Journal of Data Science and Modeling, Vol. 2, No. 1, 61-77, December 2023

Research Manuscript

# Automatic evaluation of record linkage methods in the expert system of producing statistical registers: string metric approach

Vadood Keramati*[1], Ramin Sadeghian[2], Maryam Hamedi[3], Ashkan Shabbak[4]

[1] PhDstudent,Department of Payame Noor University, POBox3697-19395,Tehran,Iran.

[2] AssociateProfessor,Department ofIndustrialEngineering, Payame NoorUniversity, P.O. Box 3697-19395, Tehran,Iran.

[3] AssistantProfessor, Department ofIndustrialEngineering, Payame NoorUniversity, PO Box 3697-19395, Tehran, Iran.

[4] Associate Professor of Statistics Department, Research Institute of Statistics, postal code 1413717911, Tehran, Iran.

Received: 18/11/2023 Accepted: 03/06/2024

**Abstract:**

Record linkage is a tool used to gather information and data from different sources. It is used in activities related to government, such as e-government and the production of register-based data. This method compares the strings in the databases, and there are different methods for record linkage, such as deterministic and probabilistic assumptions. This paper presents a proposed expert system for record linkage of data received from multiple databases. The system is designed to save time and reduce errors in the process of aggregating data. The inputs for this system include several linked fields, thresholds, and metric methods, which are explained along with the evaluation of the used method. To validate the system, inputs from two databases and seven information fields, comprising 100,000 simulated records, were used. The results reveal a higher accuracy of possible record linkage compared to deterministic records. Furthermore, the highest linkage was achieved using five fields with varying thresholds. In assessing the various metric methods, it was found that metric methods with less than 80% accuracy

*Corresponding Author: v.keramati@student.pnu.ac.ir

and the Winkler metric method with over 86% accuracy were utilized. These findings demonstrate that the implementation of the proposed automated system significantly saves time and enhances the flexibility of selection methods.

**Keywords:** Record linkage, metric string, expert system, assessment.
**Mathematics Subject Classification (2010):** 62-XX, 62-04.

# 1. Introduction

In today's world, we are witnessing a significant increase in the number of electronic services and information systems. This has led to the creation of several databases in organizations and companies. However, since these information systems have been created using different methodologies, the data obtained from them is not uniform, especially when it comes to string values such as names, surnames, and addresses. Consolidating data from these systems is crucial for expanding their reach. Therefore, the need to consolidate diverse information and data from databases has become a priority.

It is important to note that using different data sources may cause some data to be missed, while combining data could result in duplication. To address this issue, record linkage can be used as a systematic and manual method for linking and aggregating data. However, merging data from various sources to create a unified statistical database comes with its own set of challenges, including:

- The number of records of the target population across different databases is unknown.
- The available data may contain errors, such as incorrect and repetitive identification codes, as well as inconsistencies between the identification number and personal information.
- Some datasets may have missing values.
- Existing records may appear in the same or different identifying fields across multiple datasets.
- There may be affixes, prepositions, and similar character lines (such as "I" and "j" in English and "ک" and "گ" in Persian) in the data that need to be considered.

This paper discusses how to compare letter strings in different databases to facilitate record linkage. It also addresses the inconsistency in some fields, particularly the identification number. Table 1 provides an example of this topic in both Persian and English languages.

In Table 1, the second set of three rows in Persian corresponds to a single person. However, upon comparing characters and identification numbers, it became evident that there were three distinct entities. The research system's design aims to reduce errors in matching records for Persian datasets. Users have two methods

Table 1: An example of the same strings with different spellings and wrong identification numbers in different databases

| No. | English | | | |
|---|---|---|---|---|
| | Father_Name | F_Name | Name | PIN |
| 1 | Jame | Smith | Jan | 1111111111 |
| 2 | Jane | Smith | Jon | 1234567890 |
| 3 | Jane | Smith | Jon | 1111111111 |
| | Persian | | | |
| 1 | حمیدرضا | حسن زاده | داوود | 1111111111 |
| 2 | حمید رضا | حسنزاده | داود | 1234567890 |
| 3 | حمدرضا | – – – | دادود حسن زاده | 23456789 |

available to match records with assurance and potential. They can select from various options to compare two strings in information fields using well-known string metric methods like the Winkler metric, Lonstein metric, and q-gram. To assess the effectiveness of each aggregation method, we analyzed two identity information databases containing seven fields: identity identification number, first name, last name, father's name, date of birth, gender, and postal code. We explored different scenarios and presented the resulting outcomes. However, each conformity method and metric has its own set of errors, including non-conforming conformants and non-conformant non-conformants. To tackle this issue, each conformity method and metric establishes specific threshold limits based on F-Measure, Recall values, and Precision evaluation methods.

