Software Impacts 4 (2020) 100016

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

Improving Bayesian statistics understanding in the age of Big Data with the *bayesvl* R package

Quan-Hoang Vuong [a,b], Viet-Phuong La [b,c], Minh-Hoang Nguyen [b,c], Manh-Toan Ho [b,c], Manh-Tung Ho [b,c,d,*], Peter Mantello [e]

[a] Université Libre de Bruxelles, Centre Emile Bernheim, 1050 Brussels, Belgium
[b] Centre for Interdisciplinary Social Research, Phenikaa University, Yen Nghia Ward, Ha Dong District, Hanoi 100803, Viet Nam
[c] A.I. for Social Data Lab, Vuong & Associates, 3/161 Thinh Quang, Dong Da District, Hanoi, 100000, Viet Nam
[d] Institute of Philosophy, Vietnam Academy of Social Sciences, 59 Lang Ha St., Hanoi 100000, Viet Nam
[e] Ritsumeikan Asia Pacific University, Beppu City, Oita Prefecture, 874-8511, Japan

ARTICLE INFO

Keywords:
Bayesian network
MCMC
Ggplot2
Bayesvl
Big data

ABSTRACT

The exponential growth of social data both in volume and complexity has increasingly exposed many of the shortcomings of the conventional frequentist approach to statistics. The scientific community has called for careful usage of the approach and its inference. Meanwhile, the alternative method, Bayesian statistics, still faces considerable barriers toward a more widespread application. The *bayesvl* R package is an open program, designed for implementing Bayesian modeling and analysis using the Stan language's no-U-turn (NUTS) sampler. The package combines the ability to construct Bayesian network models using directed acyclic graphs (DAGs), the Markov chain Monte Carlo (MCMC) simulation technique, and the graphic capability of the *ggplot2* package. As a result, it can improve the user experience and intuitive understanding when constructing and analyzing Bayesian network models. A case example is offered to illustrate the usefulness of the package for Big Data analytics and cognitive computing.

Code metadata

Current Code version	*v0.9.5*
Permanent link to code/repository used of this code version	https://github.com/SoftwareImpacts/SIMPAC-2020-6
Legal Code License	*GPL (≥ 3)*
Code Versioning system used	*For example svn, git, mercurial, etc. put none if none*
Software Code Language used	*R*
Compilation requirements, Operating environments & dependencies	*No*
If available Link to developer documentation/manual	https://osf.io/w5dx6/
Support email for questions	phuong.laviet@phenikaa-uni.edu.vn

Software metadata

Current software version	*v0.9.5*
Permanent link to executables of this version	https://github.com/sshpa/bayesvl
Legal Software License	*GPL (≥ 3)*
Computing platform/Operating System	*OS X, Microsoft Windows*
Installation requirements & dependencies	*R v3.5.1 or more recent*
If available Link to user manual—if formally published include a reference to the publication in the reference list	https://cran.r-project.org/web/packages/bayesvl/bayesvl.pdf
Support email for questions	phuong.laviet@phenikaa-uni.edu.vn

* Corresponding author at: Centre for Interdisciplinary Social Research, Phenikaa University, Yen Nghia Ward, Ha Dong District, Hanoi 100803, Viet Nam.
E-mail addresses: qvuong@ulb.ac.be, hoang.vuongquan@phenikaa-uni.edu.vn (Q.-H. Vuong), phuong.laviet@phenikaa-uni.edu.vn (V.-P. La), hoang.nguyenminh@phenikaa-uni.edu.vn (M.-H. Nguyen), toan.manhho@phenikaa-uni.edu.vn (M.-T. Ho), tung.homanh@phenikaa-uni.edu.vn (M.-T. Ho), mantello@apu.ac.jp (P. Mantello).

https://doi.org/10.1016/j.simpa.2020.100016
Received 12 April 2020; Received in revised form 20 April 2020; Accepted 23 April 2020

