FACULDADE
DE ECONOMIA
UNIVERSIDADE
DO PORTO

Record Linkage for Official Statistics: Establishment Level Matching in Portuguese Administrative Data

Daniela Filipa Simões Lopes

Dissertation

Master in Modeling, Data Analysis and Decision Support Systems

Supervised by:

Professor Fernando Silva

Co-supervised by:

Professor Marta Silva

2026

Acknowledgements

I would first like to express my deepest gratitude to my parents for their unconditional support, which made this journey possible. Thank you for always encouraging me to dream and for believing in me. Without you, this path would have been far more difficult.

To my sister, thank you for always being there for me, for giving me strength in difficult moments, and for your constant support. And, of course, for the delicious desserts that provided a little extra motivation along the way.

To my family, thank you for your encouragement and for always supporting me throughout this journey.

To my friends, thank you for your understanding when I had to skip plans, and for continuing to motivate and support me regardless.

I would also like to express my sincere gratitude to Banco de Portugal for welcoming me and for providing the opportunity to work in Official Statistics and contribute to their development. I am particularly thankful to the BPLIM team (Ana, Diogo, Gustavo, Joana, Marta, Miguel, Professor Paulo, Rita, and Rute) for making this journey so rewarding. Thank you for your constant support and for making my first professional experience both professionally enriching and personally fulfilling. I am also grateful to my internship colleagues for their motivation, collaboration, and companionship throughout this experience.

Finally, I would like to express my deepest gratitude to my supervisor, Professor Fernando Silva, and my co-supervisor, Professor Marta Silva, for their time, patience, and invaluable guidance. Thank you for helping me overcome my fears and for your constant enthusiasm and positivity throughout this process. I would also like to extend my sincere thanks to Professor Paulo Guimarães for proposing this research topic, for helping me understand it, and for his support during the early stages of this work.

Abstract

This thesis, developed as part of an internship at Banco de Portugal Microdata Research Laboratory (BPLIM), implements a record linkage framework to construct a stable, establishment level identifier for firms reporting to the Portuguese Simplified Business Information System (IES). The work is motivated by a fundamental limitation of the existing data: the reported establishment identifier available in Annex R, the variable institutionally designated for tracking establishment identity, proves insufficiently reliable for longitudinal analysis, with accuracy declining from 84.5% in 2008 to 73.5% in 2016. This study uses detailed establishment level microdata on non-financial corporations, covering the period 2007 to 2023. The sample is restricted to firms operating more than one establishment, for which identification and tracking across years are especially challenging.

Two proposed approaches were considered to construct a primary key for establishment through record linkage pipelines. The *empirical pipeline* approach operates in three sequential steps that progressively resolve matches and non-matches from high-confidence to ambiguous cases, combining deterministic and fuzzy string similarity measures over spatial and categorical features, a hybrid TF-IDF and fuzzy-matching approach for address comparison with globally optimal assignment via the Hungarian algorithm, and Bayesian threshold optimization to calibrate five threshold hyperparameters. The *semantic pipeline* replaces the second and third steps of the empirical pipeline with a fine-tuned BERT Cross-Encoder, trained on hard-negative pairs generated by a Bi-Encoder ensemble of three Portuguese-language Transformer models, which classifies ambiguous partial matches as either a continuation of an existing establishment or a new one.

Both pipelines were evaluated against a semi-automated annotated ground truth spanning 2007 to 2016 and comprising 462,370 establishment observations across multi-establishment firms. Then, applied to unsupervised data from 2017 to 2023. The empirical pipeline achieved accuracy ranging from 97.6% in 2008 to 93.1% in 2016, with a precision of 92.4%, recall of 93.7%, and F1-score of 93.0%. The semantic pipeline achieved comparable but slightly lower accuracy, from 96.9% in 2008 to 91.9% in 2016, and also produced duplicate identifiers within the same year and firm.

By producing a harmonized establishment identifier, this thesis enables the construction of establishment level longitudinal data within the IES microdata. This allows one to reliably track the same establishment over time, and study establishment dynamics, such as entry, exit, and relocation.

Keywords: Annex R, Bayesian optimization, BERT, Data linkage, Deep learning, Deterministic matching, Establishment dynamics, Establishment level identifier, Fuzzy matching, Hungarian algorithm, Linkage algorithm, Microdata, Multi-establishments, Natural Language Processing, Official statistics, Panel datasets, Probabilistic record linkage, Record linkage, Simplified Business Information

Resumo

Esta dissertação, desenvolvida no âmbito de um estágio no Laboratório de Investigação com Microdados do Banco de Portugal (BPLIM), implementa um sistema robusto de ligação de registos para construir um identificador estável ao nível do estabelecimento para as empresas que reportam ao sistema de Informação Empresarial Simplificada (IES). O trabalho é motivado por uma limitação fundamental dos dados existentes: o identificador de estabelecimento reportado, disponível no Anexo R, a variável institucionalmente destinada a rastrear a identidade dos estabelecimentos, revela-se insuficientemente fiável para análise longitudinal, com a sua acurácia a diminuir de 84,5% em 2008 para 73,5% em 2016. Este estudo utiliza microdados detalhados ao nível do estabelecimento de sociedades não financeiras, cobrindo o período de 2007 a 2023. A amostra é restrita a empresas que operam mais do que um estabelecimento, para as quais a identificação e o rastreio ao longo dos anos é mais desafiante.

Duas abordagens foram consideradas para construir uma chave primária para o estabelecimento através de pipelines de ligação de registos. A abordagem do *pipeline empírico* opera em três etapas sequenciais que resolvem progressivamente correspondências e não correspondências, de casos de elevada confiança a casos ambíguos, combinando medidas de similaridade determinísticas e difusas (*fuzzy*) sobre características espaciais e categóricas, uma abordagem híbrida de TF-IDF e *fuzzy-matching* para comparação de endereços com atribuição globalmente ótima através do algoritmo Hungarian, e otimização Bayesiana de limiares para calibrar cinco hiperparâmetros. O *pipeline semântico* substitui a segunda e a terceira etapas do pipeline empírico por um BERT Cross-Encoder *fine-tuned*, treinado em pares de exemplos negativos difíceis (*hard-negative*) gerados por um conjunto *ensemble* de Bi-Encoders composto por três modelos Transformer em língua portuguesa, que classifica as correspondências parciais ambíguas como continuação de um estabelecimento existente ou como um novo estabelecimento.

Ambos os pipelines foram avaliados face a *ground truth* anotada semi-automática, abrangendo o período de 2007 a 2016 e compreendendo 462.370 observações de estabelecimentos em empresas multi-estabelecimento. Posteriormente, foram aplicados a dados não supervisionados de 2017 a 2023. O pipeline empírico obteve uma acurácia entre 97,6% em 2008 e 93,1% em 2016, com uma precisão de 92,4%, sensibilidade de 93,7% e F1-score de 93,0%. O pipeline semântico obteve uma acurácia comparável mas ligeiramente inferior, entre 96,9% em 2008 e 91,9% em 2016, tendo também produzido identificadores duplicados dentro do mesmo ano e empresa.

Ao produzir um identificador de estabelecimento harmonizado, esta dissertação permite a construção de dados longitudinais ao nível do estabelecimento no âmbito dos microdados (IES). Permitindo acompanhar o mesmo estabelecimento ao longo do tempo e estudar a dinâmica dos estabelecimentos, nomeadamente entradas, saídas, relocalização.

Palavras-chave: Anexo R, Otimização Bayesiana, BERT, Ligação de dados, Aprendizagem profunda, Correspondência determinística, Dinâmica dos estabelecimentos, Identificador ao nível do estabelecimento, Correspondência aproximada, Algoritmo Hungarian, Algoritmo de

ligação, Microdados, Multi-estabelecimentos, Processamento de Linguagem Natural, Estatísticas oficiais, Dados em painel, Ligação probabilística de registos, Ligação de registos, Informação Empresarial Simplificada

Contents

Acknowledgements ii

Abstract iii

Resumo v

1 Introduction **1**

1.1 Motivation 2

1.2 Problem Context 3

1.2.1 Simplified Business Information 3

1.2.2 Problem definition 6

1.3 Main Goals and Contributions 7

1.4 Dissertation Structure 8

2 Literature Review **9**

2.1 Record linkage problem definition 9

2.2 Traditional approach of record linkage 10

2.3 Evolution of the traditional record linkage approach 11

2.3.1 Data preprocessing 11

2.3.2 Blocking/ Indexing 13

2.3.2.1 Phonetic Encoding methods 14

2.3.2.2 Clustering-based methods 15

2.3.2.3 Machine Learning methods 17

2.3.3 Record Pair Comparison 19

2.3.3.1 String comparison metrics 19

2.3.3.2 Machine Learning models for pairing 21

2.3.4 Evaluation metrics . . . 23
2.3.5 Similar Works in the context of Official Statistics . . . 24

3 Data Description and Processing **25**

3.1 Dataset Description . . . 25
3.2 Preprocessing and Data Cleaning . . . 28

4 Model Implementation **32**

4.1 String Matching Pipeline Model . . . 32
4.2 Record Linkage Empirical Pipeline Implementation . . . 34
4.2.1 Blocking . . . 34
4.2.2 Identification of Possible Duplicated Records . . . 35
4.2.3 Record Pair Comparison . . . 38
4.2.3.1 Step 1: Matches Identification . . . 38
4.2.3.2 Step 2: Partial Matches from Step 1 . . . 42
4.2.3.3 Step 3: Partial Matches from Step 2 . . . 44
4.3 Record Linkage using Semantic Analysis Implementation . . . 44
4.3.1 Solving Partial Matches using Deep Learning . . . 45
4.3.1.1 Preprocessing . . . 45
4.3.1.2 Cross-Encoder Fine-Tuning . . . 47
4.3.1.3 Threshold Optimal Selection . . . 48
4.3.1.4 Establishment ID Assignment . . . 50

