Application of Record Linkage Techniques for Road Infrastructure Data Integration: A Study Using R and Real Data

María José Ginzo Villamayor[1]

Abstract

The integration of heterogeneous databases in the field of road infrastructure represents a significant challenge due to inconsistencies in nomenclature, typographical errors, and the lack of unique identifiers. This paper presents a practical application of *Record Linkage* techniques to link records from various sources related to urban roads, using the R programming language. Deterministic, probabilistic methods (based on Fellegi and Sunter's theory), and fuzzy matching techniques were applied to real data from municipalities in Galicia (Spain), employing tools such as *RecordLinkage*, *fuzzyjoin*, and *stringdist*. The process included text normalization, blocking by municipality, and comparison based on similarity distances (Levenshtein, Jaro-Winkler), achieving a robust unification of records. The results show a substantial improvement in data quality, integrity, and usability, achieving precision levels above 90% and recall greater than 85% in the best scenarios. This approach has direct applications in infrastructure planning, road maintenance management, territorial analysis, and smart city development. The conference will include visual demonstrations and practical cases, displaying how the application of Record Linkage techniques can contribute to building integrated urban databases, essential in the context of Society 5.0. **Keywords:** *Record Linkage,* road infrastructure, fuzzy matching, data quality, data integration.

Introduction

The growing availability of urban databases presents significant challenges for their effective integration. In the field of road infrastructure management, it is common to find multiple sources describing the same entities—such as streets, roads, or highways—but in different formats, with varying levels of accuracy and even errors. This heterogeneity makes consolidating information difficult and can compromise the quality of subsequent analyses.

In this context, Record Linkage emerges as a statistical and computational solution aimed at linking records that refer to the same entity, even in the absence of common unique identifiers. The ability to connect records from diverse sources has become essential in the digital age, where data is not only abundant but also varied and structurally inconsistent. The Record Linkage process, also known as record reconciliation, allows for the identification and matching of observations that belong to the same person, company,

[1] Universidade de Santiago de Compostela (USC) and Centro de Investigación y Tecnología Matemática de Galicia (CITMAGA).

address, or other type of entity, thereby improving the integrity and utility of the data. This process aims to identify and join records referring to the same entity (person, company, product, etc.) across two or more databases, even when no exact common identifier exists.

This paper proposes the application of advanced Record Linkage techniques, implemented in the R programming environment, for real urban road data from different municipalities. The combination of deterministic, probabilistic, and fuzzy matching methods is presented as an effective strategy for improving the quality of integrated data, laying the foundation for more robust analyses within the framework of smart cities and Society 5.0.

Objective

The aim of this paper is to describe and apply various Record Linkage methods focused on integrating urban road data in the context of smart cities. A comparison is made between deterministic, probabilistic, and fuzzy matching methods, evaluating their advantages and limitations in complex urban data scenarios. Additionally, the effectiveness of specific tools from the R ecosystem in the processes of cleaning, comparison, and record linking will be analyzed, which are essential for ensuring data quality and interoperability in Smart City environments. Finally, a real-life case study will be presented to illustrate the improvements achieved in the quality of the integrated data.

Material and Method

For urban data integration, two main sources were used: a cadastral dataset containing the names and characteristics of urban roads, and a road maintenance database consisting of operational records on the sections that were intervened. Both datasets presented inconsistencies in fields such as *road_name*, *road_type*, and municipality, making a prior data cleaning and normalization step necessary. Key fields like name, birth date, and address were standardized, among other actions. The analysis was carried out in R, using various packages such as *RecordLinkage* (Sariyar and Borg, 2022), *fuzzyjoin* (Robinson, 2020), *stringdist* (Van Der Loo, 2014), *dplyr* (Wickham et al., 2023), and *tidyverse* (Wickham et al., 2019), selected for their ability to handle tasks such as approximate string comparison, calculation of distances or similarities between text strings (e.g., Levenshtein (Levenshtein, 1966) or Jaro-Winkler (Winkler, 1990)), and differences in dates or numbers, as well as for efficient data manipulation. **Table 1** shows the definition and an example of use for these distances and in **Figure 1** the explanation of the examples.

