International Journal of Open Source Software and Processes
Volume 12 • Issue 1 • January-March 2021

A Survey of Open Source Statistical Software (OSSS) and Their Data Processing Functionalities

Gao Niu, Bryant University, USA
Richard S. Segall, Arkansas State University, USA
Zichen Zhao, Yale University, USA
Zhijian Wu, New York University, USA

ABSTRACT

This paper discusses the definitions of open source software, free software and freeware, and the concept of big data. The authors then introduce R and Python as the two most popular open source statistical software (OSSS). Additional OSSS, such as JASP, PSPP, GRETL, SOFA Statistics, Octave, KNIME, and Scilab, are also introduced in this paper with function descriptions and modeling examples. They further discuss OSSS's capability in artificial intelligence application and modeling and Popular OSSS-based machine learning libraries and systems. The paper intends to provide a reference for readers to make proper selections of open source software when statistical analysis tasks are needed. In addition, working platform and selective numerical, descriptive and analysis examples are provided for each software. Readers could have a direct and in-depth understanding of each software and its functional highlights.

KEYWORDS

Artificial Intelligence, General Public License, Graphical User Interface, Integrated Development Environment, Machine Learning, Open Source Software, Python, R, Statistical Software

1. INTRODUCTION

In this paper, the authors discuss the most popular open source statistical software with its creation history, target practitioners, and statistical usage examples. Although Programming languages such as Java, C++ can also perform statistical analysis with intensive coding, the authors limit discussion to the software specifically designed for statistical analysis.

The motivation of this paper started from a research insight book of Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities (Segall & Niu, 2020), edited by the first two authors of this book. The book introduces OSSS, presents multiple applications and discusses research opportunities. This paper summarizes the information, extends the discussion to

DOI: 10.4018/IJOSSP.2021010101

a broader statistical processing functionality such as machine learning and artificial intelligence. We first introduced an artificial intelligence (AI) techniques categorization, and surveyed the popular OSSS designed for AI applications.

The objective of this paper is to create a reference for the readers and guide them to make proper selection of open source software when a statistical analysis task is in demand. The discussion includes the background information, research areas that the software designed for, and the overview of how to use the software.

First section of the paper introduces the definition OSSS, several similar type of software such as Free Software and Freeware are compared, both traditional software and open source software development are summarized. Second section of the paper presents multiple popular OSSS, such as R, Python and etc., designed for statistical applications are presented. Third section of the paper presents OSSS designed for AI are presented. Popular AI techniques are categorized and briefly described, and OSSS designed for AI processing are presented. The authors focus on creating an overview of all open source statistical software in this paper.

The article introduces current machine learning data processing platform. Readers can be benefitted from this short reference of Open Source Statistical Software.

2. BACKGROUND

2.1 How Open Source Software, Free Software, And Freeware Differ

2.1.1 Open Source Software (OSS)

Open Source Software (OSS) is a type of computer software in which source code is released under a license in which the copyright holder grants users the rights to study, change, and distribute the software to anyone and for any purpose. (Wikipedia (2019a))

For software to be considered "Open Source", it must meet ten conditions as defined by the Open Source Initiative (OSI). Of these ten conditions, it's the first three that are really at the core of Open Source and differentiates it from other software. These three conditions are according to the Open Source Initiative (2007):

1. **Free Redistribution**: The software can be freely given away or sold.
2. **Source Code**: The source code must either be included or freely obtainable.
3. **Derived Works**: Redistribution of modifications must be allowed.

The other conditions are: (Open Source Initiative (2007))

4. **Integrity of The Author's Source Code**: Licenses may require that modifications are redistributed only as patches.
5. **No Discrimination against Persons or Groups**: no one can be locked out.
6. **No Discrimination against Fields of Endeavor**: commercial users cannot be excluded.
7. **Distribution of License**: The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.
8. **License Must Not Be Specific to a Product**: the program cannot be licensed only as part of a larger distribution.
9. **License Must Not Restrict Other Software**: the license cannot insist that any other software it is distributed with must also be open source.
10. **License Must Be Technology**: Neutral: no click-wrap licenses or other medium-specific ways of accepting the license must be required.