Navigation menu

Guidance

Splink: MoJ's open source library for probabilistic record linkage at scale

Updated 16 July 2021

Contents

1. Introduction

The Ministry of Justice (MoJ) has received funding from ADR UK for an ambitious programme of work called Data First, which aims to improve the quality of the department's data to enable better research. As part of this work, a new team has been established to develop and implement cutting edge approaches to data linkage at scale.

2. Business problem

The MoJ and its agencies have numerous administrative data systems. These systems were developed at different times for different purposes, and there is no consistent person identifier that is used across systems.

This results in challenges when analysts and researchers need to perform analysis that spans multiple systems, such as understanding journeys through the justice system, or repeat users of justice services. Improvements to linked data have the potential to unlock important new insights - for example improved research into the effectiveness of justice system interventions.

3. Deciding on an approach

Overall, the data from these systems amounts to tens of millions of distinct records each of which refers to an individual but lacks a consistent identifier.

The new data linking team began by investigating the various data sets and working with their customers to understand their user needs. They found that their approach to data linking needed to:

- be fast enough to link up to around 100 million records
- have a transparent methodology that could be explained by the team, and understood by users
- have as high accuracy as possible given data quality
- be flexible enough to accommodate a wider variety of input data and link multiple data sets together - including data sets that needed to be both de-duplicated as well as linked

To determine an approach, the team started with desk research into data linking theory and practice, and a review of existing open source software implementations.

One of the most common theoretical approaches described in the literature is the Fellegi-Sunter model. This statistical model has a long history of application for high profile, important record linking tasks such as in the US Census Bureau and the UK Office for National Statistics (ONS). The model takes pairwise comparisons of records as an input, and outputs a match score between 0 and 1, which (loosely) can be interpreted as the likelihood of the two records being a match. Since the record comparison can be either two records from the same data set, or records from different data sets, this is applicable to both deduplication and linkage problems, including linking an arbitrary number of data sets.

An important benefit of the model is explainability. The model uses a number of parameters, each of which has an intuitive explanation that can be understood by a non-technical audience. The relative simplicity of the model also means it is easier to understand and explain how biases in linkage may occur, such as varying levels of accuracy for different ethnic groups.

Direct and generalisable comparisons of the accuracy of different approaches are rare in the literature, so it is difficult to be sure of how well the Fellegi-Sunter model performs against

alternatives. Nonetheless, the team concluded that it is likely some of the more sophisticated models in the recent literature would have higher accuracy, at the expense of higher complexity, lower explainability and often longer processing times. In addition, the potential for improvement in accuracy over the Fellegi-Sunter approach was probably relatively modest.

After settling on the Fellegi-Sunter approach, the team reviewed the available free and open source software that could be used to estimate the model, concentrating on packages available in R, Python and Apache Spark - the tools which are available within the MoJ's Analytical Platform.

The best package the team could find was the fastLink package in R, which stood out because it is accompanied by a rigorous paper written by academics at Harvard and Princeton that describes the theoretical model implemented by the package. This paper also assesses its performance against alternatives available in Python and R. The following figure (Figure 3 in the fastLink paper) illustrates these comparisons:

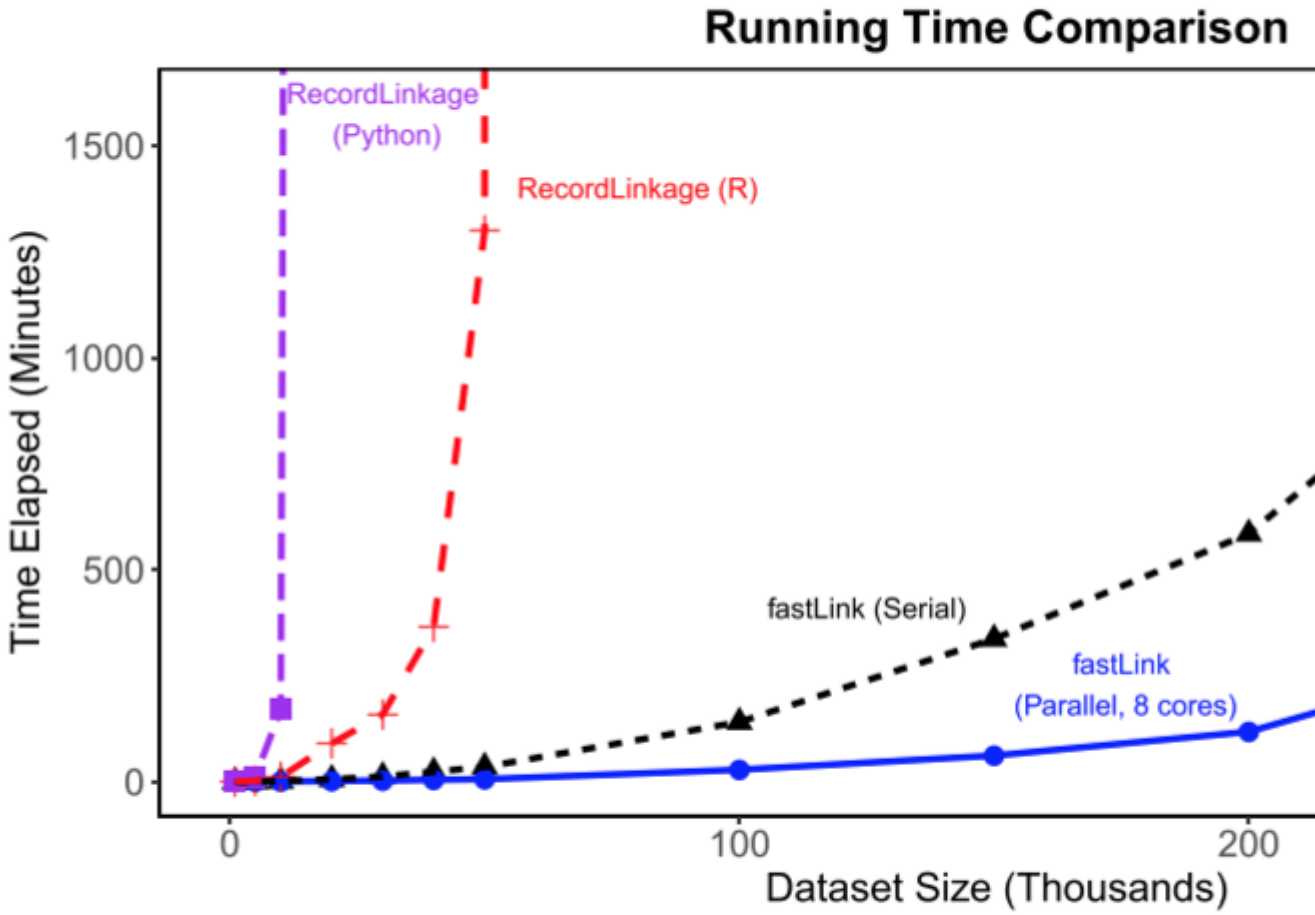

Figure 1: Comparison of linkage packages

This figure shows that fastLink lives up to its name, with substantially faster performance on large data sets than alternatives in Python and R. However, it is also clear that it is not fast enough to perform record linkage on data sets amounting to millions of records.

After reading the paper and fastLink's source code, the MoJ's data linking team realised the problem is well suited to distributed computing frameworks like Apache Spark which are able to parallelise calculations across multiple computers. This would enable the approach to be applied to linking problems involving millions, and possibly billions, of records. The team therefore set about developing a record linking package called Splink.

4. Introducing Splink

Splink is a PySpark package that implements the Fellegi-Sunter model of record linking, and enables parameters to be estimated using the Expectation Maximisation algorithm.

The package is fully open source and can be found on GitHub. It is accompanied by a set of interactive demos to illustrate its functionality, whereby users can run real record linking jobs in their web browser.

The package closely follows the approach described in fastLink. In particular it implements the same mathematical model and likelihood functions described in the fastLink paper (see pages 354 to 357), with a comprehensive suite of tests to ensure correctness of the implementation. In addition, Splink introduces a number of innovations:

- Comprehensive graphical output showing parameter estimates and iteration history make it easier to understand the model and diagnose convergence issues.
- An 'intuition report', which can be generated for any record pair, which explains the estimated match probability in words.
- Support for deduplication, linking, and a combination of both, including support for deduplicating and linking multiple data sets.
- Greater customisability of record comparisons, including the ability to specify custom, user defined comparison functions.
- Term frequency adjustments on any number of columns.
- It's possible to save a model once it's been estimated - enabling a model to be estimated, quality assured, and then reused as new data becomes available.

A companion website provides a complete description of the various configuration options, and examples of how to achieve different linking objectives. In addition, the source code is fully documented using Google's recommended documentation style.

So far, the MoJ has used it to tackle record linkage problems up to around 15 million records with a runtime of less than an hour, but it is anticipated that the approach can scale to substantially larger data sets.

A clerically labelled data set has been used to assess the performance of Splink. This has shown Splink has improved accuracy (that is, an unambiguously better ROC curve) compared to the previous rules-based approach used by the MoJ.

5. Open Source collaboration

As free and open source software, Splink is available for anyone to use. The MoJ data linking team encourages other government departments to try it out and give feedback to the authors, raise bug reports, or contribute improvements.

The team have found that open sourcing the software has enabled much more effective collaboration across government departments and beyond. For example:

- the ONS generously offered to peer review the work. To do so, they were able simply to download the code from GitHub and use the open source demo notebooks to understand how to use the software. They were therefore able to use the software without additional guidance from the MoJ. They contributed significant insights, including finding an important bug which has subsequently been fixed
- it has enabled more effective collaboration with the academic advisors to the Data First project, who have been providing expert advice on data linking theory and practice. Academics are always able to see and use the latest version of the software, with none of the usual frictions and frustrations associated with sharing closed-source, and potentially sensitive code
- it also has great benefits for the transparency of the Data First work, enabling users of MoJ linked data to understand exactly how the data linking was done. This enables analysts and researchers to properly account for the linking methodology in any subsequent uses of the linked data
- it has greatly improved the authors' ability to communicate their methodology to the wider data linking community, since other government departments can freely download and try out the code.

↑ Back to top

Help us improve GOV.UK

Services and information

Benefits
Births, death, marriages and care
Business and self-employed
Childcare and parenting
Citizenship and living in the UK
Crime, justice and the law
Disabled people
Driving and transport
Education and learning
Employing people
Environment and countryside
Housing and local services
Money and tax
Passports, travel and living abroad
Visas and immigration
Working, jobs and pensions

Government activity

- Departments
- News
- Guidance and regulation
- Research and statistics
- Policy papers and consultations
- Transparency
- How government works
- Get involved

Support links

- Help
- Privacy
- Cookies
- Accessibility statement
- Contact
- Terms and conditions
- Rhestr o Wasanaethau Cymraeg
- Government Digital Service