RUFORUM Working Document Series (ISSN 1607-9345), 2021, No. 19 (1):725-732.
Available from http://repository.ruforum.org

Research Application Summary

Virtual Training in Scientific Data Management for Post-Graduate Students Using R Programming Language

Tumwebaze, S.B. [1] & Namawejje, H.[2]

[1] Department of Forest, Biodiversity and Tourism, Makerere University, P. O. Box 7062, Kampala, Uganda
[2]Department of Statistical Methods and Actuarial Sciences, Makerere University, P.O. Box 7062, Kampala, Uganda
Corresponding author: susantumwebaze@gmail.com

Abstract

The Scientific Data Management (SDM) training for Post-Graduate Students Using R was conducted online, due to the COVID-19 challenges. This was the first of its kind with the aim of enhancing the ability of academic participants to analyze data with R programming language and report research findings in a format that would ensure wide dissemination of information in peer reviewed journals and also inform policy. R programming language was chosen because it is a free statistical programming language majorly used among statisticians and data miners. Prior to the training, a needs assessment was conducted and results indicated that participants had various difficulties regarding data management and statistical data analysis using R. When the training was advertised online, there were 1797 participants who registered for the training. About 957 University staff, researchers and postgraduate students attended the first day of training, number of participants kept on oscillating between 700 to 875 that participated in R SDM training. This indicated the need for the training and also that SDM can be conducted virtually to facilitate graduate research and publication. The training had a blend of theory and practical with more emphasis placed on hands on computer exercise using R programming Language. Participants were in agreement that the training came at the right time and the course content covered and approaches used by facilitators during the training were appropriate and relevant. However, the time was not enough to cover many areas and therefore it was suggested by participants that there was a need for a follow up training. The participants appreciated SDM training offered virtually and also indicated they would put the knowledge learnt into practice during their research work. Based on suggestions provided by participants on how to improve the course and facilitator's observations as well as past experience, a follow up SDM training using R was recommended.

Keywords: R programming, RUFORUM, Virtual training

Resume

Un cours de gestion des données scientifiques (SDM) utilisant R pour les étudiants du troisième cycle a été conduite en ligne, en raison des défis la Covid-19. C'était le premier du genre dans le but d'améliorer la capacité des participants à analyser les données avec le langage de programmation R et à présenter les résultats de la recherche dans un format qui garantisse une large diffusion d'informations dans des journaux examinés par des pairs et d'informer les politiques gouvernementales. Le langage de Programmation R a été choisi car il s'agit d'un langage de programmation statistique gratuit préférablement utilisé parmi la communauté statistique. Avant la formation, une évaluation des besoins a été réalisée et les résultats ont indiqué que les participants avaient diverses difficultés liées à la gestion et à une analyse de données statistiques à l'aide de R. Lorsque la formation a été annoncée en ligne, il y avait 1797 participants qui se sont inscrits à la formation. Environ 957 membres du personnel de l'université, des chercheurs et des étudiants de troisième cycle ont assisté au premier jour de formation, et le nombre de participants qui a continué à participant à la formation a oscillé entre 700 et 875. Cela indiquait la nécessité de la formation et que la formation peut être conduite en ligne pour faciliter la recherche et la publication des participants. La formation consistait en un mélange de théorie et de pratique avec plus de focus sur l'utilisation pratique de R pour résoudre des exercices informatiques. Les participants étaient d'accord sur la pertinence de la formation et que le contenu couvert et les approches utilisées par les facilitateurs lors de la formation étaient appropriées. Cependant, le temps n'était pas suffisant pour couvrir de nombreux domaines et il a donc été suggéré la nécessite de faire un suivi de la formation. Les participants ont hautement apprécié la formation et ont également indiqué qu'ils mettront en pratique les connaissances dans leurs travaux de recherche. Sur la base de suggestions fournies par les participants sur la manière d'améliorer la formation, ainsi que les observations faites par les animateurs de la formation et aussi au vu des expériences passées, une formation complémentaire été recommandée.

