Hybrid record matching: creating a modern Business Index.

Isabela Breton
Dana Seman-Bobulska
Joanne Sheppard

Office for National Statistics UK

Abstract

Situation:
The Office for National Statistics has a successful product called the Inter Departmental Business Register (IDBR). This product is a statistical business register consisting of various administrative and statistical data sources ie HMRC, ONS surveys. The product is very successful; however, it is based on legacy technology, clerically intensive, and thus limited in its data sources.

Task:
Our task is to replace the IDBR with a new Business Index: a more efficient approach (utilising a much smaller production team) on a modern platform and extending the number of data sources to go beyond the existing administrative sources. Business Index is not the complete product replacement of IDBR but nevertheless a core component. Discussions of additional components of the IDBR and combining indexes are out of scope of this paper.

Product:
We have used a hybrid of probabilistic record linkage and defined rules – based on subject matter expertise. Our pipeline updates the Business Index daily for births, deaths and amends. Making it an up to date (near real time) resource for researchers and policy makers - with 99.9% accuracy against truth data. At the time of writing this paper we have created over 7 million matched entities, which we term legal units. The final product will be used as a stand-alone resource or in combination with other indices.
Our work demonstrates efficiency and accuracy of using a hybrid approach. The hybrid consists of:

- **Scalable machine learning**: Splink: a machine learning implementation of Fellegi-Sunter approach;
- **Rules based approach** based on exploratory data analysis and subject matter experts (whom have extensive knowledge from working with the data for many years).

Development approach:
What we have learnt during the production of the index will be of use for others working on record linkage. Core recommendations:

- **Well developed existing approach**: Splink produces weight match scores and is based on a well understood matching process.
- **Dedicated subject matter (SME) expertise:** is high value for identifying exceptions, edge cases and general understanding. This is facilitated by (code) notebooks and a daily log.
- **Daily log:** we deployed this early in our development. This enables rapid feedback from SMEs and has enabled quick product evolution.
- **Utilising the existing Indexes:** (in our case the IDBR) as a truth data. This sped up our development: by utilising developed / existing matching knowledge.
- **Early output deliveries for QA:** We regularly deliver outputs for customers' and SMEs inspection: Speeds up improvements and corrections.

Introduction

Matching business records is a crucial task in maintaining accurate and up-to-date business registers. Inaccurate or untimely business records impact government agencies, private companies in accurately understanding our economy. The UK Office for National Statistics (ONS) currently relies on the Inter-departmental Business Register (IDBR) to maintain its business register. This product is a statistical business register consisting of various administrative and statistical data sources ie HMRC VAT, PAYE and Companies House. The product is very successful; however, it is based on legacy technology, clerically intensive, and thus limited in its data sources and scalability.

To address these challenges: we are creating a new Business Index using machine learning methods in combination with rule-based methods to improve timeliness, reduce clerical reliance, expansion of data sources, and increase scalability.

Key challenges:

- **Less reliance on clerical** correction
- **More frequent updates**, essentially daily updates to the business register
- **Add extra data sources:** such as the Financial Conduct Authority, Charities Data and GLEIF
- **Increase scalability of matching solutions**: making use of modern distributed computing

This paper discusses the use of machine learning, rule-based and hybrid methods in business record matching to address these challenges. The paper is set out as follows:

- **Literature Review:** where we briefly discuss matching methods. We focus on the advantages and disadvantages of a wide range of approaches under four categories: rules based, unsupervised, supervised and hybrid methods.
- **Methods and Data section:** where we discuss the data itself and hybrid method we have used. In addition, we discuss the daily logging and reporting approach for maintenance and improvements and finally the agile approach of early deployment and iteration.
- **Results:** we demonstrate accuracy of our approach with reference to the existing IDBR. This is used as a key benchmark. We also present the continuous improvements to our approach and demonstrate how accuracy has increased through time – fundamentally due to daily logs and frequent feedback to the data science team.
- **Conclusion:** we bring together key findings on product, ways of working and next steps for development.

