**Development of training in the open source statistical software package “R”**

**Ria Sanderson, Duncan Elliott (Office for National Statistics)**

**Summary**

This report describes a Quality Improvement Fund (QIF) project carried out at ONS to develop training in the open source statistical software “R”. Two introductory training courses were developed and piloted to members of the GSS. Feedback from participants was very positive and this feedback was used to review and further enhance the courses. The courses are now complete and the intention is to offer these courses to the GSS in the future on a regular basis.

1. **Introduction**

In 2013, a Quality Improvement Fund (QIF) project was initiated with the aim of developing training in the open source statistical software package “R” for the benefit of the GSS. The project had two main objectives:

> to provide the GSS with basic skills in R, building a platform from which GSS departments could individually assess whether R meets their statistical needs

> to address the current gap in the provision of training in R that is tailored to GSS requirements

The project was completed at the end of March 2014. The project received QIF funding to develop, pilot and review two one-day introductory training courses in R. The first of these courses (“Introduction to R”) was designed to give an introduction to R for those with little or no experience of R. The second course (“Introduction to graphics using R”) was designed to give those with some basic knowledge of R the tools and skills to create and start to customise graphics using R. The desired project outcomes were to develop the course material, to pilot each course once by offering them with no course fee to the GSS and to review the course material based on participant feedback.

There were a number of motivating factors driving the bid for QIF project funding to develop R training. A group of methodologists and IT specialists at ONS carried out a number of pilot projects in R between 2011 and 2012, which concluded that it was a beneficial tool for research. The use of R is on the increase and the additional functionality it provides (through packages contributed by a large user community) is growing. Its use has been adopted by a number of National Statistics Institutes, so it appears to be an emerging tool. As R is open source software, there are no associated license costs, so it has the potential to offer cost savings to government departments. The open source nature of R could also lend itself to the development of common tools and packages, which could lead to greater collaboration across the GSS. This ultimately has the potential to increase the capability of the GSS. However, there is obviously an associated cost in training the GSS in the use of R. The project proposal was therefore developed with the aim of addressing this training need.

1. **Developing the course material**

The course materials were developed by two methodologists with a number of years of experience in using R. The course materials were developed through the summer and early autumn of 2014. The development of the material started with brainstorming which topics were thought to be of relevance to those who had not used R before. These were categorised into “must know”, “should know” and “could know” to allow some prioritisation amongst the topics. These topics then formed the basis of the material for the “Introduction to R” course. The course materials were developed by focussing on each of the “must know” topics and a selection of the “should know” topics. Both trainers had previous experience of developing course material in R and running R training. This prior knowledge was used both to choose between topics to include and to help develop material to explain some of the more difficult concepts.

The starting point for developing the graphics course was to consider what graphical techniques are available in R. It was decided to spend time explaining the basic graphics functions in R, but also to spend some time introducing specific graphics packages. The motivation for doing this was to provide the participants with a feel for what is possible in R with the intention that they could follow up on those areas that appeared most relevant to their day to day work.

The courses were designed to be interactive. The course slides were constructed so that they contained both some descriptive material but also some R code that could be run by participants. In addition, a series of exercises were developed for both courses to test participants’ understanding after each main topic.

The “Introduction to R” course consisted of sessions that covered: the R environment; getting started with R; using and creating data in R; summarising data; data manipulation; exporting data and practical applications of R for statistical analysis.

The “Introduction to graphics using R” course consisted of session that covered: a brief overview to the R environment (designed to act as a refresher); an introduction to base graphics; an introduction to producing scatter plots, bar charts and histograms; layouts and plotting devices; an introduction to superposition and conditioning using the “lattice” package and a tutor-led introduction to the “ggplot2” package.

**2. Gathering interest for pilot courses**

A note to gather expressions of interest in attending the pilot courses was published on the GSS website in the summer of 2013 and was also circulated via GSS Heads of Profession. This resulted in 34 expressions of interest covering 14 government departments, executive agencies and devolved administrations. In an effort to help tailor the development of the course material to the GSS, those expressing an interest were asked for some additional information on their programming experience, their experience of using statistical software and the types of data and analysis that they routinely carry out. The information collected gave a useful insight into the types of analysis being carried out by those interested in attending the courses, which was very valuable in helping to shape the courses.