This article proposes a new automated evaluation method for Persian data record-linking in expert systems that generate statistical records. Our approach uses a string metrics method, where we compare the similarity of records based on their string representations. This approach is different from traditional record-linking methods that typically rely on probabilistic or deterministic algorithms to match records based on their attributes. Our proposed method is more efficient and scalable because it is automated. It can also handle missing or incomplete data, which is a common challenge in record-linking applications.

We evaluated our proposed method against several traditional deterministic and probabilistic methods using various metrics and threshold limits. By calculating evaluation values and a confusion matrix for each method using labeling, we were able to demonstrate that our proposed method outperforms others in terms

of accuracy, precision, recall, and F-score.

The article also discusses the use of expert systems for record linkage and evaluation. An expert system imitates the decision-making processes of human experts in a specific field. In the context of record linkage and evaluation, an expert system can identify and connect related records across different datasets. It can also evaluate the quality and relevance of those records, leveraging advanced algorithms and domain-specific knowledge to navigate through vast amounts of data, establish meaningful connections, and make informed assessments. This approach automates and accelerates the record management process while enhancing the accuracy and consistency of evaluations, leading to more reliable and insightful results. Expert systems use various techniques such as rule-based systems, machine learning, and natural language processing. Rule-based systems rely on a set of if-then rules as a knowledge base to make decisions and are often used in expert systems to encode the knowledge of human experts.

## 2. Literature Review

Kevin O'Hare and his colleagues in 2019 (O'Hare and Jurek-Loughrey (2018)) in an article titled "A New Technique of Selecting an Optimal Blocking Method for Better Record Linkage," introduced record linkage as a process to identify and link pairs of records that represent a real entity. They (O'Hare and Jurek-Loughrey (2018)) have shown this operation in Figure 1.

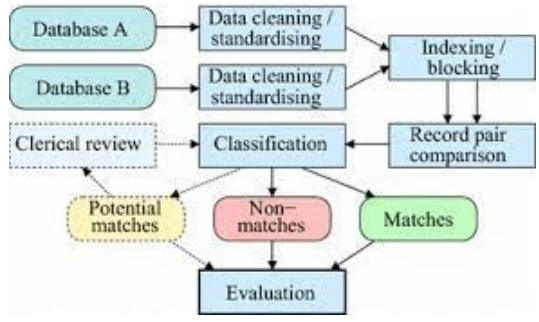


Figure 1: Overview of the record matching process (Kevin O'Hare, et al., 2019).

The overall record linkage process, as shown in Figure 1, consists of several main steps. These steps include cleaning and standardization, blocking and numbering, measuring records in the same blocks using the desired metrics, and forming similarity vectors (match, non-match, possible match). Now suppose we have two

datasets A and B that have m and n records respectively. Each pair of records, with one record from each of the two databases, can correspond to a correct match (called matched) or an incorrect match (non-matched) (Nanayakkara and Christen (2022)).

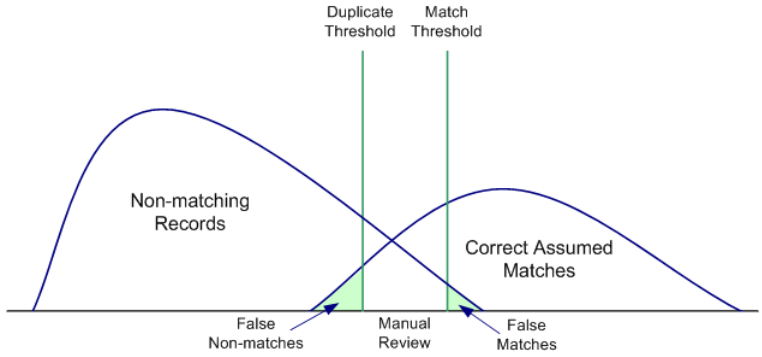


Figure 2: Distribution of match, non-match, and possible match with threshold selection (Tromp et al. (2011))

To evaluate the quality of the predicted clusters, traditionally precision and recall are used where these assess the correctness of the compared record pairs. Each record pair appearing in the same predicted cluster is considered as a positive link prediction, whereas a record pair belonging to two different predicted clusters is considered as a negative link prediction. The counts of true positive, false positive, true negative, and false negative are obtained concerning how record pairs appear in the true clusters, as shown in the error or confusion matrix (Hand and Christen (2018)) in Table 2.