1. Introduction

The emergence of Big Data analytics in recent years is characterized by a great volume and a wide variety of data, high velocity of data collection, huge potential value, and questions over the veracity of data [1]. In one estimate, the amount of text data online generated daily by Twitter alone equals to 50 GB, as compared to the total of a couple of terabytes in 1997 [2]. Capturing the value of the increased quantity of data depends on how researchers solve the problems of the veracity of data. Here, data visualization techniques play a very critical role in this process. Good data visualization can help researchers quickly identify errors in the data [3] and point them toward possible causal/correlational structures in the data. Another essential aspect of maximizing the captured value of data mining is to ensure proper investigation of the predictive models. The Bayesian network modeling method is very suitable in this regard as a Bayesian network has a natural visual presentation of its graph structure, which allows intuitive understanding and probing of the causal and correlational structures in the data [2,4].

However, as Bayesian statistics, in general, and Bayesian network modeling, in particular, are highly computational methods, it is hard to create a software program for beginners of statistics and machine learning as well as researchers who are used to the frequentist approach. The lack of intuitive and open programs for Bayesian statistics is unfortunate for the Big Data analytics movement in two senses. First, with an intuitive program, many more researchers can contribute to solving many components of the Big Data movement that are until now seen as highly inscrutable would more likely be solved. There have been many cases of black-box algorithms, powered by Big Data, making undesirable decisions [5,6], which suggests the importance of having more people understanding the basics of these new technologies. As Big Data analytics is increasingly influencing our decisions in business, entertainment, and politics [7–9], the more people participate in this movement, the better. Second, given that an enormous value to Big Data remains untapped and many questions for the reliability of Big Data still unanswered, the general population would benefit from an improved capacity to investigate causal and correlational structures. It is clear that a better dialogue between the technical world and the public will be fruitful to for the development of technologies that are built on the basis of Big Data.

Hoping to contribute a meaningful solution to the abovementioned problems and to mitigate the risk of mismanaged data, we have built a software that enhances the intuitive understanding of statistical model construction and the Bayesian approach to data analysis. This software package is called ***bayesvl***, which runs on the open-source R program. In this paper, we will briefly introduce the core functions of ***bayesvl***, its impacts, and a brief demonstration of its functions.

2. The *bayesvl* R package

The ***bayesvl*** project was launched in 2017 following a global trend in employing the R statistical programming environment [10,11]. It has been published in the Comprehensive R Archive Network (CRAN) [12] and Github [13]. It is built in a climate where the conventional frequentist approach increasingly falls under scrutiny [14–16], and the popularity of Bayesian statistics is on the rise [17]. Moreover, we believe the combination of the capability of R to generate beautiful graphics, the causality and uncertainty inherent in Bayesian Network modeling [1], and simulated data using Markov Chain Monte Carlo (MCMC) method not only make social science research in the age of Big Data more scientific but also visually appealing to the intuition of readers [18]. Hence, to capitalize on all the trends, the ***bayesvl*** R package combines the powerful ability for data simulation—Hamiltonian Monte Carlo method of ***rethinking*** [19] and ***rstanarm*** [20]; the ability to construct Bayesian network by ***bnlearn*** [21,22]; the capacity of generating beautiful graphics by ***ggplot2;*** detailed model comparison capability enabled by ***loo*** [23,24]. The following sections illustrate the model fitting procedure and the utilities of the ***bayesvl*** package through a case example. The case example investigates the perceived economic pressure on medical patients conditioned on (i) whether they have health insurance and (ii) whether they have a residence near their hospital. Here, we use a dataset of 1,042 observations on health care, medical insurance, and economic destitution, which is deposited in an open database in 2019 [25,26].