5 Results **51**

5.1 Confusion Matrix Interpretation . . . 51
5.2 Heuristic initialization approach . . . 52
5.3 Impact of Similarity Measures Choices on the Matching Performance . . . 54
5.4 Bayesian Threshold Optimization . . . 55
5.4.1 Experiment Regarding the Year 2008 . . . 56
5.4.2 Experiment Regarding a Combination of Years . . . 57

5.5 Record Linkage Empirical and Semantic Pipeline Identifier vs Reported Establishment Identifier . . . 59

5.6 Record Linkage Empirical and Semantic Pipeline Compared with Ground Truth . . . 61

5.7 Record Linkage Empirical and Semantic Pipeline - Unsupervised Dataset . . 68

6 Conclusion **72**

6.1 Limitations and Future Work . . . 73

References **75**

A Appendix **i**

A.1 Annex R . . . i

A.2 Portuguese multi-establishment distribution by level. . . . ix

A.3 Number of partial matches remain and accuracy obtained, using the average of the combine years (2008, 2010, 2012, 2013,2014 and 2016) . . . x

A.4 Metrics resulting from the threshold using the average of the combine years (2008, 2010, 2012, 2013,2014 and 2016) . . . x

A.5 Number of partial matches remain and accuracy obtained, using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016) . . . xi

A.6 Confusion Matrix on firm level, resulting from the threshold using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016) . . . xi

A.7 Metrics resulting from the threshold using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016) . . . xi

A.8 Duplicated Establishments per year. . . . xii

A.9 Dropped establishment, multi-establishment and establishments that in some year become multi-establishments . . . xii

Glossary

AFSM Adaptive Fuzzy String Matching. 21

BdP Banco de Portugal. 4

BPLIM Banco de Portugal Microdata Research Laboratory. 2, 7, 25, 27, 30, 51

CaCl Canopy Clustering Method. 15

CRFs Conditional Random Fields. 12

EU European Union. 5

HMM Hidden Markov Model. 12, 13

HNSW Hierarchical Navigable Small-World. 17

HVTB High-Value Token-Blocking. 16

IES Simplified Business Information. iii, 2–4, 6, 72

INE National Statistics Institute. 4

KNN K-nearest neighbors. 18

LLM Large Language Model. 24

LSH Locality Sensitive Hashing. 16

MSE Mean Squared Error. 37

NLP Natural Language Processing. 44

POC Official Accounting Plan. 5

pp Percentage Points. 59, 60

RNNs recurrent neural networks. 22

SFA Stochastic Finite Automata. 13

SHAP Shapley Additive Explanations. 17

SNC Accounting Standardization System. 5

SoNe Sorted Neighbourhood. 15

SRG Stochastic Regular Grammar. 12

TF-IDF Term Frequency-Inverse Document Frequency. 15, 72

TPE Tree-structured Parzen Estimator. 55

List of Figures

1.1 Structure of the Annex R. 5

2.1 General structure of a record linkage process. 10

3.1 Missing Data Matrix, from 2007 to 2023. 30

4.1 Record linkage pipeline using probabilistic matching. 35

4.2 Precision, recall and F1 vs. decision threshold (BERT). 49

5.1 (A) New & closed establishments, (B) Net change, from 2007 to 2016 (ground truth (GT) vs Empirical Pipeline vs Semantic Pipeline). 63

5.2 Survival rate difference (ground truth − pipeline) by cohort × years since birth. 64

5.3 Comparison between (A) First entry (B) Re-entry (C) Final Exit (D) Re-exit establishments yearly. 65

5.4 Gap size distribution (%), comparison between the pipeline and the ground truth.) 67

5.5 New & closed establishments and its net change, from 2007 to 2023. 70

5.6 Survival rate difference by cohort x years since birth across the pipelines. . . 71

List of Tables

2.1 Soundex, phonetic codification 14

3.1 Qualitative variables available. 26

3.2 Number of missing values present in each dataset. 29

3.3 Standardization examples of the street address field. 31

4.1 Pairs of Positive and Highly Negative Matches generated by the Bi-Encoder (illustrative). 47

4.2 Parameters used in the fine-tune of the cross-encoder. 48

5.1 Number of new establishments, partial matches remaining and accuracy obtained for the *ad hoc* experiment. 53

5.2 Confusion matrix and evaluation metrics on *ad hoc* experiment. 53

5.3 Number of partial matches and accuracy across experiments, for years 2008 to 2010. 54

5.4 Threshold resulting from the hyperparameterization, regarding the year 2008. 56

5.5 Number of partial matches and accuracy obtained, with thresholds optimized for 2008. 56

5.6 Confusion matrix and evaluation metrics, for threshold optimized for 2008. . 57

5.7 Threshold resulting from the hyperparameterization, regarding the years of 2008, 2010, 2012, 2013, 2014, and 2016. 57

5.8 Threshold resulting from the combination of years (2008, 2010, 2012, 2013, 2014 and 2016). 58

5.9 Performance comparison between the accuracy of the reported establishment identifier vs the developed algorithm, based on the ground truth for multi establishments. 60

5.10 Confusion matrix and evaluation metrics for the reported identifier and algorithm based identifiers. 61

5.11 Number of new establishment and partial match remain obtained for the final pipelines. 62

5.12 Comparison of attribution between the pipelines and the ground truth at the establishment level. 67

5.13 Number of new establishment, partial match remain. 69

A.1 Evaluation metrics. x

A.2 Evaluation metrics. xi

A.3 Number of duplicated establishments dropped per year. xii

A.4 Number of dropped establishments, multi-establishments, and establishments that in some year become multi-establishments, Empirical vs. Semantic Pipeline. xii

Chapter 1

Introduction

In recent years, the exponential growth in data collection and storage has created unprecedented opportunities for analysis, monitoring, and evidence-based decision-making across a wide range of sectors. Administrative records, surveys, and transactional databases now provide highly granular information on economic agents and their activities. However, despite the increasing availability of data, its effective use remains constrained by fundamental challenges in data pre-processing and, most critically, in record linkage. In many contexts, the absence of unique, stable identifiers prevents the reliable identification of records referring to the same real-world entity across different datasets or over time. As a result, data becomes fragmented across systems, leading to duplication, inconsistencies, and incomplete representations of entities.

High-quality record linkage is also a foundation of modern official statistics. Statistical authorities increasingly rely on integrated microdata systems to produce accurate, detailed, and timely indicators. By linking data from administrative, statistical, and register-based sources, it becomes possible to improve data completeness, reduce duplication, and enhance internal and external consistency across statistical domains. Such integration supports the measurement of dynamic phenomena, including business demography, productivity, employment flows, and structural change, while maximizing the value of existing data and minimizing the reporting burden on economic agents.

Micro-databases, particularly those based on transaction-level or establishment level reporting, are inherently rich and comprehensive within their respective domains. However, their full analytical potential can only be realized when they are viewed not as isolated repositories but as components of an integrated, high-granularity data system. The use of common identifiers for economic agents enables the combination and cross-referencing of information across sources, strengthening data quality management and ensuring closer alignment with statistical concepts, especially in the case of administrative data originally collected for non-statistical purposes.

The ability to track establishments consistently over time is essential for economic analysis that seeks to exploit the advantages of panel data. Longitudinal identification makes it possible to distinguish persistent establishment-specific characteristics from genuine economic change, thereby enabling the control for unobserved time-invariant heterogeneity and the

analysis of within-establishment dynamics. Without robust record linkage mechanisms, empirical work is confined to isolated cross-sectional snapshots that obscure adjustment processes, persistence, and causal relationships. This constraint limits the use of panel-data methods, weakens inference on growth, productivity, and structural change, and undermines the evaluation of economic performance and policy effects. Reliable linkage of records referring to the same establishment across time and data sources is therefore a fundamental prerequisite for constructing panel datasets that support rigorous longitudinal and causal economic analysis.

In this context, the Simplified Business Information (IES), introduced in 2007, provides comprehensive administrative reporting through a single electronic submission that simultaneously fulfils accounting, fiscal, statistical, and legal reporting obligations. This reporting consolidates information submitted to multiple public institutions—including the Ministry of Finance, the Ministry of Justice, Statistics Portugal, and the Bank of Portugal. The IES reduces duplication, minimizes reporting burden, and enhances data consistency.

The Banco de Portugal plays a central role in ensuring the production of high-quality monetary, financial, foreign exchange, and balance of payments statistics. At the same time, the dissemination of statistics derived from micro-level information requires strict adherence to confidentiality and data protection rules. Safeguarding the confidentiality of reporters is essential to maintaining trust in the statistical system and ensuring the continued availability and quality of elementary data. Consequently, effective record linkage must be accompanied by robust governance frameworks that balance analytical potential with the fundamental principles of statistical confidentiality (Agostinho & Miguel, 2017). It is within this context that linked microdata are made available through the Banco de Portugal Microdata Research Laboratory (BPLIM), which provides secure access to confidential datasets under controlled conditions, thereby enabling advanced economic research while fully preserving statistical confidentiality.

1.1 Motivation

Annex R provides detailed establishment level information for non-financial corporations. The reported establishment identifier is assigned by the firm without a validation mechanism that ensures the consistent tracking of the same establishment over time. This is precisely the kind of fragmentation discussed above: without a stable identifier, establishments cannot be followed over time, which undermines the ability to study establishment dynamics such as entry, exit, relocation, and structural change.

This work addresses that gap by constructing a unique, stable identifier for each establishment of a firm, thereby enabling the construction of an establishment level panel dataset. This identifier uniquely identifies a specific establishment by combining the entity's tax identification number (NIF/NIPC) with an establishment specific sequential number assigned by the firm. The empirical analysis is based on the data reported in Annex R of IES, and is restricted to firms operating more than one establishment, since these are the cases for which identification and tracking across years are most challenging and where the limitations of the reported establishment identifier are most consequential.