In an article entitled "Estimating Precision and Recall for Deterministic and Probabilistic Record Linkage," Wiley Online Library, International Statistical Review, Chipperfield et al. (2018) showed that linking administrative, survey, and census records to enhance dimensions such as time and breadth or depth of detail is now common. Because a unique personal identifier is often not available, records belonging to two different entities (e.g., individuals) may be incorrectly linked. Estimating the proportion of correct, so-called exact matches is difficult because, even after semantic checking, there will be uncertainty about whether a link is correct or incorrect. Precision measurements are useful when deciding whether two files are worth matching, when comparing alternative correlation strategies, and as a quality measure for adaptive file-based estimates (Chipperfield et al. (2018)).

"The Sortal Concept in the Context of Biomedical Record Linkage" is the title

Table 2: Confusion matrix or error matrix to evaluate the quality of record linkage

| | | Ground-truth | |
|---|---|---|---|
| | | Matches | Non-matches |
| Prediction | Positive Link | **True Positives (TP)**-Record pairs that appear in the same cluster both in the ground-truth and in the prediction. Known as true matches. | **False Positives (FP)**-Record pairs that appear in the same cluster in the prediction but in different clusters in the ground-truth. Known as false matches. |
| | Negative Link | **False Negatives (FN)**-Record pairs that appear in the same cluster in the ground-truth but in different clusters in the prediction. Known as false non-matches or missed matches. | **True Negatives (TN)**-Record pairs that appear in different clusters both in the ground-truth and prediction. Known as true non-matches. |

of the research published by Miletic and Sariyar (2022). In this research, they have focused on the application of record linkage in the data matching of patients in different respiratory states. In this study, while checking the data before and after the experiment, they used record linkage to match the data related to the same people (Miletic and Sariyar (2022)).

In an article entitled "Errors in reported ages and dates in surveys of adult mortality: A record linkage study in Niakhar (Senegal)," Masquillaire et al. (2021) using the probabilistic record linkage technique, age and date reporting errors have matched and evaluated with statistical techniques in sibling histories collected during a validation study in the health and population surveillance system of Niakhar (Senegal) (Masquillaire et al. (2021)). Mahmud et al. (2021) presented an article titled "Effect of Race and Ethnicity on Influenza Vaccine Uptake among Older US Medicare Beneficiaries: A Record-Linkage Cohort Study" in which a historical record-linkage cohort study using Medicare databases (a US national health insurance program) included all elderly (over 65 years) enrolled in Medicare during the study period (July 1, 2015, to June 30, 2016) (Mahmud et al. (2021)).

"Psychological distress, life expectancy, and all-cause mortality in the United States: results from the 1997-2014 NHIS-NDI record linkage study," is the title of the article by Lee and Singh (2021). In this paper, they describe the linkage process, which includes identifying and qualifying participants from NCHS surveys,

creating and merging the submission record, which combines identifying data such as Social Security Number or first and last name, with NDI data, and performing and checking the match. They have presented their data and analysis in the form of Cox regression and standard life table method. "Estimating Pneumococcal Vaccine Coverage among Australian Indigenous Children and Children with Medically At-Risk Conditions Using Record Linkage" is the title of the research by Kabir et al. (2021). In this study, they analyzed data from a retrospective cohort of 1.3 million children born between July 2001 and December 2012 in two Australian states - NSW and WA - which together represent 42% of the total Australian population, and annually make up 125,000 births. All singleton live births recorded in both state-based birth registration and perinatal datasets were entered and most likely matched with ACIR, death, and hospital admission datasets using name, date of birth, residential address, and gender, with an estimated matching accuracy of 99% (Kabir et al. (2021)).

