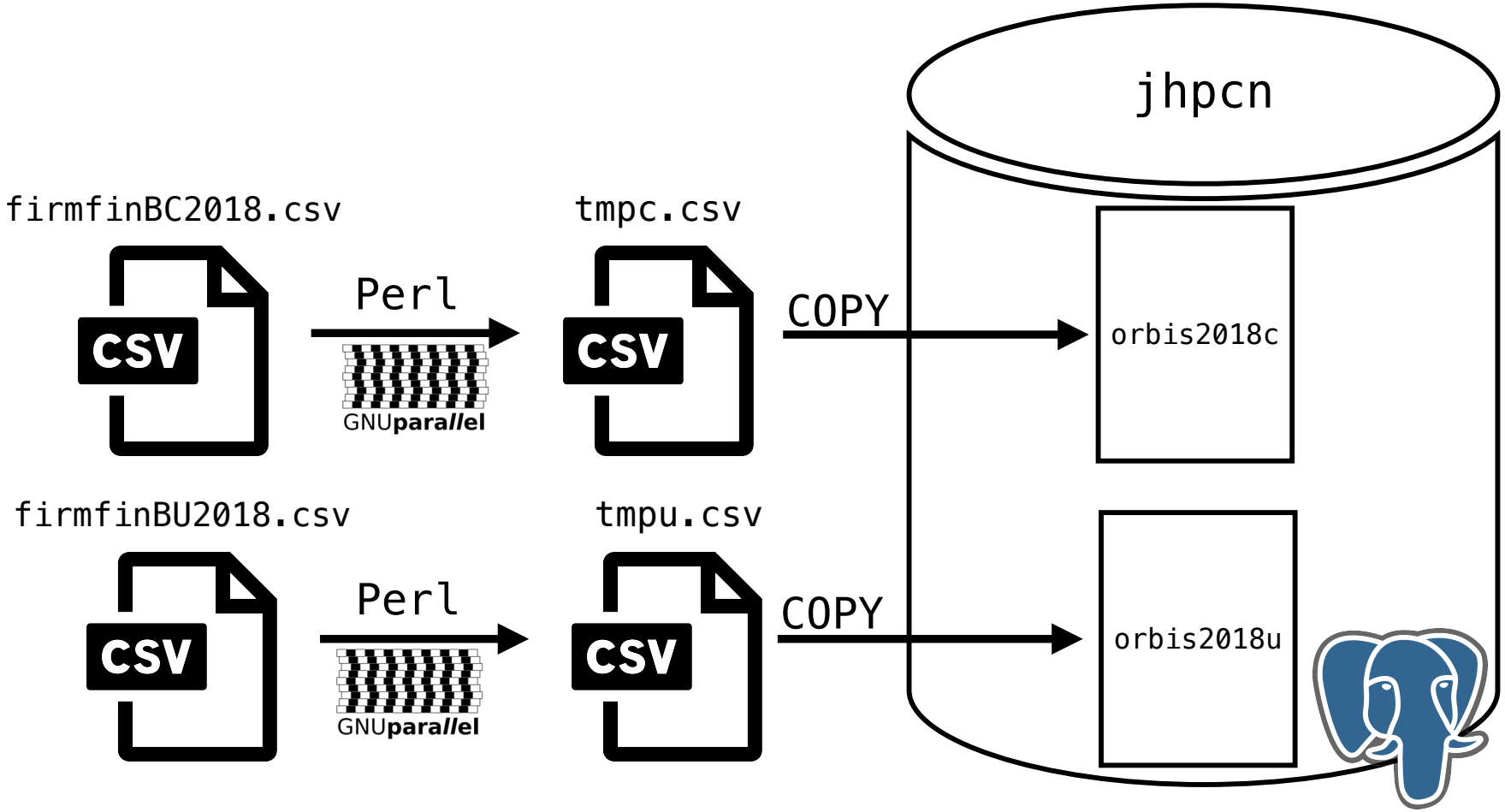
SELECT NAME FROM DB WHERE POINT > 60

通常はCPUで処理

負荷がかかると
GPUで高速化

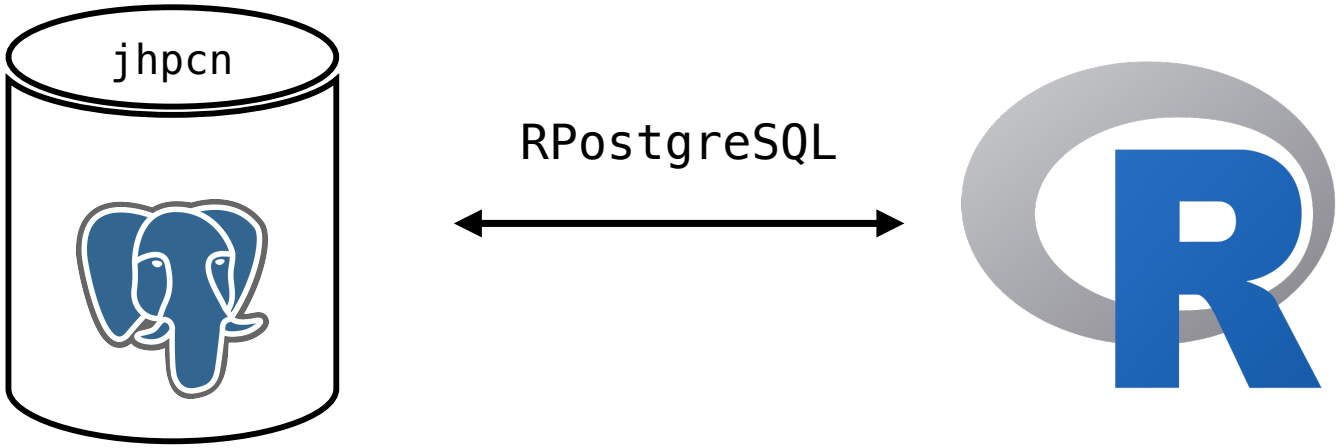

PostgreSQL
PG-Strom
GPU

Create DB `jhpcn` and Tables `orbis2018c, orbis2018u`



Dell PowerEdge R740

Connect to PostgreSQL Server from R by RPostgreSQL Package

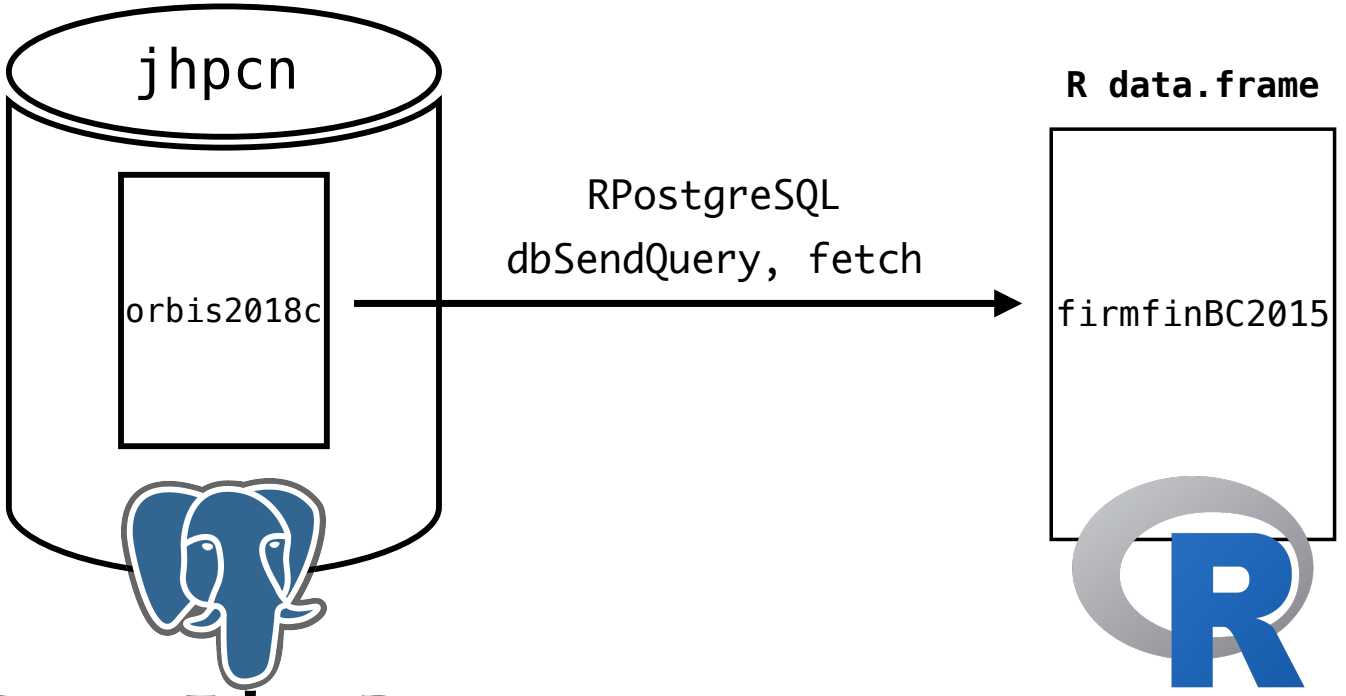


Dell PowerEdge R740

FENNEL

```
> #install.packages('RPostgreSQL')
> library(RPostgreSQL)
> drv <- dbDriver("PostgreSQL")
> con <- dbConnect(drv, host="133.11.235.6", port=5432, user= "masa", password="*********", dbname="jhpcn")
> dbListTables(con)  # テーブル一覧取得
```

Data Wrangling with PostgreSQL + R + RPostgreSQL



Dell PowerEdge R740

FENNEL

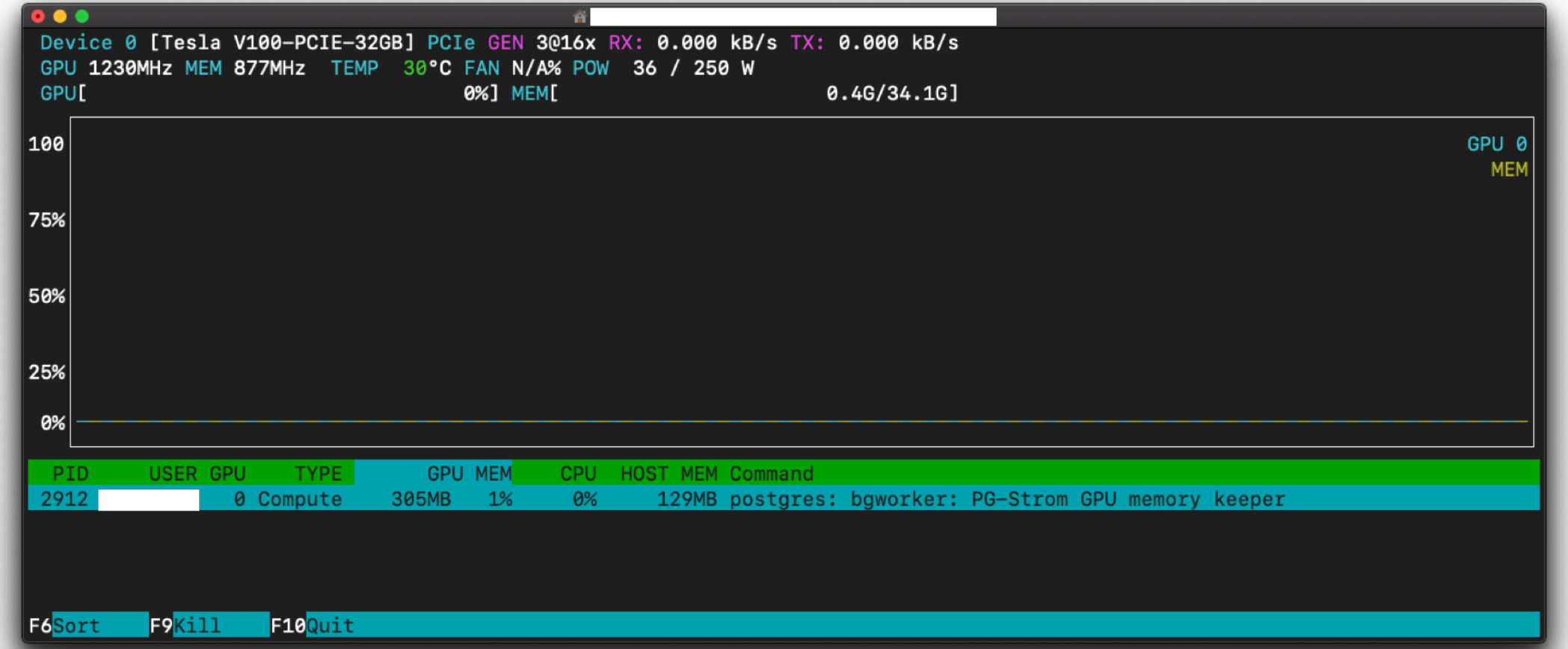
```
> sql <- "select firmID, country, cons, listed, exchange, InfoProv, sales, employees, assets_total
  from orbis2018c
  where year = 2015 and sales > 0 and employees > 0 and assets_total > 0 and month = 12"

> rs <- dbSendQuery(con, sql)

> firmfinBC2015 <- fetch(rs, n=-1)
```

ちなみに...

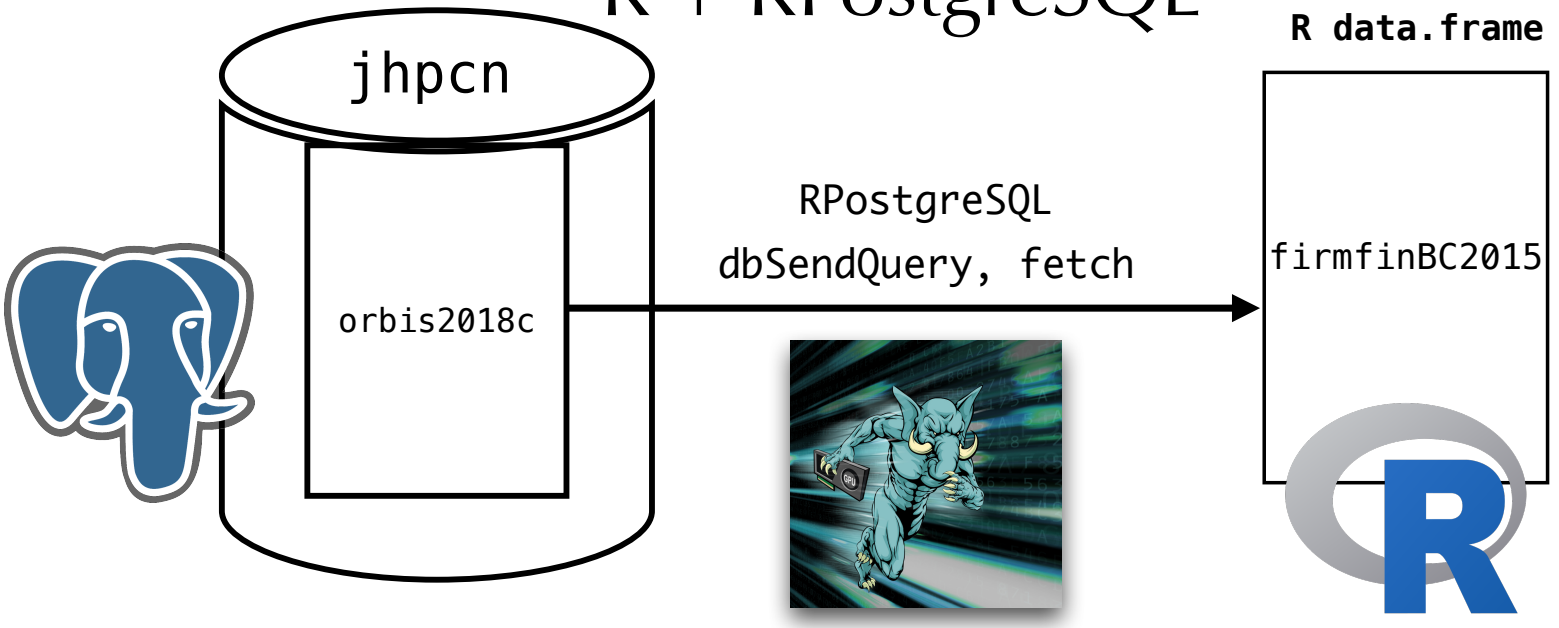
- 同一マシンでGPGPUを利用しない(CPUのみ利用)の結果は...

```
> library(tictoc)

> tic()

> sql <- "select firmID, country, cons, listed, exchange, InfoProv,
  sales, employees, assets_total from orbis2018c where year = 2015 and
  sales > 0 and employees > 0 and assets_total > 0 and month = 12"

> rs <- dbSendQuery(con, sql)

> firmfinBC2015 <- fetch(rs, n=-1)

> toc()

640.057 sec elapsed # (10分40秒)
```

`nvtop` Command: Calm!



Data Wrangling with PostgreSQL + PG-Strom + R + RPostgreSQL

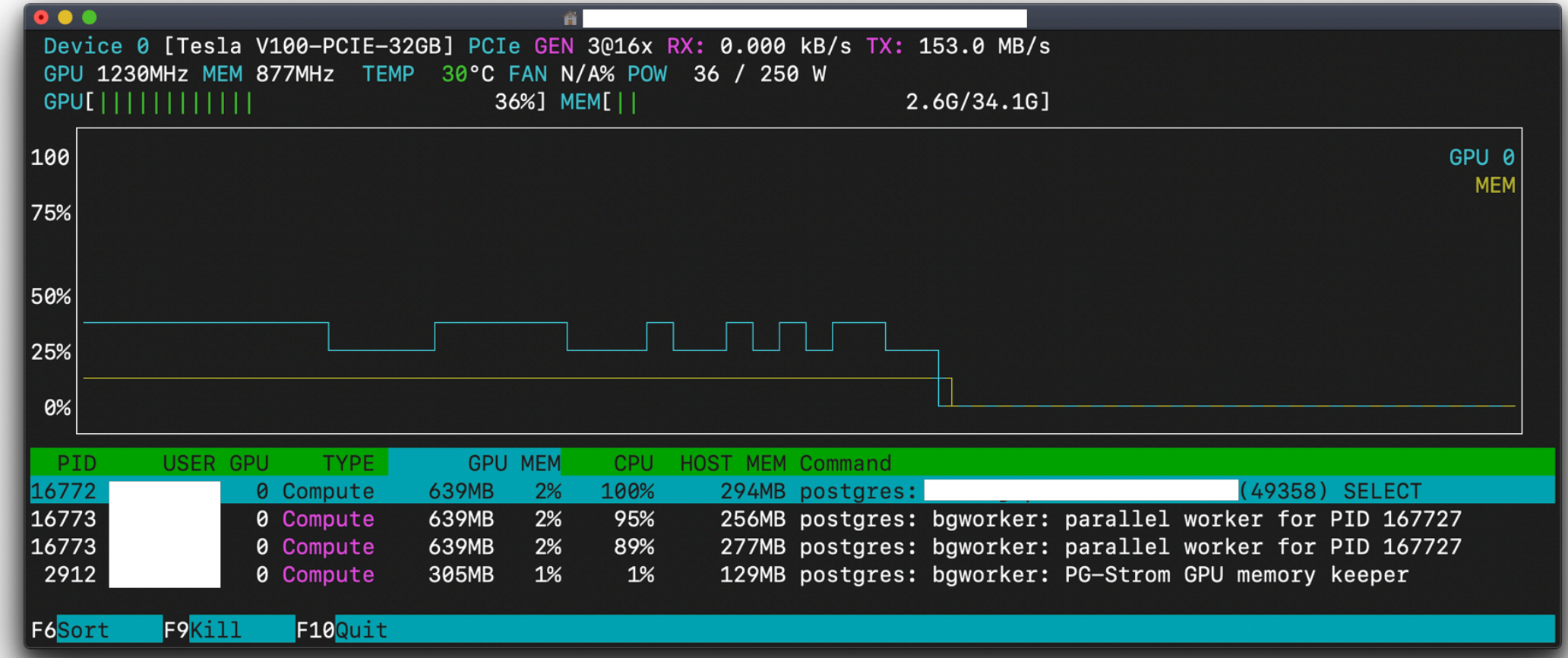


Dell PowerEdge R740

FENNEL

```
> sql <- "select
  firmID, country, cons, listed, exchange, InfoProv, sales, employees, assets_total
  from orbis2018c where year = 2015 and sales > 0 and employees > 0 and
  assets_total > 0 and month = 12"

> rs <- dbSendQuery(con, sql)

> firmfinBC2015 <- fetch(rs, n=-1)
```

`nvtop` Command: GPU works!!



Time of Data Wrangling with PostgreSQL + PG-Strom + R + RPostgreSQL

```
> library(tictoc)

> tic()

> sql <- "select firmID, country, cons, listed, exchange, InfoProv,
  sales, employees, assets_total from orbis2018c
  where year = 2015 and sales > 0 and employees > 0 and assets_total >
  0 and month = 12"

> rs <- dbSendQuery(con, sql)

> firmfinBC2015 <- fetch(rs, n=-1)

> toc()
  86.098 sec elapsed (約1分半!)
```

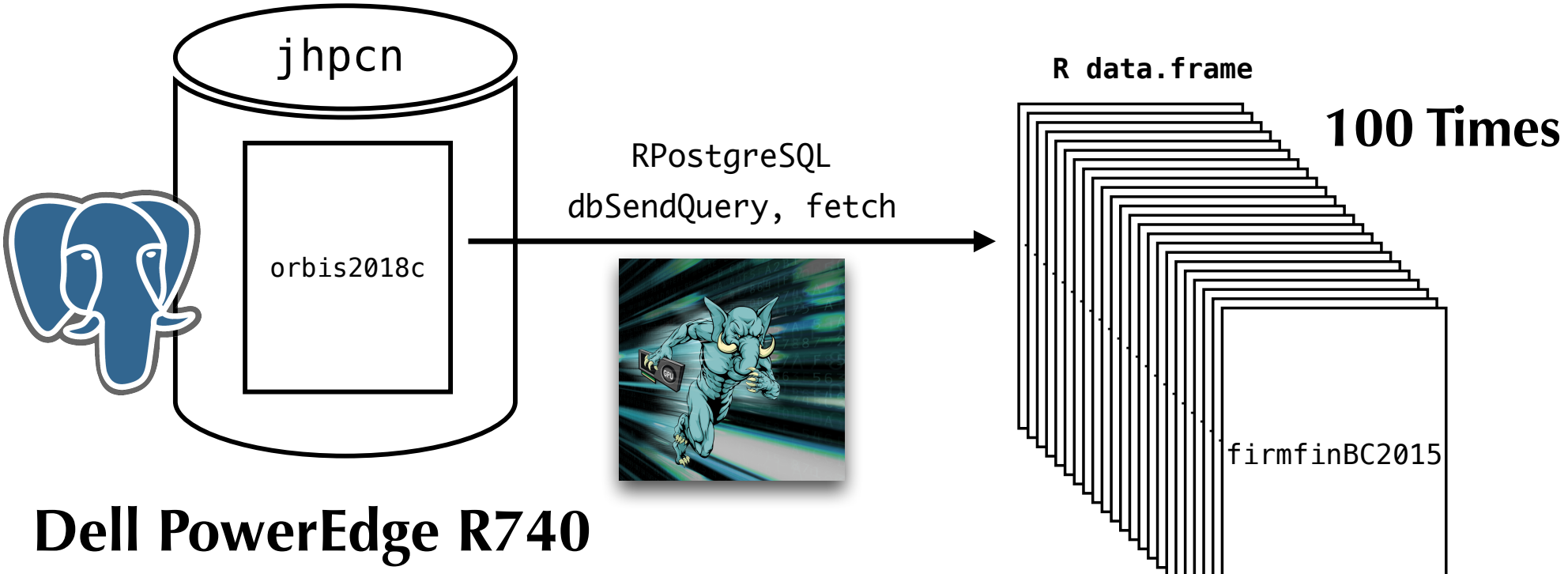
Object `firmfinBC2015` by PG-Strom

```
> head(firmfinBC2015)
                                                                 firmid country cons
1       FIVE ELEMENTS FILMS MEDIA PRODUCTIONS GMBH AT9090196104 Austria   U1
2 ING. SIEGFRIED STICHAUNER KUNSTSTOFFTECHNIK GMBH AT9090196698 Austria   U1
3                          MS HOTELBETRIEBS GMBH AT9090183182 Austria   U1
4              ERJ ELEKTROMECHANIK & TRAFOBAU GMBH AT9090183295 Austria   U1
5                                GATOM HANDELS GMBH AT9090183614 Austria   U1
6                                     SANCRET GMBH AT9090194586 Austria   U1
    listed exchange             infoprov sales employees assets_total
1 Unlisted Unlisted Creditreform Austria   305         2          190
2 Unlisted Unlisted Creditreform Austria  1524         4          599
3 Unlisted Unlisted Creditreform Austria  2286        45         4447
4 Unlisted Unlisted Creditreform Austria  2177        18          630
5 Unlisted Unlisted Creditreform Austria   109         3           64
6 Unlisted Unlisted Creditreform Austria   218         5           24
```

Benchmark Function

```
> bm <- function(){

require(RPostgreSQL)

con <- dbConnect(PostgreSQL(), host="133.11.235.6", port=5432, user=
"masa", password="********", dbname="jhpcn")

sql <- "select firmID, country, sales, employees, assets_total from
orbis2019 where year = 2015 and sales > 0 and employees > 0 and
assets_total > 0 and month = 12"

rs <- dbSendQuery(con, sql)

firmfinBC2015 <- fetch(rs, n=-1)

dbDisconnect(con)

}
```

Execute 100 Times Benchmarks!

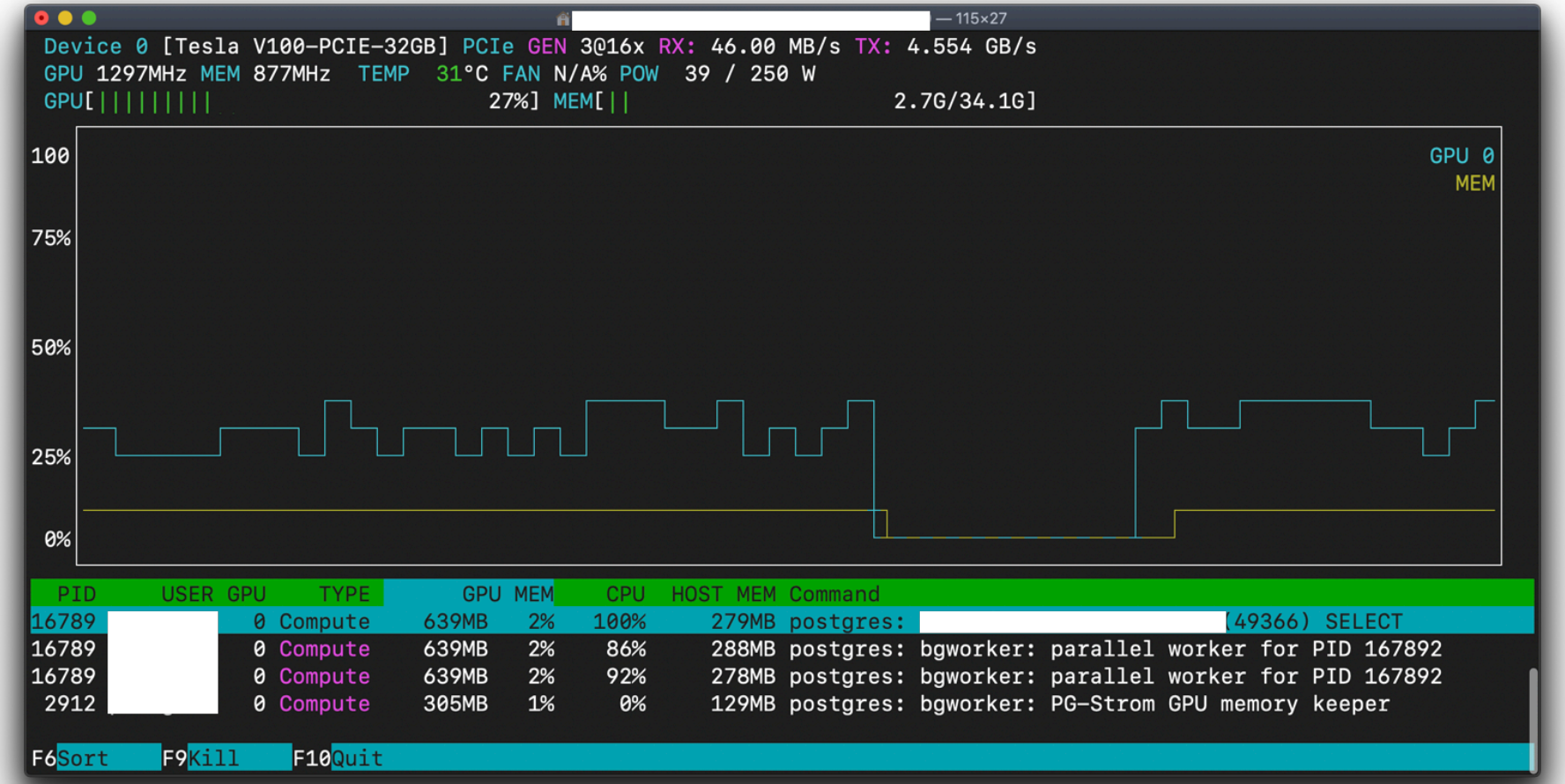


Dell PowerEdge R740

```
> queue <- matrix(nrow=100, ncol=3)
> for (i in 1:100){t1 <- proc.time(); bm(); t2 <- proc.time()-t1
    queue[i,1]<-t2[1]; queue[i,2]<-t2[2]; queue[i,3]<-t2[3]
}
```

FENNEL

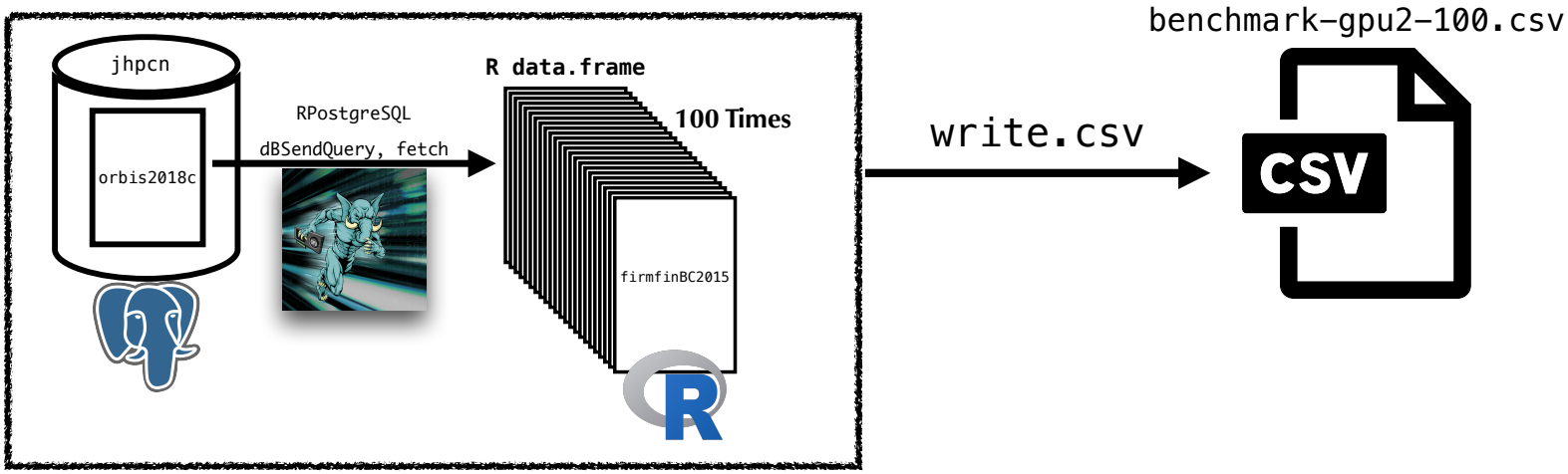
nvtop Command



Output Result of Benchmarks

```
> write.csv(queue, "benchmark-gpu2-100.csv")
```