The proposed approach consists of developing two algorithms, a rule-based and a learning-based approach, and comparing their results to perform record linkage at the establishment level, ensuring that the same establishment of a firm is consistently assigned the same identifier across different reporting periods. Comparing these two paradigms makes it possible to assess whether a learned semantic similarity model can match or outperform a carefully tuned deterministic approach and can be used to assign unique identifiers in unsupervised periods. At the same time, each algorithm is designed to detect the creation of new establishments and assign them a new sequential number, preserving the stability of previously assigned identifiers over time. This approach allows each establishment to be consistently distinguished within the same legal entity. By introducing a stable and harmonized establishment identifier, this work aims to support accurate record linkage, facilitate the construction of panel datasets, and improve the integration of microdata for statistical and analytical purposes at the level of the establishment.

1.2 Problem Context

1.2.1 Simplified Business Information

In the context of this work, it is important to distinguish the concepts of firm and establishment. A firm is a legal entity (either an individual or a collective person) that represents an organizational unit engaged in the production of goods and/or services, possessing a certain degree of decision-making autonomy, especially regarding the allocation of its current resources (Instituto Nacional de Estatística, 1994a). An establishment is a firm or part of a firm (factory, workshop, mine, warehouse, shop, depot, etc.) located at a topographically identified place, where, or from which, economic activities are carried out and for which, as a rule, one or several persons work (possibly part-time) on behalf of the same firm (Instituto Nacional de Estatística, 1994b). The firm's headquarters should be considered an establishment. A firm may therefore operate several establishments, each of which performs part or all of the entity's economic activities (Ordem dos Contabilistas Certificados, 2025).

The IES was created by the Decree-Law no. 8/2007, from January 17, which consists of a mandatory annual declaration of an accounting, tax, and statistical nature that Portuguese entities must submit in a fully dematerialised and electronic manner. The information reported in the IES refers to the previous fiscal year (Autoridade Tributária e Aduaneira, Instituto dos Registos e do Notariado, Instituto Nacional de Estatística, & Banco de Portugal, 2015).

The IES has been in production since 2007 and contains detailed economic and financial information at the firm level for the reference year 2006 and subsequent years, as well as more limited information at the establishment level. With the introduction of IES, firms comply with the following three legal obligations through a single act (Ordem dos Contabilistas Certificados, 2025):

1. Tax obligations: submission of the annual accounting and tax information statement to the Ministry of Finance.
2. Commercial obligation: registration of accounts with the commercial registry office.

3. Statistical obligations:

    i. Provision of statistical information to the National Statistics Institute (INE).

    ii. Provision of information relating to annual accounting data for statistical purposes to the Banco de Portugal (BdP).

    iii. Provision of statistical information to the Directorate-General for Economic Activities.

The IES consists of a cover page and 21 Annexes. The obligation to submit the IES applies to taxpayers who are required to file at least one specific annex. These taxpayers are divided into two groups: corporate income tax (IRC) taxpayers and personal income tax (IRS) taxpayers (Ordem dos Contabilistas Certificados, 2025).

The IES microdata provides standardized data on Portuguese firms. This thesis will focus on non-financial corporations which are required to submit Annex A with firm level data and Annex R with establishment level data.

Annex A

Decree-Law no. 8/2007 of 17 January establishes that Annex A must be submitted together with Annex R by resident firms whose main activity is commercial, industrial, or agricultural, as well as by non-resident firms with a permanent establishment in Portuguese territory (Ordem dos Contabilistas Certificados, 2025).

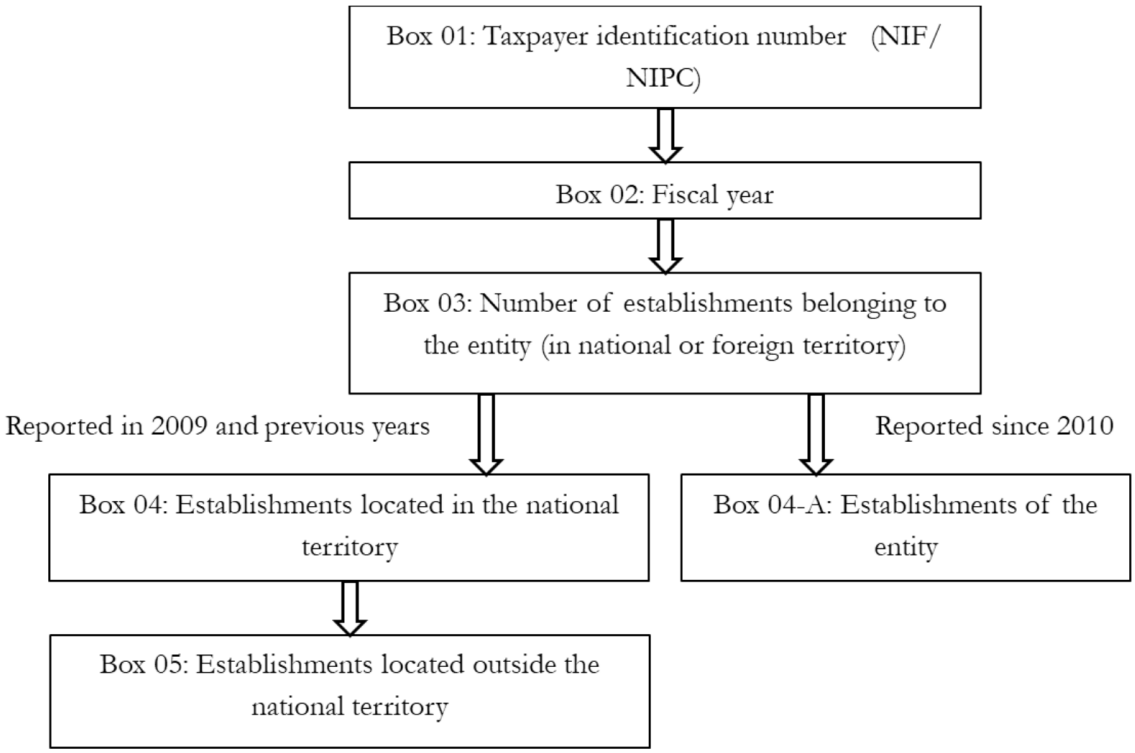
Annex A reports information about the unconsolidated annual accounts of the firm. The purpose is to provide information for compliance with two of the three legal obligations arising from the submission of the IES. Accordingly, Annex A is primarily associated with tax obligations and commercial obligations.

Annex R

Annex R (Appendix A.1) contains qualitative and quantitative information on each establishment associated with a firm and serves to fulfil the corresponding legal obligations for statistical reporting. The financial values reported at the firm level in Annex A correspond to the sum of the values for each establishment as reported in Annex R.

Annex R has undergone changes over time, and Figure 1.1 illustrates those changes. Up to 2009, firms are required to complete Box 04 for each establishment located in Portuguese territory (see Appendix A.1). If the firm has one or more establishments located abroad, firms are required to report the aggregate information in Box 05 (Autoridade Tributária e Aduaneira, 2025). Since 2010, firms report detailed information for each establishment, irrespective of its location, in Box 04-A (Autoridade Tributária e Aduaneira, 2025). In boxes 04 and 04-A, entities must report qualitative information for each establishment, including address, contact details, status, whether it is the headquarters, and the main economic activity.

Figure 1.1: Structure of the Annex R.



Each establishment operated by the firm must be assigned a unique order number. The headquarters must always be assigned order number 1, and the remaining establishments must be numbered sequentially (2, 3, 4, …) up to the total number of establishments. According to the instructions in Annex R, the establishment order number must remain consistent across future IES filings as long as the establishment continues its activity. If an establishment ceases its activity, its order number must not be reused. When a new establishment is opened, it must be assigned the next consecutive order number, immediately following the highest number previously assigned.

For each establishment, quantitative information is also required, such as the number of employees, turnover and other financial variables. For single establishment firms, the quantitative information is not reported as it is equivalent to that reported in Annex A.

Changes in the Portuguese accounting system that changed the structure of Annex R

Annex R was reformulated in 2010 to reflect changes in the Portuguese accounting system. Until 2009, entities reported financial information in accordance with the Official Accounting Plan (POC). From 2010 onwards, they have been required to follow the Accounting Standardization System (SNC). This reform was necessary to ensure that Portugal's accounting framework became comparable with European Union (EU) accounting directives and regulations. The introduction of the SNC enabled Portugal to align with the modernization of

accounting practices across the EU (Assembleia da República, 2009).

The modifications to Annex R involved the discontinuation of two boxes and the creation of a new one. Specifically, in 2010, Box 04-A was introduced, combining the information previously reported in Boxes 04 and 05, which were discontinued.

Box 04-A requires more detailed and disaggregated information. The values related to sales and services, which had previously been reported as aggregated amounts, must now be reported in separate fields, as well as the values related to the *cost of goods sold and the materials consumed and the provision of external services*. In addition, seven new variables became mandatory: *purchases; acquisition of biological assets; acquisition of tangible fixed assets; acquisition of tangible fixed assets in buildings and other constructions; acquisitions of investment properties; acquisitions of investment properties in buildings and other constructions; equity or equivalent capital.*

With the discontinuation of Box 05, information relating to foreign establishments must now be reported separately for each establishment in Box 04-A. To support this change, Box 04-A introduced specific fields identifying the country in which each establishment is located.

1.2.2 Problem definition

The submission of the Annexes of the IES is important not only for verifying compliance with the three legal obligations, but also because it represents one of the main economic and financial data sources at the firm and establishment level. For this reason, it is crucial that the information reported is reliable, validated, and free from systematic bias, so that it accurately reflects economic reality.

As discussed above, Annex A provides information at the firm level. Consequently, it does not allow for an analysis of individual establishments, nor does it provide information on their geographic distribution. This aggregation substantially limits its usefulness for studies focused on regional economic performance or the spatial distribution of economic activity.

In contrast, Annex R reports information separately for each establishment, making it possible to more accurately analyze regional economic performance and business activity. This annex is therefore particularly valuable for research aimed at understanding sectoral and regional dynamics within the Portuguese business landscape.

However, the information reported in Annex R is not subject to the same validation procedures or systematic quality controls as Annex A. As a result, the data may contain errors and inconsistencies. For example, as discussed previously, the establishment order number within a firm should be unique and consistent over time. After examination, this is not always the case.