# 3. Proposed Model

A Python-based system has been designed for this research, which requires input files in .csv format. The system consists of four main parts:

1. Dataset entry section
2. Determining the record linkage method section, which includes:
    - Deterministic record linkage
    - Probabilistic record linkage, which involves determining the minimum number of linkage fields, probability thresholds, string metric method, and blocking variable(s).
3. Evaluation section, which includes determining the percentage of test data to evaluate and calculating the values of F, P, and R.
4. Output section, which includes the number and file list of each linkage pair, the time of each linkage, the linkage percentage of each method, and the values of F, Recall, and Precision.

In addition, the system knowledge base includes cleaning algorithms for Persian datasets, string metrics, and record linkage.

## 3.1 Input Section

In this section, users can upload files in .csv format to the system with no limit on the number of records. The system has a fixed number of fields, with seven

fields in total. These fields include Name, Family, FatherName, Sex, DateOfBirth, NationalCode, and PostalCode. Once the files are uploaded, the system will clean the data based on the rules in the system's knowledge base.

## 3.2 Determining the Record Linkage Method Section

In this section, you can choose from two methods provided by the system to perform the record linkage action: deterministic and probabilistic record linkage. If you choose the deterministic method, the system will match the datasets with each other and provide the result definitively. On the other hand, if the probabilistic method is selected, you will need to select four components as follows:

1. Threshold of the probability of linkage to the record (a number between 0 and 1)
2. Minimum number of linkage fields
3. String metric method (Winkler, Lonstein, Jaro, and more)
4. Determining the blocking variables for classification (default variables include national number, gender, and province)

Additionally, you need to determine the percentage of test data to evaluate and calculate the F-Measure.

## 3.3 Knowledge Base Section

This section has four subsections, which are described below:

### 3.3.1 Cleaning Algorithm Focusing on Persian Datasets

There are four subsections in this section, and each of them is described below:

1. All Arabic letters and numbers are converted to Farsi.
2. Any non-numeric character, except the letter 'o' (which can be used instead of zero), is removed from the national code.
3. The letter 'o' is converted to zero if it appears in the national number.
4. If the length of the national number is less than 10, one or two zeros are added to the beginning of the number.
5. The character '–' in the date of birth is converted to '/'.
6. The name of the month is converted to a number, such as "January" to "01".

7. The validity of the national identity number is checked for compliance with the national number algorithm, and any incorrect numbers are removed.

### 3.3.2 String Metric Algorithm

This system offers various string metric methods that can be utilized to match the potential record. The choice of method used is based on the user's preference and includes the following options: Levenshtein, Jaro-Winkler, Jaro, Damerau-Levenshtein, q-gram, Smith-Waterman, Longest Common Subsequence (LCS), and Cosine distance.

### 3.3.3 Record Linkage Algorithm

In the linkage process, first, a table is created for two databases by subtracting the number of block variables from 14. The fields of the first database are then selected to create the final result. Any records from these two databases that were not selected in the linkage process are collected in another database. Based on the provided inputs such as threshold, minimum number of fields, metric method, block variables, and percentage of test data, a dataset is created from the community of matched records and non-matched residuals. Finally, the system provides the consumption time, the value of F, and the degree of linkage.

## 3.4 Implementation of Probabilistic Record Linkage

In this section, the system performs record matching between two datasets. Firstly, the selected databases are compared, regardless of the selection method. Then, the matched records are considered, and the record associated with the first database is treated as the base. Any records that do not match any other database are added to the final file as residuals. At last, the records that are common to all databases are appended to the matched database records, and duplicate records are removed using the deduplication technique. These adaptations are done based on the user's requirements, and for each feature, we indicate good with 1 and bad with 0. The minimum number of adaptations can be 4, 5, 6, or 7, depending on the scenario. This method examines the number of possible matches of the remaining records in each dataset leading to a definitive match.

## 3.5 System Output

This section combines the resulting datasets based on the record linkage results from the previous steps, using the algorithm in the knowledge base. The final

dataset is then presented in a .csv file format. Additionally, the following calculations are provided:

- The percentage of possible matches for each pair of data
- The percentage of actual matches for each pair of data
- The system consumption time for each compliance
- The evaluated F value and other related values for each compliance method and metric.