Mots-clés: Langage de programmation R, RUFORUM, cours, formation, virtuel

Introduction

Scientific data management enhances the capacity of postgraduate students and other researchers to meaningfully engage in conducting quality research by developing appropriate research proposals, design of studies, data collection, and analysis of data for meaningful reporting. PhD and MSc students are heavily involved in large scale experiments or surveys that sometimes lead to complex designs and to subsequent messy data. Figuring out how to handle data resulting from such experiments/surveys takes time, and getting appropriate assistance is difficult. The students are also constrained on how to effectively analyze data using appropriate statistical software, interpret the results and communicate well to the target audience. In response to these shortcomings, this course was structured to encompass broad biometrical needs to equip the postgraduate students with skills required in conducting their research efficiently and effectively. The content incorporated in this course was drawn from broader topics ranging from planning of experiments/surveys, designing and implementing experiments, conducting data analysis for qualitative and quantitative data. The participants were also exposed to R programming language for data management, data analysis and reporting. R is a free statistical programming language majorly used among statisticians and data miners. According to TIOBE index, R is popularly ranked as 8th among scholarly users throughout the world (Nayeemuddin, 2019). R is not only an open source tool used in data analysis, but it is a very flexible language that even allows other tools if required to be used like C, C++ etc. R has numerals number of advantages that support anyone who is interested in data analysis and any user can quickly learn R whether a data scientist or not. R-programming has an effective, coherent and integrated collection of tools for data analysis, provides graphical facilities for data analysis and display, widely used for statistical computing and design especially in big data and data analytics, has field-specific advantages such as great data visualization features among other benefits (Imarticus, 2019). Therefore, as a strategy for strengthening graduate students' capacity to do research competently during this COVID 19 crisis, a virtual scientific data management (SDM) training course using R programming language was organized by the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM), in collaboration with Carnegie and BADEA. SDM training course was aimed at providing the scientists involved in agricultural and other biological research systems with additional skills in R programing language, proper planning of data collection, management and analysis, interpretation and presentation of results in a scientific manner. The ultimate aim was to improve the efficient flow of agricultural and biological information and research efficiency. This primary objective falls in line with the missions and goal of the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) which aims at fostering "innovativeness and adaptive capacity of universities engaged in agricultural and rural development to develop and sustain high quality in training and impact oriented research and collaboration".

The course had a blend of theory and computer based practical with more emphasis placed on the practical using R. Each session was opened with a brief theoretical overview and recap of previous work, before moving to computer. Every day there were new participants who had not installed the required software on their computers and this slowed down some sessions especially on the first two days of the training. On the fifth and sixth day, the participants requested for more time and did not get time to discuss their studies in reference to the skills gained. Two resource persons (Assoc. Prof. Susan Balaba Tumwebaze and Dr. Hellen Namawejje) with a wide experience in research methodologies, data management and various statistical packages facilitated the training using R. The training course covered most aspects of research data, i.e., from data collection methods through data analysis to write-up and presentation of results in different media. Practical sessions explored most facilities of R programming language that are very useful for data management, exploration and analysis. Participants spent a lot of time on various research data sets throughout the training.

The specific objectives of the training were to:

a) Equip the participants with skills to download R, RStudio, R packages and install them on their computers
b) Equip the participants with skills and knowledge to import their datasets, for example from excel, txt, into R and export R datasets into excel.
c) Enhance participant's ability to manipulate their field data in R before doing any data analysis using interactive commands since R supports matrix arithmetic and data structures such as vectors, arrays, data frames and lists.
d) Enhance participant's skills and knowledge to use different techniques used in data visualization applied in R programming language that they can relate to their datasets as well as being able to produce publication-quality graphs they can use in writing up their manuscripts.
e) Equip the participants with skills to use R programming language to analyze data using all inferential statistics tools such as correlation & regression analysis; categorical data analysis techniques and generalized linear models.

