Co-funded by the European Union

STRCUU

STRCUU Scientific Data Management Training for Post-Graduate Students and Young Researchers Using R Programming Language

Dates: 24-26 February 2026

Time: 1400-1800 EAT

Registration Link: https://bit.ly/4q4fx1M

TRAINING OPEN TO GLOBAL PARTICIPATION

NDEJJE UNIVERSITY
Fear of God brings Knowledge and Wisdom

SOROTI UNIVERSITY
A University of Professionals and Innovators

RAE

UNIVERSITÀ DI PAVIA

P.PORTO

www.strcuu.org

Co-funded by the European Union

STRCUU

Funded by the European Union under Grant No. 101237180

(ERASMUS -EDU-2025-CBHE-STRAND 1)

Project Title: Strengthening Teaching and Research Capacities in Recently Established Uganda Universities to Deliver Climate Resilience and Green Energy Solutions to Farming Communities

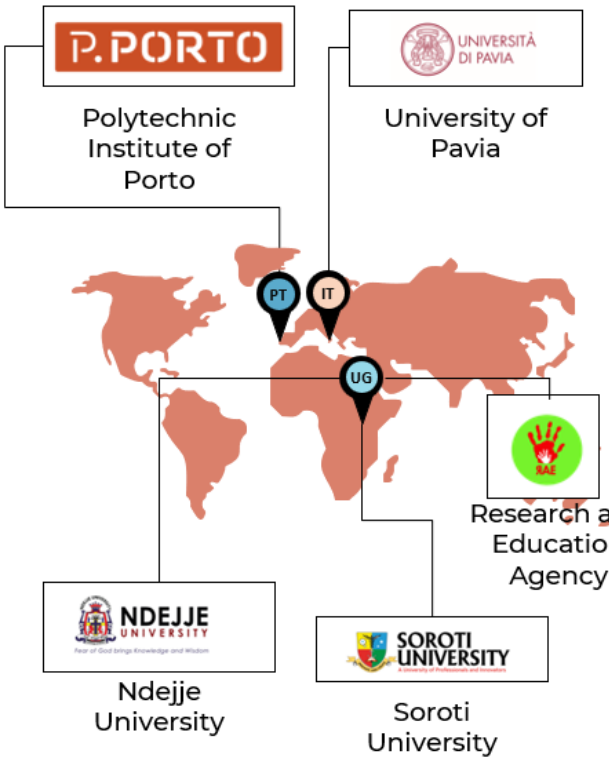
Acronym: STRCUU

Grant Call: ERASMUS -EDU-2025-CBHE-STRAND 1

Grant Number: 101237180

Lead Institution: Ndejje University, Uganda

Partner institutions



NDEJJE UNIVERSITY
Fear of God brings Knowledge and Wisdom

P.PORTO

Co-funded by
the European Union

STRCUU

Table of content

Co-funded by
the European Union

STRCUU

1.0 BACKGROUND

Scientific data management is a set of practices, procedures, techniques, and tools that enable scientists, in particular PhD, MSc students and young researchers, to manage data rationally in order to produce quality scientific reports. In the same way, its good mastery enhances their ability to meaningfully engage in conducting quality research by developing appropriate research proposals, designing of studies, collecting and analyzing data. The purpose is to preserve data in a consistent, accessible, secure and uncluttered form. Without data management, data analysis will produce incorrect results and lead to biased decision making. Indeed, it is currently observed with the technological development and the concern to have more precis and accurate results, experiments or surveys are performed on a large scale sometimes leading 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 also do not know how to effectively analyze the data using appropriate statistical software, interpret the results, and communicate properly with the target audience. Given these shortcomings, this training is structured to cover the general needs of biometrics and to provide post-graduate students and young researchers with the skills and abilities to conduct their research effectively and efficiently. The content of the modules in this training focuses on the techniques of processing, exploration, inferential analysis and modelling of data (qualitative and quantitative) most commonly encountered in the literature to solve real-life problems. Students and young researchers will also be exposed to the R programming language for data management, analysis and reporting.

R is a free, open source statistical programming language primarily used by statisticians and data miners. According to the PYPL index in 2020, R is popularly ranked 7th among scholarly users worldwide (https://darly.solutions/the-most-popular-programming-languages-in-2021/). It is a very flexible language in performing tasks and anyone

interested in data analysis and any user can quickly learn it, whether they are a data scientist or not. R is accessible to a large number of persons and meets a need of students and researchers in developing countries, most of whom cannot afford to spend large sums of money each year to renew licenses for commercially available statistical software.