# 4. System Performance Evaluation

To evaluate how well this system performs, two datasets in .csv format were created. The first dataset has 80,000 records and the second dataset has 70,000 records. These datasets were randomly selected from a centralized database that contains 100,000 records. The datasets have been noised and presented as input to the system to produce an output. By comparing the percentage of records obtained from the original file, we can determine how well each method performs. However, it should be noted that the cleaning algorithm may not be able to fully resolve all of the presented noises.

## 4.1 Deterministic Record Linkage Output

In this section, both datasets were matched using various metric methods based on all seven available fields. The results were obtained by combining the common records of both datasets and their non-common records, as described in Table 3:

Table 3: Deterministic Record Linkage Results

| | Cosine | LCS* | Smith-Waterman | q-gram | Damerau | Jaro | Jaro–Winkler | Levenshtein |
|---|---|---|---|---|---|---|---|---|
| Matched | 45190 | 45170 | 45174 | 45301 | 45268 | 45169 | 45193 | 45205 |
| Dataset1 non-matched | 34628 | 34647 | 34642 | 34521 | 34550 | 34655 | 34642 | 34619 |
| Dataset2 non-matched | 25210 | 25230 | 25226 | 25099 | 25132 | 25231 | 25207 | 25195 |
| Number of final file records | 105028 | 105047 | 105042 | 104921 | 104950 | 105055 | 105042 | 105019 |

After analyzing the results, it appears that the q-gram metric has a higher accuracy compared to the other methods. On the other hand, the metric method has a lower accuracy than the rest due to the final number of records being closer to 100,000.

*longest common subsequence

### 4.2 Probabilistic Record Linkage Output

Adaptation has been performed on two datasets in this section. The thresholds of 70%, 80%, and 90% along with the minimum number of fields of 4, 5, and 6 fields and all the string metrics of the field have been taken into consideration. The results obtained from these situations are described in the following section.

The percentage of linkage in each scenario is provided based on the threshold percentage, the minimum number of compliant fields, and the metric method. The last line of each table shows the time required for compliance for all considered states. A comparison chart is also provided, which presents the evaluation values of the tested scenarios. To prepare the data related to the evaluation part after the linkage operation, a random cut of 10% of the first dataset was selected. The evaluation values were calculated by adding noise again and matching with the second dataset based on the specified labeling.

## 5. Summary of System Performance Results for Test Data

The records of the first and second datasets were matched based on a threshold of 70 to 90 percent for the number of fields 4 to 6 using different metric methods. The results of this matching are summarized in Tables 4, 5, and 6, which include the percentage of compliance for each method and the time consumed by the system memory. To better understand the comparative behavioral process of different methods, a diagram is presented below.

Table 4: Summary of System Test Results for Linking Datasets 1 and 2 with a Threshold of 70%

| | Cosine | LCS | Smith-Waterman | q-gram | Damerau | Jaro | Jaro–Winkler | Levenshtein |
|---|---|---|---|---|---|---|---|---|
| 4 | 96.3 | 95.67 | 95.97 | 95.57 | 96.21 | 96.94 | 97.06 | 95.53 |
| 5 | 95.8 | 94.89 | 95.11 | 94.44 | 94.93 | 96.16 | 95.87 | 94.45 |
| 6 | 76.1 | 62.13 | 60.73 | 51.02 | 50.68 | 83.12 | 81 | 49.59 |
| Time(sec) | 5.6 | 3.22 | 8.31 | 2.64 | 2.31 | 2.08 | 2.06 | 2.26 |

Upon reaching the threshold of 70%, it was observed that the Smith-Waterman method took longer to complete when compared to the other methods in the system. The consumption times of the remaining methods, except for the Cosine method, were almost the same. At this limit, the threshold behavior of the metrics remained consistent for up to 5 fields, but differed for 6 fields. At this threshold, two categories of metric methods were identified - Lonstein and q-gram - which displayed similar behavior but differed from the others with a greater drop than LCS

and Smith-Waterman. They also faced a greater drop in compliance percentage when compared to the others.

