Twinning Project

Contract: GE 16 ENI ST 06 18

Strengthening the Capacity of the Georgian Statistical System

Component 1: Development of External Sector Statistics

Sub-component 1.4: "Statistical Programming in "R" – Basic training"

MISSION REPORT

Activity: 1.4.B "Statistical programming in R – Basic training II"

Mission carried out by
Anne Vinkel Hansen, Statistics Denmark
Joen Petur Jacobsen, Statistics Denmark

October 19th – October 22nd 2021

Version: Final

Expert contact information

Ms Anne Vinkel Hansen
Statistics Denmark,
Copenhagen, Denmark
Tel: +45 40405841
Email: aih@dst.dk

Mr Joen Petur Jacobsen
Statistics Denmark,
Copenhagen, Denmark
Tel: +45 24762306
Email: jpj@dst.dk

Table of contents

1. General comments

This mission report was prepared within the EU Twinning Project "Strengthening the Capacity of Georgian Statistical System". This was the second mission within the sub-component 1.4: "Statistical Programming in "R" – Basic training provided". The mission was carried out in Tbilisi in the period from October 19th to October 22nd 2021. The mission was devoted to training in the use of the software R for statistics production.

The purposes of the mission were to further train staff in R programming, building on the introductory course given in February 2021.

The consultants would like to express their gratitude to the Geostat staff for their dedicated participation in the training.

The views and observations stated in this report are those of the consultants and do not necessarily correspond to the views of the European Union, Geostat, Statistics Denmark, or other statistical institutions involved in the implementation of the project.

2. Assessment and results

The consultants have prepared a set of lectures (including both slides and exercises) which were presented to the participants. At least for the duration of this project the material will be available online.[1] The content of the lectures and the exercises have given the participants an introduction to programming in R and how to use the extended set of possibilities provided by RStudio.

The topics include (but are not restricted to):

- The concept of tidy data, and using R to clean data for analysis
- Formulating, running and evaluating statistical models of various types in R, and predicting from a previously fitted model on new data.
- Working with statistical distributions in R
- Reading in data from external sources and making it technically correct, by operations such as recovering date and numeric valuables coded as characters
- Systematized error detection and correction
- Visualizing missing data patterns, and imputing missing values by methods such as median imputation, regression imputation or random forest imputation
- Using the ggplot2 package for the production of gaphics
- Producing reports in R using Rmarkdown
- Resources for further learning

The consultants were impressed by the enthusiasm and effort displayed by the participants. Participants working on their own laptops posed some issues, but all participants were able to work on a functioning copy of R/Rstudio.

The purpose of the course was to introduce a broad range of techniques, rather than give a firm grounding in a smaller number of techniques, and thus some of the material covered will need more study in order to be implemented. However, the consultants believe that the participants are in a position to acquire the further knowledge needed to implement those of the methods they deem to be of use in their daily work.

[1] https://annevinkel.github.io/R-course-part-II/

3. Conclusions and follow up

- The participants should be allowed time to review course material and extra material linked on the course web page as relevant to their work
- The participants should try to replicate in R work already done in Excel or otherwise;

Annex 1. Terms of Reference

EU Twinning Project GE 16 ENI ST 06 18

October 18th – 22nd 2021

Component 1: Development of External Sector Statistics

Sub-component 1.4: Statistical Programming in "R" – Basic training

Mandatory results and benchmarks for sub-component 1.4:

- Statistical Programming in "R" – Basic training provided

Indicators of Achievement (baseline and targets):

- Theoretical knowledge and practical use of the statistical software package "R".
  - **Baseline:** 2020 – Use of "R" in the statistical production process is very limited in Geostat.
  - **Target:** June 2021 – at least 2 IT-staff and 10 statistical staff can perform basic statistical procedures in "R".

Activity 1.4.B: Statistical programming in R – Basic training II

1. Purpose of the activity

The course follows up on outstanding issues and focuses on the below mentioned subjects

- Visualization of distributions
- Standard parametric and non-parametric statistical tests
- Linear models with normal and binary data
- Model fit diagnostics and assumption checks
- Reading in data from external sources
- Error detection and correction
- Transforming character variables to numeric and dates
- Detecting missing values and visualizing patterns of missingness
- Regression imputation and multiple imputation

2. Expected output of the activity

- The participants have been introduced to the above mentioned topics.
- The participants have been practicing the above mentioned topics.
- Mission report written

Annex 2. Persons met

Geostat
17 employees from 4 different departments

RTA Twinning Team
Mr. Steen Bielefeldt Pedersen, Resident Twinning Advisor
Ms. Nino Grdzelishvili, Resident Twinning Advisor Assistant, Translator