Overall, this paper aims to contribute to the ongoing discussion about the use of machine learning and rule-based methods in record matching. We argue that a

combined approach has the potential to significantly improve and expand record matching and pave the way for a more effective system to replace the IDBR.

Literature Review Section

Record matching is the process of identifying and linking records that refer to the same entity in multiple databases. It is an essential task in various fields, such as healthcare, finance, and government, where data are often fragmented across different systems. Machine learning methods have been widely used for record matching tasks due to their ability to handle large-scale data and automate the matching process. In this literature review, we will discuss the advantages and disadvantages of different machine learning approaches, rules based and hybrid approaches for record matching.

	Description	Advantages	Disadvantages
Rule-based methods	Rule-based methods (RBM) are the most straightforward approach for record matching, where rules are predefined to determine the match/non-match decision. Rules are generally based on string comparison techniques, such as exact matching, tokenization, and fuzzy matching.	Advantages of rule-based methods is their interpretability and simplicity. The rules can be easily understood and modified by domain experts.	The major disadvantage of rule-based methods is their limited ability to handle variations in data, such as typos, abbreviations, and misspellings - without extensive use of clerical resolution (provided by domain experts). Domain experts are also needed in the careful definition and maintenance of rules.
Supervised learning methods	Supervised learning methods involve the training of a machine learning model on labelled data to predict the match/non-match decision. The most used supervised learning methods for record matching are logistic regression,	Advantages of supervised machine learning methods for record matching is their ability to learn from labelled data and generalise to unseen data. These methods can handle many attributes and complex feature interactions, which	However, these methods require large amounts of labelled training data, and their performance heavily depends on the quality of the training data. Additionally, they may struggle with imbalanced datasets, where

	decision trees, and support vector machines. Also, in recent years, deep learning approaches such as Siamese neural networks and transformers have also shown promising results for record matching.	can be difficult for rule-based methods.	the number of matches and non-matches is heavily skewed. They also require creation of labelled data for updating weights (in terms of maintenance).
Unsupervised learning methods	Unsupervised learning methods involve the clustering of records based on their similarities to identify potential matches. These methods cluster records based on their attribute values and use various similarity metrics to determine whether records belong to the same entity. One popular unsupervised method for record matching is Splink, a probabilistic record linkage library developed by the MOJ and the ONS, which is based on the Fergelli-Sunter approach. Other unsupervised methods for record matching include clustering algorithms such as k-means and	A key advantage of unsupervised machine learning methods for record matching is that they do not require labelled training data, which can be expensive and time-consuming to create. They can also handle large datasets and complex feature interactions, which can be challenging for rule-based methods. Methods such as Splink are also highly scalable and well used in production applications of record matching.	However, these methods may not perform as well as supervised methods when high accuracy is required and may require more manual intervention to verify potential matches. Domain expertise is often required to tune these methods appropriately.

	hierarchical clustering.		
Hybrid methods	Hybrid methods can combine multiple machine learning approaches or machine learning with rules-based matching to improve the accuracy of record matching. For example, a hybrid approach could combine rule-based methods with supervised learning methods to handle variations in data and improve interpretability. An example is a Recurrent Neural Network (RNN) model, which uses an unsupervised Siamese neural network to generate potential matches and a supervised RNN to refine the matching results. Other relevant examples concern the use of rules based matching combined with machine learning.	Hybrid methods for record matching can combine supervised, unsupervised and rules-based techniques to leverage the strengths of these approaches. These methods can be particularly effective when labelled training data is limited or when the data is highly imbalanced. Hybrid methods often use unsupervised techniques to generate candidate matches, which are then validated using supervised techniques or rules-based techniques.	However, these methods can be computationally expensive and require domain expertise to tune appropriately. Additionally, the performance of these methods may heavily depend on the quality of the training data and the chosen hybrid approach. Also, hybrid methods containing a rule-based component are reliant on domain expertise for development and maintenance.