Execute benchmark by `make`

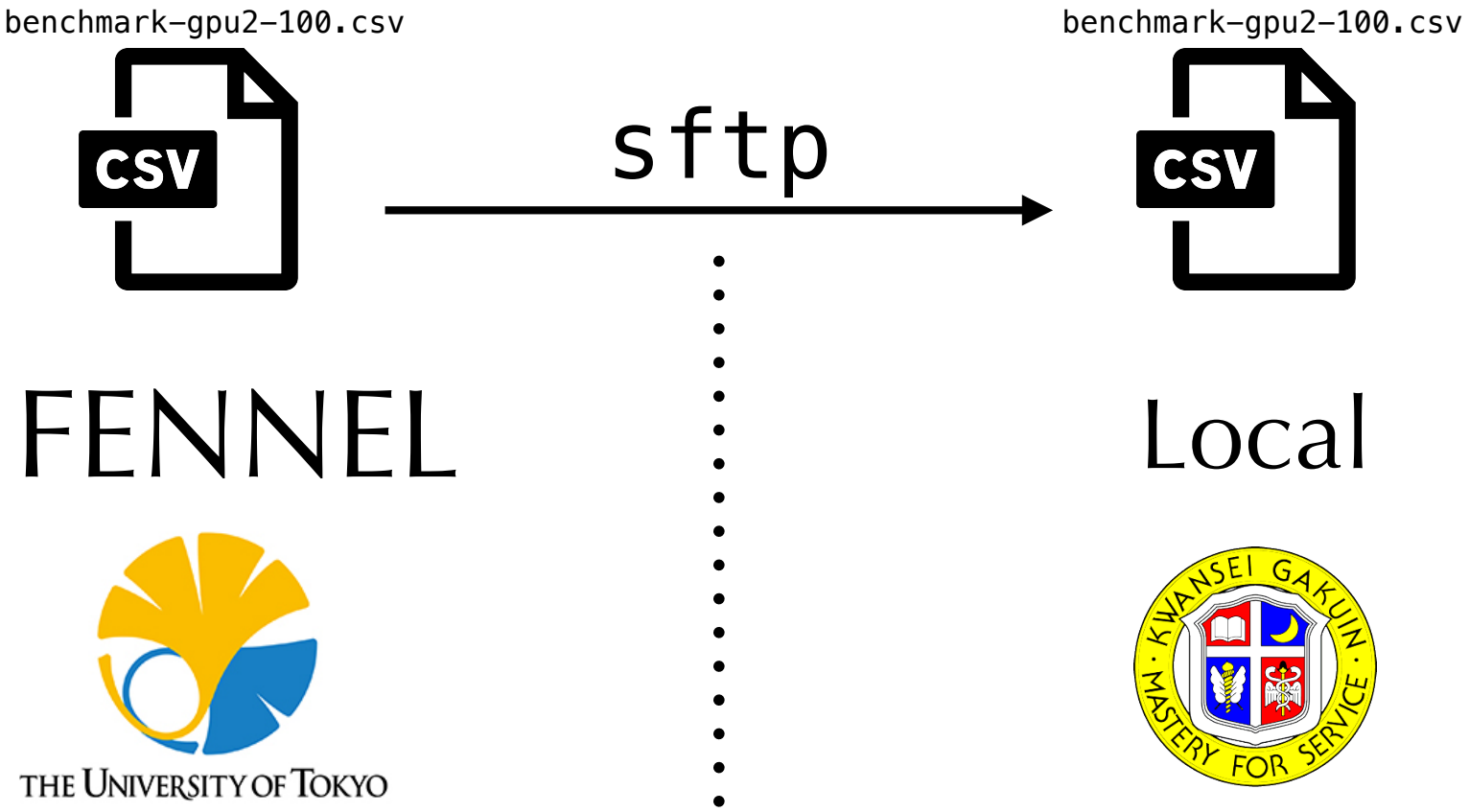



```
bm:
    date > start-bm.txt
    Rscript bm.R
    date > end-bm.txt
```

bm.R

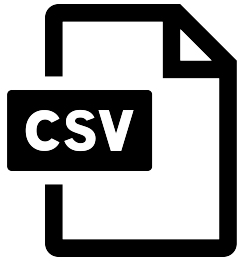
```
bm <- function(){
   require(RPostgreSQL)
   con <- dbConnect(PostgreSQL(), host="133.11.235.6", port=5432, user=
   "masa", password="***", dbname="jhpcn")
   sql <- "select firmID, country, cons, listed, exchange, InfoProv, sales,
   employees, assets_total from orbis2018c
   where year = 2015 and sales > 0 and employees > 0 and assets_total > 0 and
   month = 12"
   rs <- dbSendQuery(con, sql)
   firmfinBC2015 <- fetch(rs, n=-1)
   dbDisconnect(con)
}
queue <- matrix(nrow=100, ncol=3);

for (i in 1:100){
   t1 <- proc.time();bm();t2 <- proc.time()-t1
   queue[i,1]<-t2[1];queue[i,2]<-t2[2];queue[i,3]<-t2[3]
   cat("i=",i,",")
}
# 結果の出力

write.csv(queue, "benchmark-gpu2-100.csv")
```

Transfer CSV File from FENNEL to Local



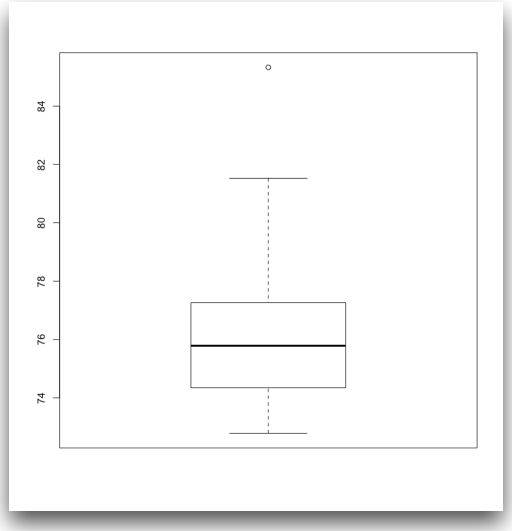
Summary and Visualize `firmfinBU2015U`

benchmark-gpu2-100.csv



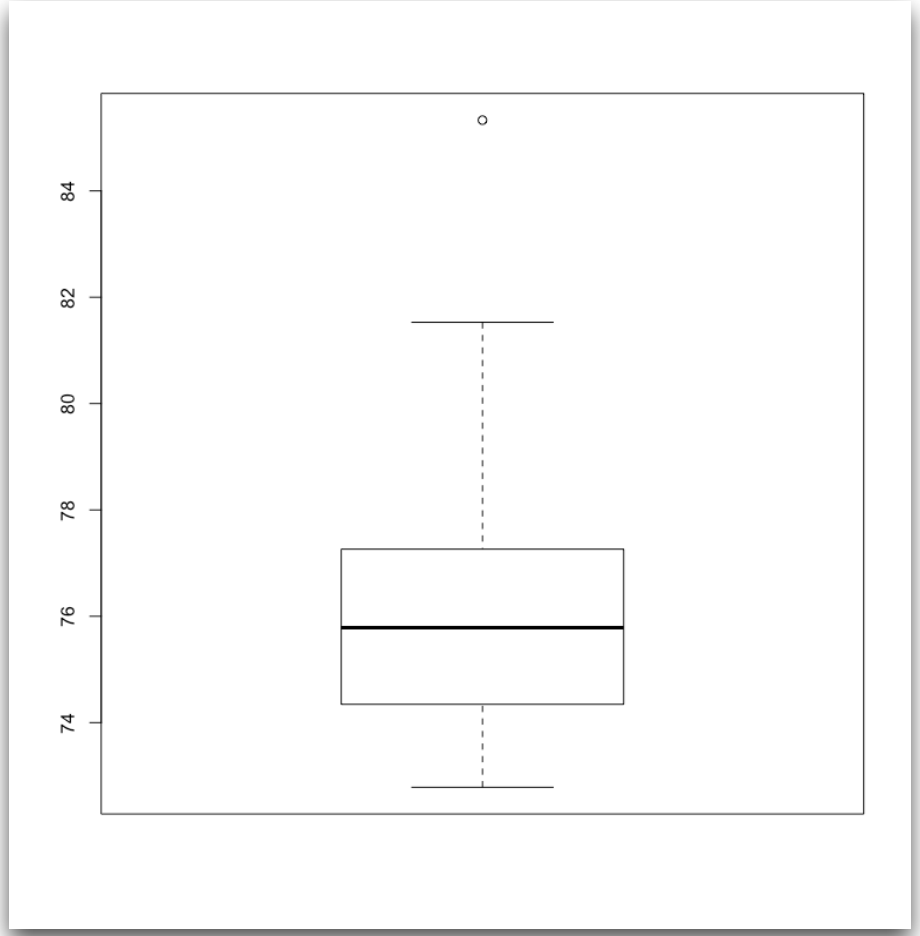
`read.csv`

`boxplot`



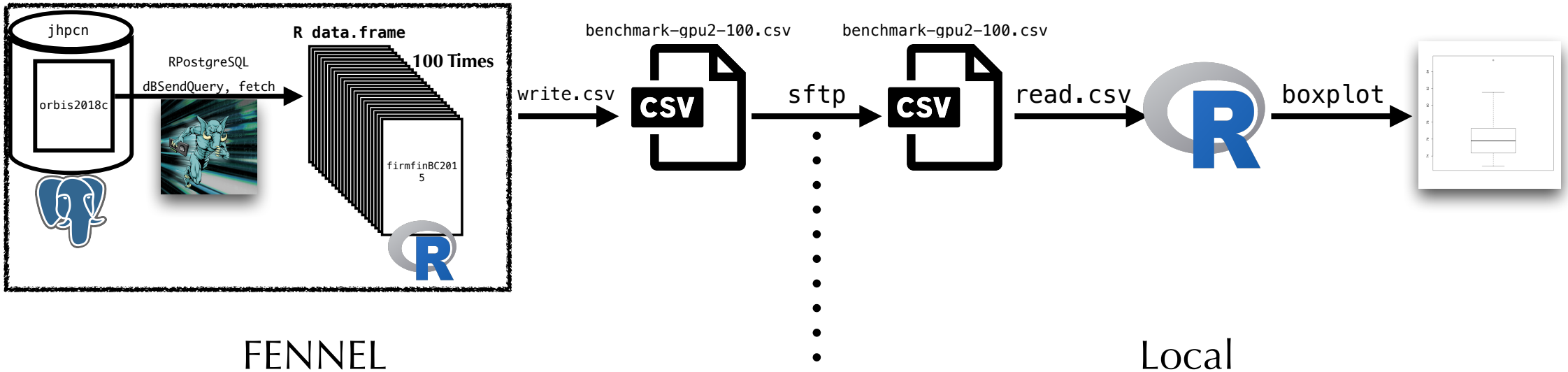
Local

Box Plot




```
> boxplot(bm100$V3)
```

Benchmark and Its Visualization

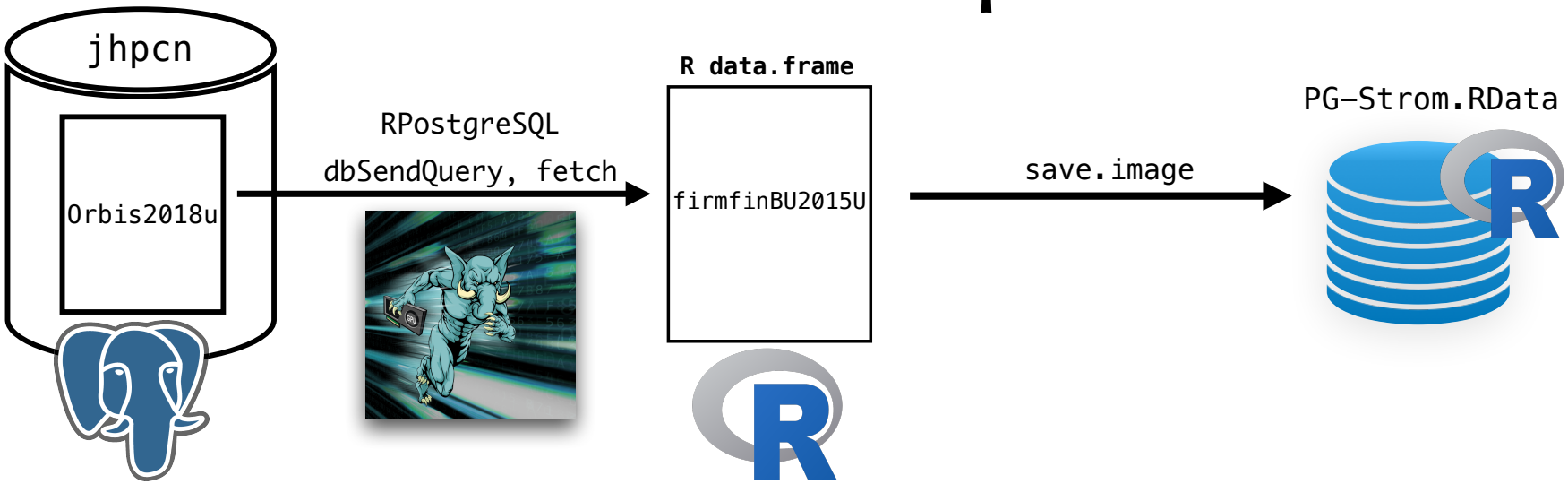


PG-Storm Data Wrangling from `orbis2018u` and Data Visualization

Data Wrangling 2015

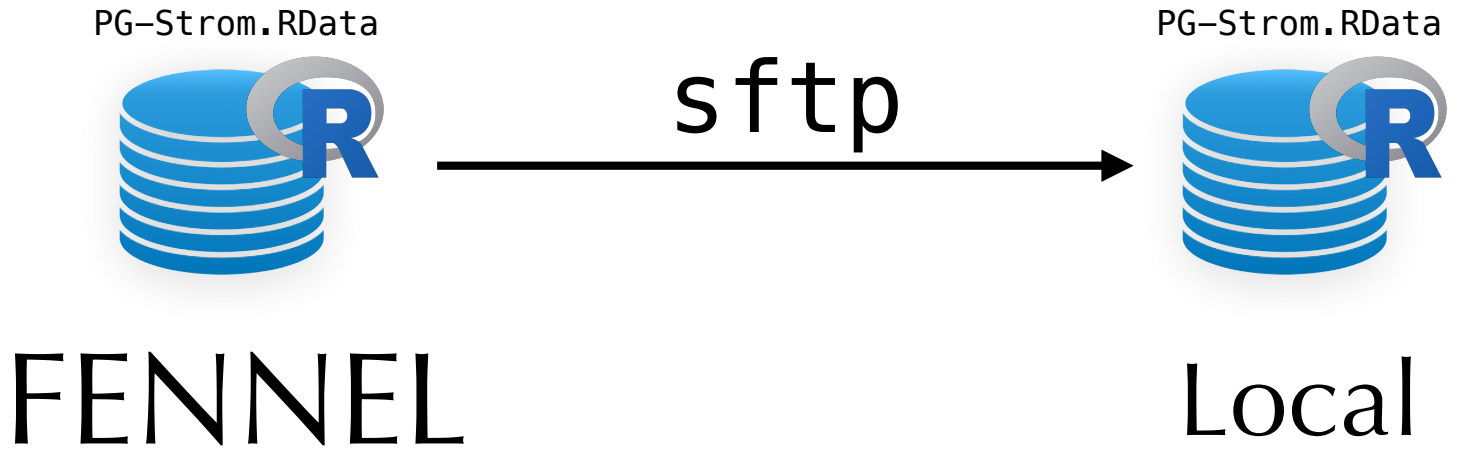
```
> tic()

> sql2 <- "select firmID, year, country, month, cons,
  listed, exchange, InfoProv, interest_paid, costs_employees,
  tax, PL_after_tax, PL_before_tax, assets_total from
  orbis2018u where year = 2015 and month = 12 and (cons =
  'U1' or cons = 'U2')"

> rs2 <- dbSendQuery(con, sql2)

> firmfinBU2015U <- fetch(rs2, n=-1)

> toc()
  176.755 sec elapsed (約3分)
```

Save Workspace




```
> save.image("PG-Storm.RData")
```

Transfer `RData` File from FENNEL to Local



Summary and Visualize `firmfinBU2015U`

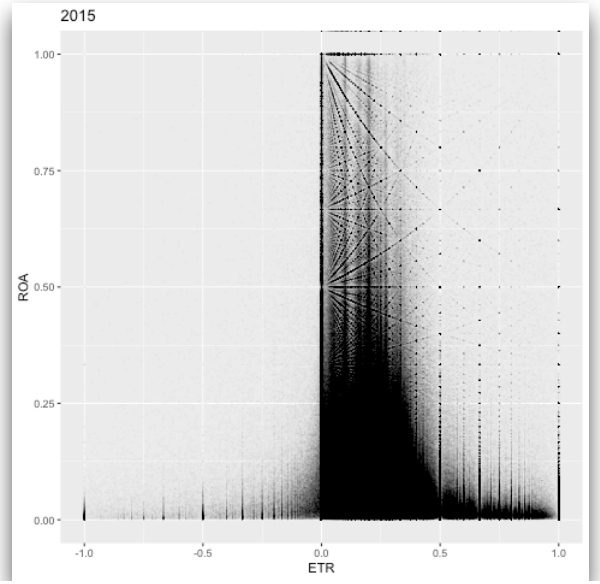
PG-Strom.RData

load

summary

```
   firmid              year       country              month
 Length:13687832    Min.   :2015   Length:13687832    Min.   :12
 Class :character   1st Qu.:2015   Class :character   1st Qu.:12
 Mode  :character   Median :2015   Mode  :character   Median :12
                    Mean   :2015                      Mean   :12
                    3rd Qu.:2015                      3rd Qu.:12
                    Max.   :2015                      Max.   :12

     cons               listed           exchange           infoprov
 Length:13687832    Length:13687832    Length:13687832    Length:13687832
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 interest_paid      costs_employees          tax             pl_after_tax
 Min.   : -45594    Min.   : -1414656   Min.   :-2288447   Min.   :-7.614e+10
 1st Qu.:      1    1st Qu.:       13   1st Qu.:       0   1st Qu.: 0.000e+00
 Median :      4    Median :       59   Median :       0   Median : 2.000e+00
 Mean   :    239    Mean   :      975   Mean   :      81   Mean   :-7.918e+03
 3rd Qu.:     18    3rd Qu.:      233   3rd Qu.:       5   3rd Qu.: 2.300e+01
 Max.   :4593225    Max.   :207705875   Max.   : 5217707   Max.   : 3.316e+07
 NA's   :10306934   NA's   :9097122     NA's   :5428452    NA's   :4418266
 pl_before_tax        assets_total
 Min.   :-7.614e+10   Min.   : -10628929
 1st Qu.: 0.000e+00   1st Qu.:        16
 Median : 3.000e+00   Median :       122
 Mean   :-7.762e+03   Mean   :      8040
 3rd Qu.: 2.900e+01   3rd Qu.:       680
 Max.   : 3.316e+07   Max.   :1468925328
 NA's   :4321555      NA's   :405095
```

ggplot



Local

firmfinBU2015U

```
> summary(firmfinBU2015U)
    firmid              year        country             month
 Length:13687832    Min.   :2015   Length:13687832    Min.   :12
 Class :character   1st Qu.:2015   Class :character   1st Qu.:12
 Mode  :character   Median :2015   Mode  :character   Median :12
                    Mean   :2015                      Mean   :12
                    3rd Qu.:2015                      3rd Qu.:12
                    Max.   :2015                      Max.   :12

     cons              listed            exchange          infoprov
 Length:13687832    Length:13687832    Length:13687832    Length:13687832
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 interest_paid      costs_employees          tax             pl_after_tax
 Min.   : -45594    Min.   : -1414656   Min.   :-2288447   Min.   :-7.614e+10
 1st Qu.:      1    1st Qu.:       13   1st Qu.:       0   1st Qu.: 0.000e+00
 Median :      4    Median :       59   Median :       0   Median : 2.000e+00
 Mean   :    239    Mean   :      975   Mean   :      81   Mean   :-7.918e+03
 3rd Qu.:     18    3rd Qu.:      233   3rd Qu.:       5   3rd Qu.: 2.300e+01
 Max.   :4593225    Max.   :207705875   Max.   : 5217707   Max.   : 3.316e+07
 NA's   :10306934   NA's   :9097122     NA's   :5428452    NA's   :4418266
 pl_before_tax         assets_total
 Min.   :-7.614e+10   Min.   : -10628929
 1st Qu.: 0.000e+00   1st Qu.:        16
 Median : 3.000e+00   Median :       122
 Mean   :-7.762e+03   Mean   :      8040
 3rd Qu.: 2.900e+01   3rd Qu.:       680
 Max.   : 3.316e+07   Max.   :1468925328
 NA's   :4321555      NA's   :405095
```

Make Objects for Scatter Plot of ROA-ETR

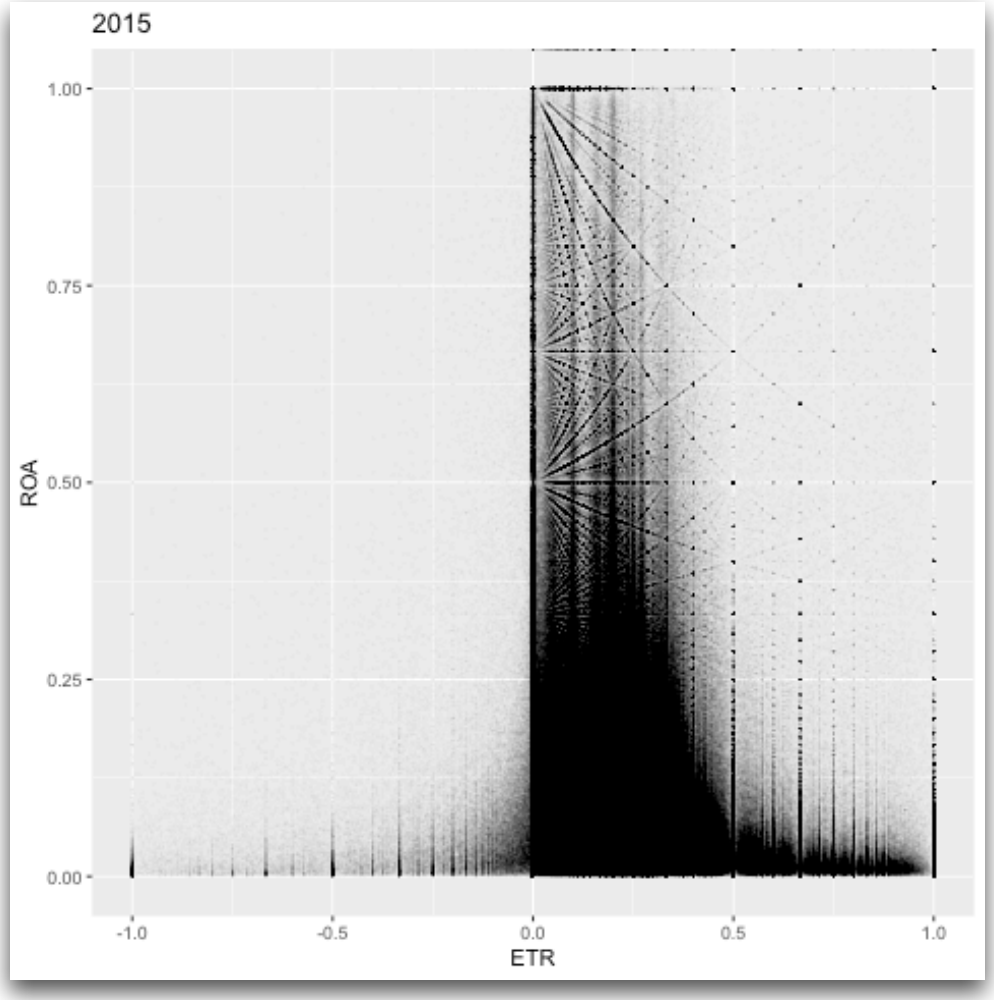
```
> firmfin.ROA.ETR.2015.firm.summary
 <- firmfinBU2015U %>%
   filter(!is.na(tax)) %>%
   filter(!is.na(pl_before_tax)) %>%
   filter(!is.na(assets_total))%>%
   filter(pl_before_tax > 0) %>%
   group_by(firmid) %>%
   summarize(ROA = pl_before_tax/assets_total,
             ETR = tax/pl_before_tax)
```

注) 総資産利益率 (Return On Asset: ROA), 実効税率 (Effective Tax Rate: ETR)

Scatter Plot of ROA-ETR

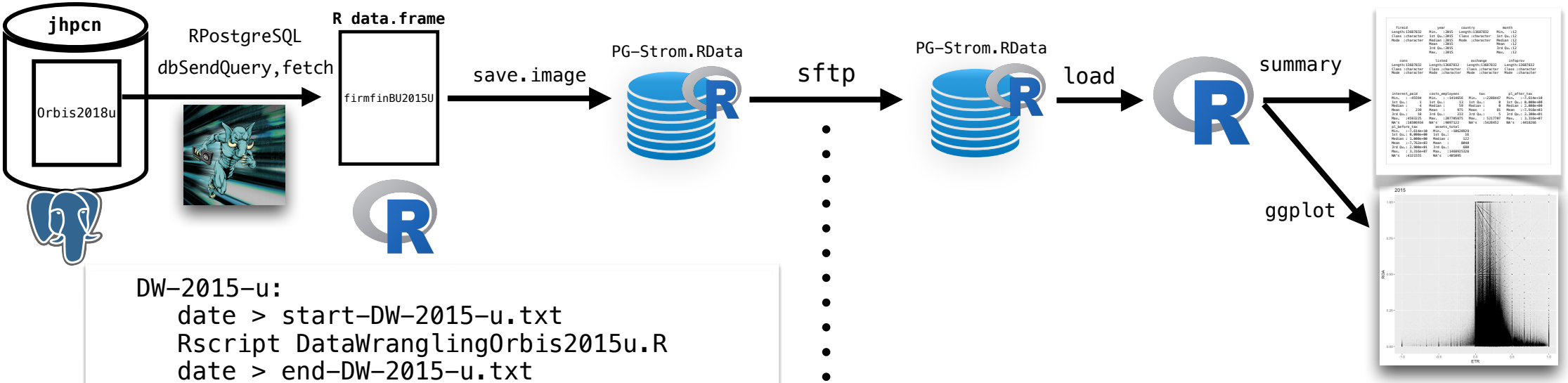
```
> p <- firmfin.ROA.ETR.2015.firm.summary %>%
   ggplot(aes(ETR, ROA)) +
   geom_point(size = 0.01, alpha = 0.01) +
   xlim(-1, 1) + ylim(0, 1) + labs(title = 2015)
> png("ROA-ETR-2016.png")
> print(p)
> dev.off()
```

Scatter Plot of ROA-ETR 2015 [-1,1]x[0,1]



Saka, C., T. Oshika, and M. Jimichi (2019) Visualization of tax avoidance and tax rate convergence: Exploratory analysis of world-scale accounting data, *Meditari Accountancy Research*, Vol. 27 No. 5, pp. 695–724, Emerald Publishing Limited.

Total Process




```
DW-2015-u:
   date > start-DW-2015-u.txt
   Rscript DataWranglingOrbis2015u.R
   date > end-DW-2015-u.txt
```

FENNEL

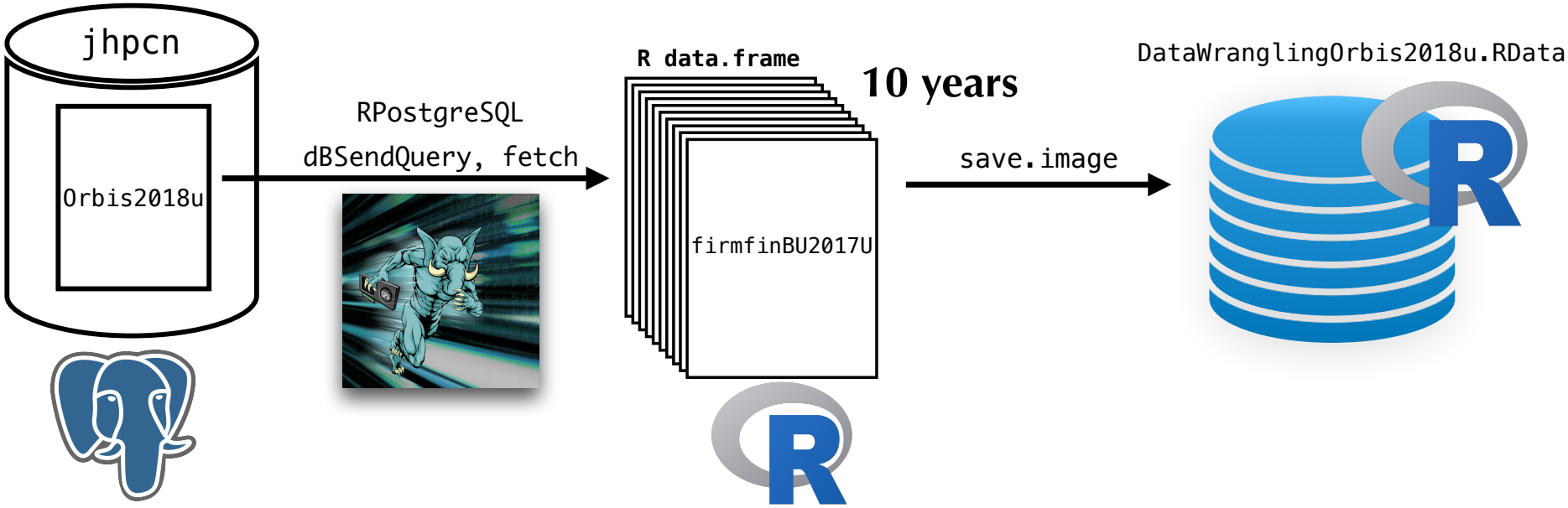
THE UNIVERSITY OF TOKYO

Local

KWANSEI GAKUIN · MASTERY FOR SERVICE ·

Data Wrangling 2008-2017 from `orbis2018u` by `make`

Data Wrangling for 10 Years with PG-Strom by `make`



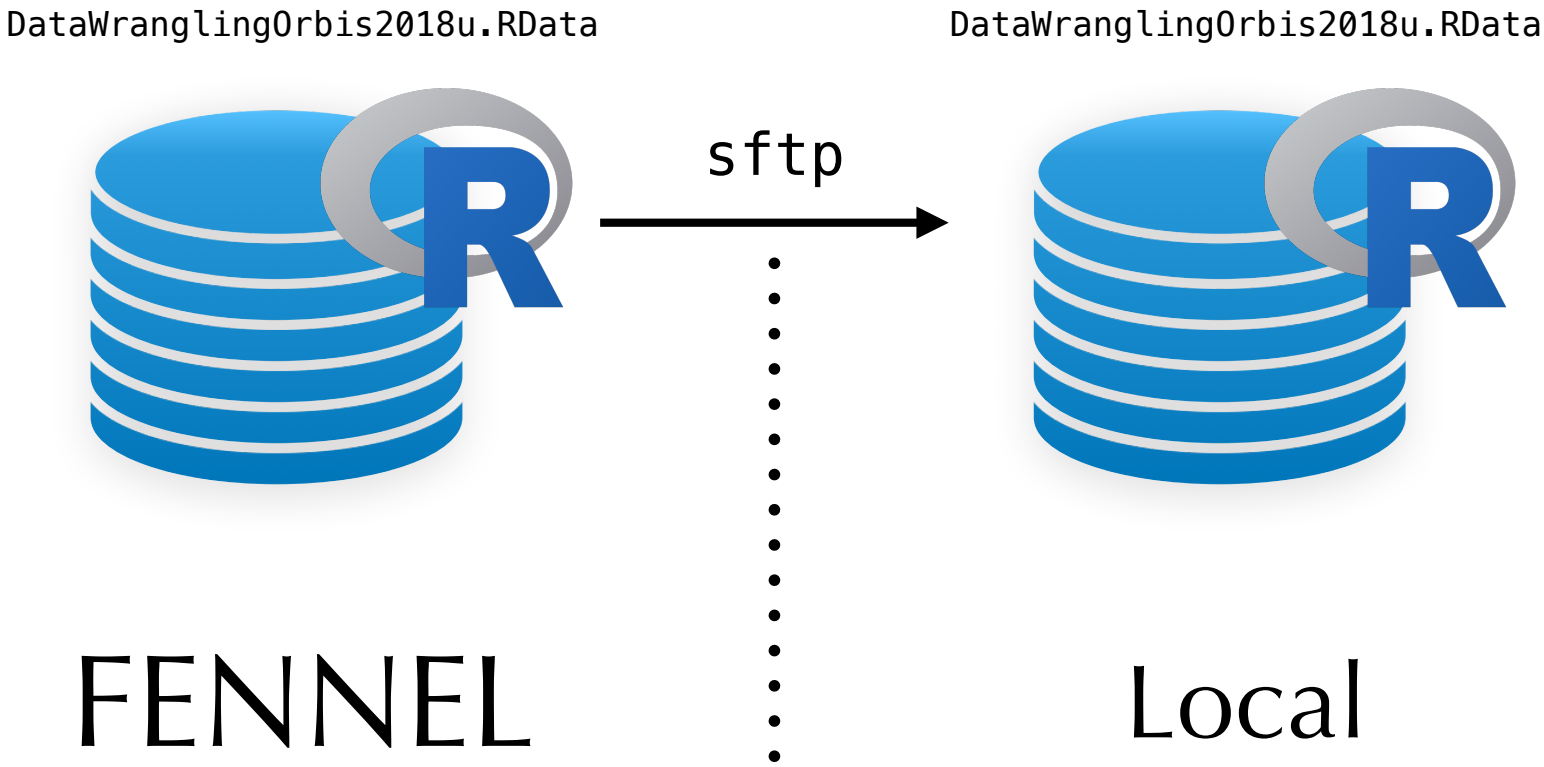

```
DW-u:
   date > start-DW-u.txt
   Rscript DataWranglingOrbis2018u.R
   date > end-DW-u.txt
```

