United Nations ECE/CES/2023/19

Economic and Social Council

Distr.: General
2 June 2023

English only

Economic Commission for Europe

Conference of European Statisticians

Seventy-first plenary session
Geneva, 22–23 June 2023
Item 3 of the provisional agenda
Moving towards open-source technologies – strategic and managerial perspective

Journey to using R – experience of the Central Statistics Office of Ireland

Prepared by Ireland

Summary

In late 2022, the Irish Central Statistics Office (CSO) made a strategic decision to use R as its main analytics tool. The paper presents how the decision was reached, what options have been considered and various aspects of the implementation project: communication, training, software environment, migration, standards and methodology. Finally, experiences to date are summarized.

The document is presented to the Conference of European Statisticians' session on "Moving towards open-source technologies – strategic and managerial perspective" for discussion.

I. Introduction

1. In late 2022, the Irish Central Statistics Office (CSO) made a strategic decision to use R as its main analytics tool. It was decided to replace its current extensive libraries of SAS code with R code over a 5-year period and be fully operational in an R environment across its statistical functions by the end of 2027. Python will also be used where needed (likely to be in specialist areas) but R is to be the primary tool used by statisticians and data analysts in their everyday work.

2. This decision was reached after examining many factors, including the functionality of tools, the costs, strategic alignment, impact on the organisation, risks, and benefits. Various options were examined including maintaining the status quo; using multiple tools and platforms in parallel; and replacing the current code and software with R over differing time periods.

3. While the consequence of this decision is a major organisation-wide programme of change impacting on every statistical area, there is a strong sense that this is the right direction for CSO and will result in a modern and economically sustainable analytics environment fit for delivering on its current and future statistical demands. The change programme is viewed not just as a major shift in technology but also as a unique opportunity to systematically review and improve the quality of code used in the statistical production lifecycle.

II. Background

A. R

4. The use of R and the demands to use it have been growing steadily in CSO in recent years and is reflective of the trend in the wider analytics industry. Both R and Python have been used in niche areas for several years in CSO. This aligns with a decision in 2019, when a maturity model approach to the introduction of R in the organisation was considered, which allowed R usage in cases only where SAS was not best suited. The R software environment, which is open-source free software, was setup in CSO for appropriate use-cases with very limited support provided to users. It was intended to revisit this decision after an interval and to continue to mature the use of R in CSO.

5. More recent progress saw the establishment of a Community of Practice for Data Science and an R User Network which includes both staff in CSO and statistical units in other Government Departments and agencies. The establishment of this group by CSO set an expectation and demand for greater usage of R.

6. It was clear that a move to R was happening organically in the organisation with some evidence of increased usage in statistical production areas which was becoming a worry as sufficient support and governance was not in place. CSO needed to take action to get ahead of this trend before it became a bigger risk to the production cycle.

7. With commercial enterprise editions of the R software environment coming on stream and maturing, a supported well-governed enterprise-scale deployment of R had become a more viable option.

B. SAS

8. At the same time, CSO was considering its options around its use of the SAS suite of products. A new licence agreement was needed which brought the timing of a decision into focus. The current SAS platform in use was expected to reach end of life by 2025 and a modernisation programme to move to a new SAS platform was required before that date. There was concern about the effort needed to shift to this new platform. With limited resources, it seemed that there was a choice of either concentrating efforts on shifting to a new SAS platform or developing an enterprise-scale R environment. This forced the question about the long-term strategic direction.

9. There was also an ongoing concern about CSO's ability to afford the software in the future. After almost 40 years of using SAS, it is heavily embedded in statistical processes. This poses a serious risk to the organisation's ability to deliver its outputs should a mismatch arise between budget and licence costs. Reducing this risk was a strategic imperative and using alternative products which would not incur the same level of risk had to be considered.

C. Comparison

1. Functionality

10. CSO had no major concerns about the functionality of either of the software options being considered. In general, its requirements in the areas of data manipulation and analysis are not hugely complex and the majority of functionality should be achievable in either SAS or R. An analysis of the current SAS code base found a small range of commonly used SAS procedures, all of which could be delivered in R. Challenges to replicate or replace certain features are expected as the products are different but there is confidence that these are not insurmountable.

2. Cost

11. The most significant costs that needed consideration were the licence costs and the costs of migration of code from SAS to R.

12. The CSO's licence costs for SAS are considerable when compared to available budget. Any sharp upward shifts in costs could result in a position where CSO could not afford the required licences. Forthcoming changes to the SAS product offering and licensing model added further uncertainty to this situation.

13. Open-source software is available for the R environment, but CSO's preference is supported commercial products. While these incur licence fees, the associated costs are a better fit to the available budget. In addition, open-source versions of the software could be considered should licence fees change significantly.