Table 1: Levenshtein e Jaro-Winkler Distance

Distance	**Example**
Levenshtein: Também conhecida como distância de edição entre duas strings, é definida como o número mínimo de operações (deleções, inserções ou substituições) necessárias para transformar uma string em outra.	Se a palavra original é "book"e a final é "back", você precisa substituir o primeiro "o"por "a"e o segundo "o"por "c". Como são necessárias duas modificações, a distância de Levenshtein entre "book"e "back"é 2.
Jaro-Winkler: Mede a similaridade entre duas strings considerando tanto o número de caracteres coincidentes quanto sua ordem relativa. Dá peso extra aos primeiros caracteres se forem iguais (prefixo comum). É uma extensão da distância de Jaro, útil em aplicações de Record Linkage e detecção de duplicatas.	Comparando "MARTHA"e "MARHTA", a distância de Jaro seria alta porque a maioria das letras coincide, embora em ordem diferente. Com o ajuste de prefixo do Jaro-Winkler, a similaridade fica ainda maior, já que "MAR"coincide no início de ambas as strings, reduzindo a distância final.
Source: GINZO VILLAMAYOR (own elaboration), 2025	

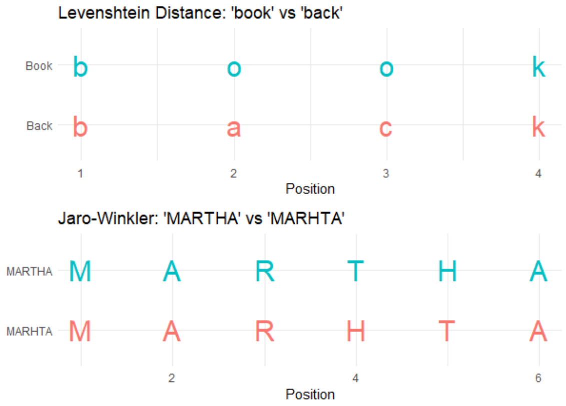

Figure 1: Top – Levenshtein Distance (“book” vs “back”). Bottom – Jaro-Winkler Distance (“MARTHA” vs “MARHTA”)

Source: GINZO VILLAMAYOR (own elaboration), 2025

Record Linkage (Christen, 2012) is a fundamental process in situations where information about the same entity is scattered across different databases or information sources. Its main objective is to identify and link records corresponding to the same entity, even when there is no common unique key, which improves the quality of data analysis. This technique is widely used in contexts such as healthcare, censuses (Winkler, 1999), government databases (Herzog et al., 2007), and urban studies related to public roads. Through record comparison methods and the assignment of matching probabilities, Record Linkage allows for the consolidation of fragmented data, providing a more accurate and comprehensive view of the analyzed entities.

There are several approaches to Record Linkage, and the choice depends on the characteristics of the available data. The deterministic approach relies on exact matching rules between key fields, such as the name or identifier of a road. Although it is faster and simpler, its effectiveness decreases when there are errors or inconsistencies in the data. In contrast, the probabilistic approach assesses the probability of a match between records, considering variations and common errors through statistical techniques. This group includes both classical approaches based on the theory proposed by Fellegi and Sunter (1969), as well as more recent techniques that employ machine learning and data mining. These newer methods not only optimize the linkage process but also allow for handling large volumes of data.

Additionally, the fuzzy matching approach is useful for detecting similar, but not identical, records, which is especially valuable when the data contains abbreviations, typos, or formatting differences (e.g., "Calle 23"and "calle veintitrés"). The combination of these methods facilitates tackling the inherent complexity of Record Linkage in urban data in a robust and efficient way.

The methodological procedure followed in this work began with a normalization phase of text, where all content was converted to lowercase, and accents and special characters were removed to standardize the character strings. Subsequently, an initial blocking by municipality was performed, reducing the number of comparisons needed and optimizing processing time. Record comparison was carried out using three complementary approaches: first, deterministic matching, based on exact matching of the name and type of road; second, fuzzy matching, using the functions *stringdist*() and *stringdist_left_join*() to detect approximate similarities between road names; and third, probabilistic comparison, implemented with the functions *compare.linkage*() and *linkage*() from the RecordLinkage package (Sariyar and Borg, 2022), adjusting similarity thresholds to maximize both precision and recall.

Figure 2 presents a visual comparison of three primary record matching techniques: Deterministic, Probabilistic, and Fuzzy Matching. The diagram outlines the key characteristics and distinctions of each approach, highlighting their underlying principles and typical applications in data linkage and comparison.