DataWranglingOrbis2018u.R

```
library(RPostgreSQL)
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, host = "133.11.235.6", port = 5432, user= "masa", password =
"*********", dbname = "jhpcn")
# Data Wrangling from orbis2018u
# 2008
sql2008 <- "select firmID, year, country, month, cons, listed, exchange, InfoProv,
interest_paid, costs_employees, tax, PL_after_tax, PL_before_tax, assets_total from
orbis2018u where year = 2008 and month = 12 and (cons = 'U1' or cons = 'U2')"
rs2008 <- dbSendQuery(con, sql2008)
firmfinBU2008U <- fetch(rs2008, n = -1)
:
:
:
# 2017
sql2017 <- "select firmID, year, country, month, cons, listed, exchange, InfoProv,
interest_paid, costs_employees, tax, PL_after_tax, PL_before_tax, assets_total from
orbis2018u where year = 2017 and month = 12 and (cons = 'U1' or cons = 'U2')"
rs2017 <- dbSendQuery(con, sql2017)
firmfinBU2017U <- fetch(rs2017, n = -1)
# dump
save.image(file = "DataWranglingOrbis2018u.RData")
```

Scatter Plots of ROA-ETR
2008-2017

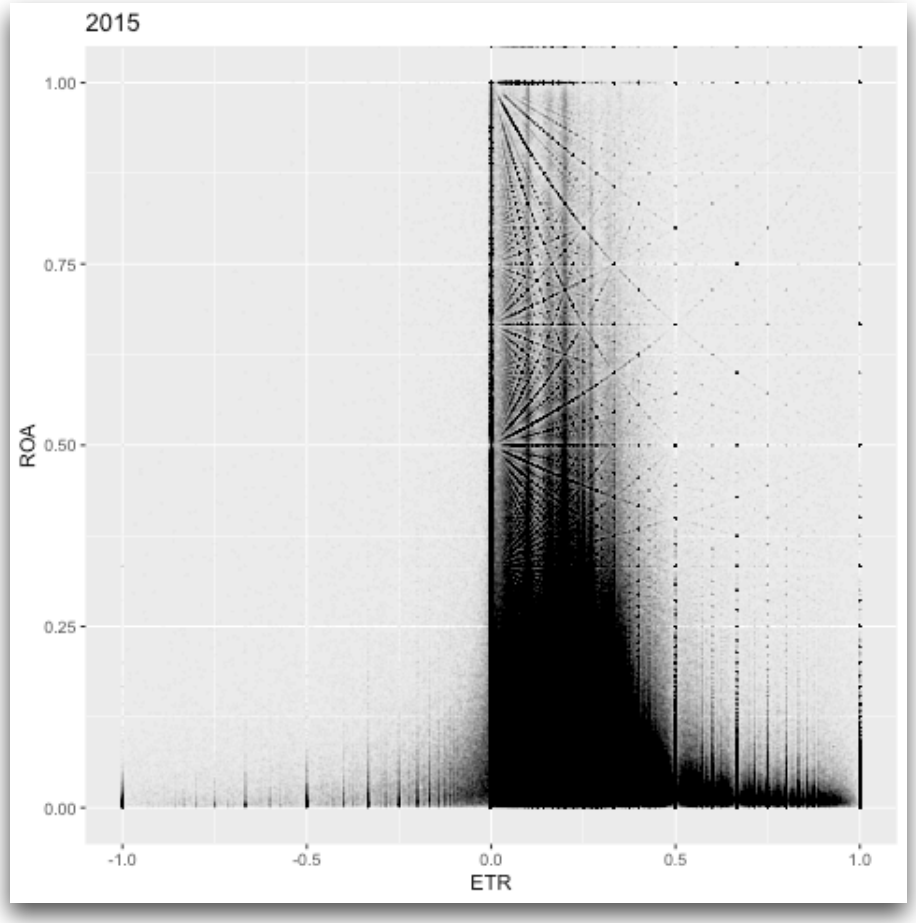
Transfer `RData` File from FENNEL to Local



Plot Function

```
> plot.ROA.ETR <- function(df)
  {
    require(ggplot2)
    require(dplyr)
    p <- df %>%
      filter(!is.na(tax)) %>%
      filter(!is.na(pl_before_tax)) %>%
      filter(!is.na(assets_total)) %>%
      filter(pl_before_tax > 0) %>%
      group_by(firmid) %>%
      summarize( ROA = pl_before_tax/assets_total,
                 ETR = tax/pl_before_tax) %>%
      ggplot(aes(ETR, ROA)) +
      geom_point(size = 0.01, alpha = 0.01) +
      xlim(-1, 1) + ylim(0, 1) + labs(title = year)
    print(p)
  }
```

Scatter Plot of ROA-ETR 2015 [-1,1]x[0,1]



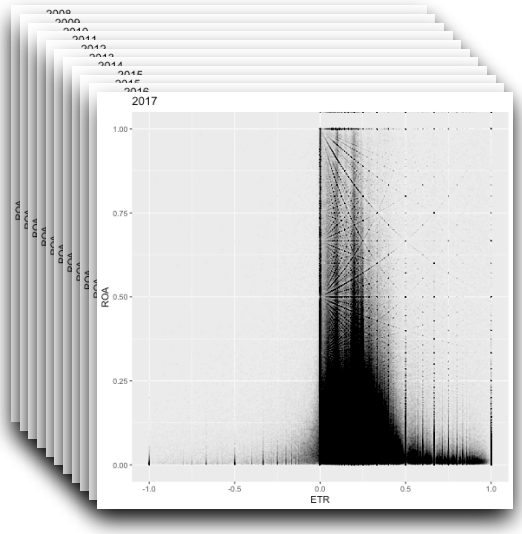

```
> plot.ROA.ETR(firmfinBU2015U)
```

Sequentially Make PNG Files

DataWranglingOrbis2018u.RData

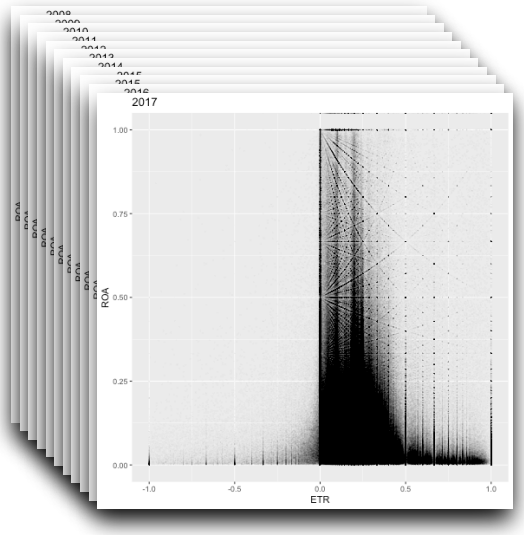
load

mkpng

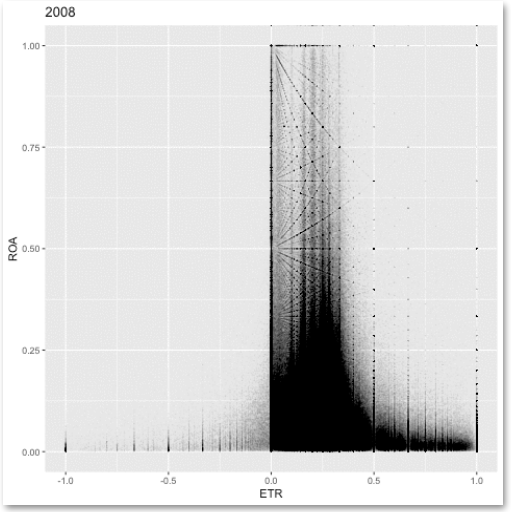


Local

Animate!

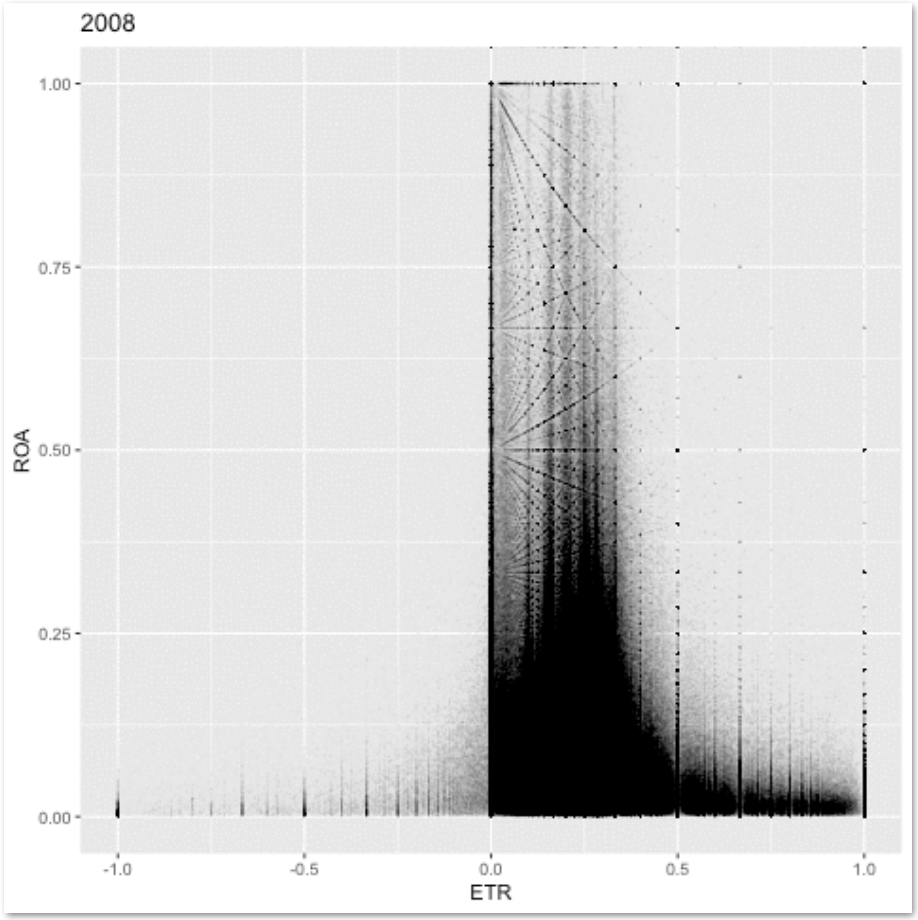


convert



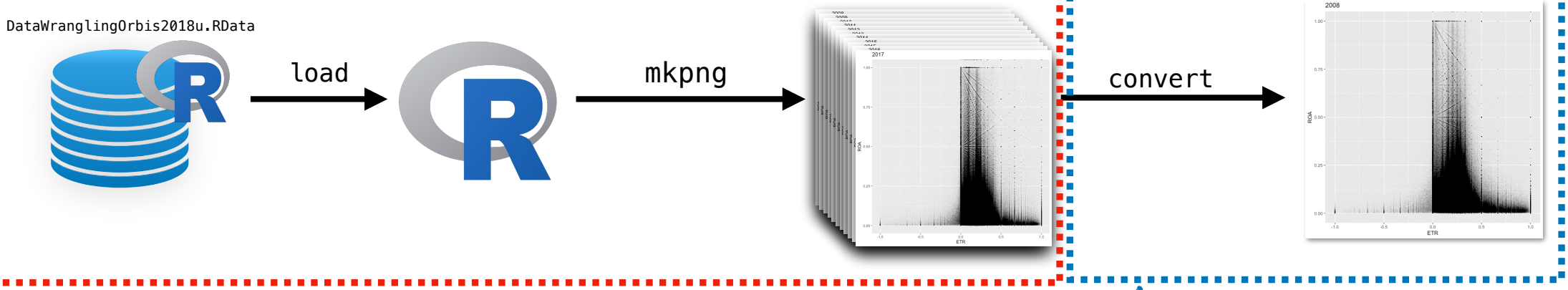
Local

Animation GIF File




```
$ convert -layers optimize -loop 0 -delay 40 ROA-ETR-????.png animation.gif
```

Automation of Sequentially Make PNG Files and Animate by `make`



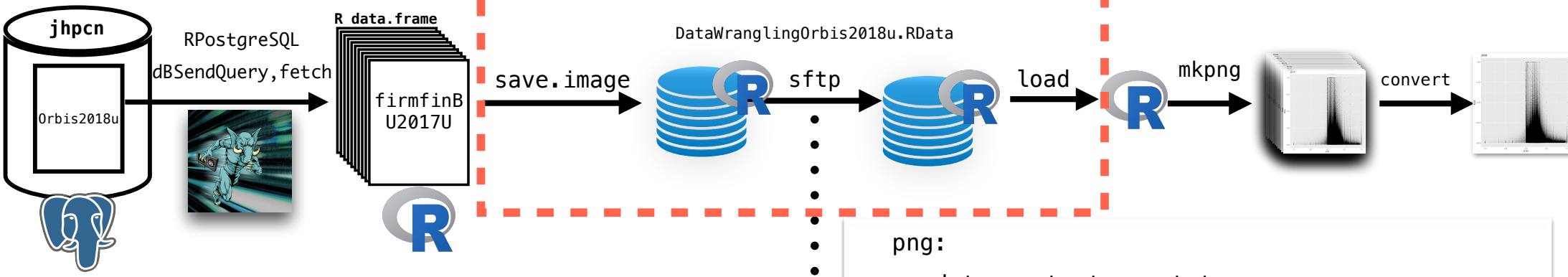
Local

```
png:
   date > start-png.txt
   Rscript makepng.R
   date > end-png.txt
animation:
   date > start-animation.txt
   convert -layers optimize -loop 0 -delay 40 ROA-ETR-????.png animation.gif
   date > end-animation.txt
```

makepng.R

```
load("./DataWranglingOrbis2018u.RData")
mkpng <- function(df, year)
{
  require(ggplot2)
  require(dplyr)
  p <- df %>%
    filter(!is.na(tax)) %>%
    filter(!is.na(pl_before_tax)) %>%
    filter(!is.na(assets_total)) %>%
    filter(pl_before_tax > 0) %>%
    group_by(firmid) %>%
    summarize(ROA = pl_before_tax/assets_total,
              ETR = tax/pl_before_tax) %>%
    ggplot(aes(ETR,ROA)) +
    geom_point(size = 0.01, alpha = 0.01) +
    xlim(-1, 1) + ylim(0, 1) + labs(title = year)
  png(paste("ROA-ETR-", year, ".png", sep = ""))
  print(p)
  dev.off()
}
mkpng(firmfinBU2008U, year = 2008)
mkpng(firmfinBU2009U, year = 2009)
mkpng(firmfinBU2010U, year = 2010)
mkpng(firmfinBU2011U, year = 2011)
mkpng(firmfinBU2012U, year = 2012)
mkpng(firmfinBU2013U, year = 2013)
mkpng(firmfinBU2014U, year = 2014)
mkpng(firmfinBU2015U, year = 2015)
mkpng(firmfinBU2016U, year = 2016)
mkpng(firmfinBU2017U, year = 2017)
```

Total Process




```
DW-u:
   date > start-DW-u.txt
   Rscript DataWranglingOrbis2018u.R
   date > end-DW-u.txt
```

```
png:
   date > start-png.txt
   Rscript makepng.R
   date > end-png.txt
animation:
   date > start-animation.txt
   convert -layers optimize -loop 0 -delay 40
   ROA-ETR-????.png animation.gif
   date > end-animation.txt
```

THE UNIVERSITY OF TOKYO

FENNEL

KWANSEI GAKUIN · MASTERY FOR SERVICE ·

Local

問題意識

1. 経済社会のサステナビリティを確保するためには、グローバルレベルでの企業活動を解明し、それが生み出す様々な社会的課題を解決することが欠かせない

2. 探索的財務ビッグデータ解析と可視化により、企業活動の実態の証拠を提示

3. 社会において存在感が高まる「企業」の2つの課題

    ①付加価値の分配：従業員 vs 投資家

    ②企業の租税回避

世界を変えるための17の目標

企業の付加価値分配

付加価値：産出面と分配面

主なステークホルダー	付加価値の構成要素
1．従業員	労働の対価としての人件費
2．債権者	借入金や社債の利息（金融費用）
3．国や地方自治体（政府）	租税公課・法人税等
4．株主	配当や社内留保として最終的に株主に分配される当期純利益