More expressions of interest were received than originally expected. Rather than have a separate registration exercise, it was decided to offer places on the courses to those expressing an interest on a first come, first served basis. It was relatively straightforward to fill all the spaces on the “Introduction to R” course. However, it was more challenging to fill all the spaces on the “Introduction to graphics using R” course. This was because it relied on some previous knowledge of R. Initially places were offered to those attending the introductory course. However, due to the requirement of course members’ departments to fund travel and subsistence to ONS Newport and the fact that the two courses were not run on consecutive days not all of the places offered this way were filled. In an effort to fill the course, places were offered to those with no previous R experience, and as part of the development of course materials, some introductory briefing material was prepared and sent to participants in advance.

**3. Pilot of “Introduction to R”**

The pilot of “Introduction to R” was held on Wednesday 15th January 2014. The course was fully booked; there were eight participants[[1]](#footnote-1) representing Department of Health, Health & Safety Executive, Ministry of Defence, Department for Education and Learning in Northern Ireland, Welsh Government and ONS. Participants travelled from different locations around the UK.

The course ran to schedule during the morning session. The first session after lunch (“Data manipulation in R”) proved to take much longer than allowed for in the schedule. This had a knock-on effect on the remainder of the afternoon session. During the afternoon, it was decided that the session on “Using R for programming” should be left out, and the final session “Practical applications of R for statistical analysis” was run as a tutor-led example.

**3.1 Feedback from participants**

To get a better understanding of what the participants wanted to get from the course, a group discussion was initiated at the start of the course. This asked participants what their main questions were on R, what they would like to be able to do with R and what they would like to get out of the training session. The responses were captured, grouped together and displayed on the whiteboard for the remainder of the course. These were used by the trainers to make sure that all questions were either covered by the course material or participants were directed to appropriate further information.

Participants on the course were asked for feedback at the end of the course using a feedback form designed for the purpose. This was compiled using existing feedback forms developed by Statistical Training Unit at ONS as a guide. Feedback forms were received from each participant. All participants rated the overall course as “excellent” or “good”. All participants “agreed” or “strongly agreed” that they would recommend the course to others.

Participants felt that the course met its objectives, was well-structured and engaging. There was positive feedback for the practical sessions. The main criticisms were related to the timing of the sessions and that some of the content was duplicated. Following the course, an electronic copy of the handouts was provided to participants. This was welcomed as the code from the course could then easily be copied and pasted into R.

**3.2 Course review**

Due to the difficulties with the timing of the afternoon session, as highlighted above, this was one of the main areas considered when reviewing the course. The question of duplication of material was also investigated as this was also noted by the tutors whilst delivering the course.

Following the course, the tutors met to review the content and the timings that were achieved on the day. This meeting consisted of looking through each of the slides, identifying any duplicated content and coming up with a revised structure to remove this duplication and some of the content. It was clear that the original course contained too much material, and this was reduced by setting the session on “Programming using R” to be a “bonus” set of slides at the end of the course material. This means that this important information would still be provided to participants, but it would not be covered during the day itself. The length of the “Data manipulation using R” session was reduced by splitting the material between other sessions and removing duplicated material.

The feedback received from participants was that the level of difficulty and the number of exercises was just right. Therefore no changes were made to the exercises during the course review.

The version of R available in the computer training facilities was updated from R2.12.1 to R3.0.2 following the courses. The code contained within the presentation slides and the exercises was checked using the new version of R to ensure that it still worked.

**3.3 Following course review**

Once the course had been reviewed, the presentation slides (including speaking notes for presenters), the exercises and solutions and the required example data sets were shared with Statistical Training Unit at ONS. A meeting was held with the head of Statistical Training Unit and it was agreed that the course would be offered again in 2014, most likely in the late summer, before becoming a regular offering.

1. **Pilot of “Introduction to graphics using R”**

The course was held on Wednesday 12th February 2014 at ONS in Newport. It proved to be more difficult to find participants for this course and the waiting list was exhausted. The course was aimed at those with some knowledge of R. This requirement was relaxed in an effort to attract participants; some additional introductory briefing material was provided for those that had no previous experience in R. By the time of the course, it was however fully booked; there were eight participants registered representing Department for Business, Innovation and Skills, Ministry of Defence, Welsh Government, Office for Rail Regulation, Ofsted and ONS. However, on the day itself, severe storms and disruption on the rail network meant that two participants were unable to travel to Newport. There was a further cancellation at short notice. In an effort to maximise the amount of feedback gathered from the course despite these cancellations, the remaining places were offered to colleagues at ONS. Two of the spaces were filled on the morning of the course, so the course went ahead with seven participants.