In data science with applications in many areas of life such as agriculture, biology, health; environment, economics, insurance, bioinformatics, statistics, etc., R plays a crucial role in pre-processing, exploration, and visualization of data (which can be massive, varied, structured or unstructured), prediction, classification, and clustering thanks to its multiple and diversified functionalities. A user of R can modify the various functions of R and create his own packages or either, he can execute code without a compiler, do many calculations with vectors. It can also be used to develop amazing web applications, has a large community support through bootcamps and R meetings. It is well maintained and R updates are always available on CRAN. Note that R is released under the GNU General Public License and there are no restrictions on its use. As for RStudio, it is an integrated development environment (IDE) for R, which also supports statistical computing, graphics and statistical models. An RStudio user can manipulate data and may be able to store used R commands for future use. The environment also provides an R markdown feature that allows work to be converted into different formats such as Word, PDF, PowerPoint,

HTML, etc. This is a boon for academics, as scholarly papers can be written directly in the R markdown environment and then published in a manuscript.

1.1 Target audience

The training is addressed to postgraduate students and young researchers in the following fields: Agricultural Economics, Agroforestry, Plant Protection, Food Science, Aquaculture and Fisheries Science, Natural Resource Management and Environmental Science.

1.2 Aim

The objective of this training is to enhance the efficiency of the flow of agricultural information and research, in particular to achieve the following objectives: to understand the different procedures related to the design of experimental/survey schemes; to implement correctly the different statistical techniques at all stages of research and to present the results in a coherent and efficient way. It will enable postgraduate students and young researchers to acquire not only the technical skills and knowledge necessary to use R programming language for data management, analysis, and presentation of results in a format that ensures their wide dissemination as peer-reviewed publications and policy formulation, but also the practical skills they need to improve the quality of their research publications.

1.3 Specific Objectives

a) Participants will be able to learn how to download R, RStudio, R packages and install them on their computer.

b) Participants will be familiar with the techniques of importing data sets in different formats such as Excel, Txt, CSV into R and exporting R data sets to Excel, Txt and CSV.

c) Participants will learn how to manipulate their field data in R before performing any data analysis using R's interactive command prompts.

d) Participants will learn to describe data sets with descriptive statistics parameters (central tendency, dispersion and shape parameters) calculated with the R programming language.

e) Participants will learn how to use R Markdown to store various R commands, write R scripts, add comments which they can then convert to pdf, HTML, Word document and Power Point. Participants will learn various effective data visualization techniques with the R programming language to produce quality graphs for use in manuscripts and technical reports. Participants will be able to use the R programming language to analyze data with inferential statistical tools such as univariate parametric and non-parametric statistical testing, categorical data analysis techniques and linear regression models.

1.4 Course Outcomes (Expectations)

At the end of the training participants would be able to:

(i) Apply key statistical concepts such as exploratory analysis, univariate parametric and non-parametric inferential statistics, categorical data analysis techniques, simple and multiple linear regression models.

(ii) Use the R programming language to manipulate, explore, analyse and model the state of a biological or agricultural system quantitatively and qualitatively in other fields such as economics and health.

(iii) Select the appropriate statistical method for the problem and the data available.

(iv) Interpret the results and graphs obtained and communicate them effectively and coherently

1.5 Delivery Method and Requirements

The delivery mode will be mixed, with very little interactive theoretical underpinning and practical work designed to complement the lecture material for real-world problem-solving purposes. The approach will be participatory, with students expected to be active learners and to engage in intensive and critical self-directed learning. The assignments will be designed to train and test critical thinking skills. Real data sets provided by the facilitators or obtained from the students before the course starts will be used in the examples and exercises. The training will be online. Each participant will need a laptop, a good connection and a dataset. The daily programme will be divided into sections giving

an overview of the topics, followed by practical computer exercises and a discussion of the statistical results. First, the basic principles, followed by examples of syntax in the R language will be presented. The participants will analyse their data using the techniques already introduced in their daily work. Discussions on the interpretation and presentation of results will take place each day during the plenary sessions. Participants will evaluate the modules on a daily basis and deficiencies will be corrected immediately. An overall evaluation of the modules will be carried out at the end of the training.

1. 6 Training Pre-requisite

This training builds on the knowledge acquired by participants during their postgraduate and undergraduate studies. It assumes understanding of basic biometrics applied to quantitative and qualitative data, and in addition, numeracy skills acquired overtime. The modules of this training provide a solid understanding of statistical techniques that relate to quantitative/qualitative aspects from application, and analytical perspective, thus balancing between theory and applied concepts.

1.7 Duration

The training program will span from **24-26 February 2026; 14:00-18:00 EAT**, allowing for comprehensive coverage of key concepts and ample hands-on practice.