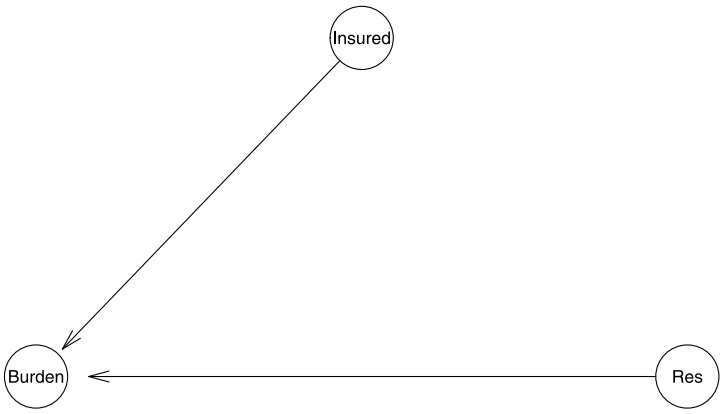

Fig. 1. A graphic representation of the model generated by the ***bayesvl*** package, which investigates whether the perceived economic pressure on medical patients (“burden”) are affected by medical insurance (“insured”) and residence status (“Res”).

3. Comparison with the state of the art

Compared to other current open-source software packages such as ***BayesPostEst*** [27], ***bayestestR*** [28], ***ArviZ*** [29], the ***bayesvl*** package has a relatively simple model fitting procedure as the Stan code is automatically generated. Before fitting a model, it is important to construct a causal diagram or a relationship tree, which characterizes the relationship of the studied variables (See Fig. 1). Based solely on two commands bvl_addNode and bvl_addArc, a relationship tree can be constructed. When creating a node with bvl_addNode, the users can choose the statistical distribution of any variable by coding it as “norm” for normal distribution, “binom” for binomial distribution, or “cat” categorical distribution, etc. The code bvl_addArc is for setting the regression relationship between two nodes: fixed-effect model (“slope”), random-slope model (“varint”), random-intercept model (“varslope”), and mixed-effect model (“varpars”). Among four statistical models, the random-intercept model (“varslope”) and mixed-effect model (“varpars”) are utilized for multilevel modeling.

In addition, while both ***BayesPostEst*** [27] and ***bayestestR*** [28] are more focused on the estimating and testing aspects of the Bayesian framework, and BMS focuses more on Bayesian model averaging and jointness [30], ***bayesvl*** offers a comprehensive tools for Bayesian network construction [22]; model fitting; model expansion and subtraction as recommended by Gabry, et al. [31]; visualization of posterior distribution and posterior predictive testing; and model selection using model weights (See Fig. 2). Compared to ***Arviz***, which is run on Python, as shown above, ***bayesvl*** offers a similar range of functionality but allows a simple code setup to construct the Bayesian network models. This aspect of the ***bayesvl*** package is advantageous for the apprentices of statistics, machine learning, or cognitive modeling. This is because the current other packages for Bayesian statistics tend to require one to code up the mathematical formula from scratch, which can be daunting for the statistical novices.