14. Migration costs, which include the costs of a migration team of CSO staff and contractors, plus the effort needed by the statistical units, would be high. Most of the costs could be attributed to the effort by statistical units – this is not an additional cost to the organisation as existing resources would be used but it is an opportunity cost as other development work or new initiatives would not be progressed. Migration costs are significant in the medium term but in the longer term become less of a consideration.

III. Options

15. Several options were examined: maintaining the status quo, supporting dual environments, and changing to R.

A. Maintaining the status quo

16. Maintaining the current situation whereby SAS is the primary analytical tool and R is to be used only where necessary had to be considered as a valid option. It had the advantage of stability with the least amount of disruption across the business. It would, however, continue to run the risk of financial viability. Additionally, it did not seem to advance the analytical capabilities of the organisation sufficiently, nor did it enable greater code collaboration across organisations.

17. This option was quickly ruled out. It was clear that CSO saw R playing a significant role in its future analytics environment. Use of R is an expectation of a modern data analytics organisation and is aligned with the skills that are available in the marketplace.

B. Dual environments

18. The main question then became whether CSO could also see SAS as part of its future analytics environment. Maintaining both a SAS and an R environment would provide statistical units with greater choice. Additionally, newer SAS products are interoperable with R and Python and so a multi-platform environment that works well together would be feasible. It would provide stability while also providing options for change. The downsides are a more complex analytics environment which would be more costly than the status quo and could have implications for the mobility of statisticians with differing skillsets becoming a constraint. While the dependency on a single proprietary software provider would be lessened, the risk of financial viability would continue.

C. Changing to R

19. The huge amount of SAS code in use in the organisation and the effort needed to migrate it was a major obstacle when considering changing from SAS to R software. It was necessary to set this legacy code issue aside and identify the CSO's ideal analytics environment in the longer term. The vision became very clear and it was agreed that an R environment, with some Python, was the preferred choice. An R platform would provide a modern flexible environment capable of a wide range of analytics and its capabilities fit well with the strategic direction of CSO and would position CSO for future data and analytics challenges.

20. With the vision set, the legacy issue came back into focus and particularly the timescale for the transition. A huge training and migration effort would be required across all our statistical units which would need to largely be absorbed within existing resources while also maintaining the statistical work programme. Several different timescales were examined ranging from 2 years to 10 years, with the former considered too ambitious and risky and latter likely to result in lack of progress and being too costly. A 5-year timeline to fully complete the migration to R was agreed on the basis that it was ambitious but achievable and should provide statistical areas sufficient time to complete the necessary work while also taking into account regular and cyclical statistical activities.

IV. Implementation

A. Initiation

21. Once the decision to move to R was reached, a small team was quickly put together to begin researching and planning for the migration. While this was a technology-led initiative, other areas in CSO saw the wider potential to improve quality and methodologies through code refactoring and rewriting and quickly came on board to support it.

22. The resultant high-level plan focussed on the following areas: communication, training, software environment, migration, standards and methodology. The project requires effort from almost all areas in the Office, and particularly requires the statistical areas to work closely with the project team to ensure a managed transition.

B. Communications

23. From the outset, communication was recognised as essential to the success of the project, with an instruction given to the project team to overcommunicate.

24. The first step was communication with senior management as ensuring buy-in from this group was considered critical. To this end, there were a series of early engagements with senior management, with strong support from the Management Board, emphasising the high priority of the work programme. These were followed by in-person meetings with senior managers and subsequently with their teams to outline the project, understand concerns, and

jointly schedule the migration for each team and division. At the same time, more general larger-scale information sessions were given to introduce the R environment.

25. A Communications Manager joined the team to ensure consistent and appropriate messaging and stakeholder engagement. A detailed communications plan has been developed. Migration champions have been appointed as liaisons between the statistical units and the team. The aforementioned R User Network acts as an expert advisory group; this provides a forum for internal experts to drive the success of the migration and provide critical skills to staff and additional support to the project team.

C. Training

26. A Training Manager has been assigned and CSO's Training Unit is working to develop a training framework with different training pathways for differing needs. With much online training available for R, CSO has chosen to use DataCamp for training staff in the basics of R coding. This will be supplemented by other training methods, including classroom training, seminars, and one-to-one supports during migration. Individual progress on training will be managed, with a skills survey to measure the skills gap in each role (this will be repeated at intervals throughout the project). Supporting documentation is being developed, including user guides and advice notes and a central R Resource Hub has been established.

D. Software environment

27. The R environment has been designed and the architecture reviewed by external experts. A support contract has been put in place with a service partner and the environment is in the process of being built. A commercial edition of RStudio, which is an integrated development environment for R and Python, is being setup.

28. A Package Manager will be installed as a repository management server to organise and centralise R and Python packages across CSO. It will be used to control the packages in use in CSO and publish and share internal packages and support reproducible results through historic snapshots.

