CONSIDERING JULIA LANGUAGE FOR TEACHING STATISTICS COURSES

Nikola Kaspříková

Abstract

The paper discusses the using of Julia programming language in teaching applied statistics and data analysis courses. Julia is similar to other languages which may be used for data analysis and statistics and which just like Julia provide an interactive session interface, such as R or Python. But in comparison with such languages, Julia may achieve better performance due to on-the-fly compilation of the code. The paper also presents the recent trends in general popularity of Julia programming language, showing data from the Google Trends or TIOBE index. Furthermore, the paper reports the results form searches for annotations including Julia language in several public university course catalogues. Finally, a case study shows how Julia can be used for solving a task in a field of statistical quality control. The operating characteristic function of an acceptance sampling plan is programmed in Julia and the function is visualized using the packages available for the Julia language.

Key words: teaching statistics, statistical software, Julia

JEL Code: C44, C80

Introduction

This paper discusses the using of Julia programming language (Bezanson et al., 2017) in teaching applied statistics and data analysis courses. This modern language has recently become one of the computing environments used for statistical calculations and data analysis. It is an open source programming language available under the MIT license. Just like in R or Python, the users may work with Julia language in an interactive session. But in comparison with such languages, Julia may achieve better performance due to on-the-fly compilation of the code. For more details, see (Shah et al, 2013), (Bezanson et al., 2018) and (Gao et al., 2020).

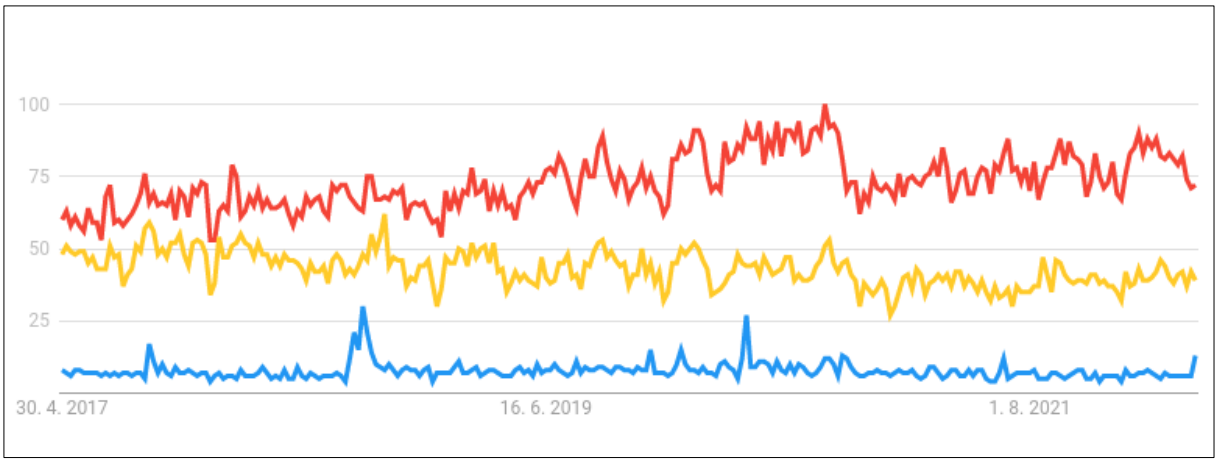
The structure remaining parts of this paper is as follows. Chapter 1 reports the recent trends in general popularity of Julia programming language, showing data from the Google Trends or TIOBE index. Chapter 2 presents the current state of using Julia for teaching data

analysis and statistics. The results of the searches for Julia in annotations in public course catalogues of Czech universities are shown. Chapter 3 presents a case study, which shows using Julia for solving a task in a field of statistical quality control. The operating characteristic function of an acceptance sampling plan is programmed in Julia and the function is visualized using the packages available for the Julia language.

1 Popularity of Julia language

The Figure 1 shows the popularity of three search terms (Python language, R language, Julia language) in Google Trends within the last five years. The maximum number of weekly searches is assigned the value 100 and the other values are shown relative to the maximum, for example value 80 reflects 80 percent of the maximum number of searches. The frequencies of the three search terms may reflect the popularity of Python, R and Julia languages. The search term "Python language" is still the most popular, "Julia language" is the last popular and the popularity of "R language" search term is approximately just in the middle between "Python language" and "Julia language".