（Freeman, 2004）　　（Riahi-Belkaoui, 1999）

(1) 付加価値のステークホルダーへの分配

付加価値分配

Stacked Area Plot

１４３カ国の

全上場企業全体

赤：債権者

緑：従業員

青：政府

紫：株主

上図：総額ベース

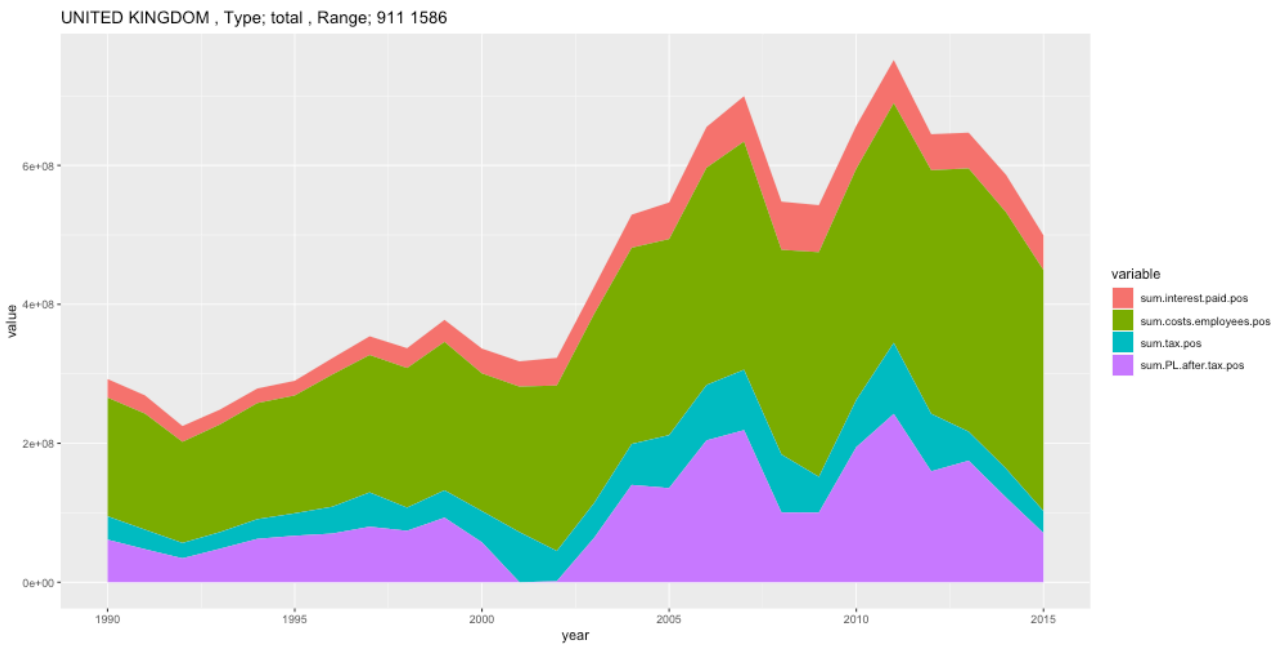
下図：構成比

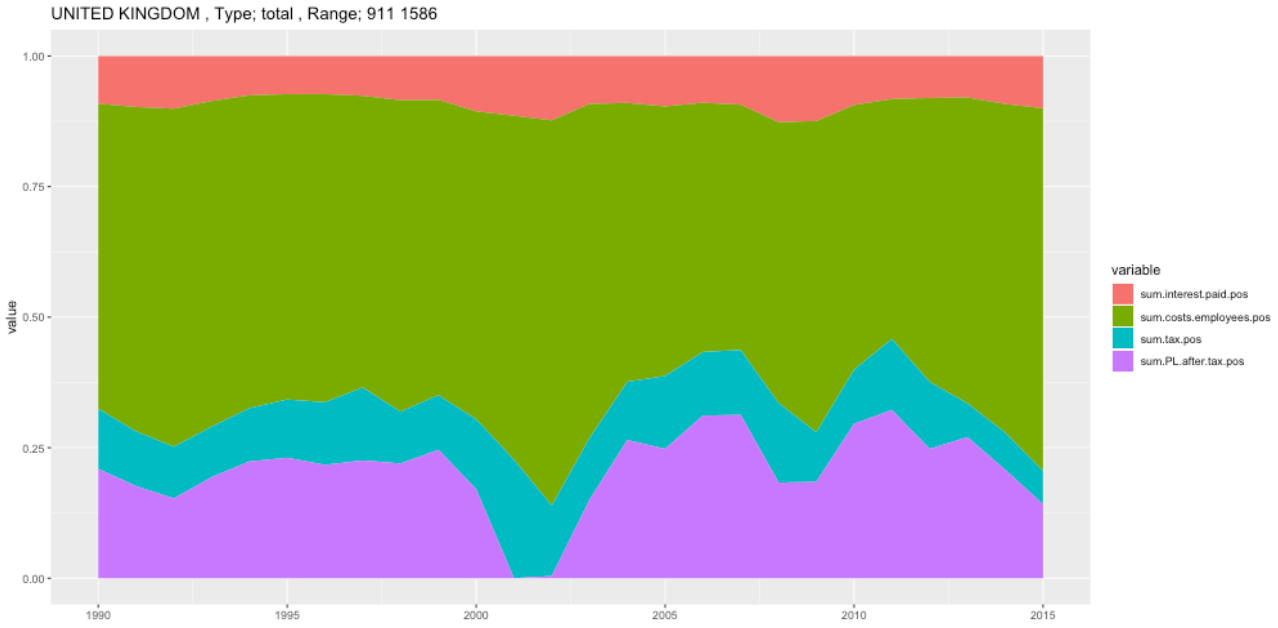
付加価値分配
Stacked Area Plot

イギリス

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比



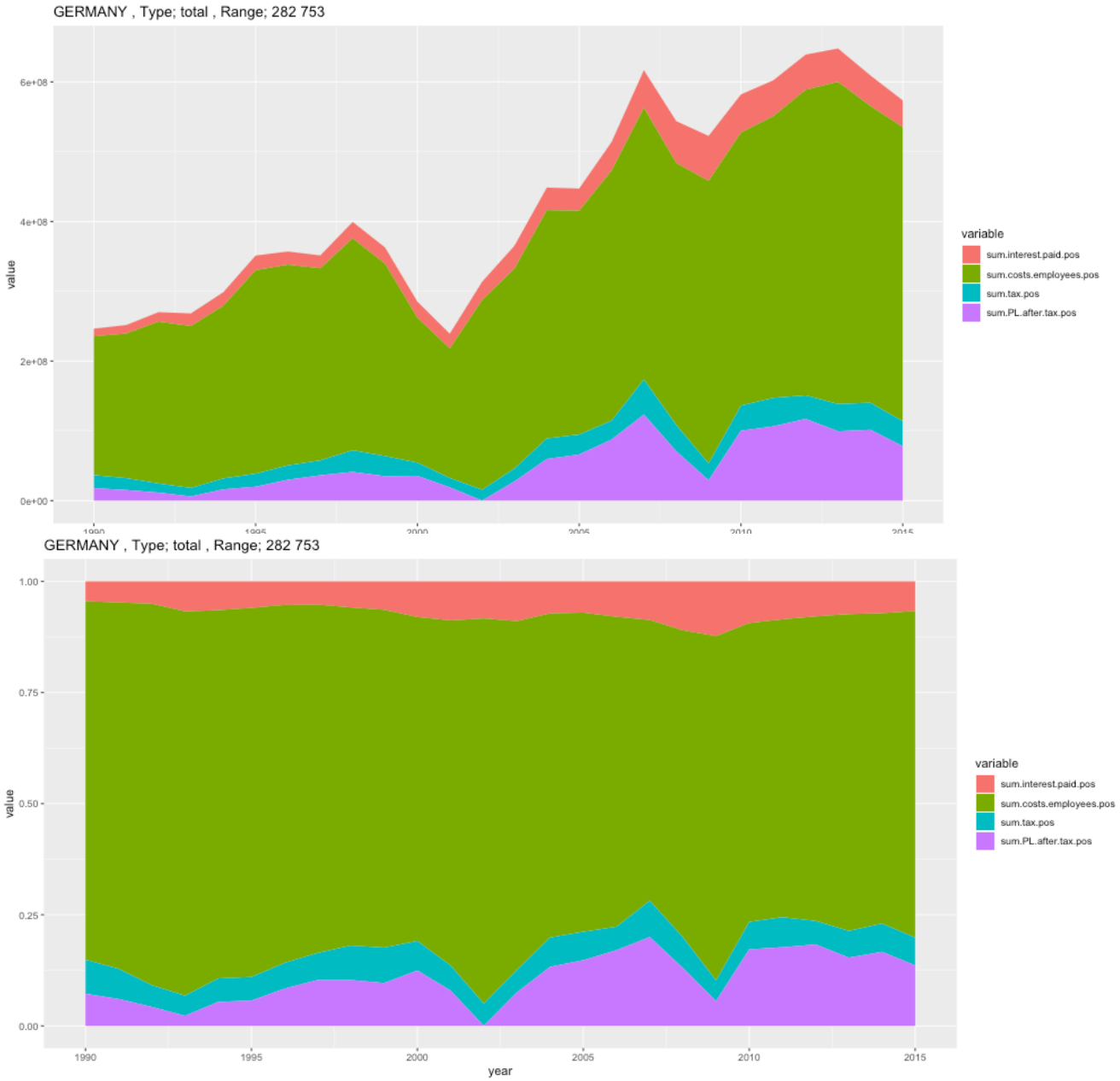


付加価値分配
Stacked Area Plot

ドイツ

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

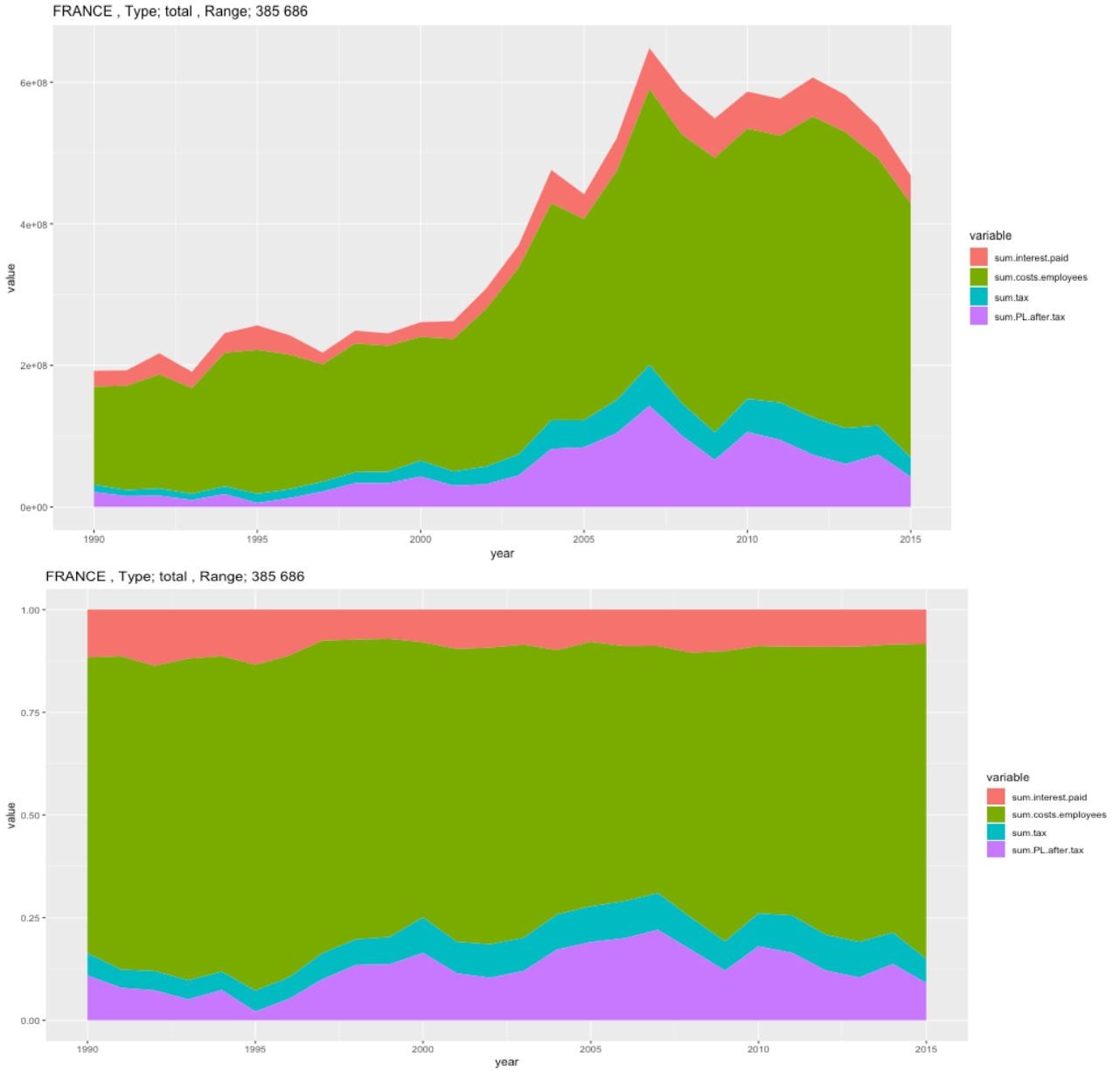


付加価値分配
Stacked Area Plot

フランス

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

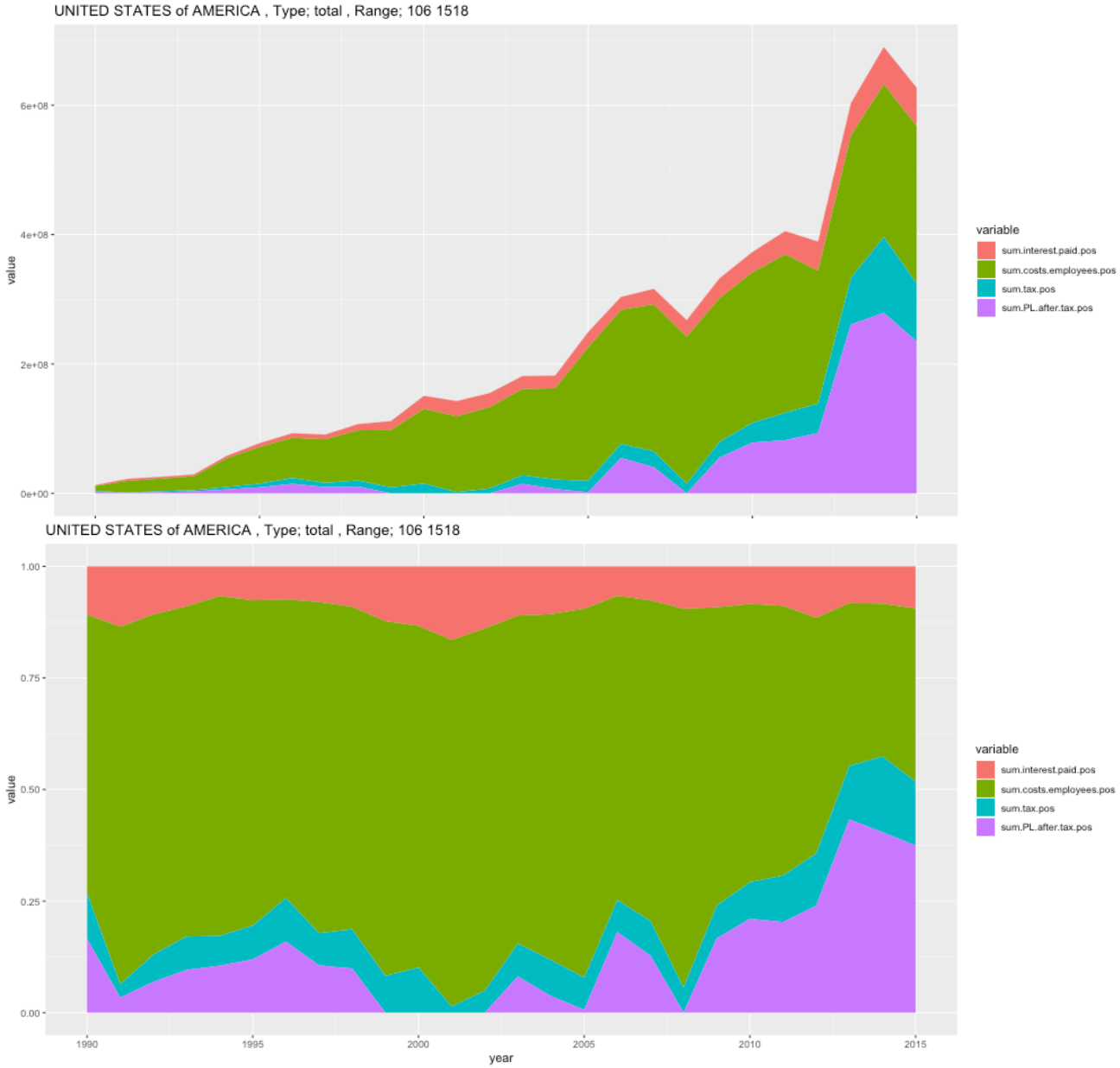


付加価値分配
Stacked Area Plot

アメリカ

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

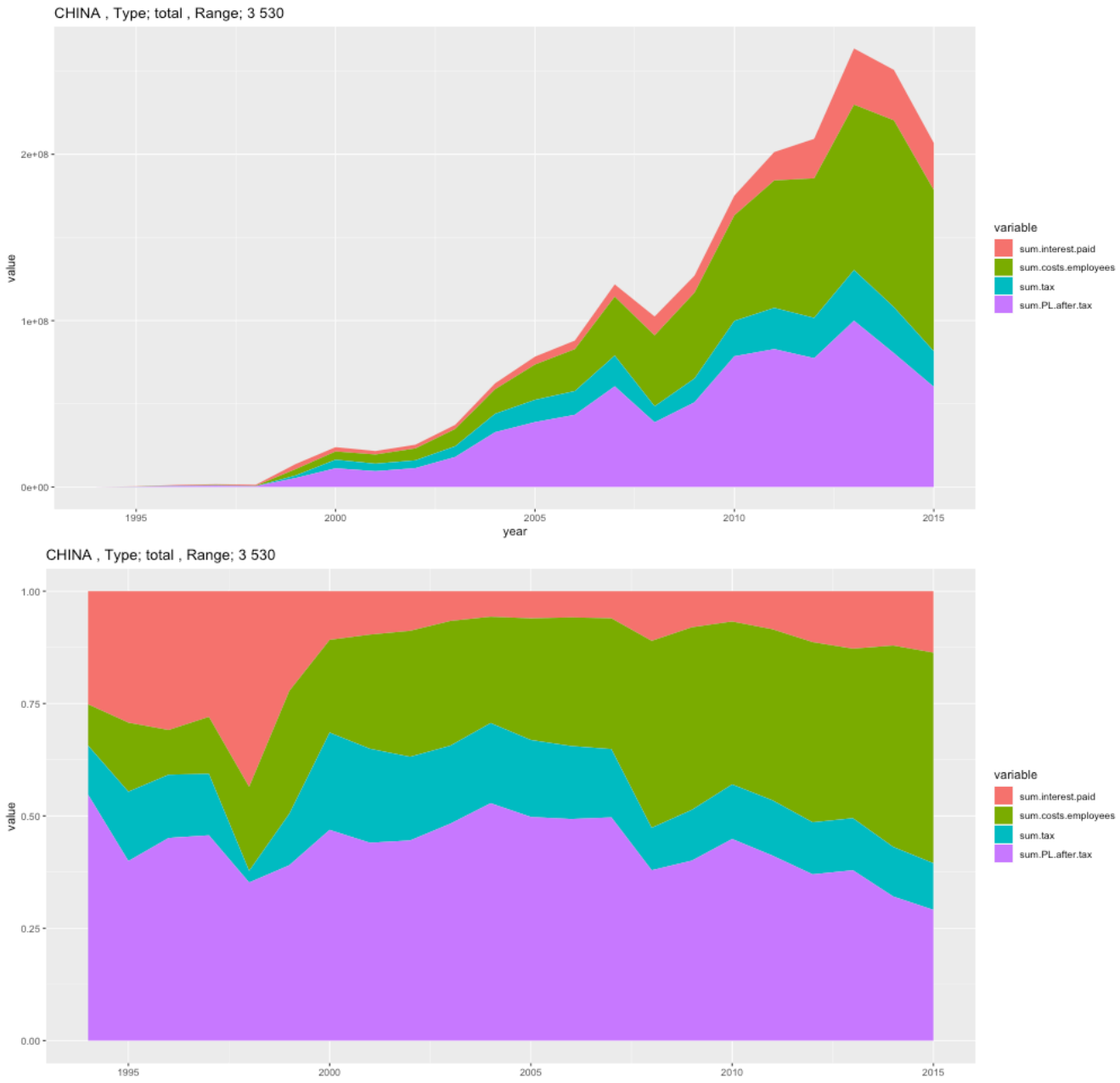


付加価値分配
Stacked Area Plot

中国

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

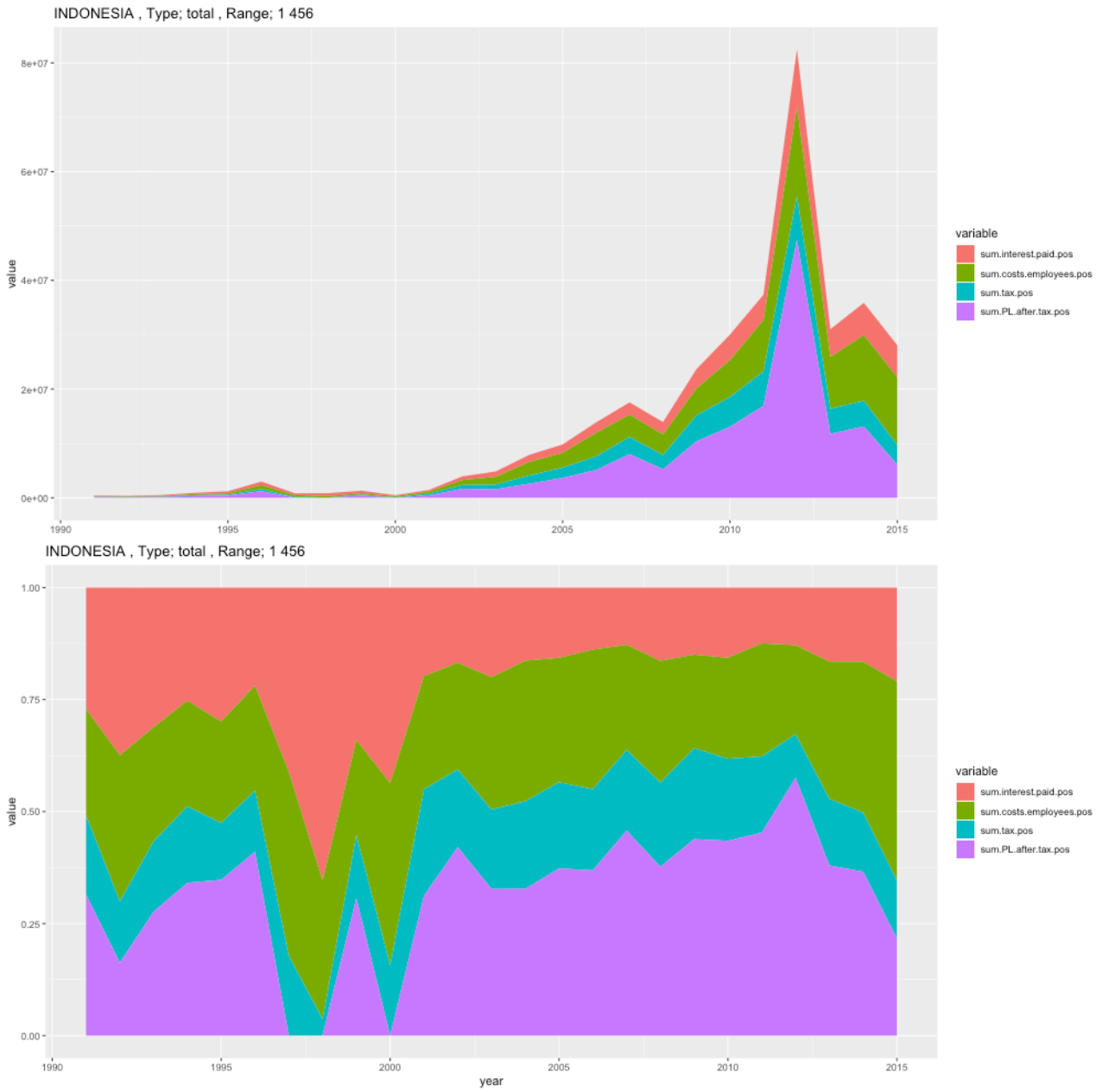


付加価値分配
Stacked Area Plot

インドネシア

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

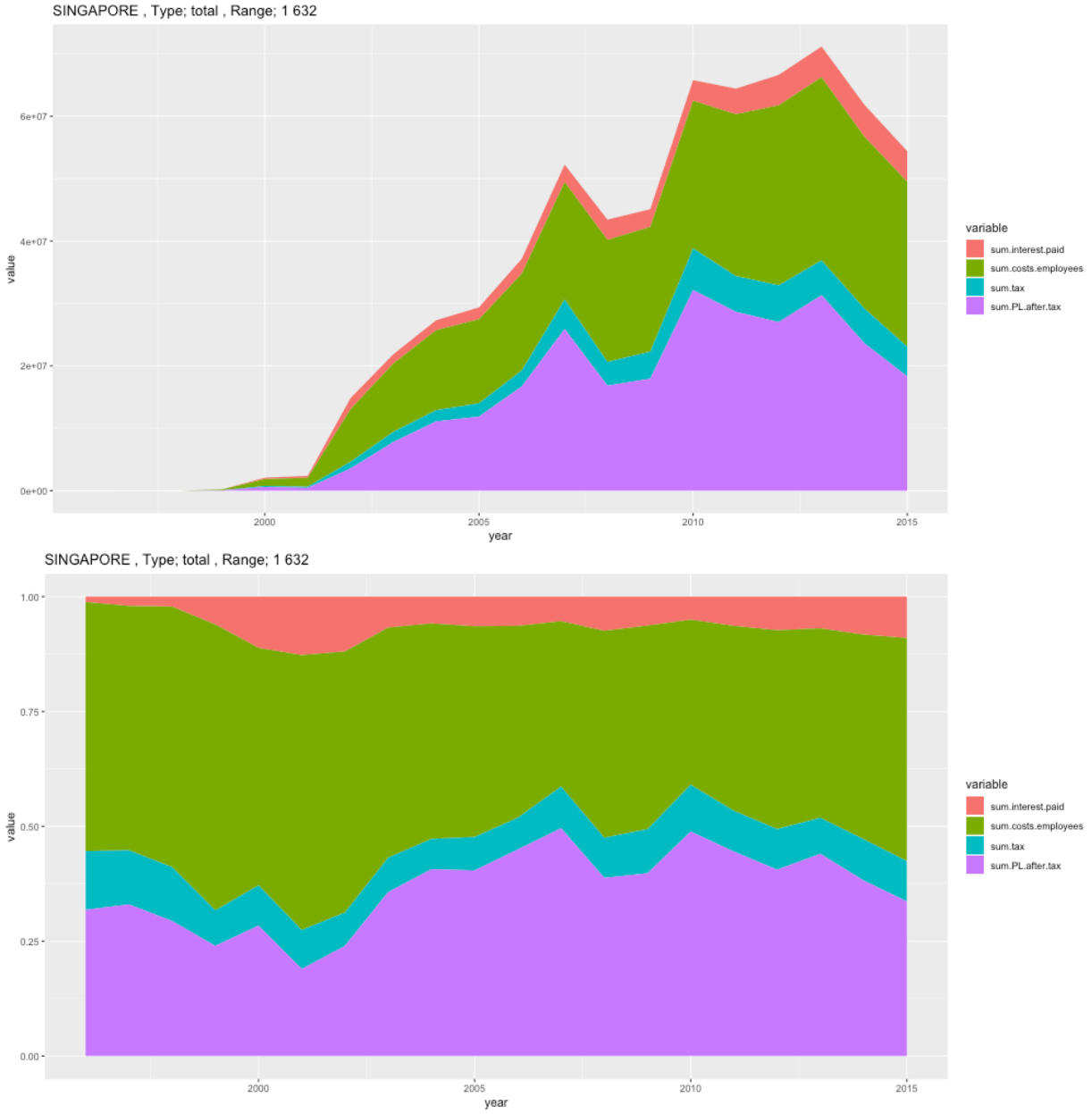


付加価値分配
Stacked Area Plot

シンガポール

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比

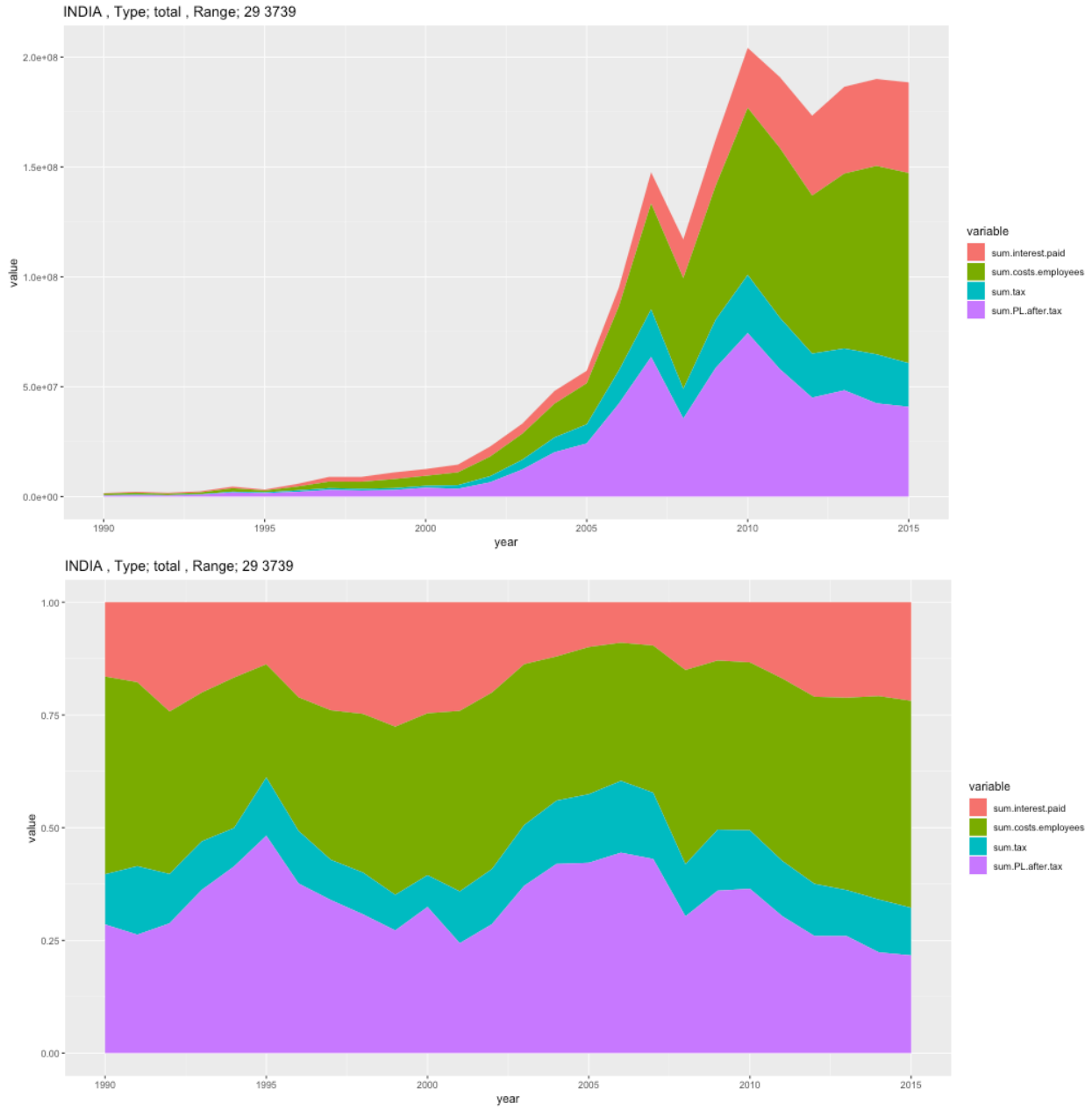


付加価値分配
Stacked Area Plot

インド

赤：債権者
緑：従業員
青：政府
紫：株主

上図：総額ベース
下図：構成比



付加価値分配
Stacked Area Plot

イラン

赤：債権者
緑：従業員
青：政府
紫：株主

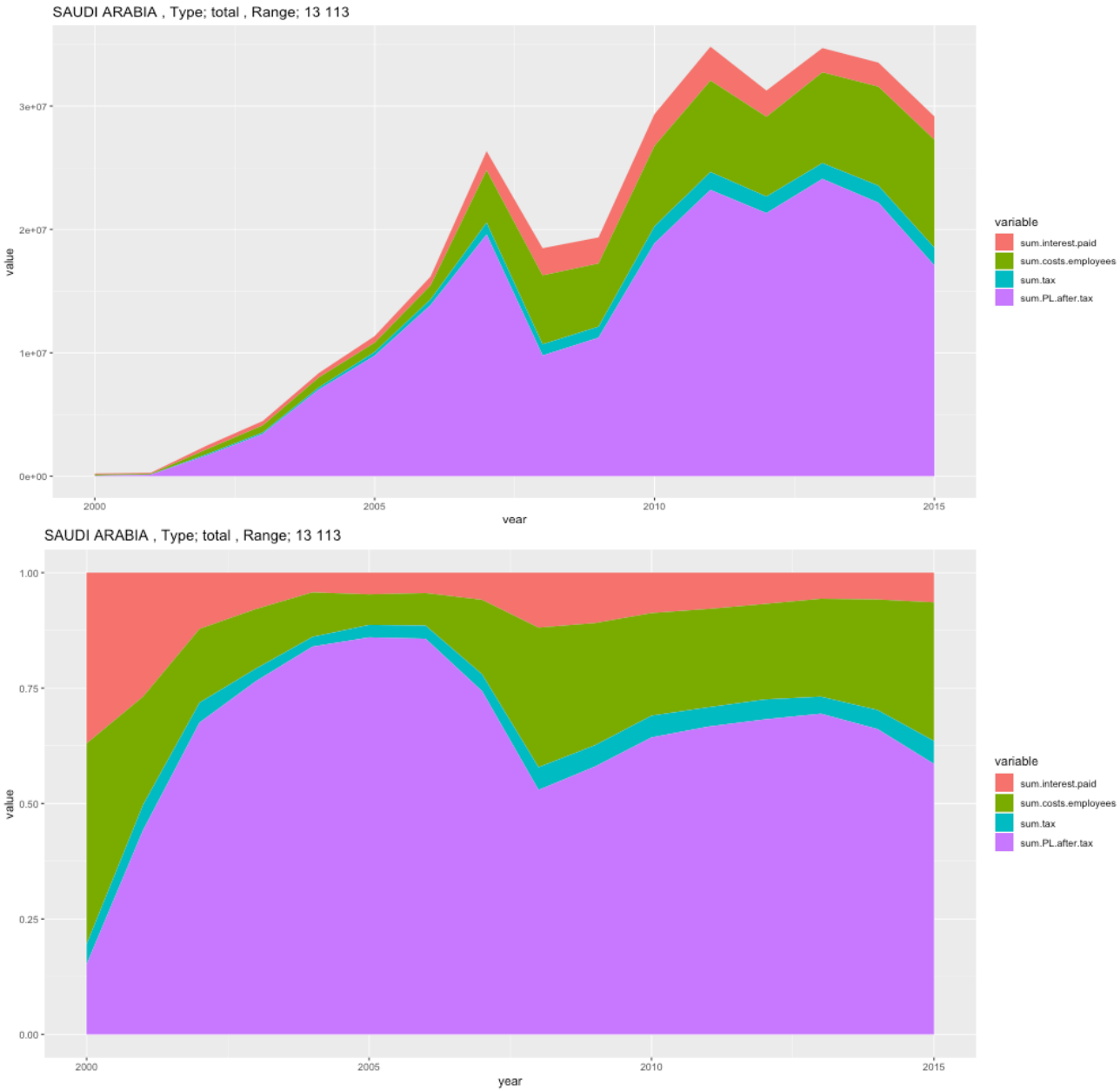
上図：総額ベース
下図：構成比

付加価値分配
Stacked Area Plot

サウジアラビア

赤：債権者
緑：従業員
青：政府
紫：株主