Materials and Methods

The basic concepts of each topic were introduced in form of power point presentations and followed by demonstrated examples using life data sets in R. During the course of every module, the participants handled data using only R programming language and facilitators guided them throughout the hands-on practical. This approach enhanced their understanding of the modules taught; analytical skills; boosted confidence to handle their own data management and analysis. Presentations were backed up by power point copies, case studies to reinforce learning and white board were used to illustrate issues were deemed necessary. This was virtual training and participants required either a laptop or a desktop to install R. Electronic version of course materials, exercises, examples and data sets were made available online using Google drive to the participants at the beginning and during the training. The level of participation by all participants was very high and the level of interaction between facilitators and participants was somewhat limited due to the virtual training. The training took place on a webinar due to large number of participants who registered and COVID 19 challenges. The first day, there were 957 participants who attended the training and on the last day we had about 700 participants.

Results and Discussion

This section shows the results of the study. Among the questions that were answered included; how productive the training was? Participants rating on the suitability of the training materials, if the training objectives were clearly communicated in advance of the meeting and if training objectives were met among others.

Table 1. Productivity of the training and suitability of the training materials

How productive the training was (n=132)	Percentage (%)	Please rate the suitability of the training materials (n=133)	Percentage (%)
Not at all productive	0.00	Not at all valuable	0.00
A little productive	0.76	Not so valuable	0.00
Neutral	1.52	Somewhat valuable	3.01
Mostly productive	55.30	Very valuable	54.89
Extremely productive	42.42	Extremely valuable	42.11

From Table 1, majority of the participants (55.30%) said that the training was mostly productive, 42.42% said that the training was extremely productive and only 0.76% said that the training was a little productive. Also, 54.89% of the participants rated the suitability of the training materials as very valuable, 42.11% said it was extremely valuable, 3.01% said the training materials were somewhat valuable. None of the participants mentioned that the training materials were not so valuable or not at all valuable. This could be explained by the fact that many participants are post-graduate students and researchers who needed the data analysis skills and the training materials to support them through proposal development, data collection, data analysis and interpretation of results.

Table 2. Training objectives, facilitators and participants opportunity to air their views freely and openly

Rating	Training objectives were clearly communicated in advance of the meeting (n=135)	Training objectives were met (n=132)	The facilitators effectively moderated the training (n=132)	I had an opportunity to air my views freely and openly (n=133)
	Percentage (%)	Percentage (%)	Percentage (%)	Percentage (%)
Strongly disagree	2.22	1.52	2.27	3.01
Disagree	0.74	0.00	0.00	1.50
Neutral	4.44	3.03	1.52	12.78
Agree	40.00	66.67	34.85	52.63
Strongly agree	52.59	28.79	61.36	30.08

From Table 2, 52.59 % of the participants stated that they strongly agree and 40.00% agree that the training objectives were clearly communicated in advance of the meeting. This could be explained by the fact that the set training objectives were aligned to the intended outcomes of the study. Only 0.74% disagreed that the training objectives were clearly communicated in advance of the meeting. Also, 66.67% agreed and 28.79% strongly agreed that the training objectives were met during the training, this is attributed to the fact that participants were able to download and install R and RStudio, do data manipulation, write R codes, do data visualizations, do descriptive statistics in R as well as design a spreadsheet for data entry for experimental and survey designs which were among the objectives of the training while none of the participants agree that the training objectives were not met. Again, Table 2 shows that 61.36% participants strongly agreed and 34.85% participants agreed that facilitators effectively moderated the training. The possible explanation for this is that both facilitators are well equipped with data analysis skills, experience in teaching, supervision and mentoring of students at university well as well as their areas of specialization that require a lot of data analysis. None of the participants disagreed that the facilitators effectively moderated the training.

In addition, 30.08% of the participants strongly agree that they had an opportunity to air out their views freely and openly, while 52.63% agree that they were able to air out their views freely and openly. Only 1.50% disagree that they never had an opportunity to air their views freely and openly. This is attributed to the fact that the training had office hours set for participants who were had some challenges during the main training were given ample time to solve their challenges and most of them were encouraged to share their work and problems were solved on a case-by-case basis.