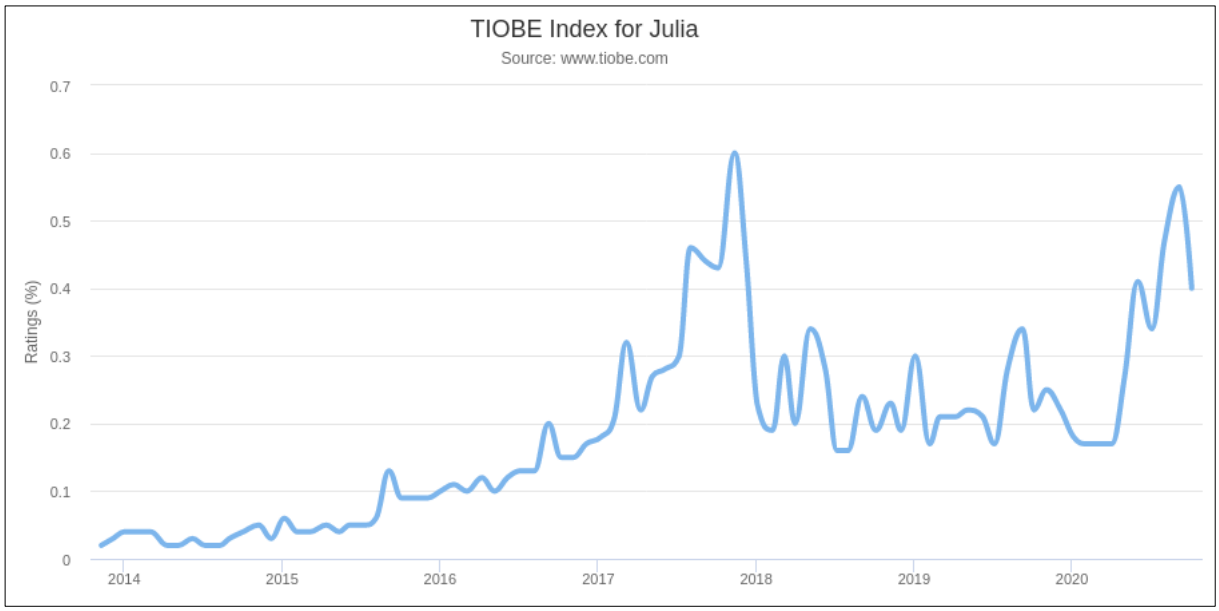
Fig. 1: Python (red), R (yellow) and Julia (blue) in Google Trends



Source: Google Trends

Another measure of popularity of the programming language which may be used is the TIOBE index (Tiobe.com, 2022). The ratings in this index are calculated based on number of hits for the search queries of the type "language programming", where "language" stands for the specific programming language in 25 search engines. The current (April 2022) Julia position in TIOBE index is 26 with rating 0.44%, which is higher than Fortran or Rust. The current position of Python is 1 with rating 13.92% and the position of R (R Core Team, 2022) is 11 (rating 1.55%). The plot of Julia TIOBE ratings in recent years is shown in Figure 2.

Fig. 2: Julia ratings in TIOBE index



Source: www.tiobe.com/tiobe-index/julia/

2 Julia in teaching statistics courses

The list Julia in the classroom (Bezanson, 2022) from the Julia project website presents a list of courses from the universities worldwide in which Julia language is used in teaching classes in various fields, including applied mathematics, statistics and data analysis.

The sample of the courses listed in the Julia in the classroom list is shown in Table 1.

The list Julia in the classroom also includes two courses taught at the Czech Technical University, namely the course Julia for Optimization and Learning and the course Scientific Programming in Julia.

The course catalogue of the Czech Technical University includes the two courses listed in the Julia in the classroom list and one course called Machine learning in Julia computing environment (course in Czech).

The search for Julia in the course catalogues of several other large Czech universities (Charles University, Brno University of Technology, University of West Bohemia) did not return any course which would refer to the Julia language in its name or syllabus.