There are cases in which a firm with two establishments assigned with order number 2 and 3 in one year, are swapped in the following year: the establishment previously identified as number 2 is reported as number 3, and vice versa. This generates substantial inconsistencies in the establishment longitudinal tracking.

These limitations constrain the use of Annex R for empirical research and underscore the need for data validation procedures that ensure consistency and accuracy over time. In response,

the main objective of this work is to construct a primary key that provides a unique and stable identifier for each establishment within a firm. This identifier enables reliable record linkage across years, thereby supporting the construction of establishment-level panel data and the application of longitudinal and panel-data methods. By ensuring temporal continuity at the establishment level, the information derived from Annex R becomes statistically consistent and more reliable. As a result, the detailed data reported in Annex R can be fully exploited for robust longitudinal analysis, yielding more credible identification of establishment and firm dynamics.

1.3 Main Goals and Contributions

This thesis was developed as part of an internship at BPLIM, with the main goal of constructing an algorithm to generate stable and consistent establishment identifiers for the period between 2007 and 2023. This identifier is defined as a primary key that combines the firm's tax identification number (NIPC) with an establishment specific sequential number, assigned through an automated algorithm designed to ensure temporal consistency while detecting the creation of new establishments. To achieve that purpose, two pipelines were developed and evaluated against a manually annotated ground truth covering the years 2007 to 2016, and subsequently applied to unsupervised data from 2017 to 2023.

This work makes five main contributions.

- The first pipeline proposes a formal three-step record linkage framework for establishment level data, combining deterministic matching rules with probabilistic and fuzzy similarity measures based on spatial, categorical, and financial features, formulated as a constrained optimization problem. In addition, the second pipeline explores the potential of deep learning approaches as an alternative representation for establishment matching. These models are used to help resolve ambiguous cases that fall within the uncertainty region identified in the first step, as defined by the classification thresholds.
- Second, from an operational standpoint, the proposed pipeline automates a process that would otherwise require manual annotation on an annual basis, reducing the burden and minimizing the risk of human error.
- Third, by improving the consistency and accuracy of Annex R data, this work contributes to enhancing the quality of official statistics and economic analysis.
- Fourth, it produces a harmonized establishment level panel dataset, accessible through BPLIM under strict confidentiality conditions, enabling the application of panel data methods and supporting longitudinal analyses of establishment dynamics in Portugal, including entry, exit, relocation, and changes in economic activity.
- Fifth, the pipeline is designed to be reusable beyond the current reference period: as new IES reporting period become available, the algorithm can be applied incrementally, allowing the dataset to be extended without requiring a full re-run. This forward compatibility ensures that the dataset remains up to date and that the linkage infrastructure serves as a long-term resource.

1.4 Dissertation Structure

The remainder of this dissertation is organized as follows.

Chapter 2 reviews the existing literature on record linkage and entity matching, with particular emphasis on methods applied to administrative and statistical microdata, and discusses the contributions of classical, probabilistic, and deep learning approaches to the field.

Chapter 3 describes the data used in this study, detailing the structure of Annex R of the IES, the coverage period, and the pre-processing steps applied to prepare the data for the linkage pipeline.

Chapter 4 presents the proposed record linkage framework, describing the three-stage pipeline: the deterministic and fuzzy similarity matching stage, the constrained optimization model for identifier assignment, and the deep learning stage based on BERT that resolves ambiguous partial matches.

Chapter 5 reports the results of the pipelines, evaluating their performance and discussing the quality of the establishment level identifiers produced.

Chapter 6 concludes the dissertation by summarizing the main findings, discussing the limitations of the proposed approach, and outlining directions for future work.

Chapter 2

Literature Review

This chapter introduces the main concepts and methods related to record linkage problems. It also summarizes related work concerning similar problems. Section 2.1 defines and clarifies the concept of record linkage. Section 2.2 presents the methodology of traditional record linkage. Finally, Section 2.3 describes the methods used to implement each stage of the traditional record linkage process, covering a range of algorithms and alternative approaches.

2.1 Record linkage problem definition

Records in data sources represent observations of entities from a given population and contain attributes (fields or variables) that help identify each entity, such as name, address, age, or gender (Gu, Baxter, Vickers, & Rainsford, 2003). A data source A with n_A records and a data source B with n_B records can each contain multiple attributes per record, where a record is an ordered sequence of attribute values describing an entity.

Record linkage is the task of comparing records from two (or more) data sources to identify which records refer to the same underlying entity. In the general case, every record in A is a potential match for every record in B, leading to $n_A \times n_B$ candidate record pairs whose match or non-match status must be determined (Gomatam, Carter, Ariet, & Mitchell, 2002). These pairs are conceptually partitioned into two disjoint sets: M, the set of true matches, and U, the set of true non-matches, both subsets of the cartesian product A × B (Tokle & Bender, 2021).

The goal of record linkage is to recover the true binary relations between A and B, that is, to identify the subset $M \subseteq A \times$ B that links records from A and B referring to the same entity. In many practical applications, records in B are assumed to correspond to distinct entities, which implies that a record in A should not match more than one record in B; violations of this constraint highlight linkage ambiguities and errors.

In simple terms, record linkage can be defined as the process of matching records across one or more databases that refer to the same entities when reliable unique identifiers are absent, so that more sophisticated linkage techniques are required (Christen, 2012b). This

concept is also widely known under many related names. Barlaug and Gulla (2021) identify a selection of these similar terms, including approximate string matching, data linkage, data matching, deduplication, duplicate detection, entity matching, entity resolution, fuzzy join, fuzzy matching, merge–purge, object identification, reference reconciliation, re-identification, similarity joins, and string matching, among others. Although some of these terms emphasize specific aspects or subproblems, they all refer to closely related formulations of the record linkage problem.

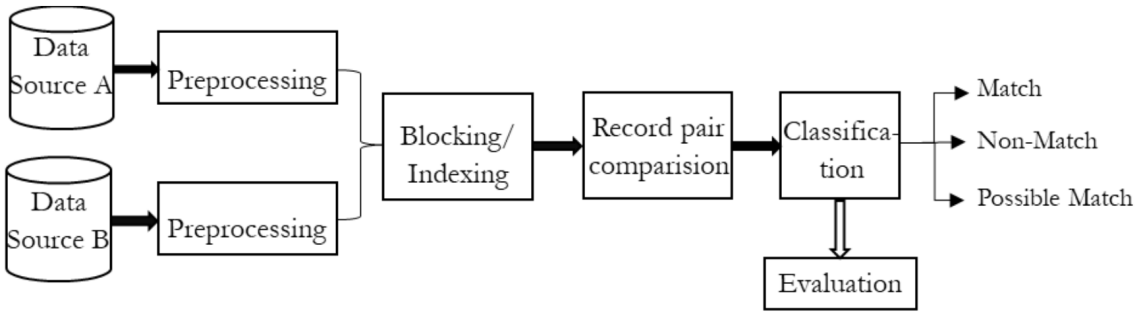
Christen (2012b) and Barlaug and Gulla (2021) note that when the process of record linkage applies exclusively to a single dataset, meaning that A = B, this process is known as deduplication or duplicate detection, because it focuses on identifying multiple records that correspond to the same entity in one source.

There are two types of record linkage problem, the deterministic and probabilistic record linkage. The deterministic approach is based on the common identifiers among the available data sources, and there is a variable or a combination of variables that can represent the majority of the data and has enough discrimination power. On the other hand, the probabilistic record linkage approach does not have a unique identifier available for the matching, but it can compensate for incomplete and erroneous information by using several variables to generate potential pairs, assigning a score to each pair, and then determining the linkage status (match, non-match, or possible match) based on that score (Statistics Canada, 2021).

2.2 Traditional approach of record linkage

Elfeky, Verykios, and Elmagarmid (2002), Christen (2012a), and Barlaug and Gulla (2021) present a general approach to describe the record linkage process. Barlaug and Gulla include one more subdivision that is Schema matching, which the authors of the first two papers consider as part of preprocessing. Figure 2.1 overviews the five main steps that must be done in order to solve a record linkage problem.

Figure 2.1: General structure of a record linkage process.



The preprocessing phase is the initial step where raw records from different data sources are cleaned and transformed into consistent formats (standardization) so they can be reliably compared for matching. Typical transformations may involve removing excess punctuation, lowercasing all letters, normalizing values, and tokenizing (Barlaug & Gulla, 2021). Its main goal is to reduce noise and heterogeneity (different encodings, formats, and errors) so that

later linkage steps, such as comparison and classification, can focus on true similarities and differences between records rather than artefacts of poor data quality.

The blocking/ indexing phase addresses the computational infeasibility of comparing every possible pair of records across two databases. To overcome this quadratic complexity, blocking techniques are commonly used to reduce the number of record comparisons (Franke, Sehili, Gladbach, & Rahm, 2018). The output is a candidate set with high recall that needs to be further compared. For the deduplication, each record in the dataset with n records must be compared a total of n*(n-1)/2 record pair comparisons, because the comparisons of two records are symmetric.

The record pair comparison is the step where each candidate record pair is compared using similarity functions that are applied on the records' attributes. A similarity function generally calculates a value between 0 (dissimilar) and 1 (exact match) that quantifies how similar two attribute values are (Franke, Christen, Christen, Rohde, & Rahm, 2024). The similarity vector contains the result of the similarity function computed for each candidate record pair comparison. This vector serves as a feature for the classification step.

The classification step decides, for each candidate record pair, whether it represents the same real-world entity (match), a different entity (non-match), or sometimes a possible match, based on the similarity vectors. The class of possible matches contains those candidate record pairs where the model was not able to make a clear decision, and these records need more elaborate solutions (Barlaug & Gulla, 2021). Many approaches rely on threshold-based classification (Franke et al., 2024).

Finally, the evaluation step is applicable only when labelled ground truth data are available. It assesses the quality of the linkage by computing performance measures such as precision, recall, and F1-score, allowing the identification of the types and magnitude of linkage errors.