Table 3. New knowledge, attitudes, skills, aspirations and motivations attained in the training

What new knowledge, attitudes, skills, aspirations and motivations have you attained	Freq.	Percent
Correlation Analysis	7	5.56
Data Analysis and Data Manipulation with R	42	33.33
Data Analysis and interpretation with R	27	21.43
Data management using R	2	1.59
Introduction to R software	39	30.95
Regression Analysis	5	3.97
Experimental design	4	3.17
Total	126	100.00

Table 3 shows that many participants got knowledge on Data Analysis and Data Manipulation with R and Introduction to R software accounting to 33.33 percent and 30.95 percent, respectively. The possible explanation for this is that the new knowledge, attitudes and skills achieved were among the set objectives of the study and this motivated the participants as most of the knowledge learnt was relevant and applicable to individual projects and various stages of research. Only 1.59 percent of the participants gained knowledge on Data Management with R.

Table 4. Tabulation of challenges faced during the training

Which challenges did you faced during this training?	Freq.	Percent
Difficulty in understanding R concepts and commands	15	17.05
None	20	22.72
Short training period	16	18.18
Sometimes Facilitators were fast	7	7.95
Unstable internet/and Power	30	34.09
Total	88	100.00

According to Table 4, frequently faced challenge among the participants was unstable internet connection or/and power accounting to 34.09% of the participants. This could be attributed to the fact that many African countries have unstable internet and power outrage. The second most frequently faced challenge was the short period of training. 18.18% of the participants mentioned training period as very short. This could be attributed to the fact that it was virtual training and using the facilitators experience, participants concertation can last between 3-4 hours per day.

Table 5: Tabulation of improving future scientific data management trainings

In your opinion what should be done to improve future scientific data management	Freq.	Percent
Diversify training; STATA, SPSS	5	5.68
Increase the Training time period	21	23.86
Introduce breaks within 3-hour sessions	6	6.82
None	12	13.64
Organize training based on Knowledge level: Beginners, Middle, Advance	3	3.41
Participants to be given short assessments	15	17.05
Punctuality in setting up the meetings	2	2.27
Several Trainings in future	16	18.1
share what to be done before training	8	9.09
Total	88	100.00

According to Table 5, many participants suggested that the training period should be increased since it was short. This accounted to 23.86% of the participants. 17. 05% of the participants also suggested that the participants should be given short assessments at the end of the session and 18.1% suggested that several trainings should be organized in future.

Conclusion

The participants appreciated scientific data management training offered virtually, all participants indicated that the course was very relevant and that they had learnt a lot of applications in R. The general impression is that participants were happy with the way the course was handled. Interaction with facilitators and use of real examples and case studies based on their experience enhance their learning.

The participants agreed that the training objectives were met and satisfied with the training material that were used during the training. The participants indicated that the new knowledge and skills learnt were Introduction to R software, data manipulation, analysis with R and interpretation. The acquired knowledge and skills would mostly be applied to research (proposal writing, thesis and dissertation). Most of the participants indicated that they needed future trainings in R and with increased training period, including breaks.

The challenges met by participants included unstable internet and short period of training (3 hours). All participants indicated that they were interested in the follow-up training and suggested topics that could be included in the training such as; data analysis and reporting, experimental designs, correlation, regression and MANOVA. Finally, most of the participants were thankful to RUFORUM for the training opportunity given to them. Based on the course objectives, content covered, evaluation and participant's expectations, it was concluded that the training was successful and met the participant's expectations. However, based on suggestions provided by participants on how to improve the course and facilitator's observations as well as past experience the following was recommended; Follow up scientific management training using R to be organized by RUFORUM. The videos for the initial training in R programming language are available at FURORUM repository and YouTube.

Acknowledgement

The authors are very grateful to Carnegie Cooperation of New York and the Arab Bank for Economic Development in Africa (BADEA) through the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM) for funding the scientific data management training to post-graduate students using R from which this paper was developed. This paper is a contribution to the Seventh Africa Higher Education Week and RUFORUM Triennial Conference held 06-10 December 2021 in Cotonou, Benin.

References

Nayeemuddin S. 2019. Introduction and importance of R-Programming language, Enterprise Information Management. Xtivia.com

Imarticus learning, 2019. Post graduate program in analytics and artifical intelligence, co-created with UCLA extension. Accessed from blog.imarticus.org.