Tab. 1: Sample of courses in data analysis and statistics making use of Julia

University	Course title
University of California, Los Angeles	Computational Methods for Biostatistical Research
University of Glasgow	An Introduction to Julia, course of Online Master of Science (MSc) in Data Analytics
Iowa State University	Topics in Statistical Computing: Julia Seminar
Northeastern University	Applied Probability and Statistics
Warsaw School of Economics	Statistical Learning Methods
TU Dortmund	One week introductory course into Julia with applications in statistics and economics
Oregon State University	Introduction to Data Science for Engineers
Rice University	Environmental Data Science
University of Canterbury	The Trustworthy Data Scientist
University of Central Florida	Statistical Learning Theory

Source: julialang.org/learning/classes/

3 Case study: Programming and visualization of an acceptance sampling plan operating characteristic

The operating characteristic of a single sampling acceptance sampling plan is the function which gives the probability of accepting a lot with proportion of nonconforming items in the lot p when sampling plan with sample size n and critical value k is used. The operating characteristic function of a single sampling plan for the inspection by variables under the assumption that the standard deviation of the normally distributed quality characteristic of interest is known is (Luca et al., 2022):

$$L(p,n,k) = \phi\left(\sqrt{n}\left(u_{1-p} - k\right)\right) \qquad (1)$$

The case of known standard deviation is discussed in (Johnson & Welch, 1940) and (Jennett & Welch, 1939).

3.1 Programming the operating characteristic function

Table 2 shows the commands needed to install an extension package called Distributions and make it available in Julia language. The string "julia>" is the prompt in all the code tables below.

Tab. 2: Installing an extension package in Julia

```
julia> import Pkg; Pkg.add("Distributions")
julia> using Distributions
```

Source: code produced by the author in Julia

The tools for package installation are loaded first using command import Pkg. Then the call to function Pkg.add is used to install the package specified as its argument. The second line of code in Table 2 makes the newly installed package ready to use, i. e. the functions defined in the package are made available for calls.

Tab. 3: Programming the operating characteristic function

```
julia> zp(p)=quantile.(Normal(), [1-p])[1]
julia> L(p,n,k)=cdf(Normal(),sqrt(n)*(zp(p)-k))
```

Source: code produced by the author in Julia

The code in Table 3 may be used to program the operating characteristic function L as a function of three variables.

3.2 Producing the plot of the operating characteristic curve

The code in Table 4 may be used to visualize the operating characteristic (OC) curve for various values of the proportion defective p and when selected acceptance sampling plan (n, k) is used.

Tab. 4: Producing an OC plot with Gadfly in Julia

```
julia> import Pkg; Pkg.add("Gadfly")
julia> using Gadfly
julia> L2(p)=L(p,26,2.7)
julia> plot(L2,0.0001,0.1,Guide.xlabel("p"), Guide.xlabel("p"))
```

Source: code produced by the author in Julia

The code includes the calls needed for the Gadfly package installation. Then the Gadfly package functions are made available and then the L2 function is defined as a single variable function. The values of sample size n and the critical value k are fixed in L2 function and the only variable is the proportion of nonconforming items p in the lot.

The Figure 2 shows the plot of the operating characteristic function produced in Julia using the call listed in Table 3.

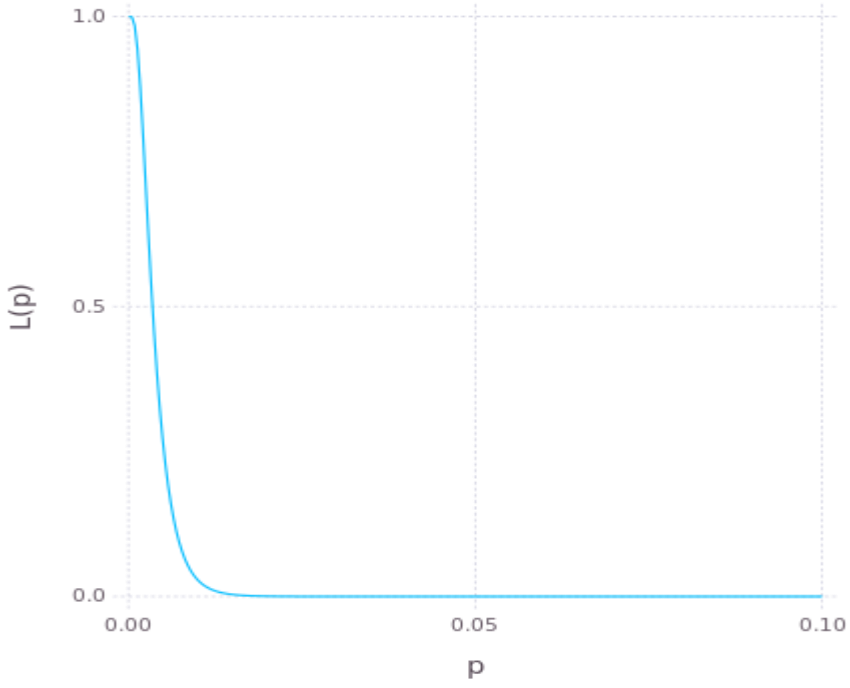
Tab. 5: Producing a plot to a PNG file

```
julia>p = plot(L2,0.0001,0.1, Guide.xlabel("p"), Guide.ylabel("L(p)"))
julia> obr=PNG("OC.png", 12cm, 12cm)
julia> draw(obr,p)
julia> p |> PNG("OC.png", 12cm, 12cm)
```