2.0 AN OVERVIEW TRAINING CONTENT

The following modules cover the fundamental aspects of computer science, data management, analysis and processing, and presentation of results, which will provide the framework for the training material. The presentation of these modules will be alternated with lectures and illustrations of real-life applications, selected from various fields, agronomic, environmental, socio-economic and health. More emphasis will be placed on the principle of statistical methods (univariate and bivariate descriptive statistics, univariate parametric and non-parametric statistical inference, and simple and multiple linear regression), when to use them and the rules for interpreting the results derived from them, rather than on technical or mathematical developments. The functions for implementing each method will be presented with an example using the R programming language. Each participant will be able to implement the applications on data related to their field or those to be made available to them, followed by presentations.

2.1 Module 1- Introduction to R programming language and data management

Basics or R uses qui inclure downloading R and R packages, arithmetic operations with R, assigning variables, basic data types in R, importing and exporting datasets, defining a directory, saving datasets in R and in excel sheets.

2.2 Module 2- *Exploratory analysis and reporting with R Markdown*

Types of data, Descriptive statistics and data visualization in R, such as histograms, scatter plots, box plots, and bar charts. With introduction to the production of a statistical report**:** Write a report including R code, the outputs of this code (results, graphs, tables) and comments with R Markdown whose documents are fully reproducible and obtainable in PDF, HTML, Word and PPT formats among others.

2.3 Module 3- *Univariate parametric and non-parametric statistical inference with R*

Univariate statistical tests (comparison of proportions and means and correlation tests – Paramatric and non parametric tests: all t-tests and ANOVA with their non parametric alternatives, chi-square tests, correlations tests [pearson, kendall, pearson, etc.]). Reporting results, interpreting statistical analyses, discussing limitations, and drawing valid conclusions.

2.4 Module 4- ***Linear regressions in R (simple and multiple)***

Principle and scripts of simple and multiple regression models, estimation and significance test of coefficients in the model, validation of the model, analysis of variance table, coefficient of determination (R^2) and adjusted R^2, Influential values, graphical and

statistical analysis of residuals; correlations between explanatory variables; use of the model in forecasting.

3.0 AN OVERVIEW OF COURSE OUTCOMES

Each of the modules to be covered during the training will result to the following outcomes:

3.1 Outcomes for Module 1 on *introduction to R programming language and data management*

- ❑ Download and Installation of R, R studio, and other R packages.
- ❑ Define a directory working and create a script in R.
- ❑ Import different datasets for example from Excel, Txt or CVS into R and export R datasets into Excel, Txt or CSV.
- ❑ Write basic data types and assigning variables in R
- ❑ Manipulate a dataset by doing sort, order, divide, stack, subset, replace, insert, extract, etc.
- ❑ Perform arithmetic operations with R.
- ❑ Save a script or all the work performed in R.

3.2 Outcomes for Module 2 on *exploratory analysis and reporting with R Markdown*

- ❑ Describe the basic statistical terms

- ❑ Calculate descriptive statistics parameters (central tendency, dispersion and shape parameters) of a dataset with R
- ❑ Give the distribution of data according to the nature of the variables with R
- ❑ Do data visualization applied in R programming language as well as produce publication-quality graphs.
- ❑ Use R Markdown to store the different R commands, write R scripts, write comments and convert R Markdown into pdf, HTML, word document, and power point.

3.3 Outcomes of module 3 on *univariate parametric and non-parametric statistical inference.*

- ❑ Perform comparison of proportions tests with R (Z test, Binomial test, McNemar test, Cochran's Q test)
- ❑ Do comparison of means tests with R: Paramatric tests- all t-tests and ANOVA with their non parametric alternatives (Wilcoxon signed rank test, Mann Whitney test and Kruskal Wallis test).
- ❑ Measure the association between two variables : correlations tests [pearson, kendall, pearson, etc.] and test of independence [chi-square, fisher exact, etc.]

3.4 Outcomes of module 4 on *linear regression (simple and multiple)*

- ❑ Distinguish correlation and regression analysis

- ❑ Recognize when to apply a regression analysis
- ❑ Build models using different scenarios of studies
- ❑ Graphical and statistical study of residuesValidate final model of the analysis
- ❑ Analyse data using correlation and regression analysis and interpret results correctly
- ❑ Make predictions from established regression models

4.0 Annexes

4.1 Timelines

Day one: Participants will use R programming language to manage data, writing the code for the different statistical methods

After day 1 of training in R programming we expect participants to have understood use of R in data management.

Day two: Participants will be described data sets using numerical summaries and graphical representations and done reporting with R Markdown.

