Exploratory Financial Big Data Analysis and Reproducible Research

M. Jimichi[a], D. Miyamoto[b], C. Saka[a] and S. Nagata[a]

[a] *Kwansei Gakuin University,* [b] *University of Tokyo*

Abstract

Jimichi *et al.* (2018) treats the financial data set which was extracted from the *Osiris* database system of the Bureau van Dijk KK in 2017 1). It contains financial information on over 80,000 listed firms around the world, and has 84 financial indices for 30 years. It is preprocessed by using Unix commands (e.g. `sed`, `grep`) and R, and is loaded to R (say, data wrangling [12]) by using Apache Spark$^{\mathrm{TM}2)}$ (e.g. [9]) and R packages SparkR. Jimichi *et al.* (2018) also treats data visualization (e.g. [11]) and statistical modeling (e.g. [3]) based on exploratory data analysis [10] with R. The dobule-log model with the skew-t error distribution (e.g. [2]) properly explains sales by employees and total assets (see [6]). It gives one contribution to predict reasonable sales, given the human resources and asset size of all listed firms in the world, and the model could be thought of as a kind of Cobb-Douglas type production function (e.g. [4]).

In this research, we use new data set which was re-extracted from the *Osiris* database system in 2018, and contains financial information on over 90,000 listed firms of 160 countries around the world. We verify that the results of Jimichi *et al.* (2018) can be reproduced, and also try to improve it in terms of velocity. The process of preprocessing the data (files) of this research is centrally managed by the shell script of Unix and R script. The process of documenting meaningful results in the process of exploratory data analysis was also generated dynamically by embedding R codes in LaTeX file using Sweave (see [8]). Furthermore, we will try to perform reproducible research by writing a script in `Makefile` and automatically executing Unix `make` command to execute all these steps (See Figure 1).

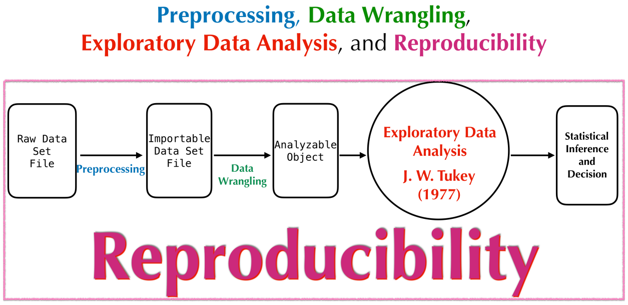
Preprocessing, Data Wrangling,
Exploratory Data Analysis, and Reproducibility
Raw Data Set File
Preprocessing
Importable Data Set File
Data Wrangling
Analyzable Object
Exploratory Data Analysis
J. W. Tukey (1977)
Statistical Inference and Decision
Reproducibility

Fig. 1 Reproducible Workflow

References

[1] Azzalini, A. (1985) A class of distributions which includes the normal Ones, *Scandinavian Journal of Statistics*, Vol. 12, No. 2, pp. 171–178.

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

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

[4] Cobb, C. W. and P. H. Douglas (1928) A theory of production, *American Economic Review*, Vol. 18, pp. 139–165.

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

[6] Jimichi, M., D. Miyamoto, C. Saka, and S. 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, pp. 347–371, `https://doi.org/10.1007/s42081-018-0019-1`

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

[8] Leisch, F. (2002) *Sweave: Dynamic generation of statistical reports using literate data analysis*, In Wolfgang Härdle and Bernd Rönz, editors, Compstat 2002 - Proceedings in Computational Statistics, pp. 575-580. Physica Verlag, Heidelberg. ISBN 3-7908-1517-9.

[9] Ryza, S., U. Laserson, S. Owen, and J. Wills (2016) *Advanced Analytics with Spark*, O'REILLY.

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

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

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

Acknowledgements

This work is partially supported by a Grant-in-Aid for Scientific Research (KAKENHI: No. 16K04022, and 19K02006) and the Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN Project ID: jh171002-NWJ, jh181001-NWJ, and jh191002-NWJ) in Japan. We would like to thank Mr. Ayumu Masuda of Bureau van Dijk KK for extracting some data set files from the *Osiris* database system.

1) `https://www.bvdinfo.com/`

2) `http://spark.apache.org/docs/latest/index.html`