The course started on schedule and the timing generally worked well. The final session, a tutor-led demonstration of the package “ggplot2”, took longer than anticipated which meant that not all of the material was covered.

**4.1 Feedback from participants**

To get a better understanding of what the participants wanted to get from the course, a group discussion was initiated at the start of the course. This asked participants what types of graphics they produce regularly, and what software tools they use to produce these graphics. The responses were captured and displayed during the course, and the discussion was re-visited at the end of the day.

Participants on the course were asked for feedback at the end of the course using a feedback form designed for the purpose. This was compiled using existing feedback forms developed by Statistical Training Unit as a guide. For those participants who also received the introductory reading material before the course, an extra feedback questionnaire specifically asking about this material was prepared. Feedback forms were received from six of seven participants. Those responding all rated the overall course as “excellent” or “good”. All of those who responded “agreed” or “strongly agreed” that they would recommend the course to others.

The majority of participants felt that the course met its objectives, was well-structured and engaging. There was in general positive feedback for the practical sessions. There were no particular criticisms, and email correspondence after the course added further positive feedback.

**4.2 Course review**

Following the course, the tutors met to review the content and the timings that were achieved on the day. This meeting consisted of looking through each of the slides. The focus of the course review was on the final tutor-led session on ggplot2, which it was not possible to complete on the day itself. The material was divided into that which was deemed “essential” and some “optional extras”, with a view that in the future, the essential material would be presented, and the optional extras would be included as additional material in the course handouts.

The feedback received from participants was that in general, the level of difficulty and the number of exercises was just right, therefore no changes were made to the exercises during the course review.

As carried out for the “Introduction to R” course, electronic copies of the handouts were provided to participants, and the course materials were tested using R3.0.2.

**4.3 Following course review**

Once the course had been reviewed, the presentation slides (including speaking notes for presenters), the exercises and solutions and the required example data sets were shared with Statistical Training Unit at ONS. A meeting was held with the head of Statistical Training Unit and it was agreed that the course would be run again in 2014, most likely in the late summer, before becoming a regular offering.

1. **Practical arrangements**

The courses were held at ONS in Newport due to the availability of computer training facilities with R installed. The pilot courses were offered to participants free of charge, although participants were required to fund travel and subsistence expenses from their own department. The requirement to travel did deter some of those offered a place on the course and this led to places being turned down. However, on the day, participants travelled from a number of locations throughout the UK. The number of attendees travelling from London remained small; this will be borne in mind when the courses are offered in the future and the most appropriate location will be selected. The start time for the pilot of “Introduction to R” (9.30am) and some travel disruption on the day meant that the start of the course was delayed by approximately fifteen minutes. When the course is run in the future, we will re-consider the starting time for the course depending on the location that it is held and where the largest number of people are travelling from.

The “Introduction to graphics using R” course assumed some basic introductory knowledge of R. This was found to deter people who had not been on the “Introduction to R” course, despite the offer of introductory pre-reading material. On the day itself those with no previous experience of R did not seem to be at a disadvantage. This should also be borne in mind when the courses are offered in the future. One solution might be to offer the course on consecutive days; this would limit the travelling costs that would be incurred by participants.

1. **Project summary**

The project was completed to time and on budget. The level of expressions of interest and the positive feedback from the pilot courses were very encouraging and indicate that there is an appetite within the GSS for training in R.

The project developed course materials that will be offered to the GSS to increase capability in using R. A number of short term benefits were highlighted from running the pilot courses, with attendees indicating that it will assist in their day to day work and also gives them the skills to start to compare different statistical software packages. The project has the potential to bring longer term benefits through increasing the capability of the GSS. It is envisaged that as more people are trained in R, it will become possible to implement new methods and to share tools between departments. There were also unexpected benefits from the project, particularly in terms of networking with GSS members in other departments which allowed information to be shared regarding experiences of installing and managing the use of R.

The courses are now complete and will be offered to the GSS in the future on a regular basis through the Statistical Training Unit[[2]](#footnote-2). It is expected that they will first be offered in the late summer of 2014 and then in the future will become a more regular offering.

1. The total number of participants was limited to 8 due to the size of the computer training facilities. [↑](#footnote-ref-1)
2. http://www.ons.gov.uk/ons/about-ons/products-and-services/training/index.html [↑](#footnote-ref-2)