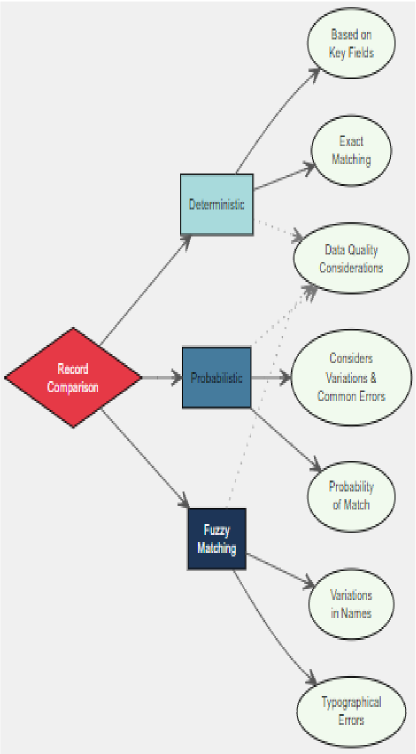

Figure 2: Comparative analysis of record matching techniques: Deterministic, Probabilistic, and Fuzzy Matching.

Source: GINZO VILLAMAYOR (own elaboration), 2025

In the R environment, various specialized tools were used to implement these techniques. The RecordLinkage package allowed for both probabilistic and deterministic comparisons using the *compare.linkage*() and *linkage*() functions, facilitating the detection of equivalent records based on matching probabilities. To address typos and variations in road names, the fuzzyjoin package (Robinson, 2020) was used, specifically the function *stringdist_left_join*(), which performs joins based on text distance using methods like Jaro-Winkler. Additionally, the stringdist package (Van Der Loo, 2014) was used to calculate similarity measures between strings, supporting the identification of fuzzy matches. Meanwhile, dplyr (Wickham et al., 2023) and the tidyverse suite of tools (Wickham et al., 2019) were used to clean, transform, and organize the data before and after the linkage process. **Table 2** presents a summary of the most used packages for Record Linkage in R.

Table 2: Common R Packages for Record Linkage.

Paquete	**Descripción**	**Autor(es)**
RecordLinkage	Complete package, allows probabilistic and deterministic linking.	Sariyar and Borg (2022)

Paquete	Descripción	Autor(es)
fastLink	Implements Bayesian models, fast and useful with large data volumes.	Enamorado, Fifield, and Imai (2023)
fuzzyjoin	Useful for joining tables with non-exact matches (similar to dplyr::join but with fuzzy logic).	Robinson (2020)
stringdist	Calculates string distances (Levenshtein, Jaro-Winkler, etc.).	Van Der Loo (2014)
deduped	More focused on record deduplication.	Gadish (2023)

Source: GINZO VILLAMAYOR (own elaboration), 2025

Record Linkage has various applications across multiple fields. In public health, it is used to integrate patient histories across different hospitals or health centres, providing a comprehensive view of a person's medical record. In the administrative sector, it facilitates the consolidation of census data, civil records, or tax data, improving the quality of governmental records. In marketing and CRM, it is employed to detect duplicate customer records and generate unified profiles, optimizing customer relationship management. In research, it is also useful for combining survey or cohort study data, allowing for a more complete and precise analysis of the studied variables.

As a practical application, this work used real data from urban roads sourced from two different databases: a municipal land registry with road names and characteristics, and a road maintenance database with operational records of interventions on specific sections. Both sources had inconsistencies in key fields such as road name, road type, and municipality, making the use of Record Linkage techniques essential for effectively integrating the information. Finally, manual validation was performed on a sample of linked pairs to calculate precision and recall metrics, ensuring the quality of the data integration process.

This comprehensive approach not only consolidated dispersed records but also highlighted the importance of applying advanced statistical and computational methods to improve data quality in urban contexts, contributing to better decision-making in infrastructure planning and management, aligned with the goals of Society 5.0 and sustainable urban development.

Practical example with road data

In this section, the Record Linkage process will be illustrated using real data from the autonomous community of Galicia (Spain). Due to the confidential nature of the data, specific details will not be presented. Let us assume we have two databases containing information about urban roads, but with some inconsistencies or variations in the names and types of roads. The data bases are as follows:

- **Data Base 1**: Land registry records with road names and types (**Table 3**).
- **Data Base 2**: Road maintenance plan records, where some names present errors or differences compared to the land registry database (**Table 4**).

Table 3: Land registry database records

id	road_name	road_type	municipality
1	Avenida Libertador	Avenida	Bogotá
2	Calle 50	Calle	Bogotá
3	Carrera 7	Carrera	Bogotá

Source: GINZO VILLAMAYOR (own elaboration), 2025

Table 4: Road maintenance plan records

id	road_name	road_type	municipality
101	Av. Libertador	Av.	Bogotá
102	Cll 50	Cll	Bogotá
103	Cra 7	Cra	Bogotá

Source: GINZO VILLAMAYOR (own elaboration), 2025

The diagram (**Figure 3**) illustrates the steps in the Record Linkage process.