2.3 Evolution of the traditional record linkage approach

2.3.1 Data preprocessing

Preprocessing is a critical step in record linkage, both for classical approaches and machine learning approaches. It ensures that data is clean, standardized, and comparable across dataset(s).

Standardization is a fundamental preprocessing step in record linkage and consists of replacing multiple spelling variants of the same word with a single canonical form. Commonly occurring terms, such as "Street", are often replaced by standardized abbreviations (e.g., "St"), which helps reduce variability caused by spelling differences and typographical errors (Winkler, 2006). The main objective of standardization is to ensure that identical entities expressed in different textual forms are converted into a consistent representation, thereby facilitating more reliable record comparisons.

Following standardization, most record linkage systems apply tokenization, which involves splitting a string into individual tokens, typically using whitespace or punctuation as delimiters. For example, in the address "123 MAIN ST., APT 5B", tokenization produces a list of tokens,

such as [”123”, ”MAIN”, ”ST.”, ”APT”, ”5B”]. This step enables subsequent processing stages to operate on individual components rather than raw strings.

Tokenization alone, however, is insufficient for effective comparison. Therefore, parsing is commonly applied to assign tokens to predefined semantic components of an address or name. Using the previous example, parsing yields the following fields:

- Street number: 123
- Street name: Main
- Street type: ST
- Secondary unit designation: APT
- Secondary unit number: 5B

Some tokens may correspond to stop words or noise components. When present, these are typically ignored or assigned lower weights during matching. By structuring the data into meaningful components, record linkage algorithms can compare corresponding fields rather than unstructured and potentially noisy strings.

Winkler (2006) provides a detailed discussion of name and address standardization and compares two distinct approaches: an earlier rule-based method (Winkler, 1993) and a more recent probabilistic approach based on Hidden Markov Model (HMM).

The rule-based approach described by Winkler (1993) relies on a set of handcrafted rules developed through extensive analysis of U.S. Census data. These rules define how strings should be split, which tokens should be treated as noise, and how abbreviations should be mapped to canonical forms. This method has been shown to perform well in person and business name matching tasks, particularly in structured datasets that follow typical U.S. naming and address conventions. Its key advantages include transparency, interpretability, and strong performance in homogeneous and well-formatted data, which contributed to its successful use in production systems at the U.S. Census Bureau.

In contrast, the probabilistic approach presented by Winkler (2006), drawing on methods proposed by Borkar, Deshmukh, and Sarawagi (2001), Churches, Christie, Lim, and Zelinsky (2002), and Christen, Churches, and Zhu (2002), models text as a sequence of hidden states (e.g., street name, city) that generate observed tokens. The model learns, from training data, the likelihood of specific tokens being emitted by particular states, as well as the transition probabilities between states. Once trained, the model can automatically segment and label address components. This approach demonstrated superior performance in complex and heterogeneous address formats, such as those commonly found in Asian and Indian contexts, where rigid rule-based systems struggle.

However, Winkler's experimental results indicate that, for name standardization tasks involving people and business names in relatively homogeneous formats, the earlier rule-based approach continued to outperform the probabilistic methods at the time of evaluation. This comparison highlights an important trade-off in record linkage preprocessing: rule-based

methods offer robustness and transparency in structured domains, while probabilistic models provide greater flexibility for handling diverse and less standardized data sources.

M. Wang et al. (2016) propose a probabilistic address parser based on a linear-chain Conditional Random Fields (CRFs), to automatically annotate semi-structured address data, and augment it with a learned Stochastic Regular Grammar (SRG) to capture higher-level segment dependencies. The CRF serves as a discriminative sequential classifier, using rich, overlapping token-wise features to label each token in an address, while the SRG operates on "lifted" label sequences, where consecutive identical labels are collapsed into macro labels, for example, the sequence "Road Road Road", "City City Postcode Postcode" will become Road City Postcode, to model segment-level patterns that cannot be easily expressed by token-level features alone, as in the HMM approach.

Semi-structured addresses often store the entire address in one field or distribute components across multiple loosely defined fields; the parser exploits this semi-structure by introducing a "field separator" feature at boundaries between fields, which helps discourage tokens in different structural segments from being assigned to the same semantic category after parsing. The learned SRG plays a role analogous to the transition matrix in an HMM, but it focuses on label transitions at the segment level rather than the token level, making it more sensitive to real address patterns observed in the training data and allowing its scores to be combined with CRF conditional probabilities to reorder and refine output label sequences. Experimental results on a postal dataset show that the implementation of CRF parsers with a field separator and Stochastic Finite Automata (SFA), which learns segment-level dependency in semi-structured datasets, outperforms (97% of accuracy) the use of HMM with a field separator and SFA (95%).

2.3.2 Blocking/ Indexing

Blocking aims to group records in blocks that share some common attributes, termed a blocking key, which are likely to indicate a potential match. Dates of birth, zip codes, or surnames are commonly used as blocking keys to partition records into blocks (Karapiperis, Tjortjis, & Verykios, 2025). However, finding an optimal blocking key definition that leads to high-quality matching results is challenging (Christen, 2012a).

The blocking key definition enables having many possible true matches, while at the same time maintaining the total set of candidate record pairs as small as possible, but a trade-off must be chosen between having a large number of blocks and a lower number of comparisons or a higher number of comparisons but a small number of blocks.

The traditional blocking record linkage approach (Fellegi & Sunter, 1969; Winkler, 2006) splits the dataset into non-overlapping blocks, where only records within the block must be compared, where the blocking key is based on a single attribute or a concatenation of several attributes. However, these approaches can be sensitive to data entry errors and variability in formatting (Beheshti, Gondara, & Zachary, 2025).

Techniques based on phonetic encoding (e.g., Soundex and Phonex), or clustering-based methods (e.g., Sorted Neighborhood, Canopy clustering, High-Value Token-Blocking), and

similarity-based approaches (e.g., Jaro-Winkler, Levenshtein distance), can implement the blocking effectively, improving the robustness against errors by considering approximate similarities rather than exact matches, while preserving accuracy (Steorts, Ventura, Sadinle, & Fienberg, 2014).

More advanced techniques are based on embeddings, which encode entity attributes into dense vector spaces where similar entities are positioned closer together. This enables more flexible and scalable blocking strategies, reducing reliance on manually crafted similarity metrics (Beheshti et al., 2025).

2.3.2.1 Phonetic Encoding methods

The Russell Soundex algorithm is a phonetic name-matching method designed primarily for English names (Lait & Randell, 1996). It converts each name into an alphanumeric code of at most four characters, which can be used to identify potentially equivalent names. Knuth (1998) formalized the algorithm based on the phonetic characteristics of letters.

The algorithm transforms an input string into a canonical representation by retaining the first letter and replacing the remaining letters with numeric phonetic codes according to predefined rules (see Table 2.1). Adjacent repetitions of the same code are reduced to a single occurrence, and all occurrences of code 0 are removed. The final Soundex code consists of the first letter followed by three digits; if fewer than three digits remain, trailing zeros are appended, and if more than three digits are produced, the excess digits are discarded from the right.

Although Soundex is more robust than simple character-based similarity measures, it has well-documented limitations (Lait & Randell, 1996). In particular, it performs poorly when name variants do not sound alike or begin with different initial letters. For example, Catarina and Citrino are both reduced to the code C365, despite being phonetically and semantically distinct. As a result, the algorithm can map dissimilar-sounding strings to the same code while also assigning different codes to genuinely similar-sounding names.

Furthermore, Soundex does not produce graded similarity scores or rankings; two strings either match or do not match. Despite these limitations, the method has demonstrated relatively high reliability in large-scale applications.

Table 2.1: Soundex, phonetic codification

Numeric code	Letters
0	a, e, i, o, u, y, h, w
1	b, p, f, v
2	c, g, j, k, q, s, x, z
3	d, t
4	l
5	m, n
6	r

Source: Letters for English Celko (2005) and the Portuguese language Marcelino (2015).

Lait and Randell (1996) created a variant of the algorithm Soundex, the Phonex algorithm.

Consisting of several modifications to the preprocessing step, where name strings are modified according to their English pronunciation before Soundex, like encodings are generated. The modifications are the following:

- All 's' characters at the end are removed.
- A 'kn', 'ph', 'wr' character sequence at the beginning is replaced with a single 'n', 'f', 'r' character, respectively.
- An 'h' character is always removed.
- If the character is a vowel (including 'y'), then it is replaced with an 'a'.
- If the character is a 'p', 'v', 'j', 'z', then it is replaced by a 'b', 'f', 'g', 's', respectively.
- If the character is a 'k' or a 'q', then it is replaced by a 'c'.

After the pre-processing step, the processed name string is encoded using the encoding created for Soundex, and it returns a code made of the initial character followed by three digits.

The percentage of true matches determined by the Phonex method is approximately 44% higher than that of the Soundex method. Nevertheless, it would likely be less effective than the Soundex algorithm when applied to data in other languages (Lait & Randell, 1996).

Other algorithms, based on the previous two, are Fuzzy Soundex (Holmes & McCabe, 2002) and Editex (Zobel & Dart, 1996).

2.3.2.2 Clustering-based methods

The Sorted Neighbourhood (SoNe) method is an unsupervised approach for structured datasets that reduces the number of record comparisons by sorting records on a domain-specific sorting key and only comparing records that fall within a moving window over this sorted order (Hernández & Stolfo, 1998). The sorting key, typically more fine-grained than a blocking key, is constructed from selected attributes (e.g., parts of names and addresses), and its design introduces a domain-specific bias because it determines which records become neighbours in the sorted list.

After computing and sorting these sorting key values, a fixed-size window of w>1 records slides one position at a time over the sorted list; at each position, all records within the window form candidate pairs, so that candidate generation depends mainly on the window size w and the ordering induced by the sorting key. When implemented as an in-memory process, key creation has a complexity of O(N), sorting has O(NlogN), and candidate generation within the window has O(wN), where N is the number of records (Hernández & Stolfo, 1998). A major drawback of the original SoNe is that a fixed window size can cause missed true matches when similar records are not close enough in the sorted key order to appear together within any window, which has motivated later extensions with adaptive window sizes (Christen, 2012a).