Table 5: Summary of System Test Results for Linking Datasets 1 and 2 with a Threshold of 80%

| | Cosine | LCS | Smith-Waterman | q-gram | Damerau | Jaro | Jaro–Winkler | Levenshtein |
|---|---|---|---|---|---|---|---|---|
| 4 | 95.67 | 95.97 | 95.57 | 96.19 | 96.29 | 96.48 | 95.53 | 95.53 |
| 5 | 94.64 | 94.76 | 94.26 | 94.73 | 95.28 | 95.56 | 94.32 | 94.45 |
| 6 | 53.14 | 51.91 | 46.39 | 46.99 | 58.44 | 72.34 | 46.35 | 49.59 |
| Time(sec) | 3.26 | 10 | 2.63 | 2.31 | 2.1 | 2.04 | 2.26 | 2.26 |

When the threshold is set at 80%, the LCS method takes longer to complete compared to other methods in the system. The consumption times of the other methods are almost the same. When the number of fields is up to 5, the behavior of the metrics is almost the same. However, there is a difference in the number of fields when it comes to 6 fields. In this scenario, the Jaro metric method has a smaller drop in compliance percentage compared to the other methods.

Table 6: Summary of System Test Results for Linking Datasets 1 and 2 with a Threshold of 90%

| | Cosine | LCS | Smith-Waterman | q-gram | Damerau | Jaro | Jaro–Winkler | Levenshtein |
|---|---|---|---|---|---|---|---|---|
| 4 | 96.28 | 95.67 | 95.97 | 95.57 | 96.19 | 96.29 | 96.17 | 95.53 |
| 5 | 94.76 | 94.16 | 94.13 | 94.13 | 94.5 | 94.85 | 94.78 | 94.22 |
| 6 | 57.9 | 47.08 | 46.38 | 46.13 | 46.61 | 47.45 | 49.47 | 46.04 |
| Time(sec) | 6.29 | 3.18 | 10.81 | 2.61 | 2.34 | 2.1 | 2.06 | 2.29 |

At the 90% threshold, much like the 70% and 80% thresholds, the Smith-Waterman method took longer than other methods in the system. The consumption times of the remaining methods, except for the Cosine method, were nearly identical. At this threshold limit for the number of 6 fields, due to the high accuracy of compliance, almost all metric methods experienced a significant drop and fell below 60%. However, in this limit, the threshold performed better for 6 cosine fields.

## 5.1 Evaluation of System Performance for Different Metric Methods in Probabilistic Record Linkage

In this section, we checked the performance of the system in record linkage for different scenarios and metrics. We used the F-Measure calculation to evaluate the parameters. To do this, we separated ten percent of the first dataset that was labeled and re-noised it. Then, we compared the records that remained from

definitive matching with the second dataset. We calculated Recall, Precision, and F-Measure values using four scales: TP, FN, FP, and TN. We have presented some of the results in the following charts.

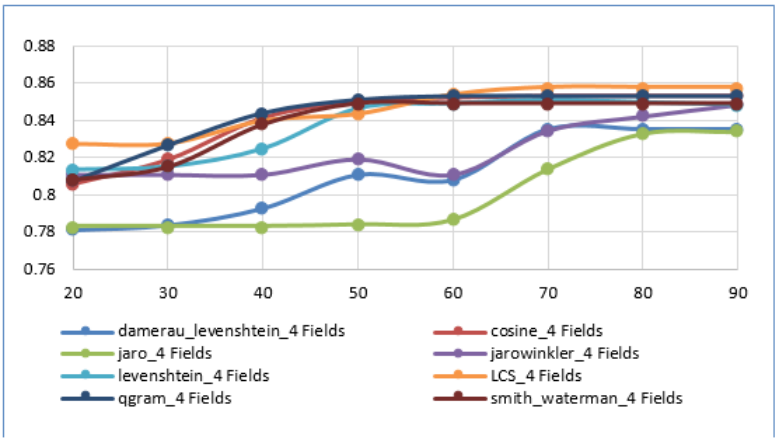


Figure 3: Comparison of evaluation values of different metric methods for 4 fields in different thresholds.

In the evaluation section of compliance metric methods, the chart in Figure 3 shows the results for different threshold ranges from 20 to 90 percent for four fields in linkage. At the 90 percent F-Measure threshold, the results are close to each other. However, Jaro's method saw a 70% increase in F-Measure with delay at the 70% threshold. For five fields, different threshold ranges from 20% to 90% showed that the obtained F-Measures are close to each other at the 90% threshold. In this case, Jaro's F-Measure increased by 70% with a delay. Figure 4 shows that for different threshold ranges from 20% to 90% for five fields in record linkage, at the 90% threshold, the obtained F-Measures are close to each other and less than 0.6 for all metrics. In this scenario (6 fields), except for Jaro Winkler's method, which had a higher evaluation than the others at the 80% threshold, the behavior of the rest of the metrics was almost the same.