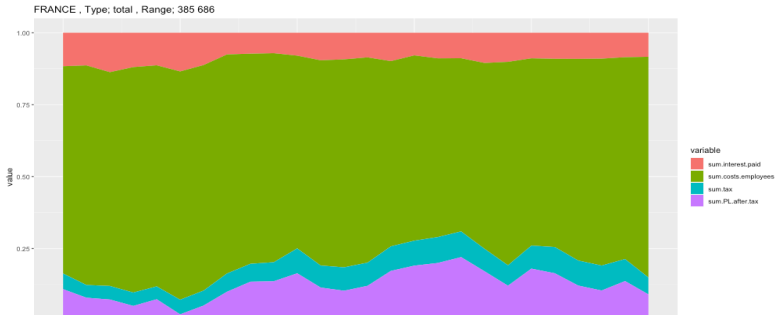
上図：総額ベース
下図：構成比



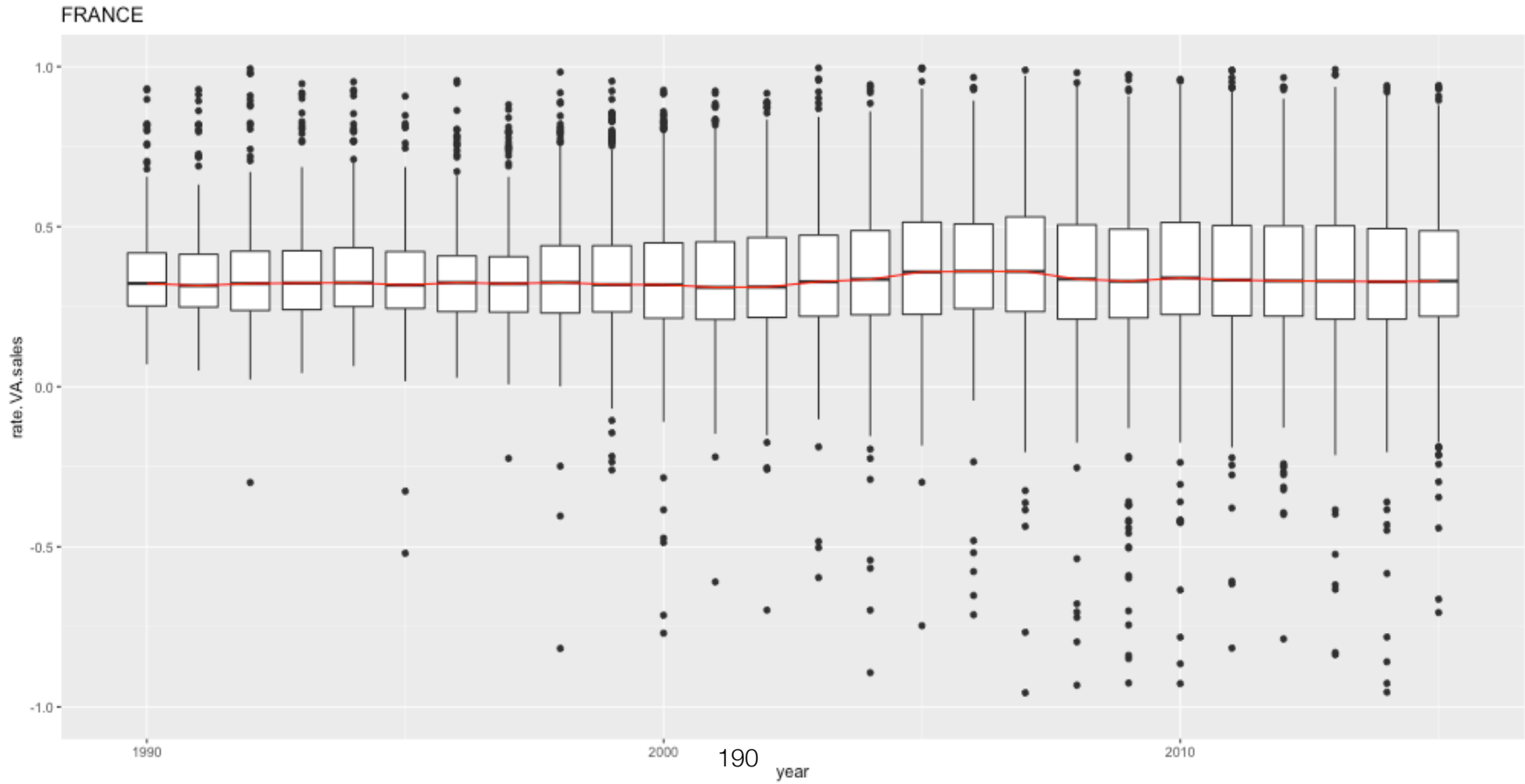
（2）付加価値率

付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計

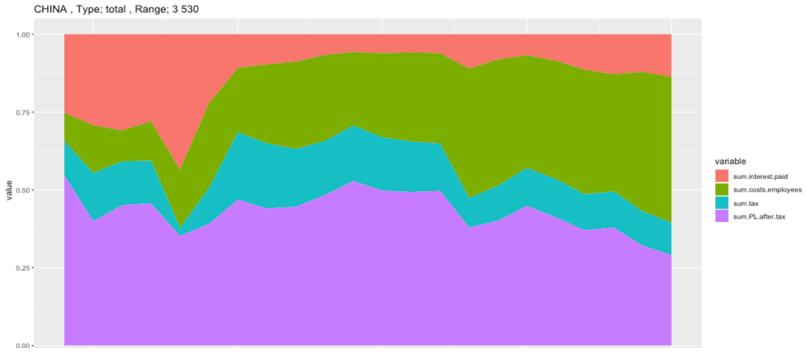


フランス

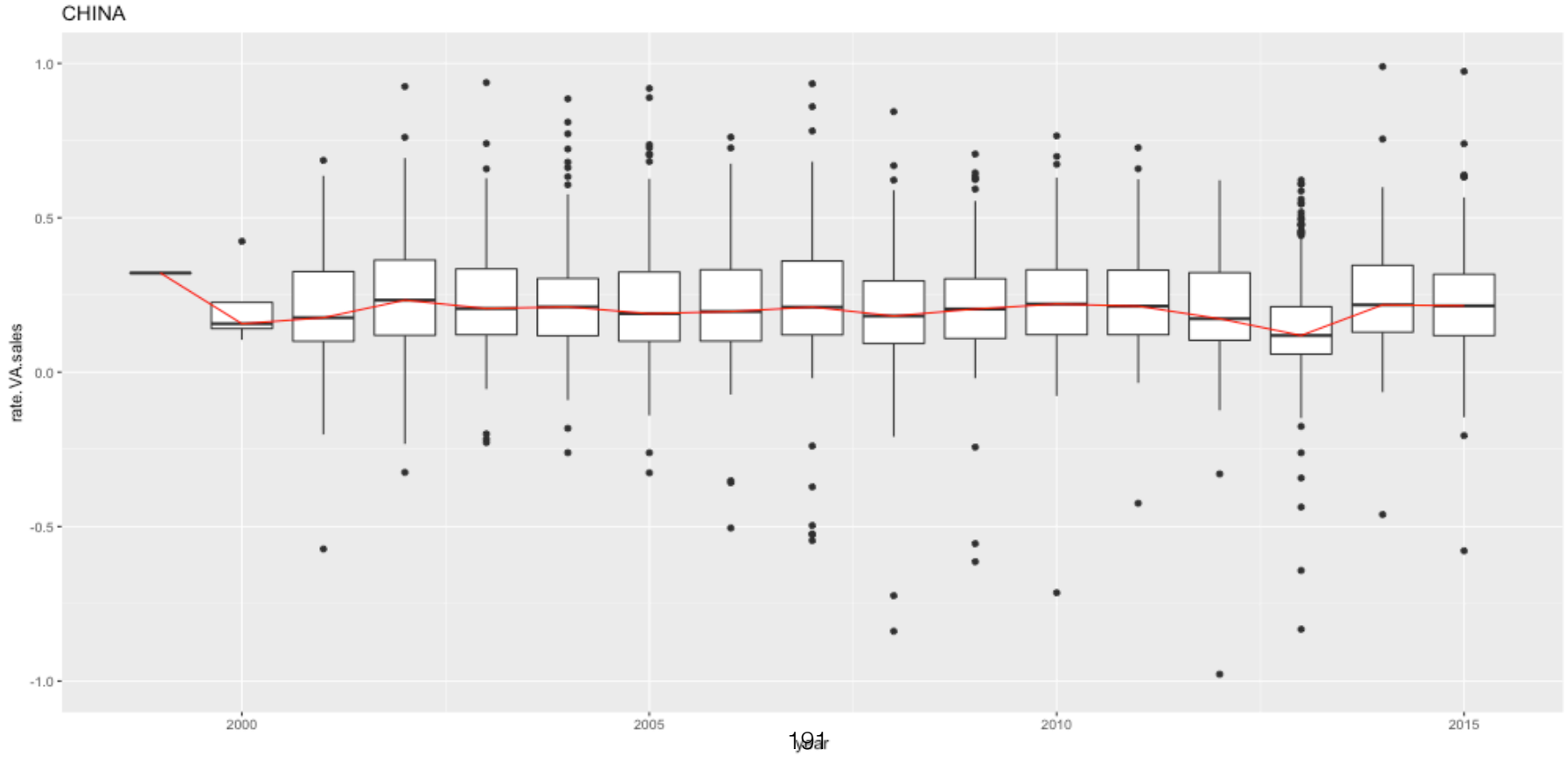


付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計

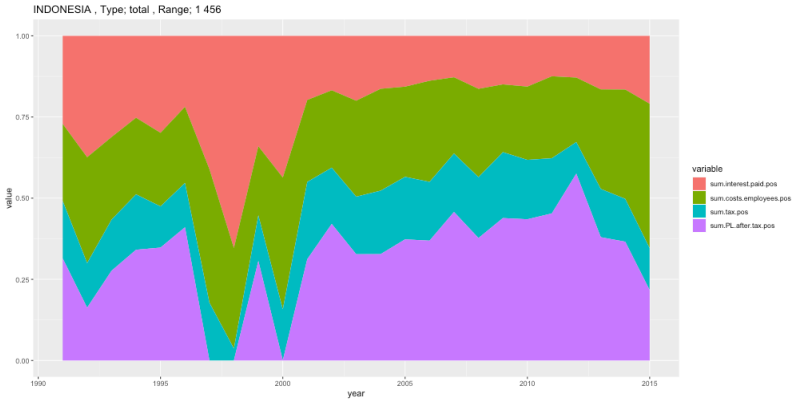


中国

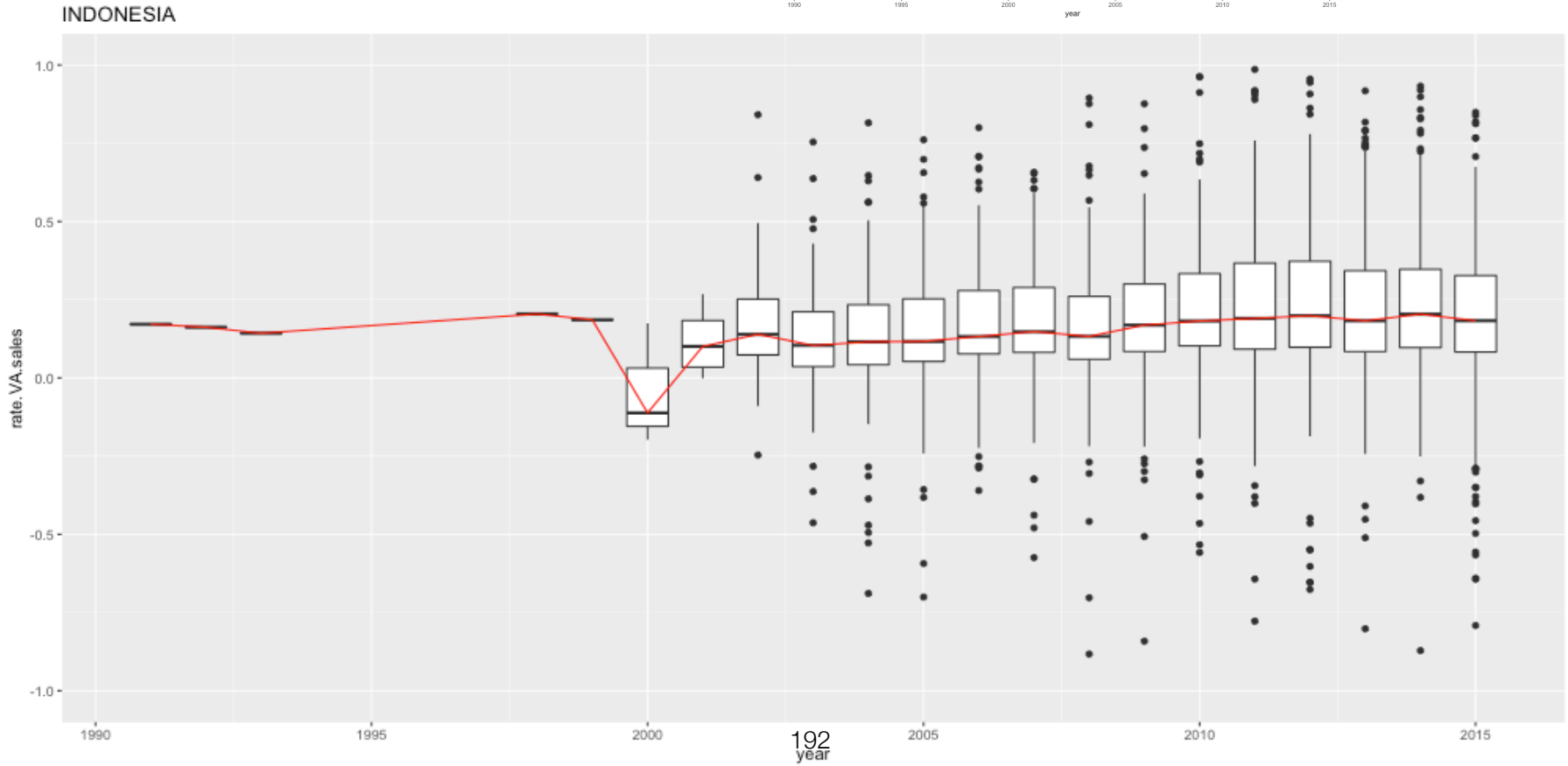


付加価値率　付加価値合計

Box Plot　　売上高合計

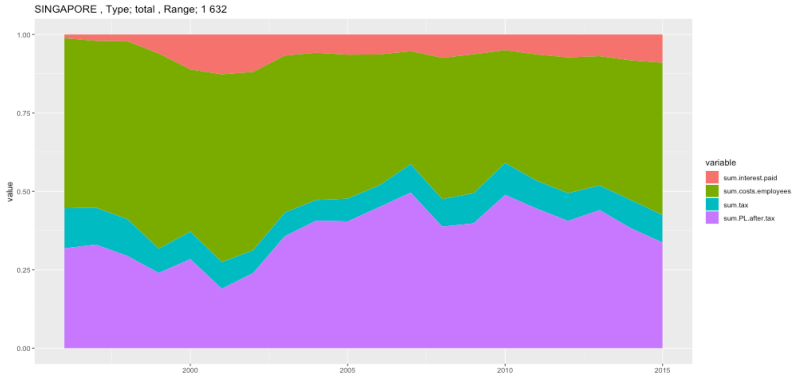


インドネシア

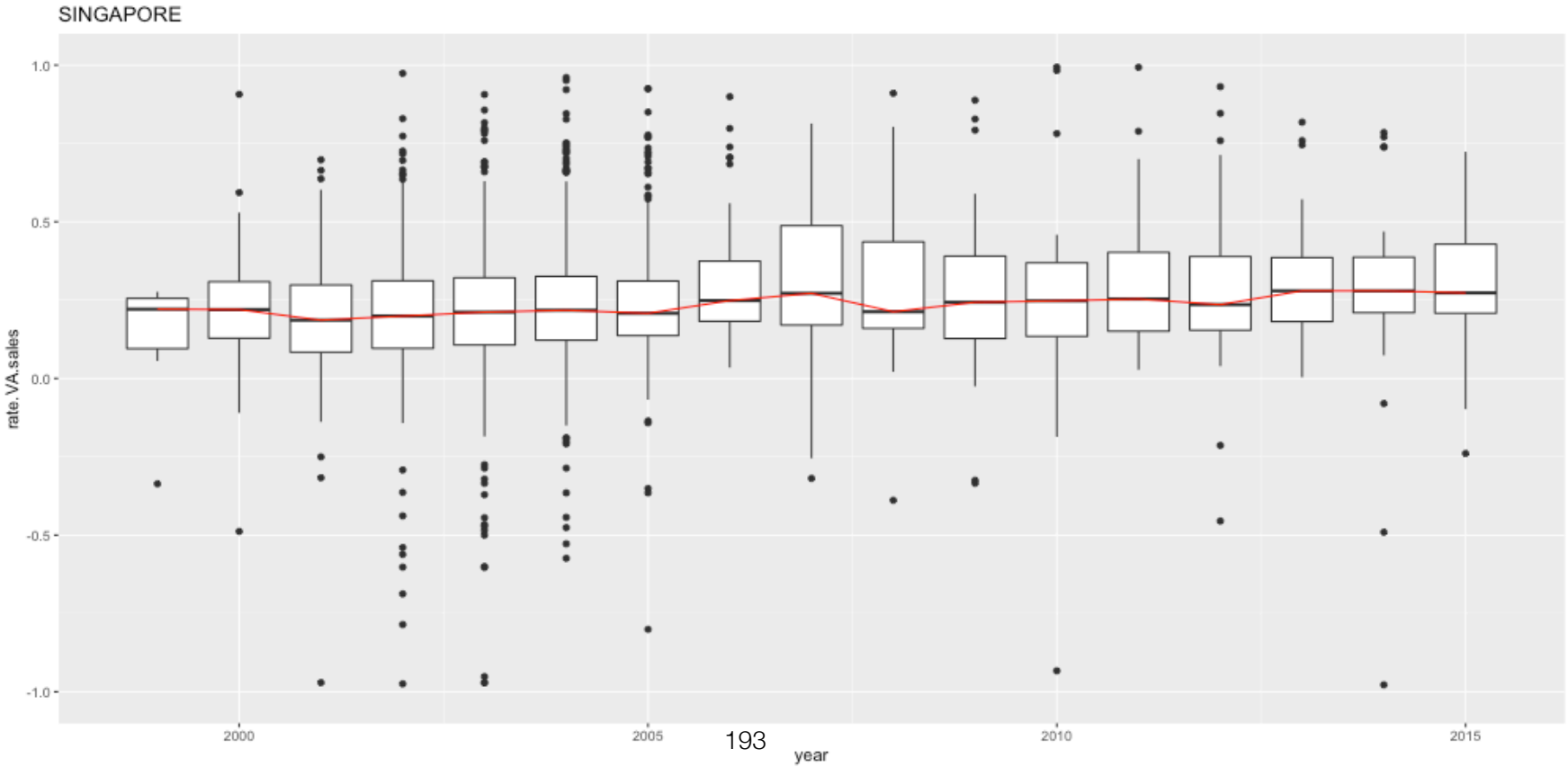


付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計

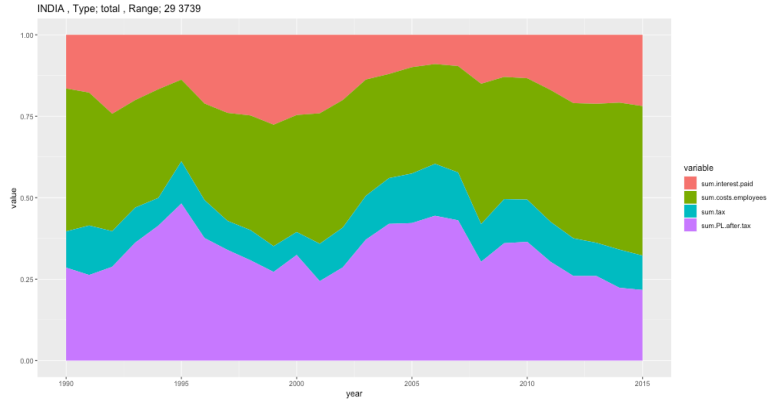


シンガポール

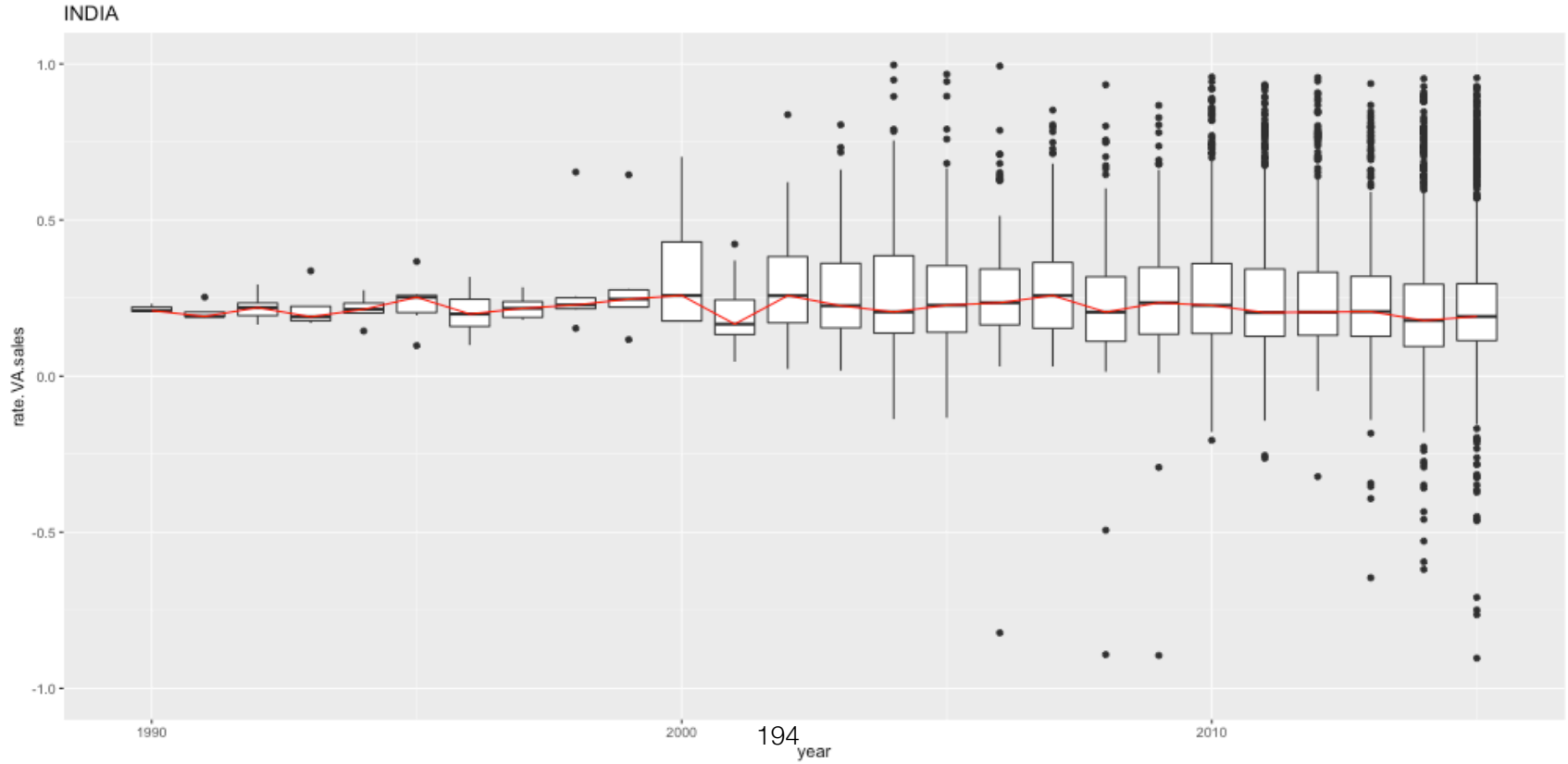


付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計

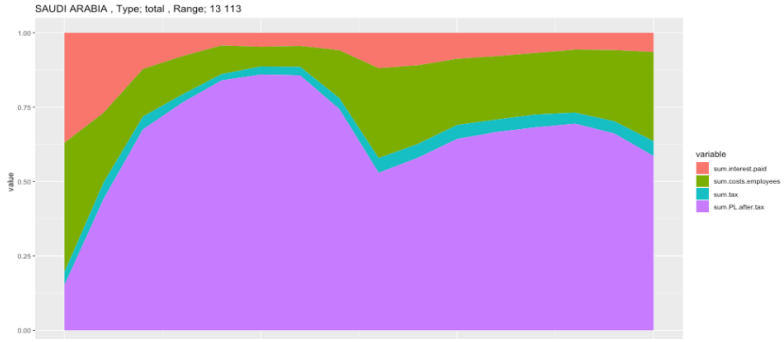


インド

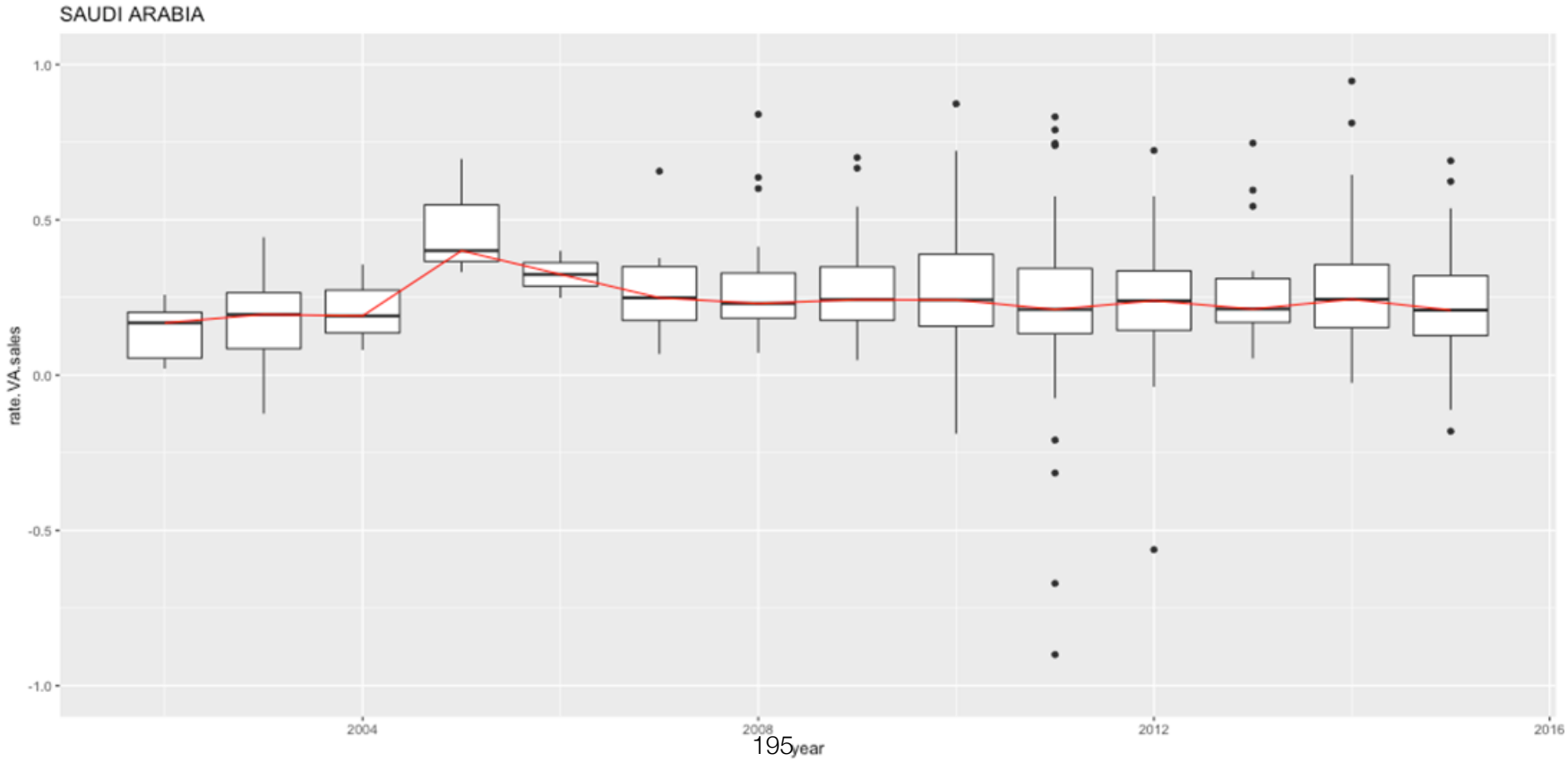


付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計

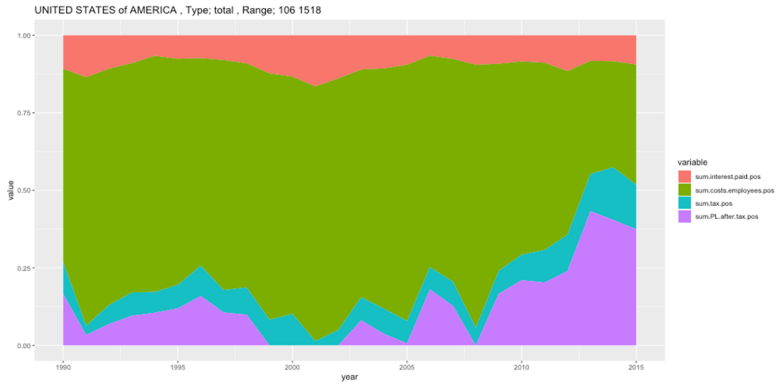


サウジアラビア

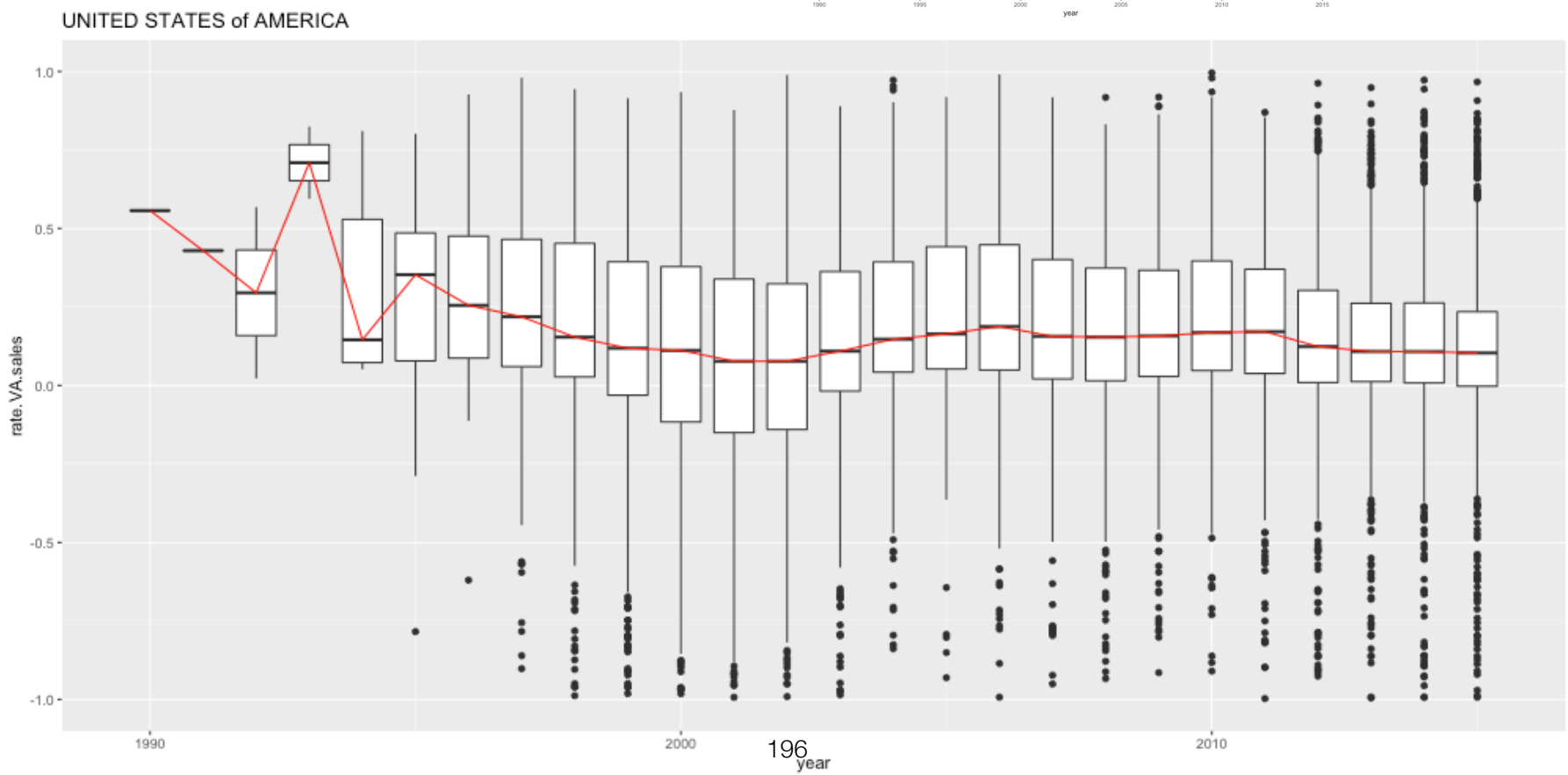


付加価値率　<u>付加価値合計</u>

Box Plot　　売上高合計



アメリカ



(3) 労働分配率

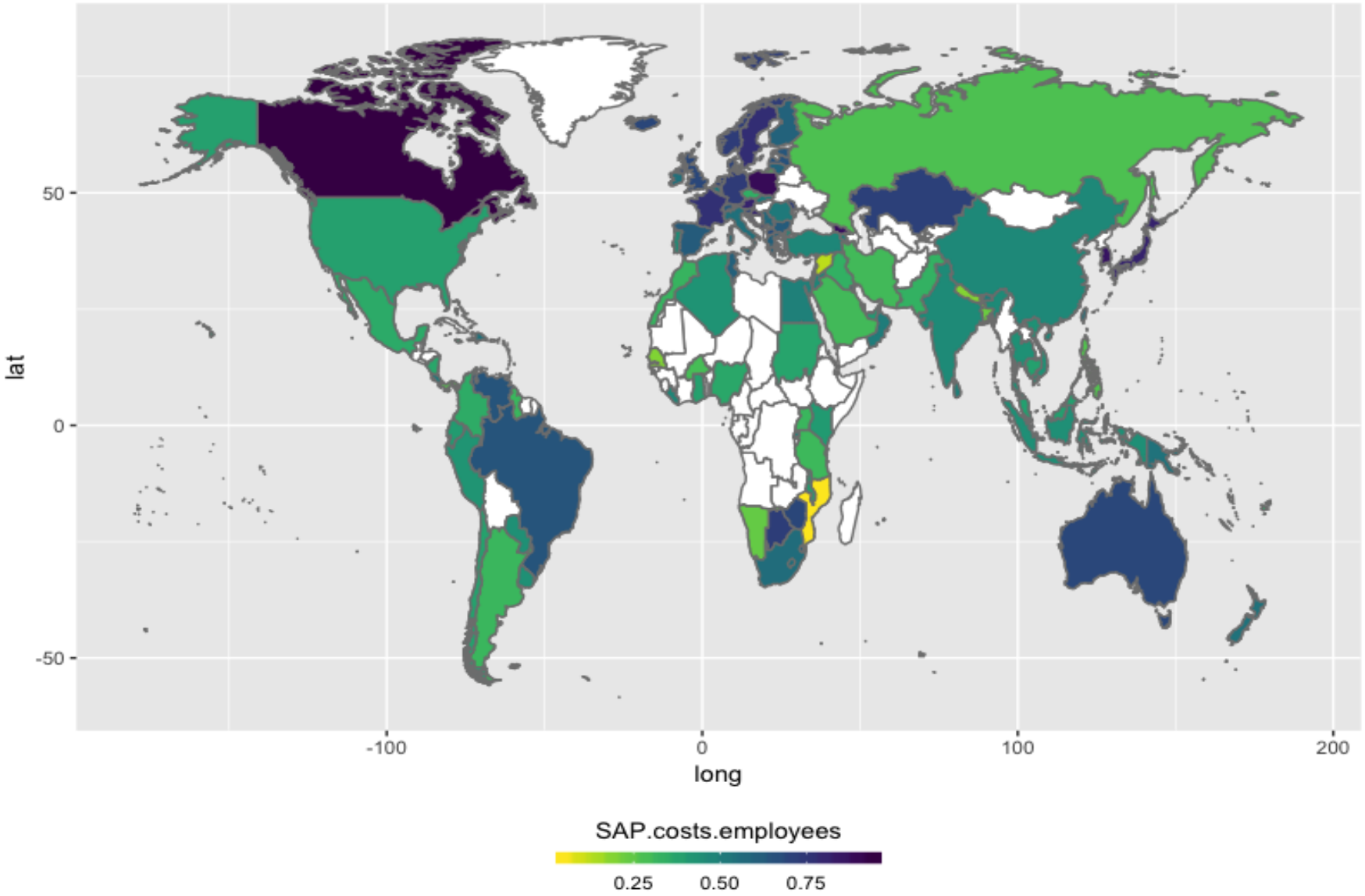
労働分配率
（国別）

従業員給付
付加価値合計

範囲 [0,1]

2015年

Map



(4) 付加価値率と労働分配率

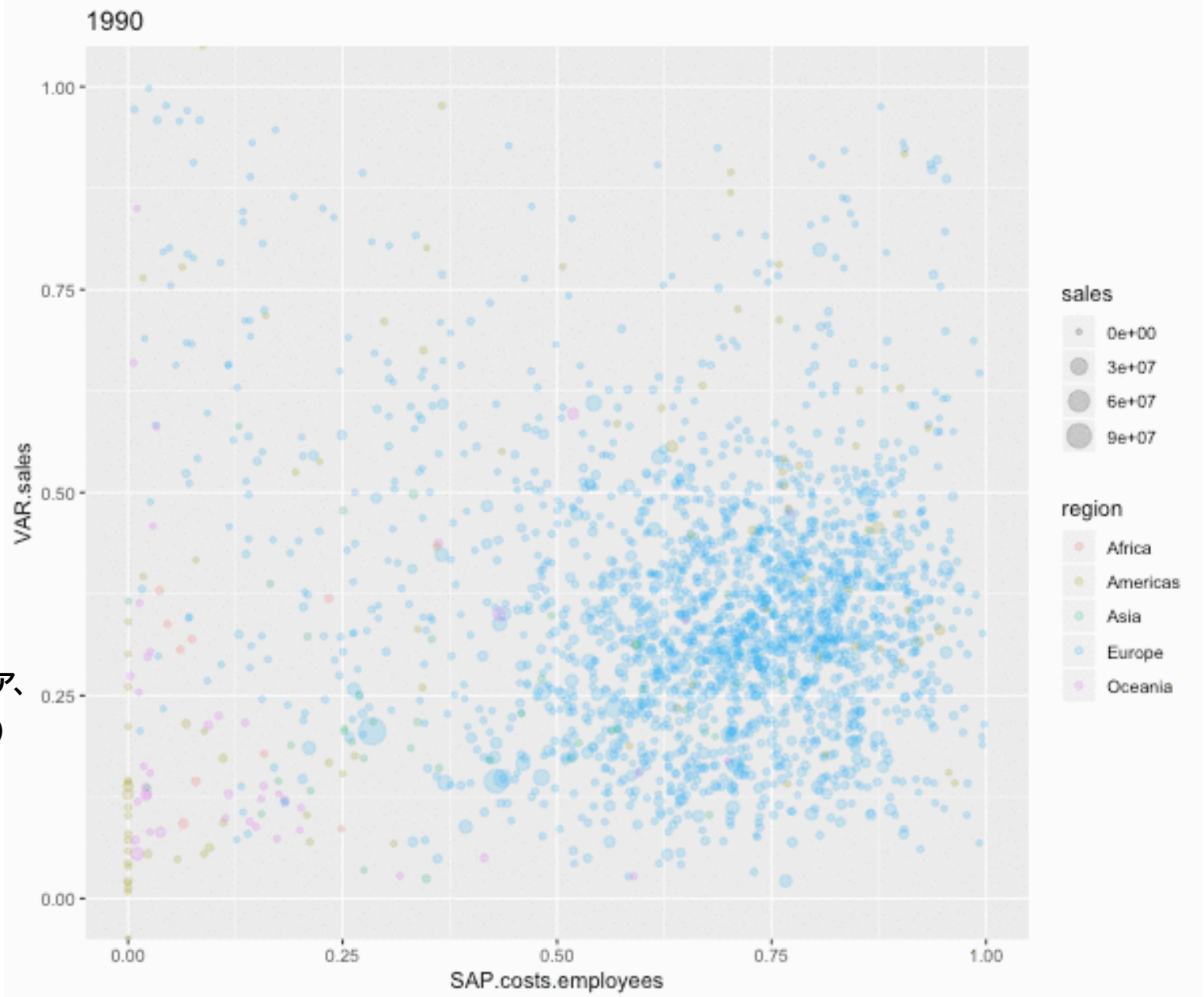
企業別労働分配率＆付加価値率

$$\text{労働分配率} = \frac{\text{従業員給付}}{\text{付加価値}}$$

$$\text{付加価値率} = \frac{\text{付加価値}}{\text{売上}}$$

Bubble Chart

x 軸：労働分配率
y 軸：付加価値率
色：地域（アフリカ、アメリカ、アジア、ヨーロッパ、オセアニア)
円：全上場企業売上高合計
143カ国、1990－2015年