The Canopy Clustering Method (CaCl) method is also an unsupervised approach for schema-agnostic/structured datasets that forms possible overlapping clusters, called canopies, of similar records using a computationally inexpensive similarity measure, and then only generates candidate pairs within these clusters rather than across the whole dataset (Babu & Santoshi, 2014). The method consists of representing records using blocking key values, tokenizing these representations (e.g., into words or q-grams), and constructing an inverted index mapping tokens to record identifiers. Using this index, simple similarities such as Jaccard or Term Frequency-Inverse Document Frequency (TF-IDF) or Cosine similarity are computed between a randomly selected centroid record and other records sharing at least one token.

Two thresholds are applied: a loose threshold to determine canopy membership and a tighter threshold to decide which records are removed from further consideration. Records below the tight threshold remain eligible for subsequent canopies, allowing overlapping clusters. Each canopy acts as a block from which candidate record pairs are generated. While overlapping canopies may propose the same pair multiple times, each unique pair is ultimately compared only once. The performance is heavily dependent on the parameter choices (similarity metric and threshold values), which crucially determine canopy sizes and thus the trade-off between reducing comparisons and recall. Experiments on the Addresses dataset show that this clustering algorithm has a recall between 90% to 95%; however, CaCl fails to capture many coreferential pairs compared to Disjunctive Normal Form Blocking proposed by Bilenko, Kamath, and Richardson (2006).

The High-Value Token-Blocking (HVTB) proposed by O'Hare, Jurek-Loughrey, and De Campos (2021) is an unsupervised and schema-agnostic method, inspired by the TF-IDF, which identifies significant ("high-value") tokens to group records that likely refer to the same real-world entity. The method is designed to handle heterogeneous data sources without requiring labeled data or fine-tuning parameters, enhancing its practical applicability.

HVTB operates to ensure efficiency and accuracy, starting by filtering and standardizing tokens, updating inverted indices, and computing TF-IDF values for each token. For each record, an average TF-IDF threshold is derived from its non-unique tokens, and the record is assigned to blocks only when token values exceed this threshold. Record pairs are indexed simply by counting how many high-value token blocks they share; pairs with below-average shared blocks are pruned, and a "least common block" criterion ensures that each pair is only inspected once, avoiding redundant comparisons.

The authors report that HVTB achieves optimal asymptotic complexity linear in both the dataset size and the number of generated pairs, while using substantially less memory than methods such as SoNe and CaCl, largely due to early token reduction and the avoidance of heavy graph structures. Regarding effectiveness, particularly on unstructured data (e.g., titles-and-abstracts datasets), HVTB significantly outperforms SoNe and CaCl when evaluated using the harmonic mean of reduction ratio and pair completeness. A non-parametric Wilcoxon signed-rank test across 11 datasets shows that HVTB is statistically superior to SoNe (p=0.005) and CaCl (p=0.001), with consistently positive median differences.

2.3.2.3 Machine Learning methods

While most neural network research in entity matching focuses on classification, Barlaug and Gulla (2021) identify two deep learning methods designed specifically for blocking: DeepER and AutoBlock. Both approaches transform records into dense vector embeddings such that similar records are located close to one another in a shared metric space.

To efficiently identify candidate matches, both methods employ Locality Sensitive Hashing (LSH) to retrieve similar records in subquadratic time. DeepER applies hyperplane LSH (Indyk & Motwani, 1998) over learned embeddings, whereas AutoBlock uses cross-polytope LSH (Andoni, Indyk, Laarhoven, Razenshteyn, & Schmidt, 2015), a state-of-the-art LSH family for cosine similarity that offers theoretically optimal query-time guarantees and efficient implementations. In both methods, cosine distance is commonly used to measure proximity between record representations.

The two approaches differ primarily in their training strategies. AutoBlock is trained explicitly for the blocking task using a custom loss function applied directly to cosine distances, maximizing separation between matched and unmatched record pairs. In contrast, DeepER learns record representations indirectly by training the network end-to-end on the downstream classification task, ensuring that the learned embeddings align with the final matching objective while remaining compatible with LSH-based blocking.

Empirical results show that AutoBlock is particularly strong compared to traditional blocking techniques, especially in settings involving dirty or noisy data. This suggests that it is well-suited to handling heterogeneous or imperfect information that conventional syntactic blocking methods may fail to capture.

LSBlock is a method proposed by Karapiperis et al. (2025) that adapts a hybrid blocking framework integrating lexical and semantic similarity search to improve record linkage accuracy. The approach aims to increase recall while maintaining high precision, addressing limitations of traditional blocking methods in jointly capturing textual similarity and contextual meaning.

The method operates by first grouping similar records using blocking keys derived from attribute importance scores computed with Shapley Additive Explanations (SHAP). SHAP is used to identify the attributes that have the greatest influence on match predictions, and these attributes are then employed to construct effective blocking keys. The resulting blocking keys are converted into MinHash vectors, which represent strings as sets of q-grams and apply the MurmurHash3 function to approximate Jaccard similarity, retaining only candidate pairs that exceed a predefined lexical similarity threshold. This process allows the system to tolerate minor variations such as typographical errors or formatting differences.

To ensure scalability, LSBlock indexes the MinHash vectors using a Hierarchical Navigable Small-World (HNSW) graph, a multi-layer structure that supports efficient approximate nearest neighbor search with roughly logarithmic query-time complexity.

After candidate retrieval through the lexical index, LSBlock applies a semantic refinement stage to capture deeper similarities in meaning and phrasing. This step uses a pre-trained Sen-

tenceBERT model to generate dense embeddings for records and computes cosine similarity between query and candidate embeddings, retaining only those pairs that exceed a semantic similarity threshold. This ensures that retained matches are not only textually similar but also semantically aligned.

Both the lexical and semantic similarity thresholds are learned by training lightweight neural network models on labeled data to maximize the F1-score. The framework is designed to be flexible: for structured datasets, the full hybrid pipeline achieves high recall (above 0.95) while maintaining moderate-to-high precision (above 0.65). For unstructured or semi-structured datasets, where attribute-based blocking is less effective, the method can omit the lexical stage and rely solely on semantic similarity while still preserving high recall.

Karapiperis et al. (2025) explore the use of transformer-based language models for blocking in record linkage by leveraging their ability to capture semantic similarity through dense vector representations. In this approach, a RoBERTa model is fine-tuned using the Sentence-Transformers (SBERT) framework (Reimers & Gurevych, 2019a) to produce embeddings in which records referring to the same entity are positioned close to one another in the embedding space, while non-matching records are pushed further apart. Before embedding generation, individual records are serialized into textual representations that concatenate identifying attributes such as name, date of birth, and sex, allowing the model to jointly encode multiple fields within a single input sequence.

Candidate retrieval is performed using a K-nearest neighbors (KNN) search over the learned embeddings, implemented with the FAISS library (Douze et al., 2024) to enable efficient similarity search at scale. Cosine similarity is used as the distance metric, and a similarity threshold is applied to filter out low-scoring pairs, ensuring that only the most semantically similar candidate pairs are passed to the subsequent matching stage. Empirical results indicate that this embedding-based blocking strategy can substantially reduce the number of candidate comparisons; in particular, the fine-tuned RoBERTa model achieved a reduction of approximately 92% in candidate pairs while maintaining near-perfect recall.

However, comparative analyses suggest that purely language model–based blocking may be less effective than hybrid or probabilistic approaches when applied to highly structured data, such as patient records. In these settings, combining traditional rule-based blocking with probabilistic linkage methods that compute an overall similarity score has been shown to yield superior performance, achieving reductions of up to 95% in candidate pairs while preserving 100% recall.

A key limitation identified for RoBERTa blocking is its sensitivity to minor typographical variations. Because RoBERTa relies on subword tokenization, small character-level differences can lead to divergent token sequences, causing semantically similar records to be mapped to distant locations in the embedding space and resulting in low similarity scores.

2.3.3 Record Pair Comparison

After candidate pairs have been reduced to a manageable set through indexing or blocking, the record linkage process proceeds with detailed pairwise comparison. In this stage, each candidate pair is evaluated by comparing multiple attributes to quantify their similarity. These comparisons produce a similarity vector composed of numerical scores that capture agreement or disagreement across attributes. By combining evidence from several fields, the comparison step provides the foundation for determining whether two records refer to the same real-world entity.

2.3.3.1 String comparison metrics

String comparison is the process of assessing the similarity or difference between two sequences of characters (strings), either to determine whether they are identical or to quantify how alike they are using a defined measure (for example, an edit-distance or similarity metric). The level of similarity is given in a range from zero to one, where zero represents no similarity, and one represents an exact match. To calculate the similarity, some metrics can be used:

- Levenshtein metric

The Levenshtein distance is a minimum edit distance, which is the minimum number of operations (insertion, deletion, replacement, and transposition) required for editing and transforming one string a to another string b, with length $|a|$ and $|b|$ respectively (Agapito & Cauteruccio, 2025). It is given by $lev(a, b)$:

$$lev(a,b) = \begin{cases} |\mathrm{a}|, & \text{if } |\mathrm{b}| = 0, \\ |\mathrm{b}|, & \text{if } |\mathrm{a}| = 0, \\ lev(tail(a), tail(b)), & \text{if } head(a) = head(b), \\ 1 + min \begin{cases} lev(tail(a), b), \\ lev(a, tail(b)), \\ lev(tail(a), tail(b)) \end{cases}, & \text{otherwise.} \end{cases}$$

However, the Levenshtein distance does not necessarily reflect the distance between the names as perceived by human developers. For example, some letters are often confused, while others are very distinct. As a result, switching among similar letters can go unnoticed (e.g. simple vs. sinple), but switching among distinct letters would be immediately identified as wrong (e.g. complex vs. compfex) (Munk & Feitelson, 2022). This metric works at character level; it misses word-level or semantic similarity; it treats all edits as equally costly; it is sensitive to small local changes, for example, an extra character at the beginning that shifts many positions will augment the distance.