## 5.2 Comparison between Systematic and Old Approaches for Linking Two Datasets

Table 7 presents a comparison between systematic and old approaches for linking two datasets using two famous metrics: Jaro-Winkler and Levenshtein. We conducted this comparison with two datasets containing 3,428,000 and 2,260,000 records, respectively. The results include the percentage of matches and evaluation

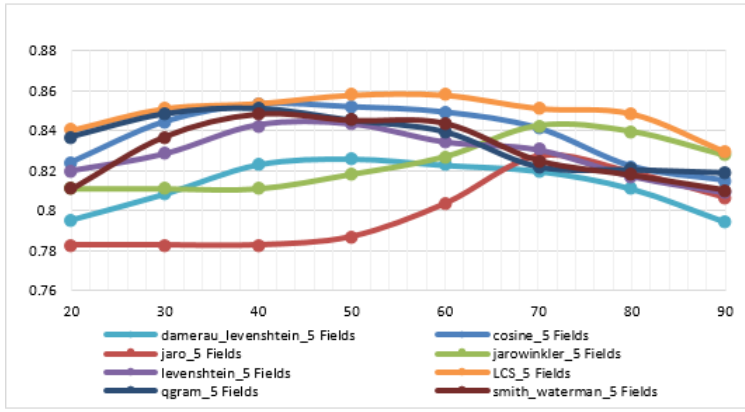


Figure 4: Comparison of evaluation values of different metric methods for 5 fields in different thresholds.

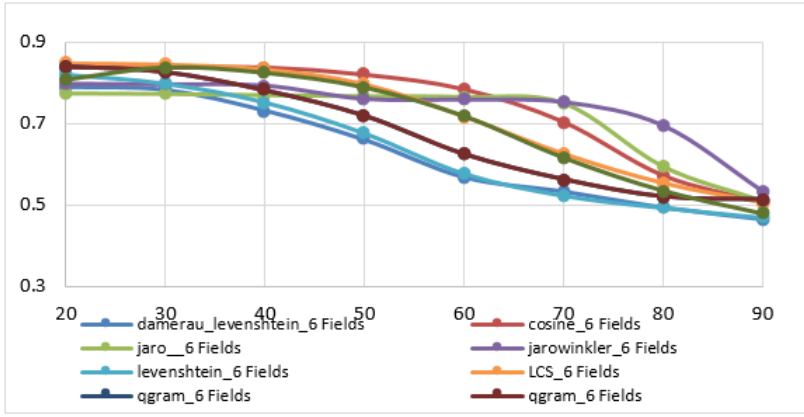


Figure 5: Comparison of evaluation values of different metric methods for 6 fields in different thresholds.

measures for different thresholds (0.5, 0.6, 0.7, 0.8, and 0.9) for varying numbers of fields.

## 6. Conclusion

In conclusion, this article proposes the development of an expert system capable of integrating data from multiple databases into a unified database. The complexity of the system lies in its ability to offer users flexibility in selecting various record linkage methods, metric methods, blocking fields, and the number of desired fields.

Table 7: Comparison between Systematic and Old Approaches for Linking Two Datasets