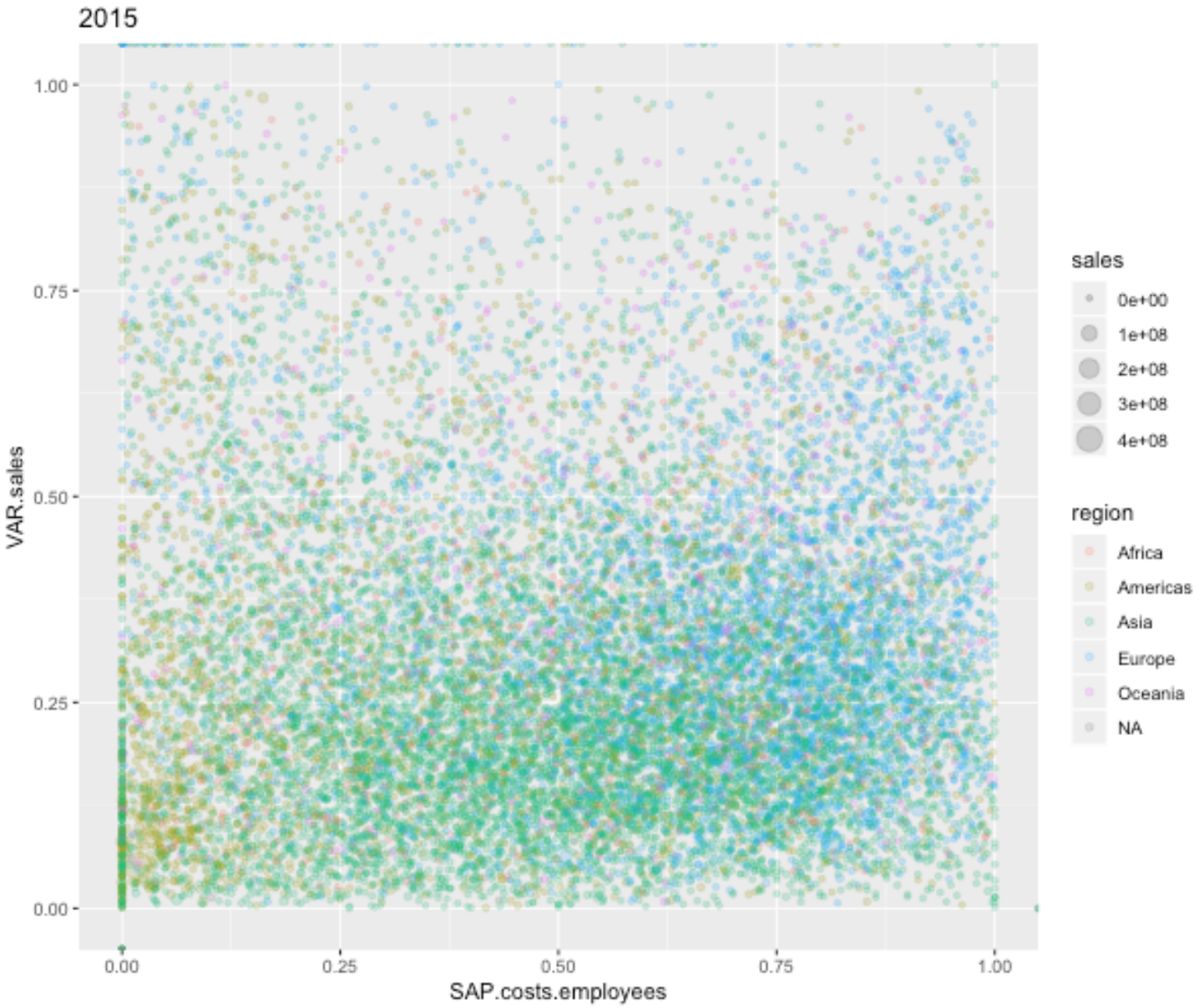
企業別労働分配率＆付加価値率

$$\text{労働分配率} = \frac{\text{従業員給付}}{\text{付加価値}}$$

$$\text{付加価値率} = \frac{\text{付加価値}}{\text{売上}}$$

Bubble Chart

x 軸：労働分配率
y 軸：付加価値率
色：地域（アフリカ、アメリカ、アジア、ヨーロッパ、オセアニア)
円：全上場企業売上高合計
143カ国、2015年



(5) 付加価値と関連する財務指標の相関

変数の相関

Heat Map Animation

付加価値合計額

4つの付加価値

（支払利息、人件費、

支払税金、当期純利益）

売上、総資産、純資産

ROA、ROE

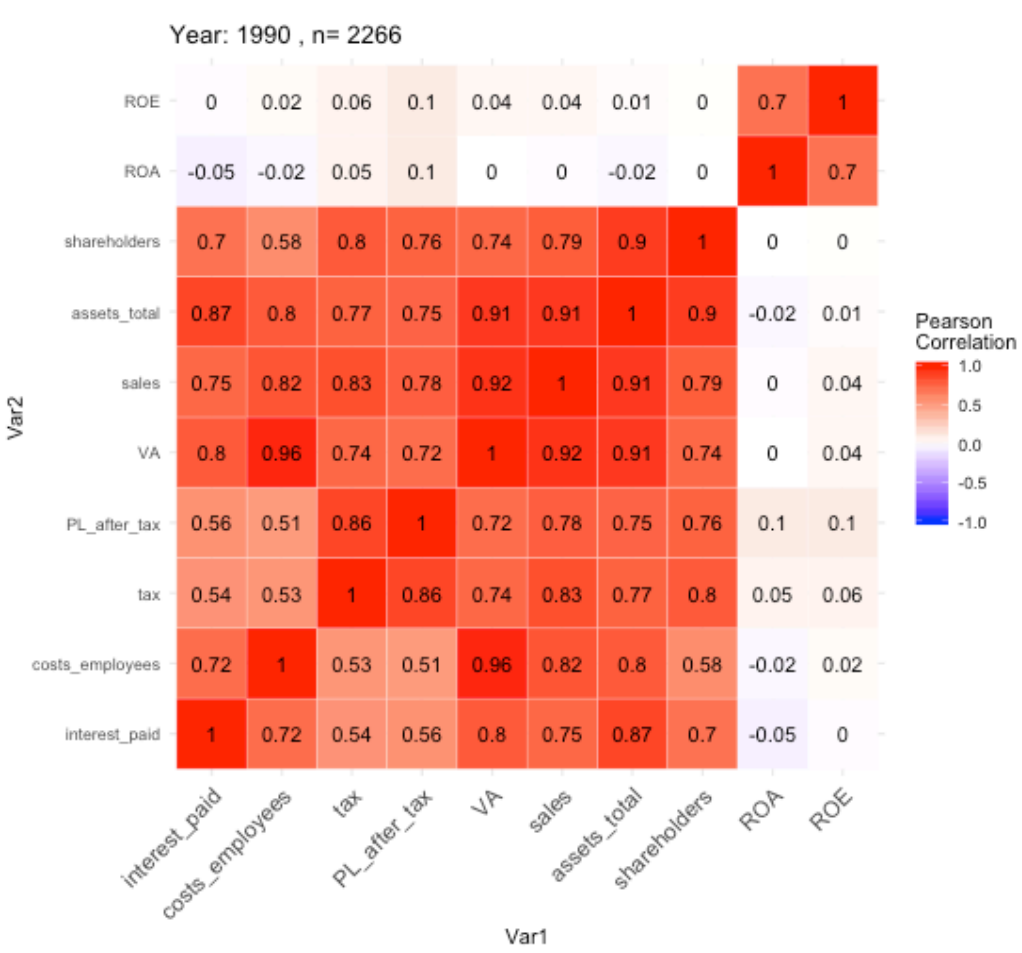
143か国

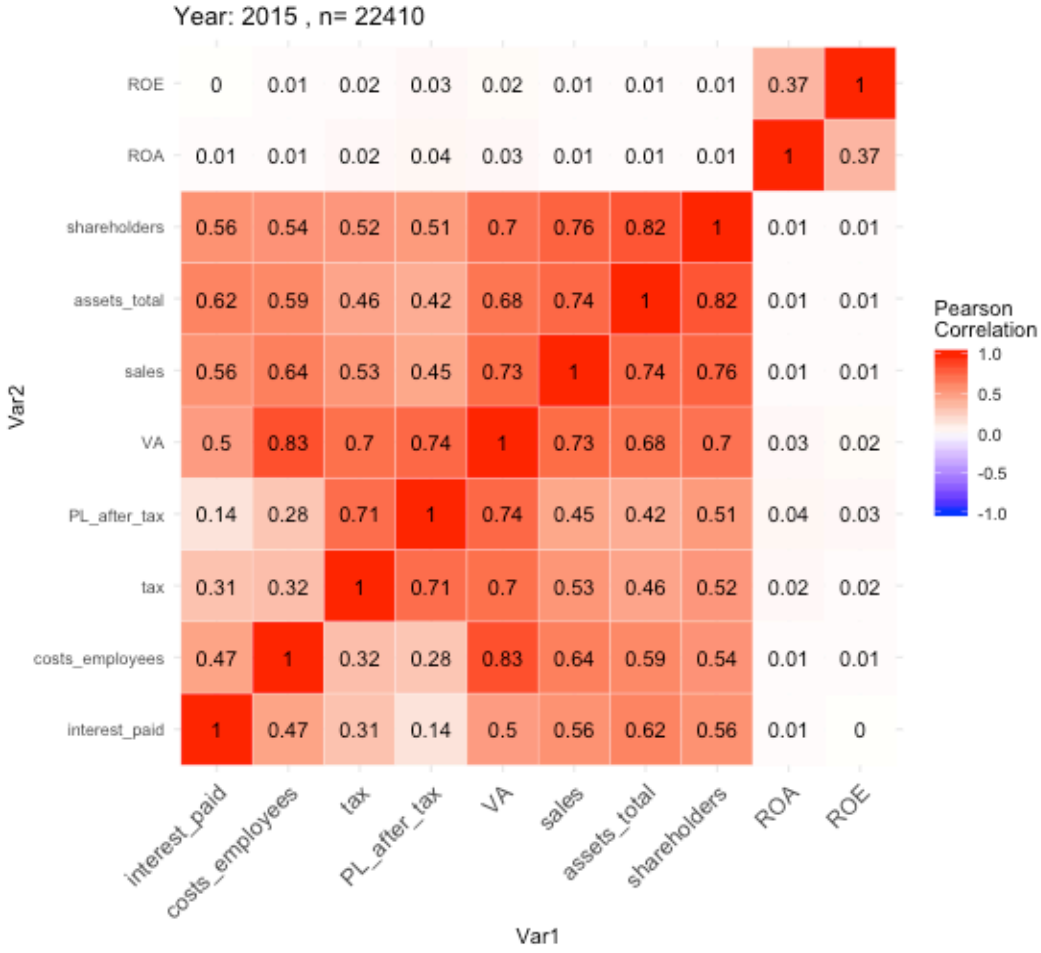
1990－2015年

変数の相関　Heat Map

付加価値合計額、4つの付加価値（支払利息、人件費、支払税金、当期純利益）

売上、総資産、純資産、ROA、ROE（143か国）





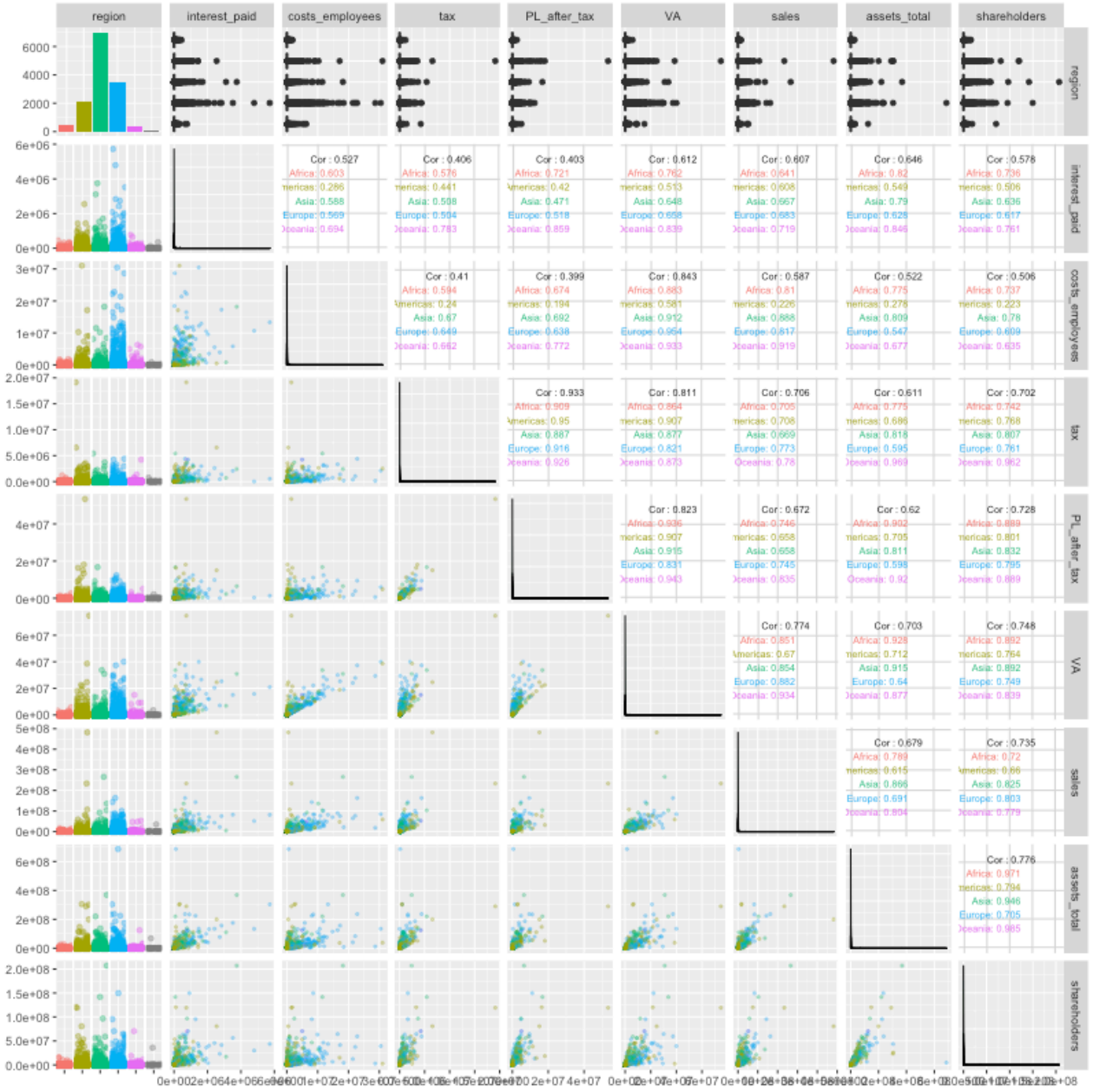
変数の相関Heat Map（143カ国、1990ー2015年）からわかること

- 売上と付加価値合計額の相関　　1990年 0.92 → 2015年 0.73
- 売上と人件費の相関　　1990年 0.82 → 2015年 0.64
- 売上と支払税金の相関　　1990年 0.83 → 2015年 0.53
- 売上と当期純利益の相関　　1990年 0.78 → 2015年 0.45
- 付加価値合計額と人件費の相関　1990年 0.96 → 2015年 0.83

5地域の相関

ggpairs

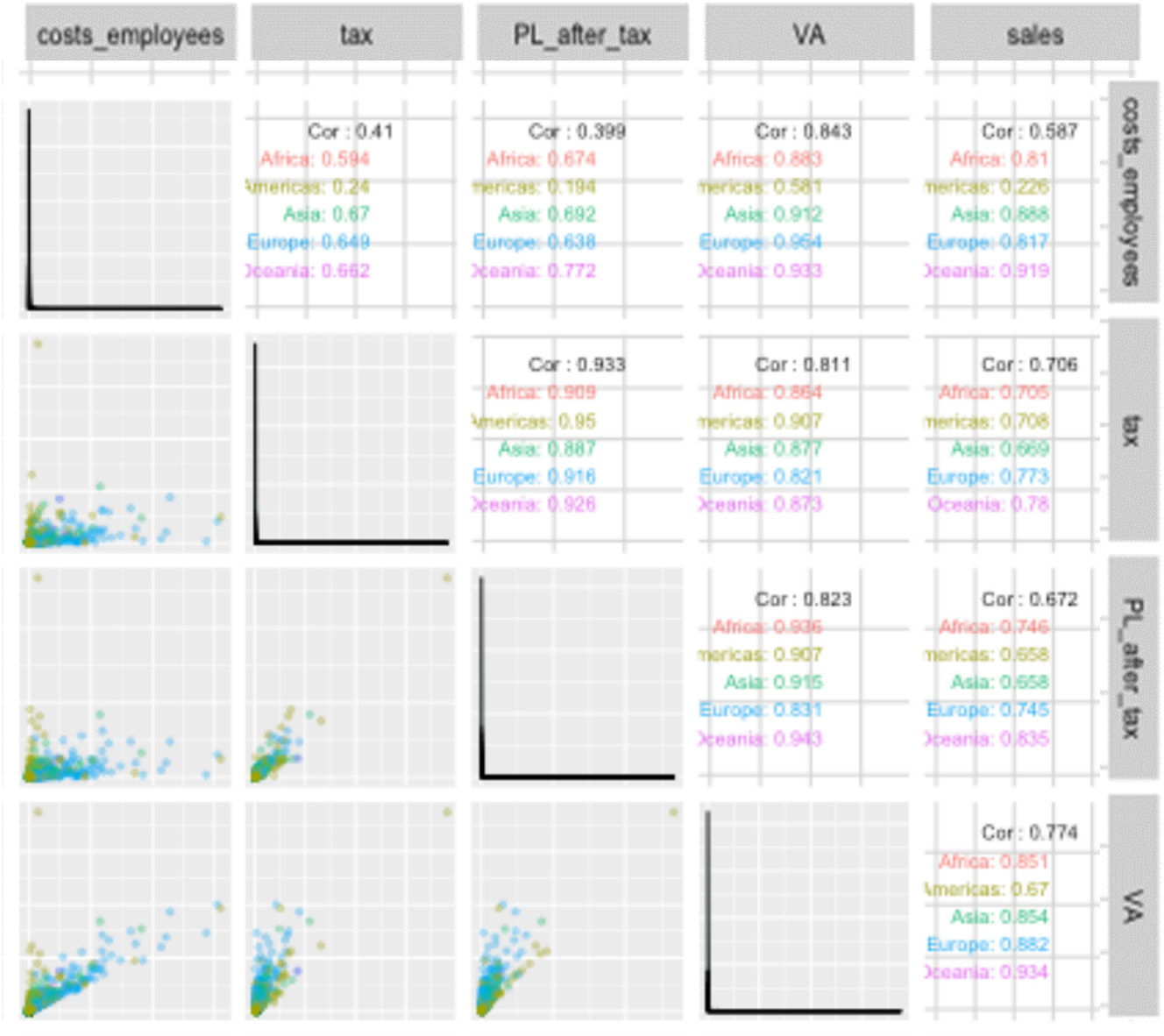
付加価値合計額と人件費の相関(ヨーロッパが高く、アメリカが低い)



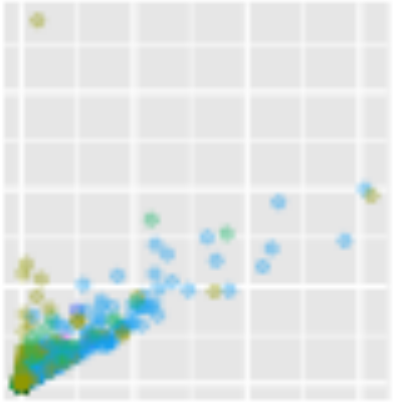
5地域の相関

ggpairs
（拡大）

付加価値合計額と人件費の相関
(ヨーロッパが高く、アメリカが低い)



人件費



付加価値

Cor: 0.843

Africa: 0.883
Americas: 0.581
Asia: 0.912
Europe: 0.964
Oceania: 0.933

企業の租税回避

租税回避の蓋然性

- 租税回避とは (Dyreng et al. 2008; Chen et al., 2010; Hanlon and Heitzman, 2010; Sikka, 2010; Lanis and Richardson, 2015)
  - Downward management of taxable income through tax-planning activities

- 指標 (Shackelford and Shevlin, 2001; Dyreng et al., 2008; Hanlon and Heitzman, 2010; Chen et al., 2010; Graham et al., 2012; Badertscher et al., 2013; Suzuki, 2014; Dyreng et al., 2017)

$$\text{GAAP ETR} = \frac{\text{Total Tax Expense}}{\text{Pre} - \text{tax Income}}$$

$$\text{GAAP ETR} - \text{Statutory Tax Rate} = \frac{\text{Total Tax Expense}}{\text{Pre} - \text{tax Income}} - \text{Statutory Tax Rate}$$

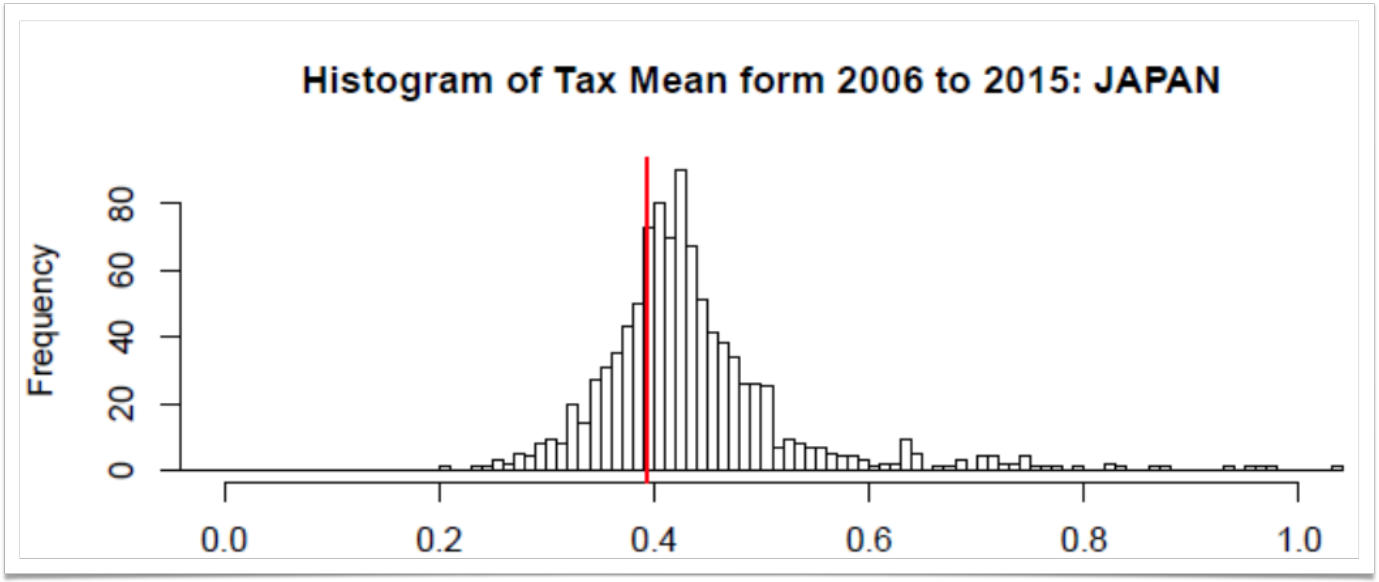
$$\text{Long term GAAP ETR} = \frac{\Sigma\ \text{Total Tax Expense}}{\Sigma\ \text{Pretax Income}}$$

$$\text{Long term GAAP ETR} - \text{Statutory Tax Rate} = \frac{\Sigma\ \text{Total Tax Expense}}{\Sigma\ \text{Pretax Income}} - \text{Statutory Tax Rate}$$

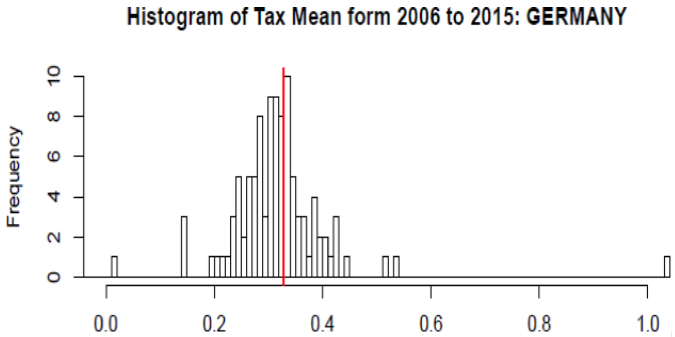
各国のETRの分布

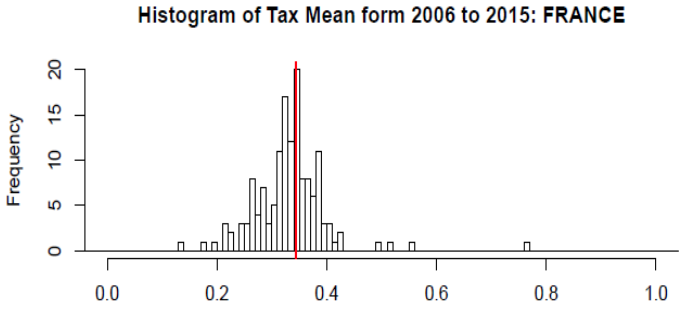
- 各国の全上場企業（下図は10年間の平均）の実効税率（ETR）分布
- Statutory Tax Rates（法定税率：赤線)

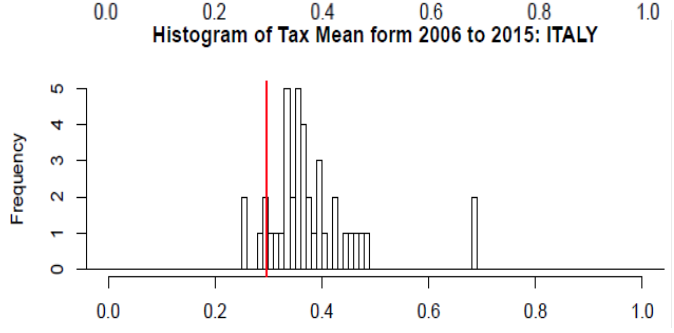
（企業数上位20カ国、2006-2015年） ➡ 租税回避行動の証拠1

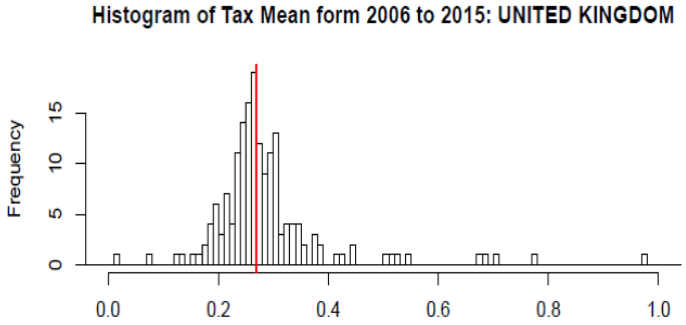


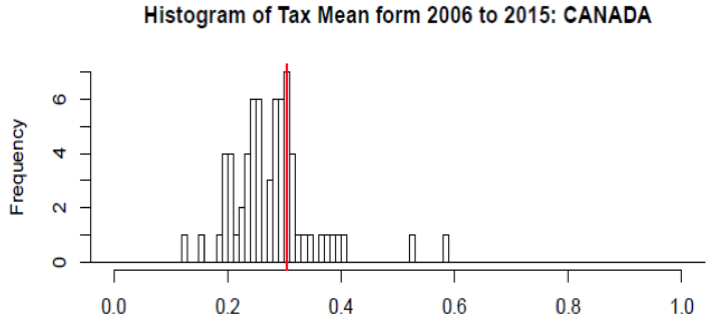
G7 countries



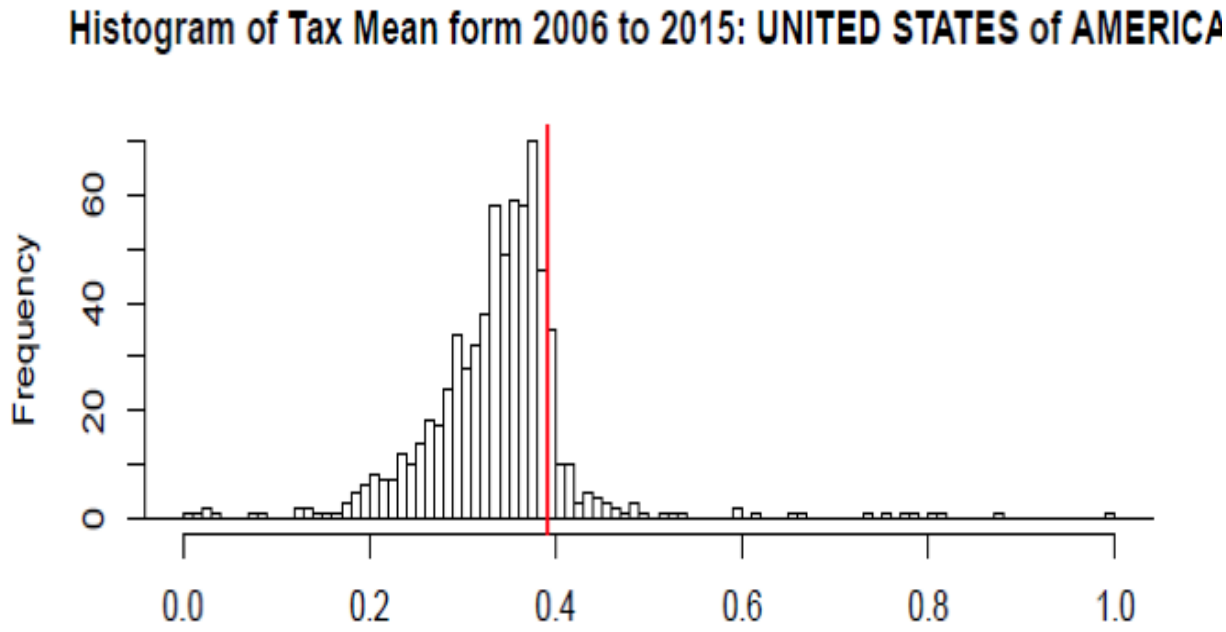




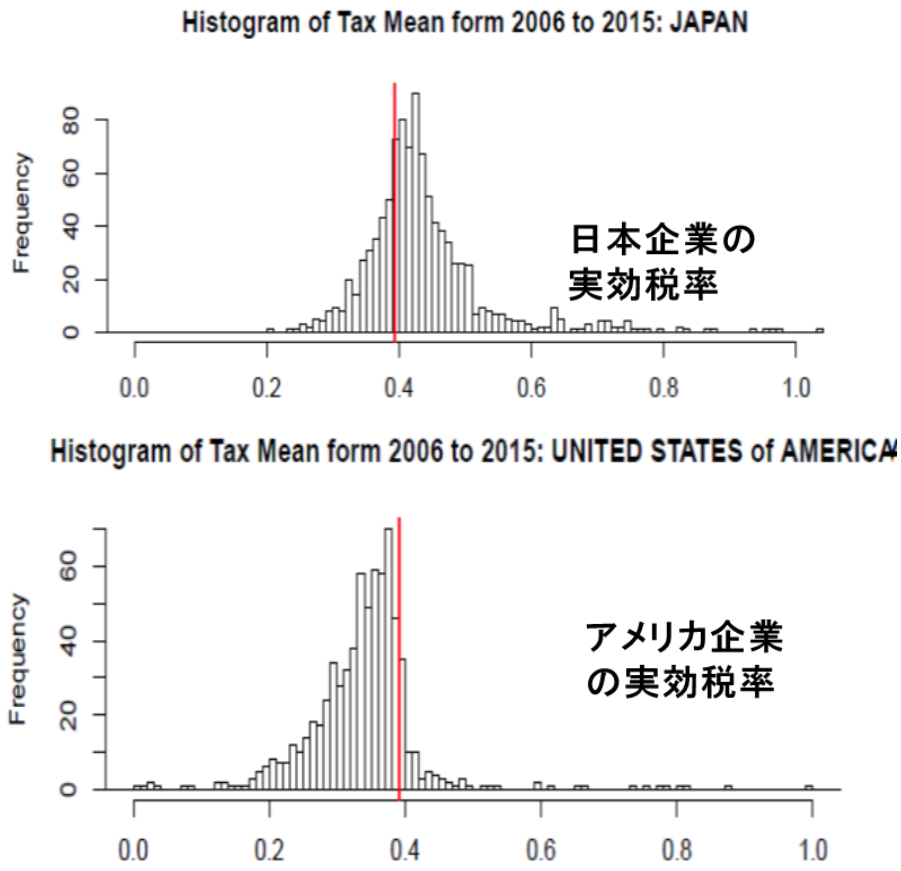




G7 countries (cont'd)

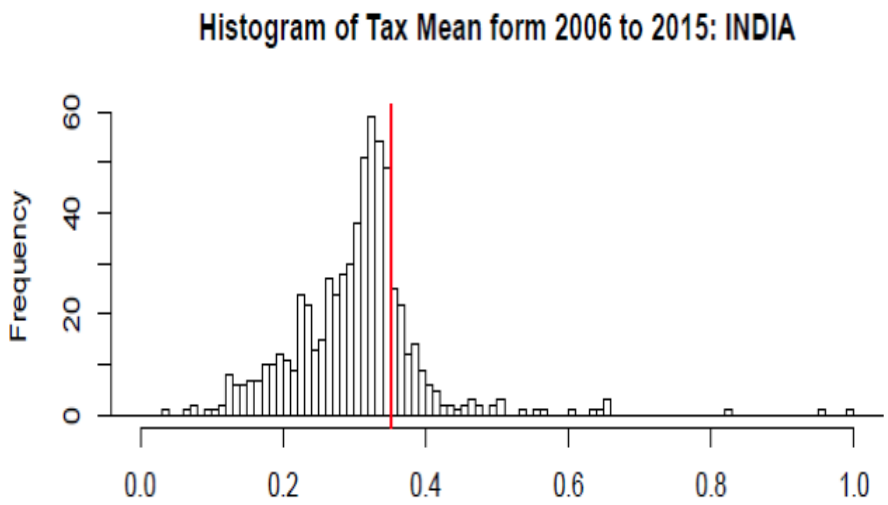


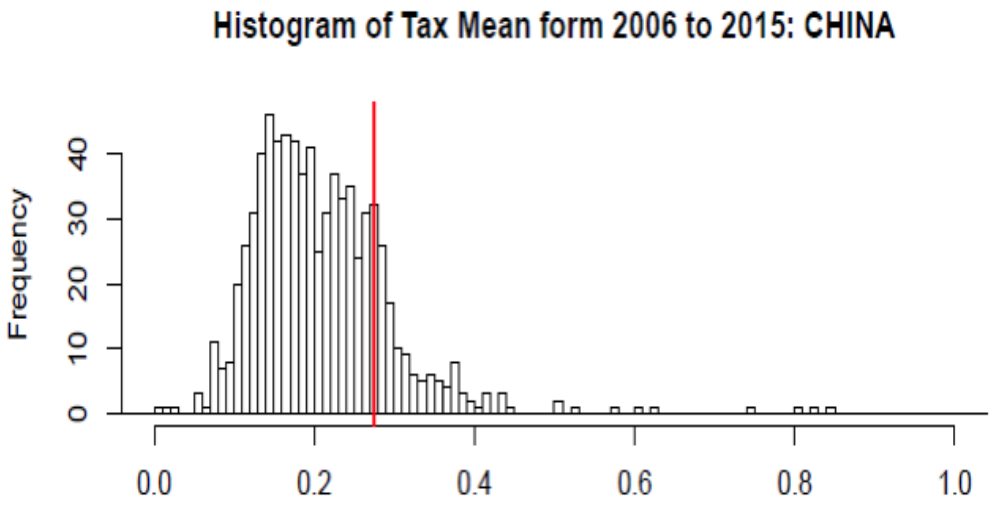
企業の10年平均の実効税率（日本、アメリカ）

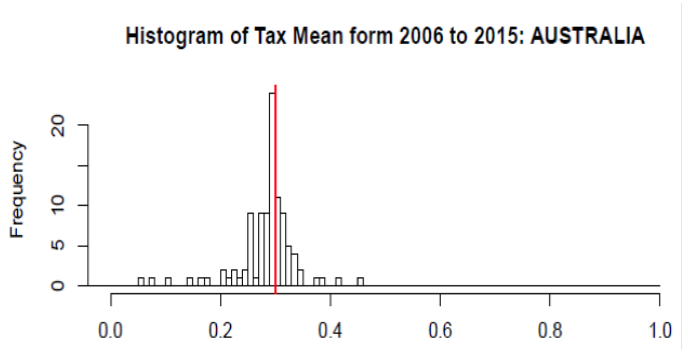


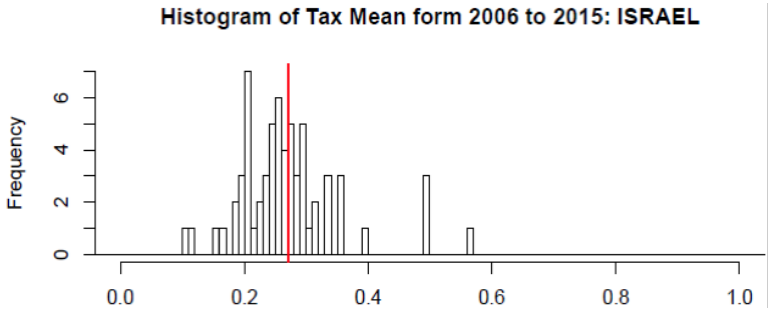
C. Saka, T. Oshika, and M. Jimichi （2019） Visualization of Tax Avoidance and Tax Rate Convergence: Exploratory Analysis of Worldscale Accounting Data, Meditari Accountancy Research