- Jaro-Winkler

The Jaro-Winkler metric gives extra weight to common prefixes of the two strings (a, and b) being compared, based on the number of matching characters (m) and transpositions (t). It adjusts the value based on the prefix scale factor; typically, if the initial characters match, the score increases (Winkler, 1990). It is given by $Jaro(a, b)$:

$$\text{Jaro}(a,b) = \begin{cases} 0, & \text{if } m = 0, \\ \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right), & \text{if } m \neq 0. \end{cases}$$

This metric also cannot cover the semantic relationship between the two strings; it can overweight common prefixes, which is useful for some matching tasks, but can overestimate similarity when many strings share a standard prefix. It is designed mainly for relatively short strings.

- Cosine metric

The cosine similarity measures the similarity between two non-zero vectors in an inner product space by calculating the cosine of the angle between the two vectors and determines whether two vectors are pointing in roughly the same direction (Han, Pei, & Tong, 2023). It is given by $cosine(a, b)$:

$$\text{cosine}(a,b) = \frac{a \cdot b}{|a| \cdot |b|}$$

As this metric requires vector representation, the quality of the similarity depends entirely on the chosen features (e.g., character n-grams, TF-IDF, embeddings, etc.). It also ignores the absolute length and magnitude, focusing only on angle, so two strings of very different length with similar token proportions can look similar, and does not inherently capture order information, only co-occurrence patterns.

- Ratcliff-Obershelp similarity

The Ratcliff-Obershelp similarity is a common string-comparison algorithm that prefers matching contiguous substrings over matching subsequences of disjoint letters, based on the (Ratcliff & Obershelp, 1988) algorithm. It is designed to compare two strings (a and b) and produce a similarity ratio that reflects how closely they match. It identifies matching character (m) sequences between the strings and calculates a similarity score using the formula:

$$\text{ratio}(a,b) = \frac{2m}{|a| + |b|}$$

While it often aligns well with human perception of similarity, it can produce asymmetric results depending on the order of comparison and may perform poorly when the sequence

of words is rearranged or when multiple equally long matches exist (Munk & Feitelson, 2022). As it focuses on character sequences and cannot distinguish between minor yet semantically important changes and innocuous edits. In Python, this similarity is computed by Sequence Matcher, a component of the *difflib* library.

- Code Names Matcher

Munk and Feitelson (2022) developed the Code Names Matcher Library to more accurately compare names in program code, such as variables and functions. Unlike traditional algorithms, it emphasizes perceptual and semantic similarity rather than purely lexical matching.

The library introduces full symmetry in its comparisons, ensuring that swapping the order of inputs yields identical results by basing the match on the substring that maximizes the similarity score rather than simply on the first maximal substring found by the algorithm.

A defining feature of the library is its word-level awareness. It interprets identifiers as collections of words, accommodating naming conventions like camelCase or underscore separation, and ensures normalization through lowercase conversion and careful word boundary handling. This prevents misleading matches between unrelated fragments of text.

The library also accounts for reordered and semantically related words. It enables unordered matching, in which word order does not affect similarity, and semantic matching based on synonymy, pluralization, and abbreviation relationships identified through a thesaurus.

To preserve and reward continuity, the authors introduced a ”glue” mechanism within the scoring formula, which increases the score for sequences that maintain the same order of adjacent words. Additionally, configurable thresholds determine when partial word matches are significant enough to count, thereby filtering out weak or ambiguous similarities.

Overall, the Code Names Matcher Library represents a more human-aligned approach to code name comparison, balancing lexical structure, logical order, and semantic meaning to produce similarity scores that better reflect programmer intent.

This library was primarily developed and optimized for English, as its semantic matching, pluralization, and default stop word features are based on English linguistic rules. However, it can also be used with Portuguese to a certain extent. While language-specific features like synonym and plural detection may not work correctly, its core lexical and structural matching functions—such as ordered and unordered matching based on character sequences or naming conventions—are language-independent. By customizing parameters like stop words and word separators, the library can be partially adapted to support Portuguese codebases effectively.

2.3.3.2 Machine Learning models for pairing

The Adaptive Fuzzy String Matching (AFSM) is a human-in-the-loop ensemble learning algorithm developed by Kaufman and Klevs (2021) to address fuzzy string-matching problems that arise when datasets must be merged using a single, imperfect identifying variable (such as

a name or address). The algorithm formulates fuzzy matching as a supervised classification task on pairs of strings.

For each candidate pair, AFSM computes a vector of similarity measures, including standard fuzzy metrics such as Levenshtein distance to capture insertions, deletions, and substitutions; Jaro–Winkler similarity to handle incomplete or slightly truncated strings; and Jaccard and cosine similarities computed over tokens or characters to account for missing substrings, reordered tokens, or rare typographical errors.

These similarity features are then passed to a supervised learning model, typically a random forest or boosted-tree-style classifier, which predicts the probability that a given pair represents a true match. The model is flexible enough to learn nonlinear interactions among similarity metrics.

A key component of AFSM is the human-in-the-loop process. Human reviewers evaluate a sample of proposed matches, labelling them as correct or incorrect, with particular emphasis on uncertain cases or potential false positives. These labels are appended to the training data, after which the model is retrained, and a new set of match predictions is generated. This iterative process continues until the model converges or the false-positive rate falls below a predefined threshold (1% in the authors' application).

Based on empirical results, AFSM generally identifies more matches with higher precision than traditional fuzzy string-matching approaches. When applied to a geographic dataset containing misspelled city names (PPP data), the algorithm achieved confidence scores above 0.95, within which all identified matches were confirmed as true matches through manual validation (i.e., 100% precision in that range). In terms of scalability, the authors report that convergence is typically achieved within a small number of iterations (approximately 1–10).

DeepMatcher, developed by Mudgal et al. (2018), is a modular deep learning framework for entity matching, designed within the Magellan project (Konda et al., 2016) with an emphasis on configurability and extensibility. It is released as an open-source Python package and is tailored for structured data.

The model pipeline begins with attribute embedding, where attribute values are transformed into vector representations. DeepMatcher typically employs character-level embeddings such as fastText, which are more robust to typographical errors and out-of-vocabulary terms than word-level embeddings.

Next, the attribute similarity representation module constructs similarity vectors for each attribute pair through two stages. First, attribute summarization aggregates token-level information within each attribute into a fixed-length vector. Second, attribute comparison applies a comparison function to the summarized attribute vectors from the two records, producing an attribute-level similarity representation.

These similarity representations across all attributes are then passed to the classifier module, which uses a multi-layer neural network based on Highway Networks to generate the final matching prediction.

DeepMatcher supports four strategies for attribute summarization (second step): Aggregate

Function (SFF), which computes a weighted average of word embeddings; Sequence-aware models, which use bidirectional recurrent neural networks (RNNs) to capture token order and contextual semantics; Sequence alignment models, which apply decomposable attention to perform soft alignment and pairwise token comparisons between attribute values; Hybrid models, which combine sequence-aware and sequence alignment approaches to jointly capture word order and cross-sequence interactions.

In terms of performance, DeepMatcher achieves results comparable to traditional machine learning systems such as Magellan, with an average F1 score of 87.9% versus Magellan's 88.8%. However, DeepMatcher has notable drawbacks, including long training times (averaging 5.4 hours) and a substantial requirement for labeled training data. Despite these limitations, DeepMatcher can significantly outperform traditional methods in scenarios where semantic similarity is critical. For example, on the Amazon–Google dataset, DeepMatcher achieves a 20.2% absolute F1 improvement by identifying synonymy in product titles that string-based similarity measures fail to capture.

2.3.4 Evaluation metrics

When the problem is supervised, and it is based on a machine learning model, it is important to understand how well the model performs. So, evaluation metrics enable the measure of effectiveness of the model.

- Accuracy

Accuracy is a metric used to evaluate the performance of a classification model. It measures the proportion of correct predictions made by the model out of all predictions. It is calculated as:

$$\text{accuracy} = \frac{Numberofcorrectpredictions}{Totalnumberofpredictions}$$

However, accuracy can be misleading when the dataset is imbalanced. For example, if one class represents only 5% of the data, a model that always predicts the majority class could still achieve 95% accuracy, while completely failing to correctly classify the minority class. Therefore, accuracy alone may not be sufficient to assess model performance in such cases.

- Precision

Precision measures the proportion of positive predictions made by the model that are actually correct. In other words, it indicates how reliable the model is when it predicts the positive class. Precision is especially useful in situations where the cost of a false positive is high. It is calculated as:

$$\text{precision} = \frac{TruePositives}{TruePositives + FalsePositives}$$

A high precision value means that most of the instances classified as positive by the model are truly positive.

- Recall

Recall measures the proportion of actual positive cases that were correctly identified by the model. In other words, it evaluates the model's ability to detect positive instances. Recall is particularly important in situations where missing a positive case (a false negative) is more costly than incorrectly labelling a negative case as positive. It is calculated as:

$$\text{Recall} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

A high recall value indicates that the model successfully captures most of the positive cases in the dataset.

- F1-score

The F1-score is a metric that combines both precision and recall into a single value. It provides a balanced measure of a model's performance, especially when dealing with imbalanced datasets. The F1-score is particularly useful when both false positives and false negatives are important and need to be considered simultaneously. It is calculated as:

$$\text{F1-score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

A high F1-score indicates that the model achieves a good balance between correctly identifying positive cases (recall) and ensuring that positive predictions are accurate (precision).