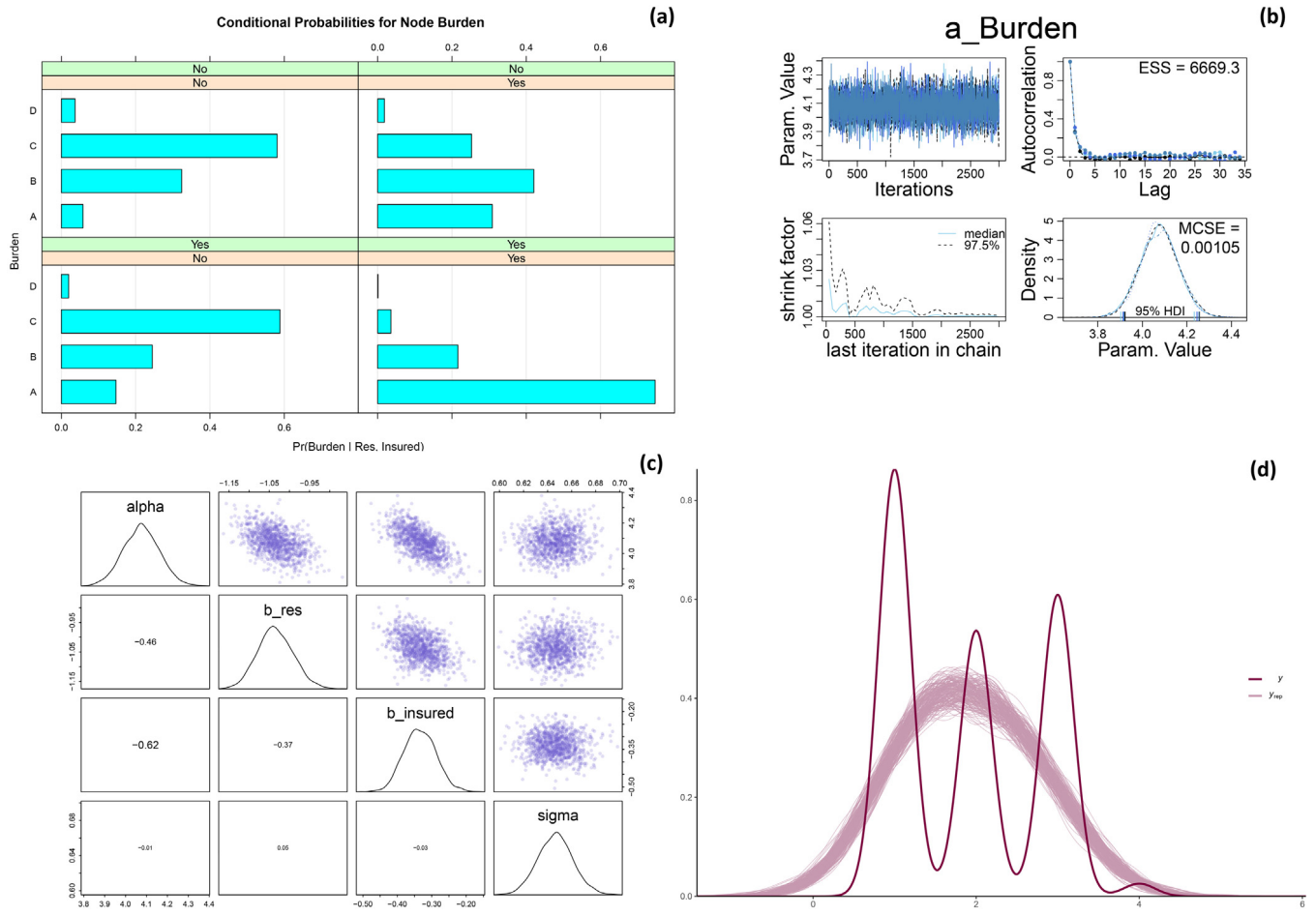

Fig. 2. (a) Conditional probabilities table of all the variables in the model. (b) The convergence diagnostics of the Markov chain property of the data after simulation. (c) Visualization of pair posterior distribution of coefficients in the model. (d) Posterior predictive test for a variable in the model.

4. Overview of impacts

The software package has enabled a wide range of publications in social sciences and humanities. In particular, it is instrumental in the investigation into the phenomenon of cultural additivity [32]; the cultural evolution of Franco-Chinese architectures [33]; the interaction of violence and lie with East Asian religious virtues in Buddhism, Confucianism, and Taoism in folktales [34]; the mental health issues and help-seeking behaviors in international students in a multicultural environment [35]; the youth's digital competencies [36]; social disparities and gender gap in STEM learning; a detailed comparison of research output among economics, social medicine, and education in Vietnam [37]; and the effects of health insurance and socio-economic status on perceived economic pressure of medical patients [25].