| Threshold | Metric | 4 fields | 5 fields | 6 fields | %Matches | Recall | Precision | F-Measure | Time(sec) | Memory(GB) |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 | Jaro-Winkler | 78.43 | 84.21 | 87.34 | 91.2 | 0.86 | 0.89 | 0.87 | 120 | 1.5 |
| 0.6 | Jaro-Winkler | 79.12 | 85.32 | 88.45 | 92.3 | 0.87 | 0.90 | 0.88 | 130 | 1.6 |
| 0.7 | Jaro-Winkler | 80.23 | 86.43 | 89.56 | 93.4 | 0.88 | 0.91 | 0.89 | 140 | 1.7 |
| 0.8 | Jaro-Winkler | 81.34 | 87.54 | 90.67 | 94.5 | 0.89 | 0.92 | 0.90 | 150 | 1.8 |
| 0.9 | Jaro-Winkler | 82.45 | 88.65 | 91.78 | 95.6 | 0.90 | 0.93 | 0.91 | 160 | 1.9 |
| 0.5 | Levenshtein | 76.54 | 82.32 | 85.45 | 89.3 | 0.84 | 0.87 | 0.85 | 110 | 1.4 |
| 0.6 | Levenshtein | 77.65 | 83.43 | 86.56 | 90.4 | 0.85 | 0.88 | 0.86 | 120 | 1.5 |
| 0.7 | Levenshtein | 78.76 | 84.54 | 87.67 | 91.5 | 0.86 | 0.89 | 0.87 | 130 | 1.6 |
| 0.8 | Levenshtein | 79.87 | 85.65 | 88.78 | 92.6 | 0.87 | 0.90 | 0.88 | 140 | 1.7 |
| 0.9 | Levenshtein | 80.98 | 86.76 | 89.89 | 93.7 | 0.88 | 0.91 | 0.89 | 150 | 1.8 |

Additionally, evaluating the effectiveness of these methods using F-Measures poses a significant challenge.

The study conducted a thorough analysis of two datasets to assess the system's performance. These datasets were randomly sampled from a larger dataset containing 100,000 records. Both probabilistic and deterministic record linkage methods were tested and evaluated. The results indicate that the probabilistic record linkage method generally outperformed the deterministic method in terms of accuracy. Among the metric methods, the Jaro metric method demonstrated superior performance, both in terms of computation time and F-Measure.

# 7. Recommendations

Based on the findings of this study, several recommendations for future research and system development emerge:

1. **Utilizing Artificial Intelligence and Machine Learning:** Future iterations of the system could leverage advanced techniques such as AI and ML to enhance performance and automate decision-making processes.

2. **Flexibility in Handling Varying Numbers of Fields:** Enhancing the system's capability to handle datasets with different numbers of fields would increase its versatility and applicability to diverse datasets.

3. **Integration of Data Mining Techniques:** Incorporating data mining techniques such as dimension reduction could improve system efficiency and scalability, particularly when dealing with large datasets.

By implementing these recommendations, the proposed expert system can evolve into a robust tool for efficient and accurate data integration across multiple databases.

## References

Kabir, A., Newall, A. T., Randall, D., Menzies, R., Sheridan, S., Jayasinghe, S., Fathima, P., Liu, B., Moore, H., McIntyre, P., & Gidding, H. F. (2021). Estimating pneumococcal vaccine coverage among Australian Indigenous children and children with medically at-risk conditions using record linkage. *Vaccine*, **39(12)**, 1727-1735.

Masquelier, B., Kanyangarara, M., Pison, G., Kanté, A. M., Ndiaye, C. T., Douillot, L., Duthé, G., Sokhna, C., Delaunay, V., & Helleringer, S. (2021). Errors in reported ages and dates in surveys of adult mortality: A record linkage study in Niakhar (Senegal). *Population Studies*, **75(2)**, 269-287.

Nanayakkara, C., & Christen, P. (2022). Locality Sensitive Hashing with Temporal and Spatial Constraints for Efficient Population Record Linkage. In *CIKM22: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 4354–4358.*

Hand, D., & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Stat Comput*, **28**, 539–547

Lee, H., & Singh, G. K. (2021). Psychological distress, life expectancy, and all-cause mortality in the United States: Results from the 1997-2014 NHIS-NDI record linkage study. *Annals of Epidemiology*, **56**, 9-17.

Chipperfield, J., Hansen, N., & Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *International Statistical Review*, **86(2)**, 219-236.

O'Hare, K., & Jurek-Loughrey, A. (2018). A new technique of selecting an optimal blocking method for better record linkage. *Information Systems*, **77**, 151-166.

Miletic, M., & Sariyar, M. (2022). The sortal concept in the context of biomedical record linkage. In *Stud Health Technol Inform*, **295**, 293-297.

Mahmud, S. M., Xu, L., Hall, L. L., Puckrein, G., Thommes, E., Loiacono, M. M., & Chit, A. (2021). Effect of race and ethnicity on influenza vaccine uptake among older US Medicare beneficiaries: A record-linkage cohort study. *Lancet Healthy Longev*, **2(3)**, e143-e153.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., & Reitsma, J.B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology*, **64(5)**, 565–572.