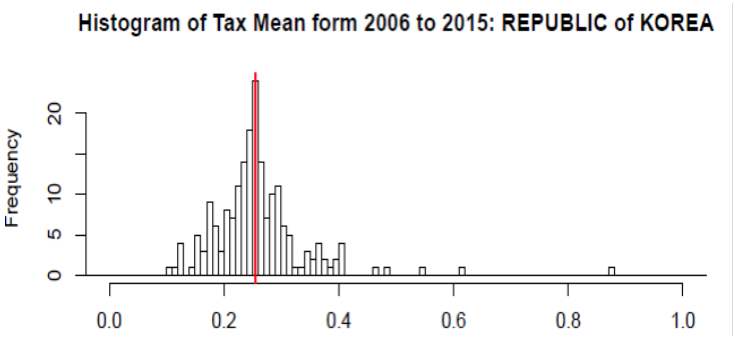
Other large countries

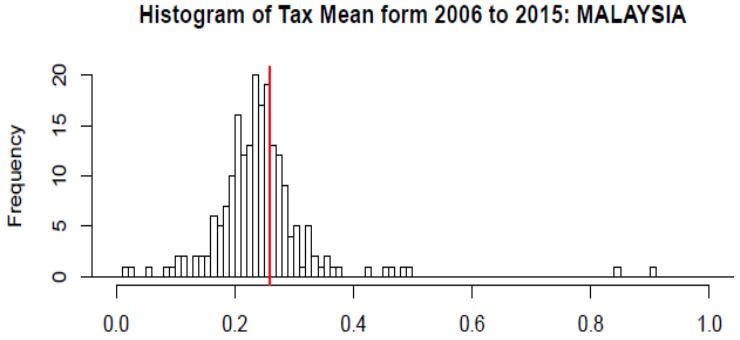


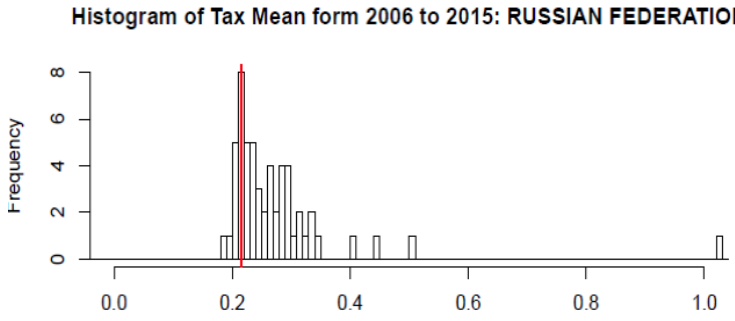


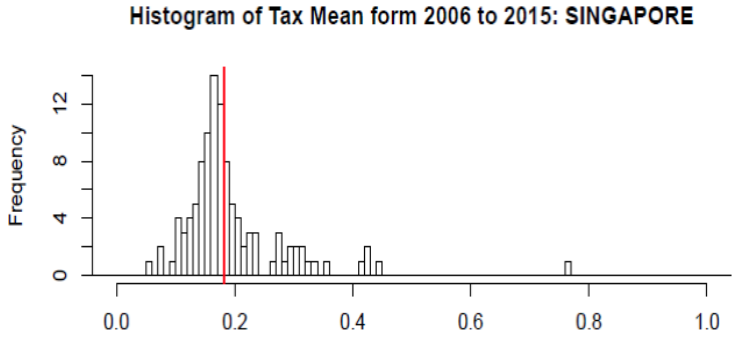

Histogram of Tax Mean form 2006 to 2015: AUSTRALIA
Frequency
0
5
10
20
0.0
0.2
0.4
0.6
0.8
1.0

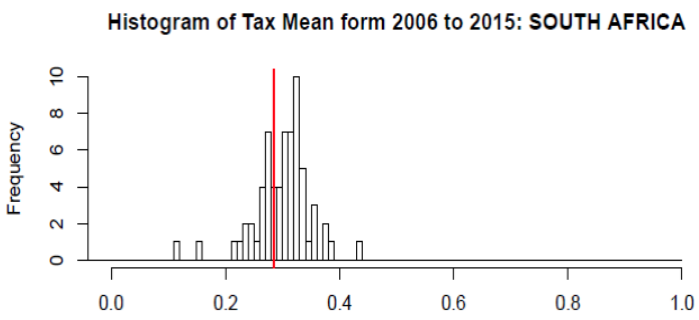

Histogram of Tax Mean form 2006 to 2015: ISRAEL
Frequency
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

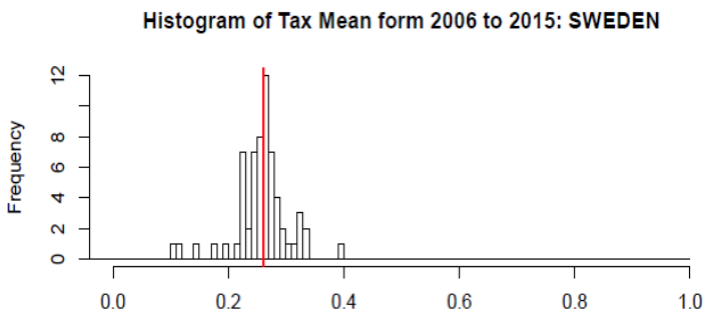

Histogram of Tax Mean form 2006 to 2015: REPUBLIC of KOREA
Frequency
0
5
10
20
0.0
0.2
0.4
0.6
0.8
1.0

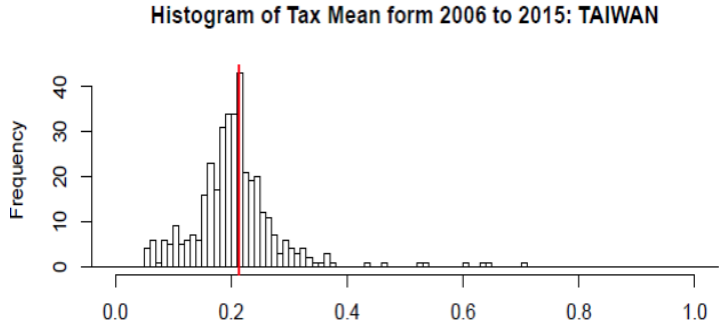

Histogram of Tax Mean form 2006 to 2015: MALAYSIA
Frequency
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0

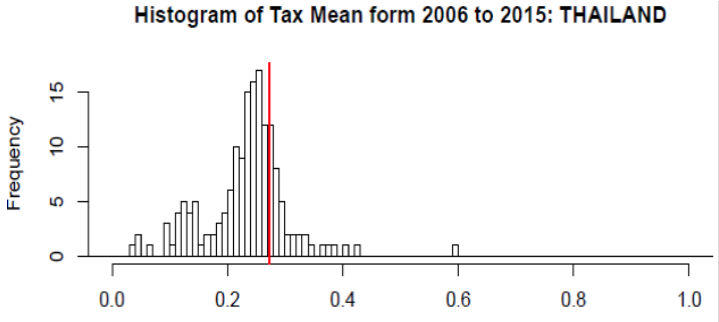

Histogram of Tax Mean form 2006 to 2015: RUSSIAN FEDERATION
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

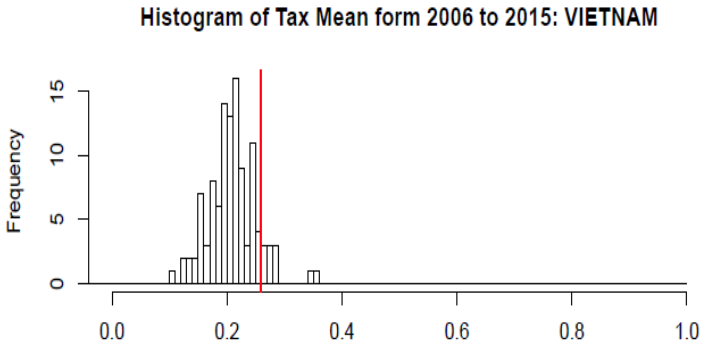

Histogram of Tax Mean form 2006 to 2015: SINGAPORE
Frequency
0
4
8
12
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of Tax Mean form 2006 to 2015: SOUTH AFRICA
Frequency
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0

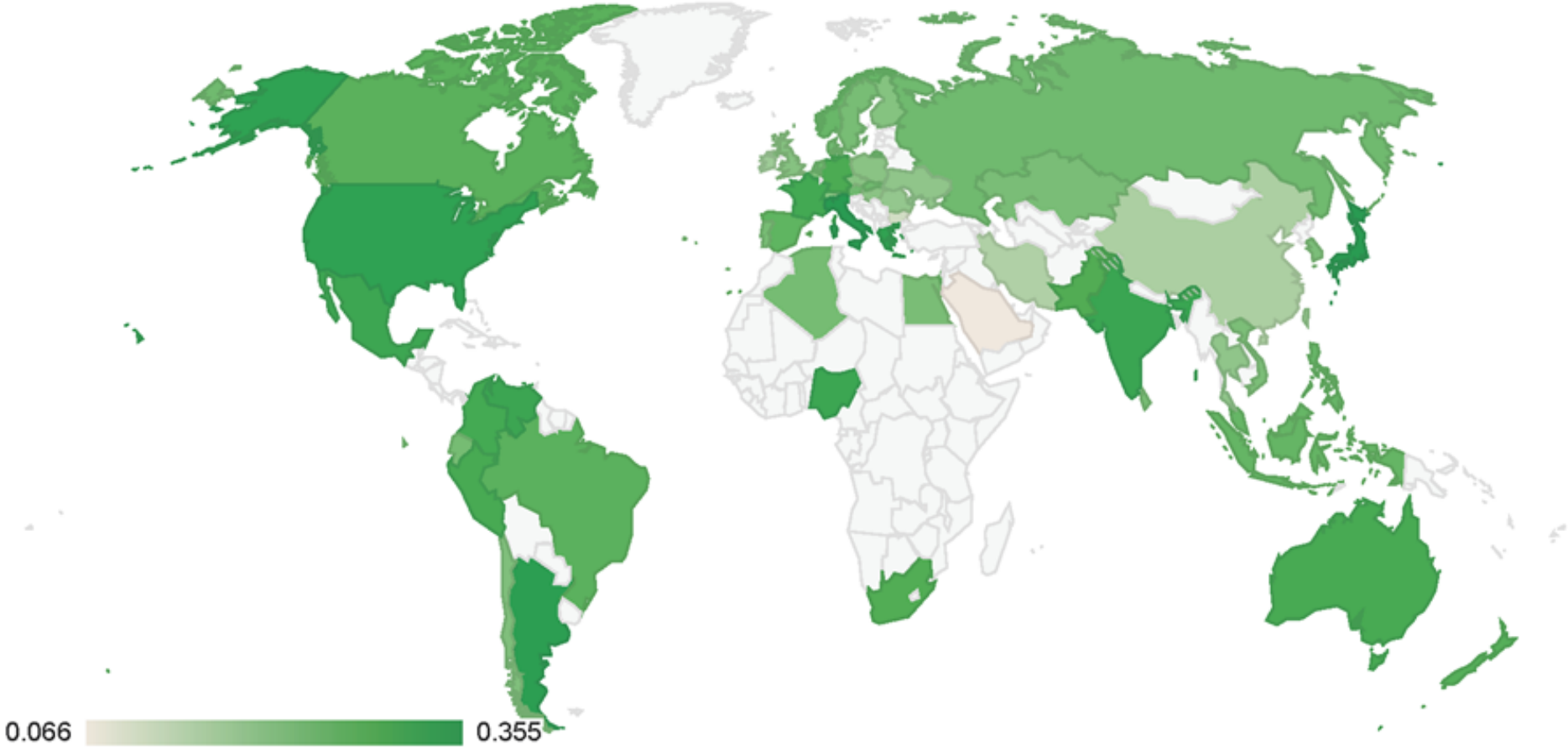

Histogram of Tax Mean form 2006 to 2015: SWEDEN
Frequency
0
2
4
6
8
12
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of Tax Mean form 2006 to 2015: TAIWAN
Frequency
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of Tax Mean form 2006 to 2015: THAILAND
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of Tax Mean form 2006 to 2015: VIETNAM
Frequency
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0

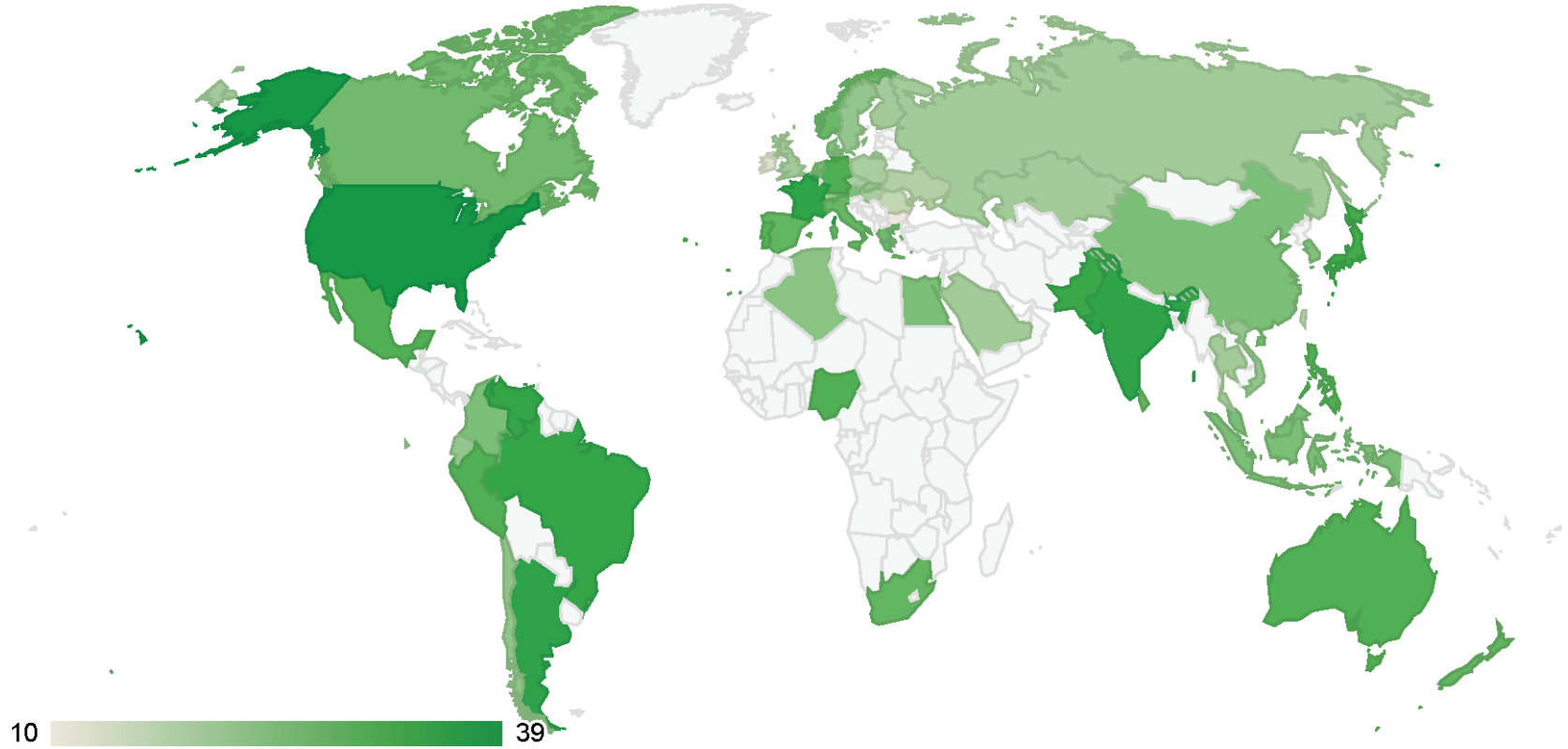
58カ国の全上場企業の実効税率（税金/利益）の中央値分布（2015年）



C. Saka, T. Oshika, and M. Jimichi （2019） Visualization of Tax Avoidance and Tax Rate Convergence: Exploratory Analysis of Worldscale Accounting Data, Meditari Accountancy Research

58カ国（法定税率のある国）の法定税率（STR）（2015年）

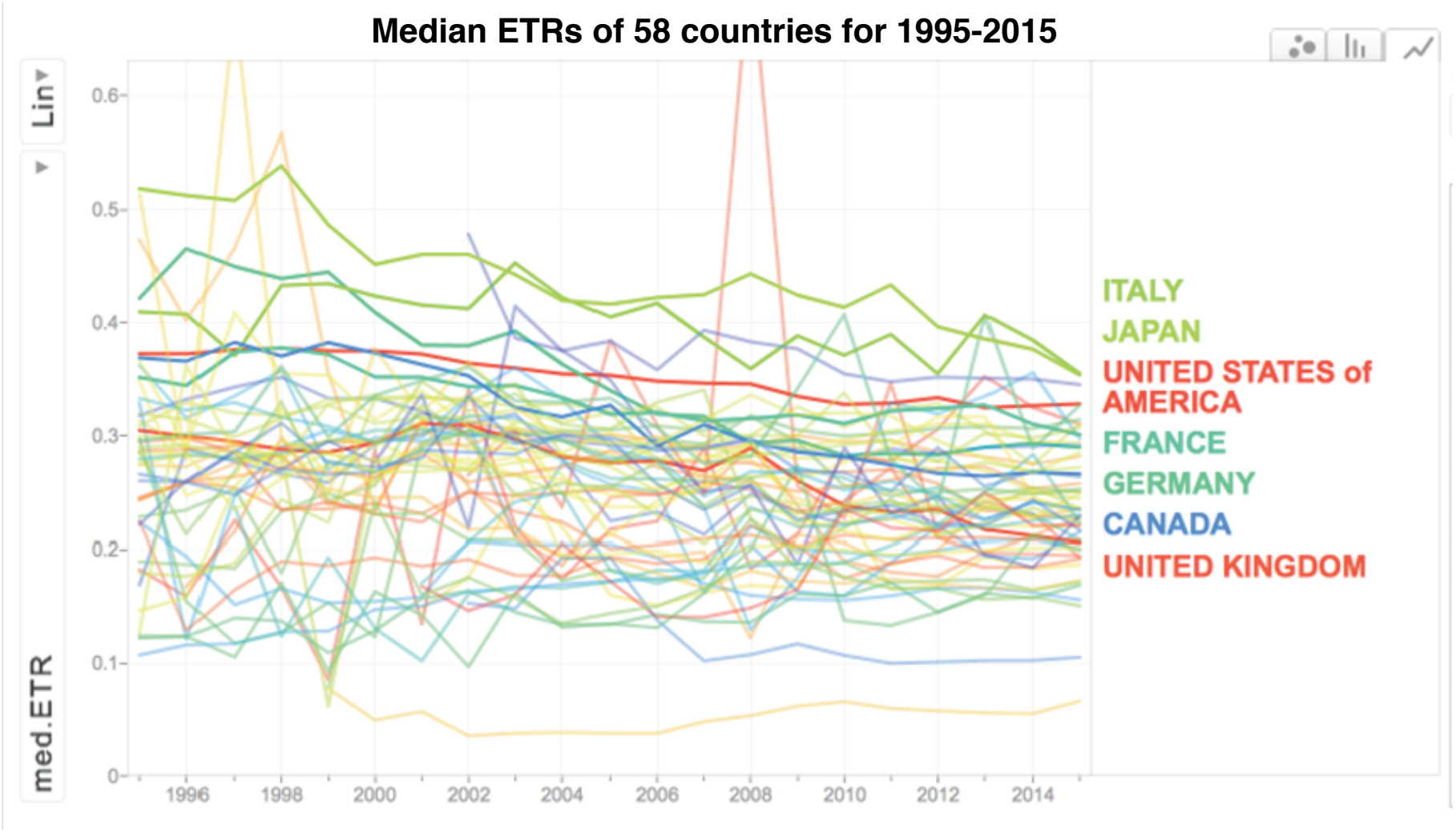
GeoChartID17369183926c6



Data: na.omit(filter(firmfin.STR.med.ETR.year.country.summary, year == • Chart ID: GeoChartID17369183926c6 • googleVis-0.6.2
Data: 2015)) • Chart ID: GeoChartID17369183926c6 • googleVis-0.6.2
R version 3.4.2 (2017-09-28) • Google Terms of Use • Documentation and Data Policy

実効税率(Effective tax rates: ETRs) ：58カ国、1995-2015年

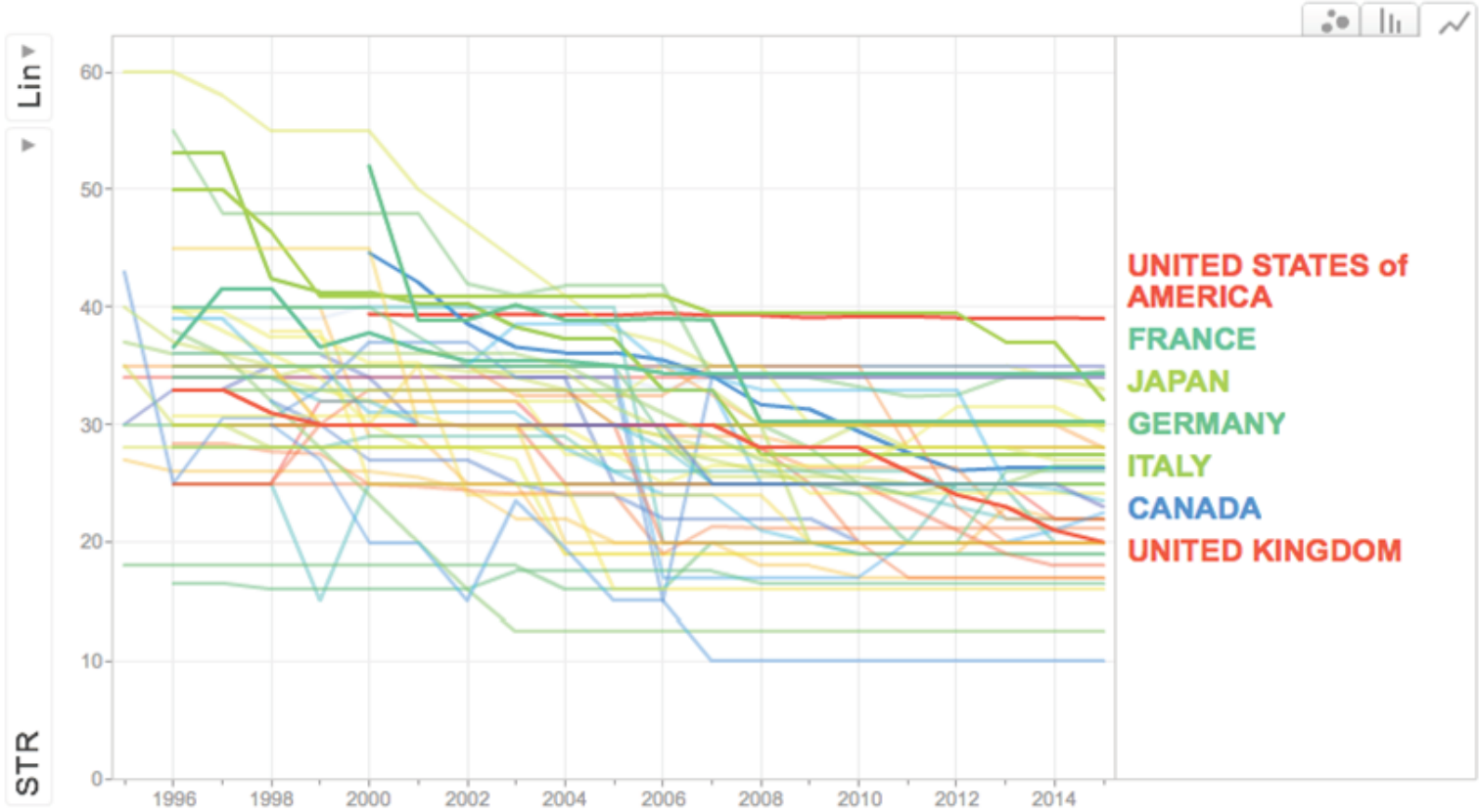
低下傾向



法定税率（Statutory Tax Rates）：58カ国、1995-2015年

➡ 低下傾向

Statutory Tax Rates of 58 countries for 1995-2015



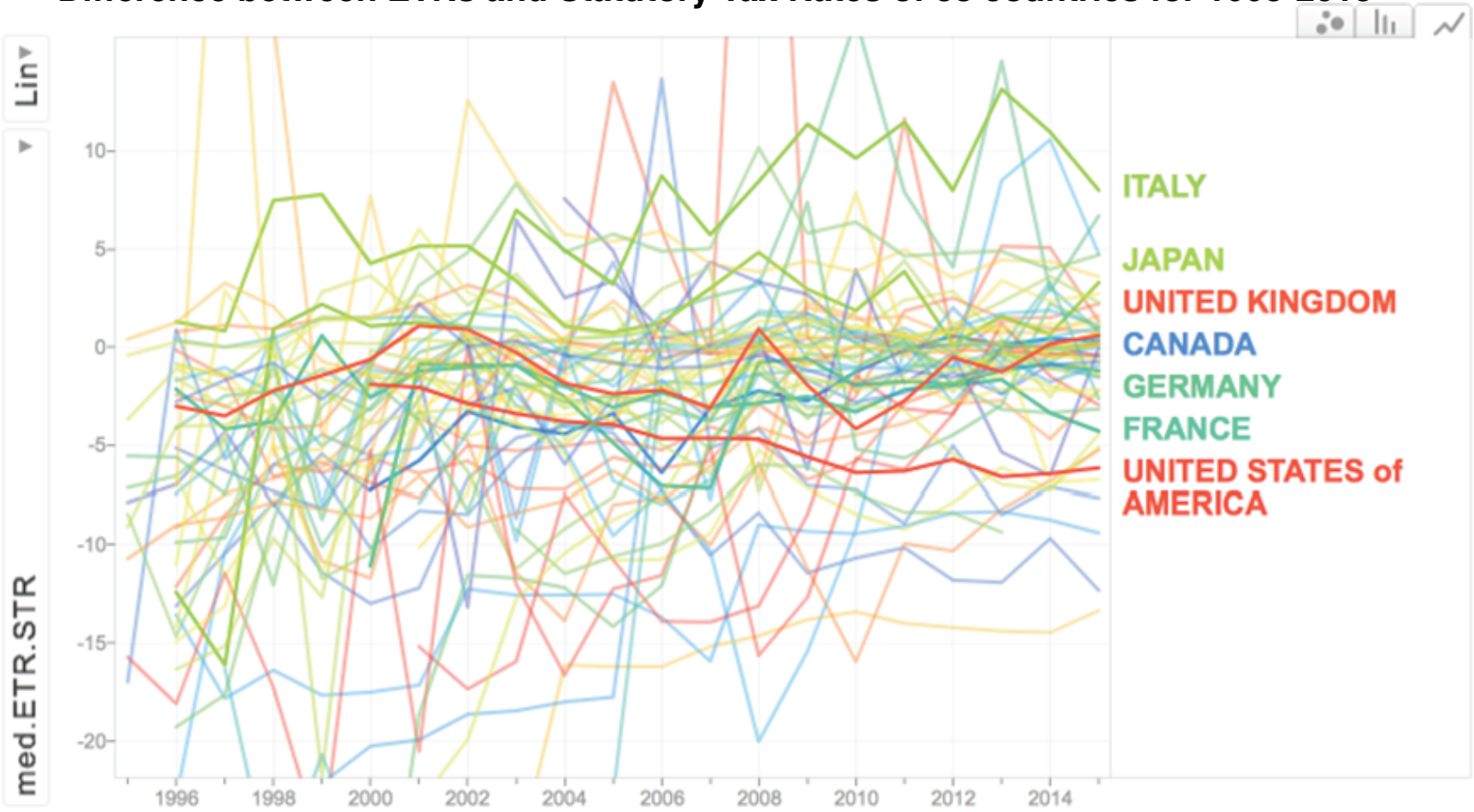
Data: EIU Market Indicators & Forecasts/Bureau van Dijk

実効税率(ETRs)ー法定税率（STRs）：58カ国、1995-2015年

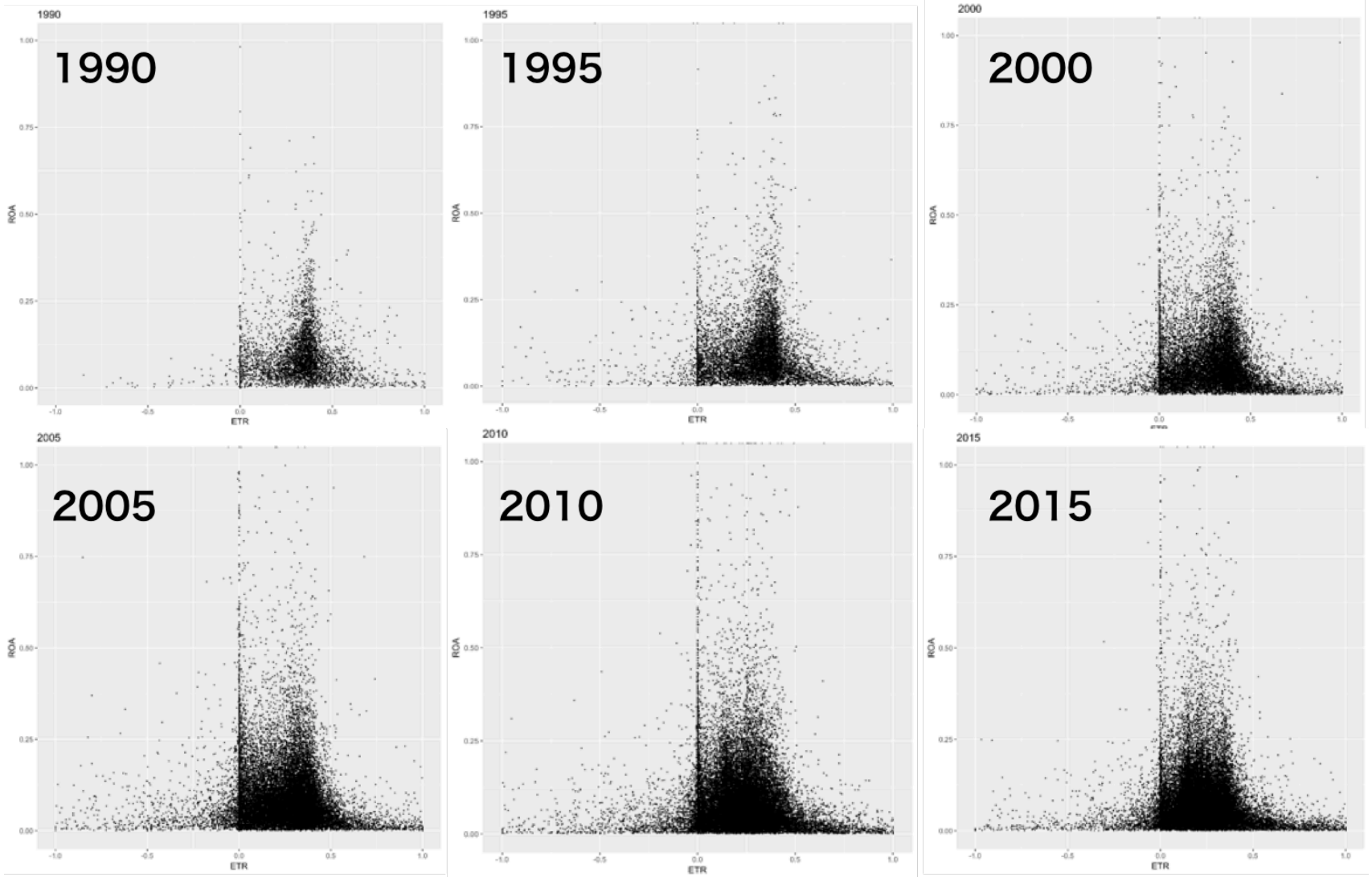
各国メディアンのメディアンは負(-1.43)、39/58カ国 (67%) のメディアンは負

➡ 租税回避行動の証拠 2

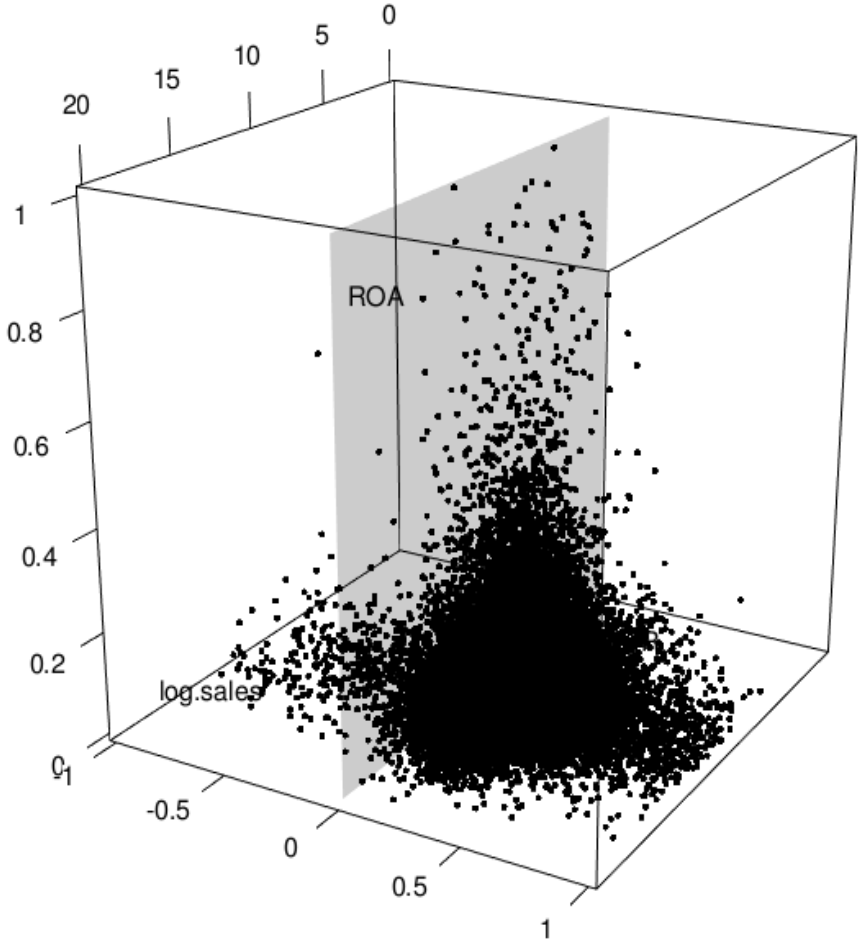
Difference between ETRs and Statutory Tax Rates of 58 countries for 1995-2015