More importantly, as demonstrated in the examples above, because the users of ***bayesvl*** can bypass the process of writing Stan code when doing the model fitting, this will also be beneficial for researchers who used to frequentist statistics to make a shift to Bayesian statistics. The ***bayesvl*** R package can also be useful for the statistical novices to start practicing model construction and running data simulation using the MCMC method. With the eye-catching graphic capability, the users can investigate the results and carry out the model comparison process with ease. The ability to visualize the model and easily code it up will make the task of investigating the causal and correlational structures of any dataset less daunting. Moreover, visualization has been shown to support four cognitive mechanism: reinterpretation, abstraction, combination, and mapping [38,39]. For this reason, we hope the wide-ranging visualization tools of ***bayesvl*** will help improve the pedagogical effectiveness and creativity when teaching and applying Bayesian analysis.

Beyond ease-of-use, and pedagogical effectiveness, we also hope that the ***bayesvl*** R package will contribute to the movement toward an established process of Bayesian inference [31,40]. The lack of an established method of Bayesian inference has been argued to limit the its spread among social and behavioral scientists [40]. Progress in this area could mean the mitigation of some problems of the frequentist statistics, such as the controversy related to interpreting the "*p*-value" [16,41]. In addition, higher appreciation of novel quantitative methodologies, we believe, will make social sciences and humanities more scientific and reproducible [16,42]. Thus it will help reduce the so-called social sciences deficit in AI and Big Data analytics [43]. Reproducibility and transparency are the two values we must uphold in the age of Big Data and obscure algorithms. Doing so will greatly reduce the cost of doing science and improve the general public's trust in science [44].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Njah, S. Jamoussi, W. Mahdi, Deep Bayesian network architecture for big data mining, Concurr. Comput.: Pract. Exper. 31 (2019) e4418, http://dx.doi.org/10.1002/cpe.4418.