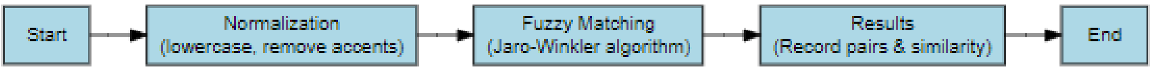

Figure 3: Record Linkage Process Flow.

Source: GINZO VILLAMAYOR (own elaboration), 2025

Procedure

The procedure to follow for applying the *Record Linkage* methodology is as follows:

Phase 1: Normalization

The first step is to normalize the data so that the strings are comparable. This involves converting all road names to lowercase and removing differences caused by accents or special characters. This process ensures that the comparisons between the road names are consistent.

Phase 2: Fuzzy Matching

Next, a fuzzy matching method is used to compare the records between the two databases. This method seeks similarities in the road names, even when they do not match exactly. The comparison is done through an algorithm based on the string distance, such as Jaro-Winkler. This type of comparison is useful for detecting records with typographical errors or small variations in road names.

Phase 3: Results

By applying the fuzzy matching method, we obtain **Table 5**, which shows the pairs of records that are considered similar, along with their similarity measure. The records "Avenida Libertador"and "Av. Libertador", "Calle 50"and "Cll 50", as well as "Carrera 7"and "Cra 7", are identified as approximate matches. The similarity between the pairs of records is calculated using a distance scale, where lower values indicate greater similarity between the names.

Table 5: Table 5 – Matching results

road_name DBase 1	**road_name DBase 2**	**Similarity**
Avenida Libertador	Av. Libertador	0.13
Calle 50	Cll 50	0.11
Carrera 7	Cra 7	0.12

Source: GINZO VILLAMAYOR (own elaboration), 2025

In **Figure 4**, a diagram is shown illustrating the Record Linkage process using two previous databases. The first database, Land Registry, contains the names and types of urban roads, while the second, Maintenance Plan, holds operational maintenance records. The process begins with Normalization, which transforms text strings into a uniform format (e.g., converting to lowercase and removing accents). Next, Fuzzy Matching is applied, where records between the two databases are compared using a string distance algorithm like Jaro-Winkler, allowing approximate matches to be found despite variations in the names. Finally, the Results display pairs of records that are considered similar, along with a numerical value indicating the degree of similarity. This allows the identification of matches between records, such as "Avenida Libertador"and "Av. Libertador, Calle 50"and "Cll 50,"and "Carrera 7"and "Cra 7,"along with their calculated similarity scores.

Results and Discussion

The results obtained showed significant differences between the methods applied for record linkage. The *deterministic approach* exhibited high precision, but its coverage was limited due to the strict requirement for exact matches between records, resulting in fewer linked records. This approach is suitable when the data is well-structured and error-free, but it is less effective in scenarios with variations or inconsistencies.

On the other hand, the *probabilistic method*, after adjusting the similarity thresholds, achieved a recall rate of over 85% and a precision close to 90%. This reflects a proper balance between the number of records retrieved and the quality of the links, standing out for its ability to handle records with errors or variations in the data. The probabilistic approach proved to be more robust in complex scenarios, as it does not rely on exact matches but instead allows for handling uncertainty and errors in the data.

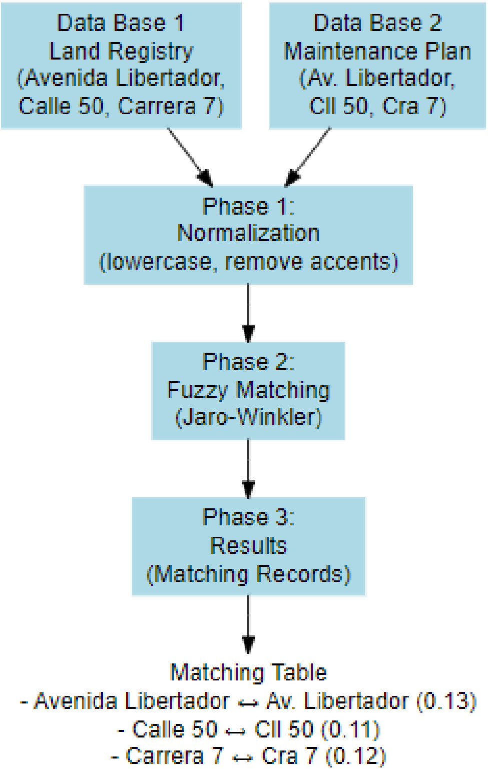

Figure 4: Record Linkage Process: Integration of Urban Road Data.
Source: GINZO VILLAMAYOR (own elaboration), 2025

Regarding the use of *Fuzzy Matching* techniques, it proved especially effective in identifying common variations in street names, such as abbreviations or typographical errors. For example, records such as "Cra"instead of "Carrera"were correctly matched, which helped overcome frequent inconsistencies in urban databases. This fuzzy matching approach is essential when the data contains common human errors or minor differences in the way names are written.