Source: code produced by the author in Julia

The code in Table 5 shows how the plot of the operating characteristic curve may be produced with a PNG file output. The last line of the code in Table 5 may be used in place of the second and third line, making use of the Julia's function chaining syntax.

Fig. 3: The operating characteristic function curve



Source: plot produced by the author in Julia

Conclusion

Regarding general popularity of Julia programming language, the data from the Google Trends as well as the data from TIOBE index suggest that Julia is still not reaching the popularity of Python or R.

Julia is already used for teaching data analysis and statistics courses at many top-class universities. The results of the searches for Julia in annotations in public course catalogues of several Czech universities have not found many cases of using Julia in teaching. The Czech Technical University has three courses related to Julia.

From a case study focused on programming and visualization of an operating characteristic of an acceptance sampling plan presented in Chapter 3 it follows that Julia can be easily used for statistical programming and for producing of the plots.

Considering the above mentioned facts, it seems that Julia as a modern and open source software may be a suitable tool to be used in teaching of statistics and data analysis courses. If the integration with other, more popular tools is needed, the Julia language can be integrated with other statistical software. For example, the JuliaConnectoR (Lenz et al., 2022) package provides an interface for integrating Julia in R.

Acknowledgement

This paper has been produced with support from the Prague University of Economics and Business.

References

Bezanson, J., Chen, J., Chung, B., Karpinski, S., Shah, V. B., Vitek, J., & Zoubritzky, L. (2018). Julia: Dynamism and performance reconciled by Design. Proceedings of the ACM on Programming Languages, 2(OOPSLA), 1–23. https://doi.org/10.1145/3276490

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. SIAM Review, 59(1), 65–98. https://doi.org/10.1137/141000671

Bezanson, J. (2022). Julia in the classroom. The Julia Programming Language. Retrieved April 20, 2022, from https://julialang.org/learning/classes/

Gao, K., Mei, G., Piccialli, F., Cuomo, S., Tu, J., & Huo, Z. (2020). Julia language in Machine learning: Algorithms, applications, and open issues. Computer Science Review, 37, 100254. https://doi.org/10.1016/j.cosrev.2020.100254

Jennett, W. J., & Welch, B. L. (1939). The control of proportion defective as judged by a single quality characteristic varying on a continuous scale. Supplement to the Journal of the Royal Statistical Society, 6(1), 80. https://doi.org/10.2307/2983626

Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central T-distribution. Biometrika, 31(3/4), 362. https://doi.org/10.2307/2332616

Lenz, S., Hackenberg, M., & Binder, H. (2022). The juliaconnector: A functionally-oriented interface for integrating Julia in R. Journal of Statistical Software, 101(6). https://doi.org/10.18637/jss.v101.i06

Luca, S., Vandercappellen, J., & Claes, J. (2019). A web-based tool to design and analyze single- and double-stage acceptance sampling plans. Quality Engineering, 32(1), 58–74. https://doi.org/10.1080/08982112.2019.1641207

R Core Team. (2022). R: A language and environment for statistical computing. [Computer software]. Retrieved from www.r-project.org

Shah, V. B., Edelman, A., Karpinski, S., Bezanson, J., & Kepner, J. (2013). Novel algebras for advanced analytics in julia. 2013 IEEE High Performance Extreme Computing Conference (HPEC). https://doi.org/10.1109/hpec.2013.6670347

Tiobe.com. (2022). TIOBE Index. Retrieved April 20, 2022, from http://www.tiobe.com/tiobe-index/

Contact

Nikola Kaspříková

Prague University of Economics and Business, Department of Mathematics

nám. W. Churchilla 4, 130 67 Prague

nikola.kasprikova@vse.cz