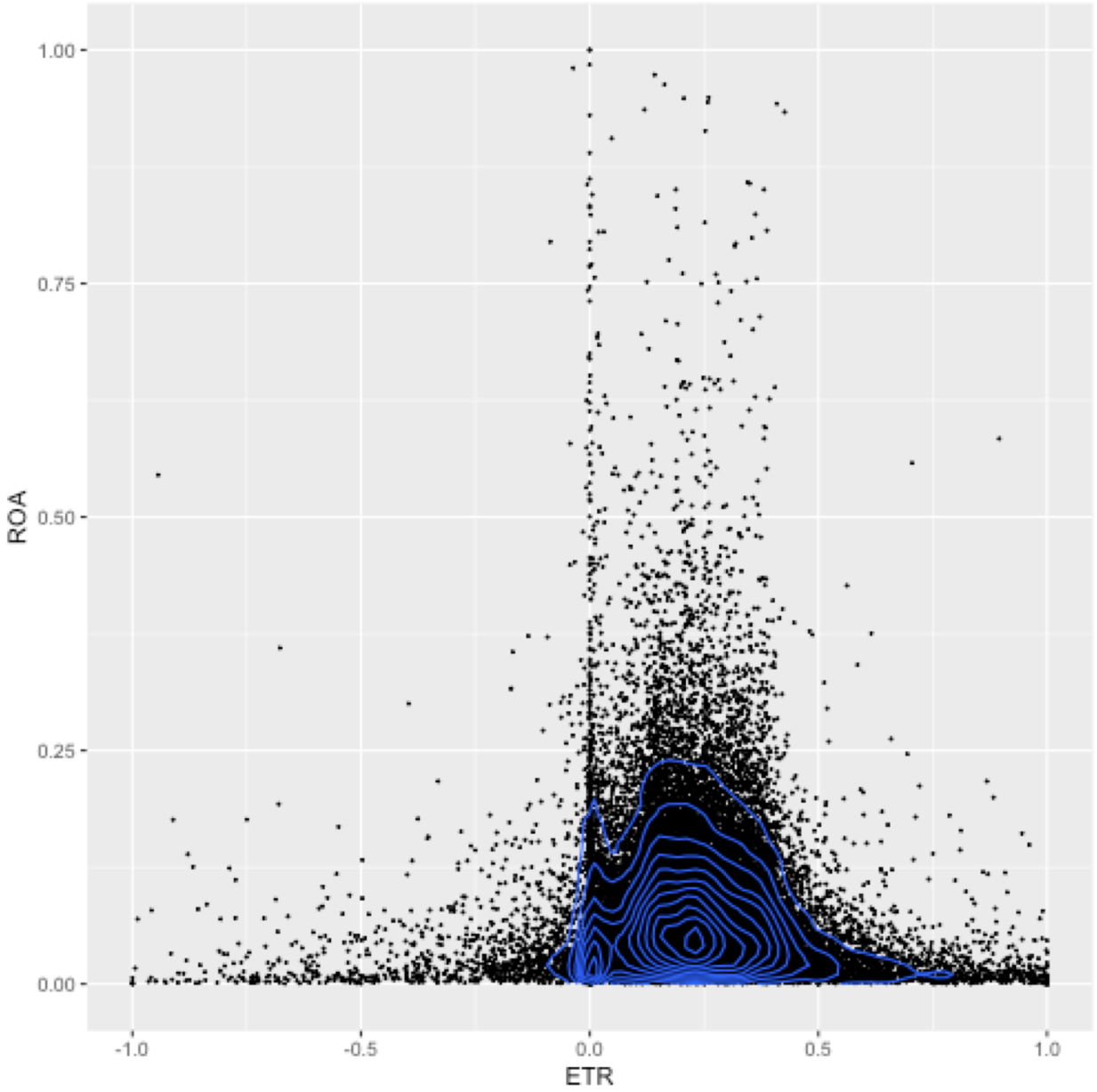


利益率（ROA、縦）、実効税率（横）の散布図（1990-2015年）



実効税率(x軸)，売上高(対数, y軸), 利益率(z軸)の散布図



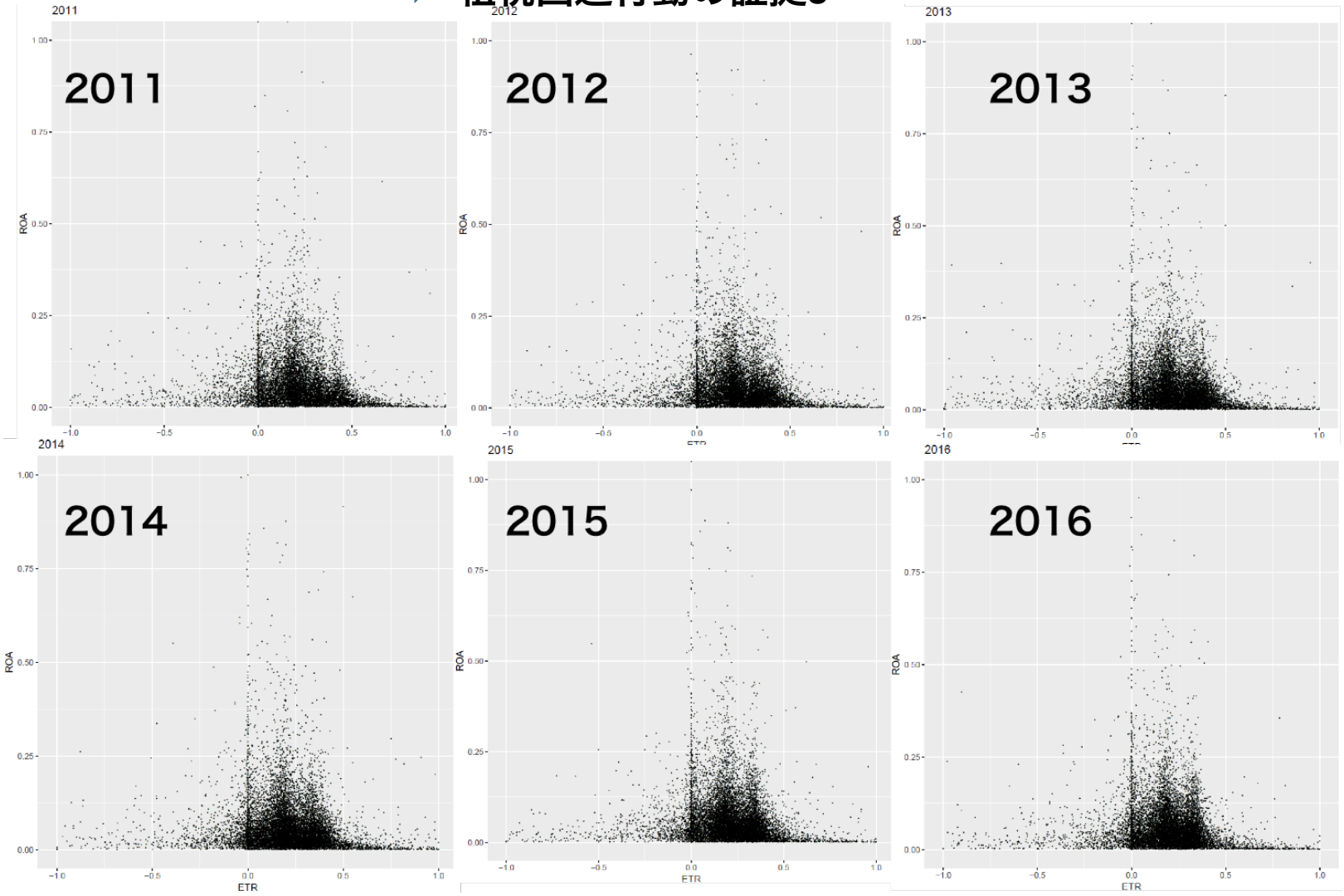


2014年 実効税率=0, 1,383社

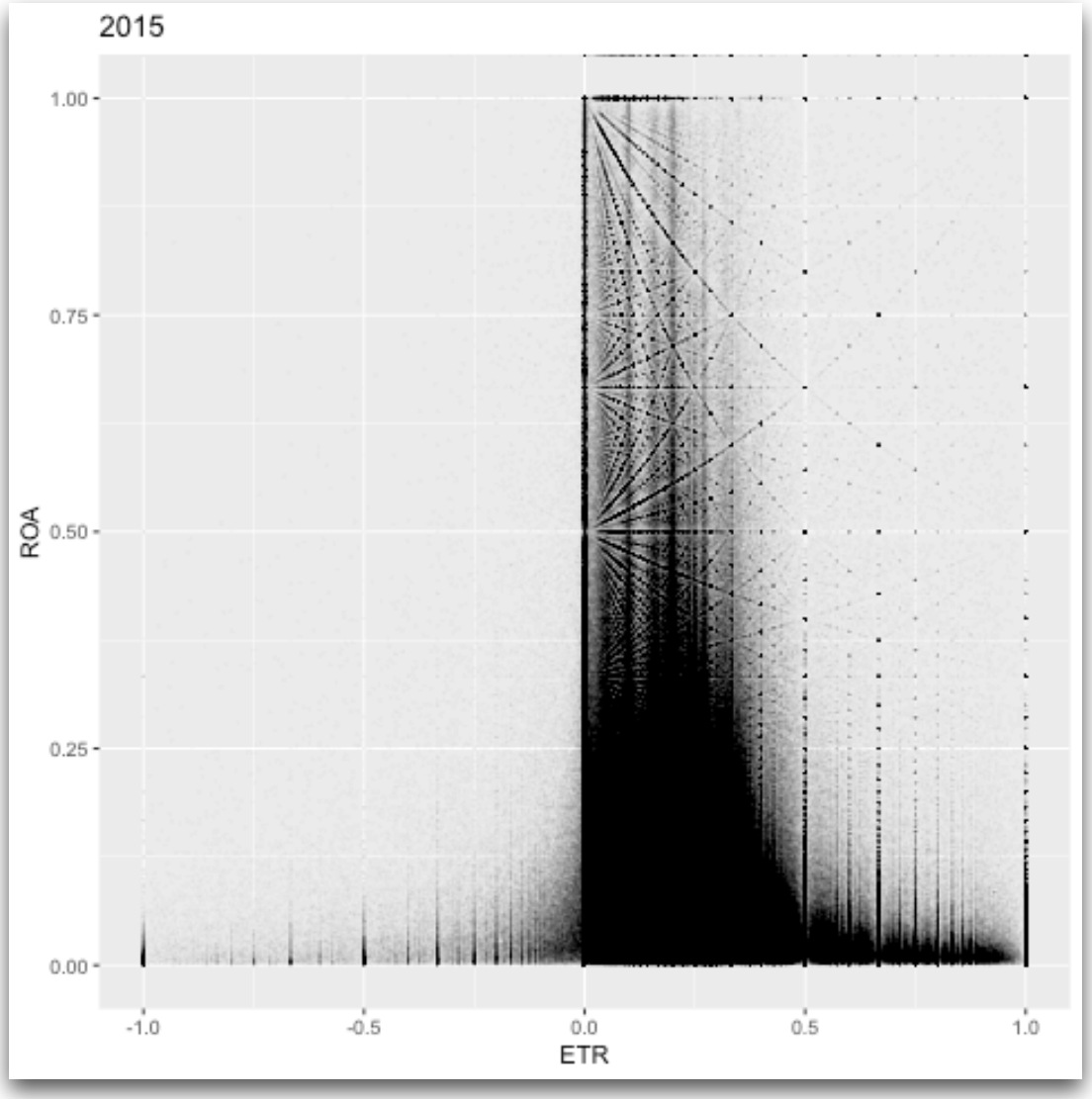
アメリカ	316社
インド	140社
イギリス	92社
カナダ	90社
オーストラリア	79社
中国	52社
ウクライナ	49社
ルーマニア	37社
イラン	30社
韓国	26社
ドイツ	25社
:	
:	
日本	0 社

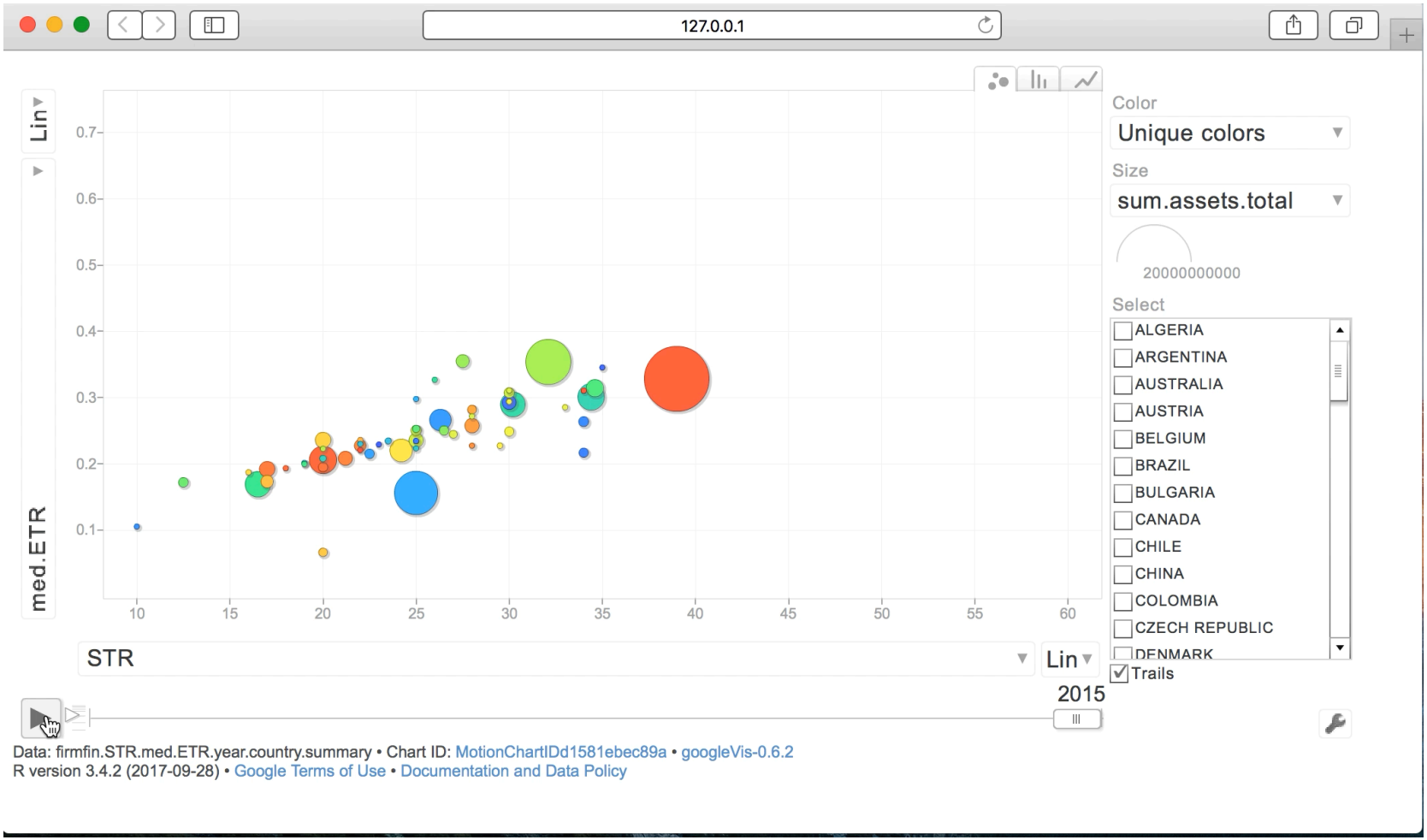
利益率（ROA）、実効税率の散布図（単体2011-2016年）

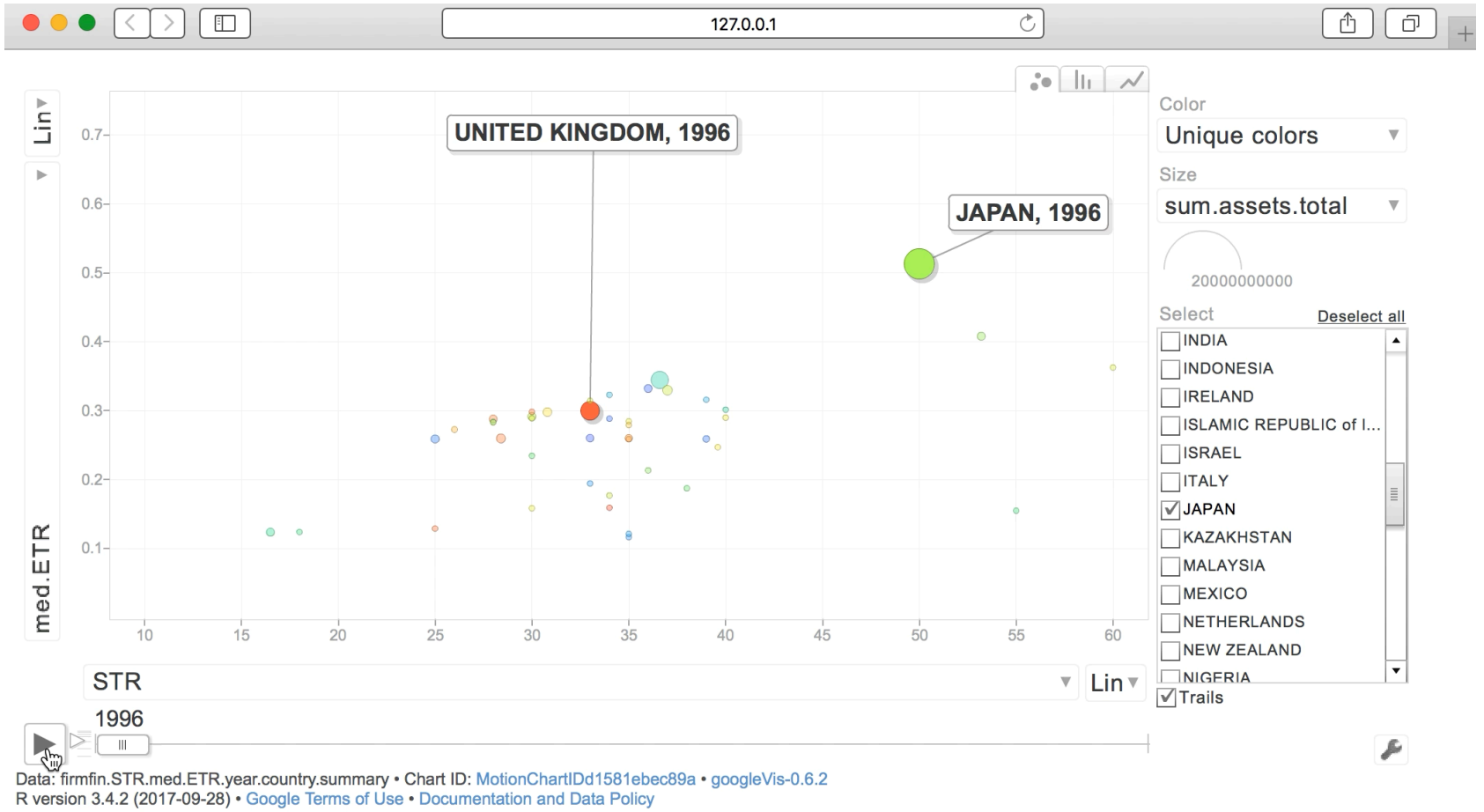
➡ 租税回避行動の証拠3



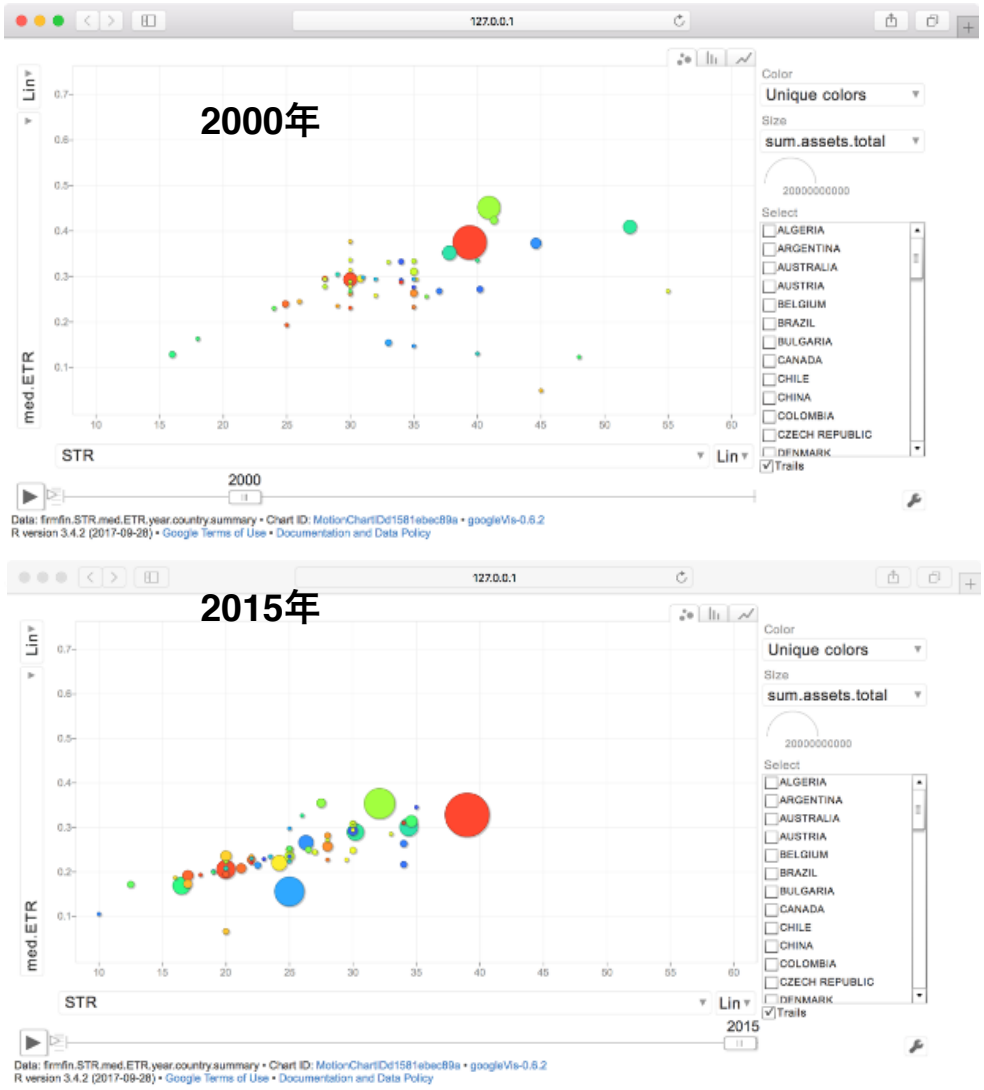
非上場企業




127.0.0.1
Lin
med.ETR
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
15
20
25
30
35
40
45
50
55
60
STR
Lin
2015
Color
Unique colors
Size
sum.assets.total
20000000000
Select
ALGERIA
ARGENTINA
AUSTRALIA
AUSTRIA
BELGIUM
BRAZIL
BULGARIA
CANADA
CHILE
CHINA
COLOMBIA
CZECH REPUBLIC
DENMARK
Trails
Data: firmfin.STR.med.ETR.year.country.summary • Chart ID: MotionChartIDd1581ebec89a • googleVis-0.6.2
R version 3.4.2 (2017-09-28) • Google Terms of Use • Documentation and Data Policy


127.0.0.1
UNITED KINGDOM, 1996
JAPAN, 1996
Lin
med.ETR
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
15
20
25
30
35
40
45
50
55
60
STR
Lin
1996
Color
Unique colors
Size
sum.assets.total
20000000000
Select
Deselect all
INDIA
INDONESIA
IRELAND
ISLAMIC REPUBLIC of I...
ISRAEL
ITALY
JAPAN
KAZAKHSTAN
MALAYSIA
MEXICO
NETHERLANDS
NEW ZEALAND
NIGERIA
Trails
Data: firmfin.STR.med.ETR.year.country.summary • Chart ID: MotionChartIDd1581ebec89a • googleVis-0.6.2
R version 3.4.2 (2017-09-28) • Google Terms of Use • Documentation and Data Policy

58か国の実効税率（y）、法定税率（x）の変化（2000～2015年）



税率の引下競争により、過去15年間に、企業の実効税率と各国法定税率が、世界規模で下方に収斂

法人課税の危機？

結果の解釈と結論

世界の企業行動の実態の可視化

1. 企業の国内・国際的格差の拡大（集中度の高まり）
2. 労働者賃金が削られ（労働分配率低下）, 投資家利益が増加
3. 企業の租税回避の横行、法人課税の危機

- 配分主体の大部分は企業
- 会計は、企業のステークホルダーへの配分の手段であるにもかかわらず、株主利益の計算システムになっている

Piketty (2013/2014)

- **「資本の民主的なコントロール」を主張**
- **「資本の民主的統制の各種形態を大きく左右するのは，参加者それぞれへの経済情報の提供だ」**
- **「企業が現在公開を求められている会計データは、労働者や一般市民が集団的な決定について意見をまとめるのには、まったく不十分なものでしかないし、まして決定に介入するほどの情報はない。」（ピケティ、2013/2014、p. 600）**

Sikka（2015）

- **会計が資本の富の増加に資する**
- **改善提案**
  1. **社会コスト会計を開発し、企業実務の社会への影響を明らかにする**
  2. **労務費を可視化する会計能力を構築し、労働者への配分が縮減している現状を明らかにする**
     → **付加価値情報開示**

- **証拠を示し、多くの人達と問題を共有して課題解決の方策を探ることは、研究者に課せられた使命**

今後の展望

- R + Apache Arrow
- R Package arrow
- Apache Parquet
- PostgreSQL + PG-Strom, Arrow_Fdw, pg2arrow

謝辞

- 本研究の一部は以下の研究費とBvDより助成を得ている. ここに感謝の意を表する.

科学研究費 基盤研究C:「グラフィカル・データ・アナリシスによる格差研究と社会環境会計による解決方法の提案」(2016年〜2018年), 課題番号: 16K04022, 研究代表者: 阪 智香

科学研究費 基盤研究C:「共有価値創造（CSV）のための社会環境会計の構築」(2019年〜2021年), 課題番号: 19K02006, 研究代表者: 阪 智香

平成30年度 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 課題: 「財務ビッグデータの可視化と統計モデリング」, 課題番号: jh181001-NWJ, 研究代表者: 地道 正行関西学院大学

平成31年度(令和元年度) 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 課題: 「財務ビッグデータの可視化と統計モデリング」, 課題番号: jh191002-NWJ, 研究代表者: 地道 正行関西学院大学

関西学院大学 研究設備費(III), 個人研究費, 図書館 図書費B

BvD 増田 歩氏

Bibliography & Information

Bibliography:
Statistics and Data Visualization

- Akaike, H. (1973) Information theory and an extension of the maximum likeli- hood principle, *Proceedings of the 2nd International Symposium on Information Theory*, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: pp. 267–281.

- Azzalini, A. with the collaboration of A. Capitanio (2014) *The Skew-Normal and Related Families*, Cambridge University Press, Institute of Mathematical Statistics Monographs.

- Chambers, J. M. and T. J. Hastie ed. (1991) *Statistical Models in S*. Chapman and Hall/CRC.

- Efron, B. and T. Hastie (2016) *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge University Press.

- Fox, J. and S. Weisbrerg (2011) *An R Companion to Applied Regression, Second edition*, Sage.

- James, G., D. Witten, T. Hastie, and R. Tibshirani (2013) *An Introduction to Statistical Learning with Applications in R*, Springer.

Bibliography:
Statistics and Data Visualization

- Jimichi, M. and S. Maeda (2014) Visualization and Statistical Modeling of Financial Data with R, Poster at *The R User Conference 2014*. `http://user2014. stat.ucla.edu/abstracts/posters/48_Jimichi.pdf`

- 地道正行 (2014)『R を利用した財務データの可視化と統計モデリング: 探索的データ解析の視点から』, 商学論究, 第61巻, 第3号, pp. 241--295.

- 地道正行 (2017)『R による対数非対称正規線形モデルによる財務データの統計モデリング』, 商学論究, 第64巻, 第5号, pp. 159-185, 関西学院大学商学研究会.

- 地道正行 (2017)『R を利用した対数非対称分布族にもとづく財務データの統計モデリング』, 経済学論究, 第71巻, 第2号, pp. 141-174, 関西学院大学経済学部研究会.

Bibliography:
Statistics and Data Visualization

- 地道正行 (2018) 『探索的財務ビッグデータ解析 –前処理、データラングリング、再現可能性-』, 商学論究 第66巻, 第1号, pp. 1--31, 関西学院大学商学研究会.

- 地道正行 (2018) 『探索的財務ビッグデータ解析 –データ可視化, 統計モデリング, モデル選択, モデル評価, 動的文書生成, 再現可能研究-』, 商学論究 第66巻, 第2号, pp. 1--41, 関西学院大学商学研究会.

- Jimichi, Masayuki, Daisuke Miyamoto, Chika Saka, Shuichi Nagata (2018) Visualization and statistical modeling of financial big data: double-log modeling with skew-symmetric error distributions, *Japanese Journal of Statistics and Data Science*, Vol. 1, No. 2.

- 地道正行 (2019)『変換による財務データの統計解析 -売上高の場合-』, 商学論究 第67巻, 第1号, pp. 27--46, 関西学院大学商学研究会.

- 地道 正行 (2020)『探索的財務ビッグデータ解析 -前処理の並列化-』商学論究 第67巻, 第3号, pp. 1--19, 関西学院大学商学研究会.

Bibliography:
Statistics and Data Visualization

- Konishi, S. and G. Kitagawa (2008) *Information Criteria and Statistical Modeling*, Springer.

- Mosteller, F. and J. W. Tukey (1977) *Data Analysis and Regression: A Second Course in Statistics,* Addison-Wesley, Reading Mass.

- Saka, C. and M. Jimichi (2017) Evidence of inequality from accounting data visualisation, *Taiwan Accounting Review*, Vol. 13, No. 2, pp. 193–234.

- Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Co.

- Unwin, A. (2015) *Graphical Data Analysis with R*, Chapman and Hall/CRC.

Bibliography:
Spark

- Karau, H., A. Konwinski, P. Wendell, and M. Zaharia (2015) *Learning Spark*, O'REILLY. (玉川 竜司訳 (2015)『初めてのSpark』, オーライリー・ジャパン.)

- Ryza, S., U. Laserson, S. Owen, and J. Wills (2016) *Advanced Analytics with Spark*, O'REILLY. (玉川 竜司訳 (2016)『Sparkによる実践データ解析』, オライリー・ジャパン.)

- 猿田 浩輔 他 (2015)『Apache Spark 入門: 動かして学ぶ最新並列分散処理フレームワーク』, 翔泳社.

- 下田 倫大 他 (2016)『詳解 Apache Spark』, 技術評論社

- Wickham, H. and G. Grolemund (2016) *R for Data Science*, O'Reilly.

Bibliography:
R, Dynamic Documents and Reproducible Research

- Gandrud, C. (2015) *Reproducible Research with R and RStudio*, *Second Edition*, CRC Press.

- Knuth, D. E. (1984) Literate Programming , *The Computer Journal*, British Computer Society, Vol. 27, No. 2, pp. 97-111.

- Leisch, F. (2002) Sweave: Dynamic generation of statistical reports using literate data analysis, In Wolfgang H¥"ardle and Bernd R¥"onz, editors, Compstat 2002 - Proceedings in Computational Statistics, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

- Xie, Y. (2015) *Dynamic Documents with R and knitr*, *Second Edition*, CRC Press.

URL

- Spark Web Page: https://spark.apache.org/
- sparklyr Web Page: https://spark.rstudio.com/index.html
- BvD Web Page: https://www.bvdinfo.com/en-gb
- 海外 浩平 (2019)『PostgreSQLは最新ハードウェアでどこまでやれるのか? ~GPUとNVMEで実現する超高速ログデータ処理基盤~』, https://www.sraoss.co.jp/event_seminar/2019/20190418_SRAOSS_seminar_PGStrom_on_ArrowFdw.pdf
- 海外 浩平 (2019)『Arrow_Fdw ~PostgreSQLで大量のログデータを処理するためのハードウェア最適化アプローチ~』20191115-PGconf.Japan https://www.slideshare.net/kaigai/20190314-pgstrom-arrowfdw