In conclusion, machine learning methods have been widely used for record matching tasks due to their ability to handle large-scale data and automate the matching

process. Different machine learning approaches have their own advantages and disadvantages, and the choice of approach depends on the specific needs and constraints of the task at hand. Rule-based methods are simple and interpretable but have limited ability to handle variations in data. Supervised learning methods can handle complex variations in data but require large amounts of labelled data. Unsupervised learning methods can handle large datasets but rely on similarity metrics that may not capture all relevant features of the data. Hybrid methods can leverage the strengths of multiple machine learning approaches and/or rules based but are more complex and difficult to interpret and potentially maintain.

Methods and Data Section

In this section we discuss the data sources and the outputs we create. We then cover the hybrid method used for record matching, and why we adopted this approach. In addition, we discuss the daily logging and reporting approach for maintenance and improvements and finally the agile approach of early deployment and iteration.

Data: sources and output

Source data: HMRC VAT

A legal unit (business) must register for VAT if its taxable turnover is more than £85,000 (2021). The Business Index takes in many frequency feeds from HMRC for VAT information which covers every VAT record. Our daily update notifies us of births, deaths and amends i.e., gives information surrounding the VAT attributes, whereas the monthly and quarterly feeds mainly focus on updates to turnover.

Source data: HMRC PAYE

PAYE is HM Revenue and Customs' system to collect Income Tax and National Insurance from employment. A legal unit only needs to register if it has employees paid £120 or more a week or gets expenses & benefits or has another job or pension. The Business Index has two updates for PAYE, that of quarterly, which is the main data delivery for any births, deaths, or amends to the PAYE attributes i.e., name, address and that of annual, which gives us information around the PAYE classification.

Source data: Companies House

Companies House stores and maintains information for every limited or limited liability partnership plus some Charities and Mutual Societies that have decided to register for a company registration number (CRN). Registering for a CRN enhances the identity and adds credibility to the business. The Business Index gets two updates, that of daily which is the main data delivery for any births, deaths, or amends to the CRN attributes and that of monthly, where the BI takes more granular information for classification and legal form.

Master tables

For every administrative data source, we onboard, the Business Index creates a 'Master Table'. This table allows us to store every delta change against every variable as well as recording the incorporation date (birth) and closure date (death) of that administrative record. In short, the Business Index gives a longitudinal view of the administrative record. However, this dates to April 2019 at the start of the Business Index. To create the master tables for VAT/PAYE/Companies House the data was taken from the ONS current business register of the Inter-Departmental Business Register (IDBR) to ensure the BI held the full population of registered data.

For all ‘other’ administrative data the BI took the data from the 1st file we could download from source.

Legal Unit Table

The BI then uses a hybrid of machine learning and rules-based methods to ‘match and link’ the administrative data to achieve the Legal Unit. This information is stored in table format to showcase what data the ‘businesses’ has registered for with any high-level attributes. The legal unit table provides the link between records on the master tables: allowing access to a combined set of features for single entity (represented by a unique business index number). At the writing of this paper there are over 7 million legal unit entities created and this is updated daily.

Business Index number	Vat reference	Paye Reference	Companies House Reference	Vat guid	Paye guid	Companies house guid	Load date	Creation date
11111111								
22222222								

Legal units can be defined in various ways: Each will only contain one companies house (CH) record and this record can only be contained within one legal unit. Each legal unit can have multiple PAYE records. Multiple VAT can occur - but these are an exception and need clerical checking. Each PAYE and VAT can only be contained within one legal unit. Patterns are explained below as a series of tuples (n meaning usually 1 but can be greater in rare circumstances):

- **(CH, VAT) – (1,1,0)**
- **(CH, PAYE) – (1,0,1)**
- **(VAT, PAYE) - (0,1,1)**
- **(CH, VAT, PAYE) – (1,n,1)**
- **(CH, VAT, PAYE) – (1,1,n)**

Legal units can also be defined by single records, however this can represent a time lag in the arrival of the different records:

- **(CH) or (VAT) or (PAYE)**

Matching and Linking:
The hybrid approach we have adopted is using Splink probabilistic record matching combined with rules-based matching. This combines the advantages of both probabilistic and rules-based matching. This approach naturally fits a key structural divide within our data: namely differences between corporate and non-corporate entities. The differences and the respective solutions will be developed in this section.