[2] C. Champion, C. Elkan, Visualizing the consequences of evidence in bayesian networks, 2017, arXiv preprint arXiv:1707.00791.
[3] Q.-H. Vuong, V.-P. La, T.-T. Vuong, M.-T. Ho, H.-K.T. Nguyen, V.-H. Nguyen, H.-H. Pham, M.-T. Ho, An open database of productivity in Vietnam's social sciences and humanities for public use, Sci. Data 5 (2018) 180188, http://dx.doi.org/10.1038/sdata.2018.188.
[4] J. Wang, Y. Tang, M. Nguyen, I. Altintas, A scalable data science workflow approach for big data bayesian network learning, in: Proceedings of 2014 IEEE/ACM International Symposium on Big Data Computing, 8-11 Dec. 2014, pp. 16-25.
[5] A. Springer, V. Hollis, S. Whittaker, Dice in the black box: User experiences with an inscrutable algorithm, in: Proceedings of 2017 AAAI Spring Symposium Series.
[6] K.J. Strandburg, Rulemaking and inscrutable automated decision tools, Columbia Law Rev. 119 (2019) 1851–1886.
[7] S. Spettel, D. Vagianos, Twitter analyzer—How to use semantic analysis to retrieve an atmospheric image around political topics in twitter, Big Data Cogn. Comput. 3 (2019) http://dx.doi.org/10.3390/bdcc3030038.
[8] H. Hassani, X. Huang, E. Silva, Big-crypto: Big data, blockchain and cryptocurrency, Big Data Cogn. Comput. 2 (2018) 34.
[9] M.T. Yazici, S. Basurra, M.M. Gaber, Edge machine learning: Enabling smart internet of things applications, Big Data Cogn. Comput. 2 (2018) 26.
[10] M.T. Ho, Q.H. Vuong, The values and challenges of 'openness' in addressing the reproducibility crisis and regaining public trust in social sciences and humanities, Eur. Sci. Ed. 45 (2019) 14–17.
[11] Q.H. Vuong, M.T. Ho, V.P. La, 'Stargazing' and p-hacking behaviours in social sciences: some insights from a developing country, Eur. Sci. Editing 45 (2019) 54–55.
[12] V.P. La, Q.H. Vuong, Bayesvl: Visually learning the graphical structure of Bayesian networks and performing MCMC with 'stan'. the comprehensive r archive network (cran), 2020, https://cran.r-project.org/web/packages/bayesvl/index.html, version 0.8.5 (accessed on 21 Apr. 2020).
[13] Q.H. Vuong, V.P. La, BayesVL package for Bayesian statistical analyses in r. github: BayesVL version 0.8.5, 2019, Available from: https://github.com/sshpa/bayesvl http://dx.doi.org/10.31219/osf.io/ya9u6.
[14] S.E. Lazic, J.R. Mellor, M.C. Ashby, M.R. Munafo, A Bayesian predictive approach for dealing with pseudoreplication, Sci. Rep. 10 (2020) 2366, http://dx.doi.org/10.1038/s41598-020-59384-7.
[15] A. Gelman, C.R. Shalizi, Philosophy and the practice of Bayesian statistics, Br. J. Math. Stat. Psychol. 66 (2013) 8–38.
[16] V. Amrhein, S. Greenland, B. McShane, Scientists rise up against statistical significance, Nature 567 (2019) 305–307, http://dx.doi.org/10.1038/d41586-019-00857-9.
[17] F.F. Nascimento, M.d. Reis, Z. Yang, A biologist's guide to Bayesian phylogenetic analysis, Nat. Ecol. Evol. 1 (2017) 1446–1454, http://dx.doi.org/10.1038/s41559-017-0280-x.
[18] Q.H. Vuong, N.K. Napier, Academic research: The difficulty of being simple and beautiful, Eur. Sci. Editing 43 (2017) 32–33.
[19] R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and Stan, first ed., Chapman and Hall/CRC, 2018.
[20] A. Gelman, B. Goodrich, J. Gabry, A. Vehtari, R-squared for Bayesian regression models, Amer. Statist. 73 (2019) 307–309.
[21] M. Scutari, J.B. Denis, Bayesian Networks: With Examples in R, CRC Press, Boca Raton, 2015.
[22] M. Scutari, Learning Bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (2010).
[23] A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, Statist. Comput. 27 (2017) 1413–1432.
[24] Y. Yao, A. Vehtari, D. Simpson, A. Gelman, Using stacking to average Bayesian predictive distributions (with discussion), Bayesian Anal. 13 (2018) 917–1007.
[25] M.-T. Ho, V.-P. La, M.-H. Nguyen, T.-T. Vuong, K.-C.P. Nghiem, T. Tran, H.-K.T. Nguyen, Q.-H. Vuong, Health Care, medical insurance, and economic destitution: A dataset of 1042 stories, Data 4 (2019) 57.