Day three: Participants will be Discovered a complete overview of statistical methods for data analysis. They will know which method to use depending on the data available and the objectives to be achieved. Participants will acquire methodological and practical knowledge of linear regression methods to obtain an explanatory analysis of a phenomenon, to confirm hypotheses, to take decisions or to make forecasts. A global synthesis of the 4 modules will be made followed by a discussion with all participants on these models. This discussion can be related to the tasks they will be given or to their own experiences in data management and analysis.

N.B: Evaluation of participant's expectation will be conducted once in three days and an overall assessment of the training.

Co-funded by the European Union

STRCUU

Time	Tuesday (24/2/26)	Wednesday (25/2/26)	Thursday (26/2/26)
14:00 -	Registration Opening Ceremony Objectives & overview of course **All**	Overview of Exploratory data analysis (What is EDA? Importance of EDA, Types of EDA) Summarizing quantitative variables-univariate using R Summarizing qualitative variables -univariate **using R** **TLO/ST/HN**	Relationship between two variables - qualitative and quantitative **Relationships between two qualitative variables** **Relationships** **TLO /ST/HN**
14:30 -	Installation of R Introduction to R programming language **HN/ST/TLO**	Relationship between two variables - qualitative and quantitative Relationships between two qualitative variables Relationships between two quantitative variables **ST/HN/TLO**	*Univariate parametric and non-parametric statistical inference with R* **(TLO/ST/HN)**
16:15 -	***Health Break***	***Health Break***	***Health Break***
16:30 -	Introduction to R programming language	Overview of data management (quality measures, integrity, and consistency checks, etc.)	**Correlation and Regression analysis**

	Introduction to R programming language and reporting with R Markdown **HN/ST/TLO**	Type of variables and measurements **TLO/ST/HN**	Model building steps, validation of assumptions linear models (ANOVA, Regression) **ST/HN/TLO**
1800-	***Closing***	***Closing***	***Closing***

4.1: Annex 1: Training Program

4.2 Annex 2: Training needs assessment tool

Pre-Course Questionnaire for Scientific Data Management Course Schedule for July, 2021

Section A: Personnel Information

Sex: ☐ Male ☐ Female

2. Name of degree being under taken

_____

3.College/School/Department_____

_____

4. Title of your research_____

_____

5. Stage of research process

_____

6. Email_____

SECTION B: Competence in Data Management and Analysis

Data Management

11. Knowledge in data management techniques: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

12. Knowledge in design of spreadsheet for data entry: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

13. Knowledge in data checking: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

14. Knowledge in importation to statistical Software: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

Basic Statistics and Interpretation of Results

15. Knowledge in some basic descriptive statistics (*measures of central tendency and measures of dispersion*):

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

16. Knowledge in some basic inferential statistics (*hypotheses testing, t-test, ANOVA, Linear regression models*) :

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

17. Knowledge in data interpretation and reporting: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

18. Knowledge in design of data collection tools: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

19. Knowledge in presentation of results: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

Statistical Software and Computer Skills

20. Which of the following statistical software you have access to and use/used?

Statistical software aware	Aware of	Have access to	Use or used
R			
Python			
SPSS			
SAS			
STATA			
GENSTAT			

MINITAB			
Any other specify:			

21. What is your level of knowledge in R?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

22. What is your level of knowledge in other statistics software (e.g.Genstat, *Python, Stata; SPSS, SAS, etc.*)?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

23 What is your level of knowledge in Spreadsheet/Microsoft Excel?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

SPECIFIC TOPICS

24. Please tick in the appropriate cell in the grid below, your assessment of your needs and capabilities in the topics listed (*VD= very deficient, ND= Not deficient, MR= Major enhancement required, LR= Little enhancement required)*

Component	**Specific Topics**	**Level of competence on this topic**					**Level of enhancement required**				
		VD				**ND**	**MR**				**LR**
		1	2	3	4	5	1	2	3	4	5

Data entry and management	Techniques of data checking										
	Data storage and retrieval										
	Data management strategy										
Data analysis and Interpretations	Exploratory data analysis										
	Tests relative to means										
	Correlation tests										
	Proportion tests										
	Independence tests										
	Non-parametric methods for experience/survey data										
	Linear regression models (simple and multiple)										
	Interpretation of statistical results										

Co-funded by the European Union

STRCUU

This form will be shared with registered participants and will be emailed to the Resource Persons:

1. Dr Thomas Odong (thomas.l.odong@gmail.com)
2. Prof. Susan Tumwebaze(susantumwebaze7@gmail.com)
3. Dr Hellen Namawejje (namawejje-hnamawejje@gmail.com)

Thank you for taking off time to complete this questionnaire