Why hybrid matching?
The hybrid method was not our first approach. Through exploratory data analysis (EDA), domain expertise and trial of earlier methods (supervised and unsupervised methods) we discovered key differences amongst corporate and non-corporate records. Fundamentally, features relating to location, industry and legal status are not reliable matching features for corporates. Corporates can have many locations, different teams completing VAT, PAYE and Companies house returns. This leads to frequent differences in returns which represent the same corporate organisation. In contrast, non-corporates have much greater consistency across location, industry and legal status: making them more effective predictors of record linkage. This leads to the need for a diversity of approach.

Matching features

	Matching features
Corporates	Primary Name, Secondary Name [trading as], *Birth [registration date], *Postcode *Only needed to confirm match
Non-Corporates	Primary Name, Secondary Name [trading as], Industry [SIC], Address, Birth [registration date], Postcode, Legal Status [such as sole proprietor]

Data preparation
Prior to record matching, all databases underwent pre-processing to standardise and clean the data. This included:

- **Removing** special characters and spaces
- **Converting** text to lowercase
- **Standardising** company/organisation names
- **Stop words** were also removed from company names including titles, connecting words and punctuation.
- **Investigating missing and null value**s with any inconsistencies fed back to data engineering in the creation of the master tables
- **Continuous improvement:** All forms of pre-processing are kept up to date, and evolve based on findings from the logs

Birth table and the master tables

Birth table presents new records within a 35-day window. In our matching process the birth table is matched to the master tables of existing records to determine match pairs. For example, when a new corporate record is created on Companies House, we match it to both VAT and the PAYE master table. The matching process runs daily and can include hundreds to thousands of births to match to existing records contained within the master tables (these contain millions of records).

Rules-based matching

We use rules-based record for corporate business records because it allows for greater control and transparency in the matching process. Rather than relying solely on automated algorithms or probabilistic matching, predefined rules are used to match records based on specific criteria.

A simplified version of the rules-based match is as follows:

1. Check if the record is a corporate or not by identifying if it contains suffixes such as 'ltd', 'plc', 'llc'.
2. Attempt to match the corporate name to a tax record [VAT, PAYE] name via exact string match.
3. If more than one company's name matches to a tax record name (many to one), output the matches to a log for further review by the clerical team.
4. The clerical team will then resolve the match by referencing the postcode and pre-cleaned name, features which indicate record update/replacement for each record.
5. Once the match is resolved, update the relevant records with the matching tax information.
6. Sequence of match matters: from analysis and domain knowledge understanding we expect that the companies house record will exist before the arrival or update of the tax records. Matches out of sequence are reported in the log.
7. Stranded corporate tax records if a VAT or PAYE reference is identified as a corporate and does not match a Companies House record is reported in the log for clerical investigation.
8. Time window can be used to ensure that records are only matched if they were created within a certain time. However, our analysis on our early pipeline line deployment have led to us expanding this window [20-year window].

Overall, rules-based record matching has proved a more effective matching approach for corporate business records [leading to significant increases in accuracy] because it allows for greater control and transparency in the matching process. It can also be tailored to specific needs and criteria, making it a flexible and reliable method for record matching.

Probabilistic matching
For records identified as non-corporates Splink is used. Splink is a probabilistic record matching method that utilises machine learning algorithms to match records based on multiple features or attributes. Some of the key technical details of Splink include the use of blocking, levels of agreement, and additional conditions on individual levels of agreement. For more information see splink

We have created multiple models based on the historic data. This is a simplified explanation of the matching models created:

1. Blocking is a common technique used in record matching to reduce the number of comparisons needed between records. Our simplest model blocks on the first 4 characters of postcode. Later models combine multiple blocking rules to increase the number of record comparisons.
2. Levels of agreement are used in Splink to determine the overall similarity between records. The records are compared based on attributes such organisation name, address, postcode, legal status and industry. Each level of agreement is assigned a "gamma" value [0,1,2], which reflects the degree of similarity between records on that attribute, 0 being the lowest level of similarity. Our modelling states that the gamma of organisation name should be 1 or above this is equivalent to >80% jaro-winkler similarity.
3. The weighted combined match threshold combines the gammas to form our threshold [55.6], as to whether a record is a match or not. We set this threshold via clerical feedback based on comparisons to the IDBR.
4. The clerical team review the output from Splink and the data science team make any necessary corrections or adjustments. This is part of the wider logging and reporting process that has been critical to our success.
5. Finally, a time window of 4 years has been used to limit the scope of the matching process to records created within a specific timeframe.

Overall, Splink is a powerful and flexible probabilistic record matching method that can be tailored to specific needs and criteria. The use of blocking, levels of agreement, additional conditions, feedback, and time window can help to ensure accurate and reliable matching of records.

Scalability of solution
We are deploying our solutions on scalable tools: namely Spark/hive/hdfs. However, solve speeds are still challenging [30 – 60 mins] due to the scale of our daily comparisons.
Daily logging and reporting
Frequent feedback between the domain experts and data scientists have been enabled using daily logging and reporting. Daily logging and reporting have been critical for the development, improvement, and maintenance of machine learning pipelines. Here are some reasons why:

1. Detecting errors and bugs: By logging the input data, output data, and any errors or exceptions that occur during the rules based or machine learning pipeline's execution, we can quickly identify errors and bugs in the system. This helps us diagnose and resolve issues before they become critical.

2. Monitoring performance: Logging and reporting enable us to monitor the performance of the rules-based matching and machine learning to improve technique and performance: such as adjustments to stop word lists, cleaning methods or blocking.

Agile approach (early pipeline deployment)
We have used an agile approach to the early deployment of a hybrid matching pipeline this has allowed for quick iterations and feedback cycles. with continuous deployment through daily logging, domain expertise, and data scientists, can lead to faster development cycles, better collaboration, reduced risk, improved performance, and greater flexibility. We deployed our original pipeline in September 2022 through effective logging and early deployment have allowed us to create a highly accurate matching pipeline.

Results Section

The hybrid approach of using Splink with a rules-based approach for record matching demonstrated high accuracy match rates across multiple datasets. The accuracy increased steadily over time due to the early deployment of the machine learning pipeline and continuous monitoring and feedback from the domain expert based on daily logs and reports.

Across the datasets, the original machine learning approach achieved an initial daily match rate of around 30%. After months of continuous monitoring and feedback, the daily match rate improved to >70 to 80%. In comparison to the IDBR across our 7.5 million matches we are 99% the same - over several years of matching. False matches have remained to a minimal [<1%] in comparison to the IDBR. The reasons for the lower daily match rate (70-80%):

- Records can come in at significant time periods apart: for example, a business may file for a vat return months/years before paye
- Some legal units are genuinely single source: for example, they are based on vat only

This success is a result of continuous tuning of the Splink algorithm and rules-based approach by the data scientists, with the supported by the analysis of the domain expert and feedback from our key customers.

Key results

matched CH/VAT – ML pipeline (8/22)	matched CH/VAT – hybrid pipeline (4/23)
304 (30.0%*)	763 (75.3%*)
309 (34.6 %*)	694 (77.7%*)
781 (34.0%*)	1726 (75.2%*)
2226 (37.3%*)	4691 (78.6%)

Conclusion

Overall, the hybrid approach of using Splink with a rules-based approach demonstrated high accuracy match rates across multiple datasets. The accuracy improved steadily over time due to the early deployment of the hybrid pipeline and continuous monitoring and feedback from the domain expert. These results demonstrate the effectiveness of the hybrid approach in matching business data accurately and reliably.

Next steps

1. **Front end development** for customer interaction
2. **Splink modelling improvements**: expanding blocking
3. **More data** FCA, GLEIF and Charities
4. **Further integration** across indexes