[26] M.T. Ho, Health Care, medical insurance, and economic destitution: A dataset of 1042 stories, in: Open Science Framework, 2019, https://osf.io/2k8nd/.
[27] S. Scogin, J. Karreth, A. Beger, R. Williams, BayesPostEst: An r package to generate postestimation quantities for Bayesian MCMC estimation, J. Open Source Softw. 4 (2019) 1722.
[28] D. Makowski, M. Ben-Shachar, D. Lüdecke, BayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework, J. Open Source Softw. 4 (2019) 1541.
[29] R. Kumar, C. Carroll, A. Hartikainen, O. Martin, ArviZ a unified library for exploratory analysis of Bayesian models in python, J. Open Source Softw. 4 (2019) 1143.
[30] S. Amini, F.C. Parmeter, A review of the 'BMS' package for R with focus on jointness, Econometrics 8 (2020) http://dx.doi.org/10.3390/econometrics8010006.
[31] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, A. Gelman, Visualization in Bayesian workflow, J. R. Statist. Soc.: Ser. A (Statist. Soc.) 182 (2019) 389–402.
[32] Q.-H. Vuong, Q.-K. Bui, V.-P. La, T.-T. Vuong, V.-H.T. Nguyen, M.-T. Ho, H.-K.T. Nguyen, M.-T. Ho, Cultural additivity: behavioural insights from the interaction of confucianism, buddhism and taoism in folktales, Palgrave Commun. 4 (2018) 143, http://dx.doi.org/10.1057/s41599-018-0189-2.
[33] Q.-H. Vuong, Q.-K. Bui, V.-P. La, T.-T. Vuong, M.-T. Ho, H.-K.T. Nguyen, H.-N. Nguyen, K.-C.P. Nghiem, M.-T. Ho, Cultural evolution in Vietnam's early 20th century: A Bayesian networks analysis of hanoi franco-chinese house designs, Soc. Sci. Humanit. Open 1 (2019) 100001, http://dx.doi.org/10.1016/j.ssaho.2019.100001.
[34] Q.H. Vuong, M.T. Ho, T.H.K. Nguyen, T.T. Vuong, T. Tran, K.L. Hoang, T.H. Vu, P.H. Hoang, M.H. Nguyen, M.T. Ho, V.P. La, On how religions could accidentally incite lies and violence: Folktales as a cultural transmitter, Palgrave Commun. 6 (2020) http://dx.doi.org/10.1057/s41599-020-0442-3, (In press) https://www.nature.com/articles/s41599-020-0442-3.
[35] M.-H. Nguyen, M.-T. Ho, T.Q.-Y. Nguyen, Q.-H. Vuong, A dataset of students' mental health and help-seeking behaviors in a multicultural environment, Data 4 (2019) http://dx.doi.org/10.3390/data4030124.
[36] A.-V. Le, D.-L. Do, D.-Q. Pham, P.-H. Hoang, T.-H. Duong, H.-N. Nguyen, T.-T. Vuong, T.H.-K. Nguyen, M.-T. Ho, V.-P. La, et al., Exploration of youth's digital competencies: A dataset in the educational context of Vietnam, Data 4 (2019) http://dx.doi.org/10.3390/data4020069.
[37] Q.H. Vuong, P.K.L. Nguyen, V.P. La, T.-T. Vuong, M.T. Ho, M.-H. Nguyen, T.-H. Pham, M.T. Ho, Mirror on the Wall: Is Economics the Fairest of Them All ?, Working Papers CEB WP 20-004, ULB, Universite Libre de Bruxelles, 2020.
[38] L. Martin, D.L. Schwartz, A pragmatic perspective on visual representation and creative thinking, Vis. Stud. 29 (2014) 80–93.
[39] J.H. Mathewson, Visual-spatial thinking: An aspect of science overlooked by educators, Sci. Educ. 83 (1999) 33–54.
[40] B. Aczel, R. Hoekstra, A. Gelman, E.-J. Wagenmakers, I.G. Klugkist, J.N. Rouder, J. Vandekerckhove, M.D. Lee, R.D. Morey, W. Vanpaemel, Discussion points for Bayesian inference, Nat. Hum. Behav. (2020) 1–3.
[41] Q.H. Vuong, How did researchers get it so wrong? the acute problem of plagiarism in Vietnamese social sciences and humanities, Eur. Sci. Editing 44 (2018) 56–58.
[42] G. D'Oca, I. Hrynaszkiewicz, Palgrave communications' commitment to promoting transparency and reproducibility in research, Palgrave Commun. 1 (2015) 15013, http://dx.doi.org/10.1057/palcomms.2015.13.
[43] M. Sloane, E. Moss, Ai's social sciences deficit, Nat. Mach. Intell. 1 (8) (2019) 330–331, http://dx.doi.org/10.1038/s41562-017-0281-4.
[44] Q.-H. Vuong, The (ir)rational consideration of the cost of science in transition economies, Nat. Hum. Behav. 2 (2018) 5, http://dx.doi.org/10.1038/s41562-017-0281-4.