Data integration using these approaches significantly contributed to improving the integrity of the resulting database, enabling a more comprehensive analysis of the intervened streets. Additionally, it facilitated the relationship of the streets with other relevant variables, such as maintenance frequency or segment lengths, crucial factors for infrastructure planning in the context of smart cities. The results underline the importance of applying advanced record linkage methods to improve data quality and, consequently, support better decision-making in urban management.

Conclusion

The use of *Record Linkage* techniques in R proves to be an effective tool for addressing issues arising from the heterogeneity of data on road infrastructure. The combination of *probabilistic methods* and *fuzzy matching* turned out to be superior to *deterministic approaches*, especially in urban contexts where multiple sources of information with variations and errors are managed. This study demonstrates that by integrating data with these advanced techniques, a significant improvement in record quality is achieved, which, in turn, enables more informed and accurate decision-making in infrastructure management. Based on the results obtained, it is recommended to apply these methodologies in projects related to road planning, maintenance management, and territorial analysis, contributing to better data management and optimizing urban decision-making.

References

CHRISTEN, Peter. Data matching: concepts and techniques for *Record Linkage*, entity resolution, and duplicate detection. Springer, 2012.

ENAMORADO, Ted; FIFIELD, Benjamin; IMAI, Kosuke. fastLink: Fast Probabilistic *Record Linkage* with Missing Data. R package version 0.6.1, `https://CRAN.R-project.org/package=fastLink` 2023.

GADISH, Or. deduped: Making "Deduplicated"Functions. R package version 0.2.0, `https://cran.r-project.org/web/packages/deduped/index.html`. 2023.

HERZOG, Thomas N.; SCHEUREN, Fritz J.; WINKLER, William E. Data quality and *Record Linkage* techniques. Springer, 2007.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8), 707–710, 1966.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation, 2024.

ROBINSON, David. fuzzyjoin: Join Tables Together on Inexact Matching. R package version 0.1.6, `https://CRAN.R-project.org/package=fuzzyjoin`. 2020.

SARIYAR, Murat; BORG, Andreas. RecordLinkage: *Record Linkage* Functions for Linking and Deduplicating Data Sets. R package version 0.4-12.4, `https://CRAN.R-project.org/package=RecordLinkage` 2022.

VAN DER LOO, Mark. "The stringdist package for approximate string matching." The R Journal, *6*, 111-122. `https://CRAN.R-project.org/package=stringdist`. 2014.

WICKHAM, Hadley; AVERICK, Mara; BRYAN, Jennifer; CHANG, Winston; McGOWAN, Lucy D'Agostino; FRANÇOIS, Romain; GROLEMUND, Garrett; HAYES, Alex; HENRY, Lionel; HESTER, Jim; KUHN, Max; PEDERSEN, Thomas L.; MILLER, Evan; BACHE, Stephan M.; MÜLLER, Kirill; OOMS, Jeroen; ROBINSON, David; SEIDEL, Dana Paige; SPINU, Vitalie; TAKAHASHI, Kohske; VAUGHAN, Davis; WILKE, Claus; WOO, Kara; YUTANI, Hiroaki. "Welcome to the tidyverse." Journal of Open-Source Software, *4*(43), 1686. doi:10.21105/joss.01686 `https://doi.org/10.21105/joss.01686`, 2019.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; MÜLLER, Kirill; VAUGHAN, Davis. dplyr: A Grammar of Data Manipulation. R package version 1.1.2, `https://CRAN.R-project.org/package=dplyr`. 2023.

WINKLER, William E. "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of *Record Linkage.*"*Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1990, pp. 354–359.

WINKLER, William E. The state of *Record Linkage* and current research problems. U.S. Census Bureau, 1999.

FELLEGI, Ivan P.; SUNTER, Alan B. A theory for *Record Linkage.* Journal of the American Statistical Association, 1969.