29. Git will be used for version control of R and Python code.

E. Standards and practices

30. A standards framework is under development. Coding standards and good practice guidelines have been drawn up. Likewise, a testing framework is being developed to introduce a more rigorous testing regime for user-developed code. Version control of code will be required so that there is traceability on production processes over time.

F. Methodology

31. The project is a good opportunity to introduce more standardised methods by using central code libraries and functions. A recommended list of such R functions has been developed for the sub-processes in the Generic Statistical Business Process Model (GSBPM). Statisticians are expected to use these standard functions unless there is a methodological reason to do otherwise.

32. The project is also an opportunity to improve statistical methods and processes. While some SAS code will simply be refactored in R to produce the same outputs, other SAS code should undergo a more in-depth examination to ensure that appropriate methods are in use and that the code is reflective of the statistical process. The extent of this quality improvement work will require careful balancing with the imperative is move to R within a 5-year timeframe.

G. Migration

33. Project leads have been assigned to each of the four statistical directorates in CSO. These project leads will act as a single point of contact for the statistical units in those directorates. They will be supported by the project team, comprising CSO staff and contractors.

34. Different migration pathways have been identified based on the level of R expertise in the statistical area and the level of complexity of the code rewrite. Training will be matched to the individual with formal training underpinning those with limited experience and 1-to-1 support provided to all. The migration will be done either by the statisticians or in combination with the project team. Different programmes of code reviews and testing reviews will be put in place for each pathway. Statisticians retain ownership and responsibility for code irrespective of the migration pathway.

35. Pilots projects in different types of statistical units have been underway for months and the learnings from these have been used to inform the approach taken for the migration. The project team has also been engaging internationally to learn from experiences in other statistical offices.

36. The migrations are commencing across all directorates with individual statistical units being scheduled across the 5-year timeframe. The schedule must be mindful to ensure sufficient year-on-year progress as well a spread of activity across the years. Only SAS code that is in use or likely to be used again will be migrated.

H. Use of Artificial Intelligence

37. In late 2022, a prototype of ChatGPT, which is an AI model which interacts in a conversational way, was launched. The project team briefly investigated ChatGPT's capabilities for generating R code, documenting existing SAS code and translating from SAS to R. Initial results were very promising and the potential to use it for translating legacy SAS code into R was recognised.

38. This led the team to the OpenAI Codex model, which is specifically trained on code, and is available through APIs. It comprises a family of AI models that translate between natural language and code in more than a dozen programming languages (although Python appears to be main language used in training). This model has since been replaced by a larger Davinci model.

39. So far, the team has been exploring the functionality of Davinci, configuring the parameters, learning to prompt engineer and developing key value pairs (small chunks of SAS code with its equivalent R code). It seems that the model has good potential in assisting CSO in translating large amounts of SAS code into R. CSO is approaching it with caution and view the model as an assistant to the migration team. All code generated will be reviewed and will be treated as the starting point before human intervention. Until more is learnt, use of this model has been restricted in CSO to the project team while use of ChatGPT has been prohibited.

40. Well coded or well understood SAS programs are likely to be good migration use-cases for the model and it is expected to be a significant labour-saving tool. For SAS programs where quality improvements are needed either in the code or methodology, then the value of the model is arguable as poor inputs are likely to result in poor outputs.

41. The next step is to start training or fine-tuning a CSO version of the model which will then be used as part of the migration effort. While there is potential to accelerate the migration to R code using AI, it needs to be balanced with rigorous testing and statisticians and the project team need to be able to keep pace.

42. It is still early days but Davinci looks like a very promising tool. It seems to have great potential, not only for code translation, but for code generation based on dialogue and could change the coding role of statisticians and data analysts in the future.

V. Experiences to date

43. Some of the big risks for the project are lack of buy-in at senior management level and resistance to change by the statisticians. A wide variety of reactions were expected by the project team during initial engagements with senior managers and their teams. There has been a much greater positive reaction to the decision to move to R than anticipated.

44. There are certainly concerns about how areas will manage their day to day work and migrate but the 5-year timeframe and the supports offered by the project team help mitigate these. There is fear and resistance to the change that is understandable as new skillsets are required along with a significant body of migration work, as well as opposition by those who do not agree with the decision. Early migration efforts will focus on those who are willing, in order to gain project momentum. There are far more enthusiastic early adopters than expected and, while this poses capacity problems for the project team, all are encouraged and accommodated in the process.

45. The project team has been engaging internationally to learn from experiences in other statistical offices and appreciation goes out to all those who have helped. In particular, the experiences of Statistics Canada, who have been very generous with their time and a couple of years ahead in their move, have been helpful in shaping the planned approach.

46. Finally, the use of AI in the migration could be a very significant development for the project. It appears to have great potential and while CSO is planning to train its own SAS to R model, there would appear to be an opportunity for wider cooperation across the statistical community to explore the wider potential of these tools.