2.3.5 Similar Works in the context of Official Statistics

Ather (2026) proposes a hybrid framework for record linkage in official statistics that combines the Fellegi–Sunter probabilistic model with selective large language model Large Language Model (LLM) classification for ambiguous record pairs. The study argues that classical string-similarity methods often perform poorly on free-text fields, abbreviations, and acronyms, while open-source LLMs can capture the underlying semantic equivalence more effectively. To preserve confidentiality and reproducibility in national statistical offices, the framework is designed for secure on-premises deployment using open-weight models rather than cloud-based systems. The key methodological contribution is a two-stage decision process: the Fellegi–Sunter model assigns a prior match probability, pairs in the gray zone are routed to the LLM, and the LLM's binary output is then combined with the prior through Bayesian updating to produce a posterior probability. This design treats the LLM as informative but imperfect evidence and allows unresolved cases to be escalated to human review. In experiments on 1,000 challenging Canadian business record pairs, the 8B Llama model achieved the best performance with an F1-score of 0.87, outperforming Jaro–Winkler and Levenshtein, although at substantially higher computational cost. Overall, the paper shows that LLMs can reduce clerical burden and improve linkage quality when used selectively within a traditional probabilistic workflow, but it also highlights the need for calibration, cost control, and careful governance.

Chapter 3

Data Description and Processing

This chapter describes the data used to develop and evaluate the record linkage pipeline, and the preprocessing steps applied to prepare it for the matching process. Section 3.1 characterizes the two datasets drawn from Annex R of IES, covering the periods 2007 to 2016 (supervised) and 2017 to 2023 (unsupervised), respectively, detailing the available variables, the scope of the analysis, and the key structural changes in the data that affect longitudinal matching. Section 3.2 describes the preprocessing and data cleaning procedures applied to both datasets.

3.1 Dataset Description

To construct a primary key that uniquely identifies each establishment within a firm, BPLIM granted access to Annex R data covering the period from 2007 to 2023 (Banco de Portugal Microdata Research Laboratory (BPLIM), 2025). Access to the data was granted as part of an internship at BPLIM, exclusively through on-site use at the Porto branch of the Banco de Portugal.

The data covers 1,447,117 establishments over a sixteen-year period, corresponding to 7,740,845 observations. Multi-establishment firms correspond to approximately 10% of the total number of observations. Each row in the data corresponds to an establishment in a given year. Although this dataset is constructed based on mandatory administrative reporting, there is a residual share of establishments with at least one gap, 6.67%. This issue must be considered during the record linkage process, since these establishments should retain the same identifiers assigned in previously reported years. Consequently, maintaining access to historical data is essential to ensure the consistency and accuracy of the final dataset.

The dataset contains twenty-three variables, of which thirteen are qualitative (described in Table 3.1), and ten are the following numerical variables: the average number of persons employed during the year (*num_persons*); *costs_compile* [1]; *sales_compile* [2]; personnel expenses (*personnel_expenses*); employee remuneration costs (*employee_costs*); change in production inventories

[1] Variable that after 2010 reports the information disaggregated: (*costs*) and (*external_supplies*)

[2] Variable that after 2010 reports the information disaggregated: (*sales*) and (*services*)

(*inventory_change*); cost of goods sold and materials consumed (*costs*); external supplies and services (*external_supplies*); sales (*sales*); services rendered (*services*).

Table 3.1: Qualitative variables available.

Variable	Notation	Definition
NIPC	nipc	Taxpayer Identification Number.
Year	year	The year that refers to the data extracted.
Country	country	The country where the establishment is located. If the country is different from Portugal, the data for the specific location (street address, postal code 4 and 3 digits, locality, district, municipality, parish) should not be filled.
Street address	street_address	The street where the establishment is located.
Postal code 4 digits	postal_code4	The first digit designates one of nine postal regions; the following two digits designate postal distribution centres; the fourth digit is 0 if it belongs to a capital of a municipality, 5 if not, or any other digit if it is a designated address (Wikipedia contributors, 2025).
Postal code 3 digits	postal_code3	Classify the building blocks and designated addresses (Wikipedia contributors, 2025).
Locality	locality	Any location belonging to a parish.
Municipality	municipality	The second-level administrative division in Portugal, which has various localities and parishes.
Parish	parish	The third-level administrative division of Portugal, which has several localities.
Activity status	status	Indicates whether the establishment is (01) awaiting start of activity; (02) active; (03) activity suspended; or (04) activity ceased.
CAE Rev. 3	cae	Indicates the code of the main activity of the establishment according to the Portuguese Classification of Economic Activities (Instituto Nacional de Estatística, n.d.).
Order number of the establishment	id_estab	Sequential number of each establishment belonging to the company, and this numbering must be kept in future IES submissions. If a given establishment ceases activity, its previously assigned sequence number must not be reused for new establishments; these must instead receive the sequence number immediately following the last highest number assigned.

Continued on next page

Variable	Notation	Definition
Headquarter	headquarter	Binary variable that indicates if the establishment is the headquarter (1) or not (0).

Among the qualitative variables, country, municipality, parish, activity status, and headquarters are selected from predefined categories, ensuring consistency in spelling and formatting; however, this does not guarantee that the information is always accurate or complete. The remaining qualitative variables are recorded as free-text fields, which may include typographical errors, inconsistencies, or variations in reporting.

Two datasets were made available. The first dataset contains reports from 2007 to 2016 and is treated as ground truth for training and evaluating the record linkage model, as the primary keys assigned to establishments have been harmonized by BPLIM. This dataset provides a valuable benchmark for training and evaluation. However, this dataset was constructed with a semi-automatic approach which is costly to implement in regular updates of the data.

The second dataset, covering the period from 2017 to 2023, consists of raw data without any harmonization. This dataset is used to apply the trained record linkage model and to assign stable identifiers to establishments across years based on the matched records from the first dataset.

The main focus of this work is on firms that operate more than one establishment, the multi-establishment firms, in which the identification and tracking of individual establishments across years is more complex. Establishments located abroad are excluded because their reporting is inconsistent over time: until 2010, data on foreign establishments were aggregated rather than reported at the individual establishment level. Moreover, after 2010, the number of foreign establishments is very small.

Ensuring that each establishment receives a stable identifier is essential for accurately tracking its evolution across reporting years. This is particularly important given that establishments may not appear in the dataset every year, may experience changes in address or activity classification, or may be affected by reporting inconsistencies. The record linkage procedure, therefore, aims to address these challenges by matching establishments across years and assigning a single unique identifier that remains constant throughout the observation period, provided that the establishment's location remains unchanged.

An additional complication arises from the fact that some categorical variables, such as the main activity and parish, may undergo revisions over time, leading to changes in their classifications. Similar considerations apply to the parish, where in 2013 there was a reorganization of the parishes in Portugal, and 1,168 parishes were aggregated from an initial set of 4,260 parishes, which led to changes in the official designations of some parishes (Assembleia da República, 2013). In 2025, new territorial changes occurred, with the disaggregation of the 302 parishes making a total of 3,258 parishes (Assembleia da República, 2025). The record linkage procedure will not be performed on 2025 data; however, the model in the future will be implemented for this year, so the model must be able to handle the changes in the classifications, ensuring that establishments are correctly matched across years, even when their

CAE or parish classifications have been updated.

From the above, the records used are from the multi-establishment firms. In the first dataset, the core population consists of 29,981 firms located in Portugal. The second dataset is composed of 21,814 firms located in Portugal.

3.2 Preprocessing and Data Cleaning

The preprocessing and data cleaning steps are crucial for ensuring the quality and consistency of the data before applying the record linkage model. These preprocessing steps have been performed on both datasets.

The quality of the data can vary across years and establishments and this may pose challenges for the record linkage process. In particular, inconsistencies in reporting, missing values, and changes in the classification of activities and parishes can affect the accuracy of the matching process.

Analysis of the data started with missing values. Table 3.2 presents the variables and the number of missing values existing in both datasets. The variables with the highest number of missing values are those related to financial characteristics, such as costs and sales. The qualitative variables related to location and activity status have relatively fewer missing values, which is important for the record linkage process since these variables are key for a primary distinction between establishments. The dataset from 2017 to 2023 has fewer missing values compared to the dataset from 2007 to 2016, which may be due to improvements in data collection and reporting processes over time.

Regarding the annual distribution of missing values (Figure 3.1), the period from 2007 to 2016 shows relatively stable patterns across most variables. For financial variables such as costs_compile, sales_compile, personnel_expenses, and employee_costs, the largest variation occurs between 2007 and 2008 (approximately 6.50%), after which fluctuations remain moderate, ranging between about -2.50% and 3.50%. The inventory_change variable also exhibits a relatively stable trend over time, although consistently at a high proportion of missing values, with an average variation of around -1.20% during this period. These financial variables are important to distinguish between a truly duplicate record and records that share the same address information but represent different establishments.

The variables *costs*, *external_supplies*, *sales*, and *services* were only reported after 2010, and therefore display missing values up to that year. Regarding the identifier id_estab, 68% of missing values are concentrated in 2007, with the remainder spread between 2008 and 2011. Similarly, 77% of missing values for headquarter occur in 2007, with the rest distributed across 2008 and 2010. The num_persons variable remains relatively stable throughout the period. Finally, variables such as cod_postal4, cod_postal3, municipality, cae, and status have missing values almost entirely concentrated in 2010, with status being the only variable that also records a minimal number of missing values in 2011.

The pattern observed in Figure 3.1 indicates a clear improvement in data quality over time.

Table 3.2: Number of missing values present in each dataset.

Feature	2007 to 2016	Percentage (%)	2017 to 2023	Percentage (%)
id_estab	280	0.05	133	0.04
postal_code4	14	0.002	0	0.00
postal_code3	14	0.002	0	0.00
municipality	25	0.004	0	0.00
cae	37	0.006	0	0.00
headquarter	245	0.04	0	0.00
num_persons	158,731	27.27	28,703	8.02
costs_compile	176,358	30.30	48,336	13.51
sales_compile	186,958	32.12	51,696	14.45
personnel_expenses	183,878	31.59	56,014	15.65
employee_costs	184,855	31.76	56,095	15.68
inventory_change	451,608	77.59	232,251	64.90
costs	345,953	59.44	90,532	25.30
external_supplies	309,332	53.15	50,785	14.19
sales	360,917	62.01	106,450	29.75
services	390,023	67.01	116,437	32.54
status	262	0.05	0	0.00

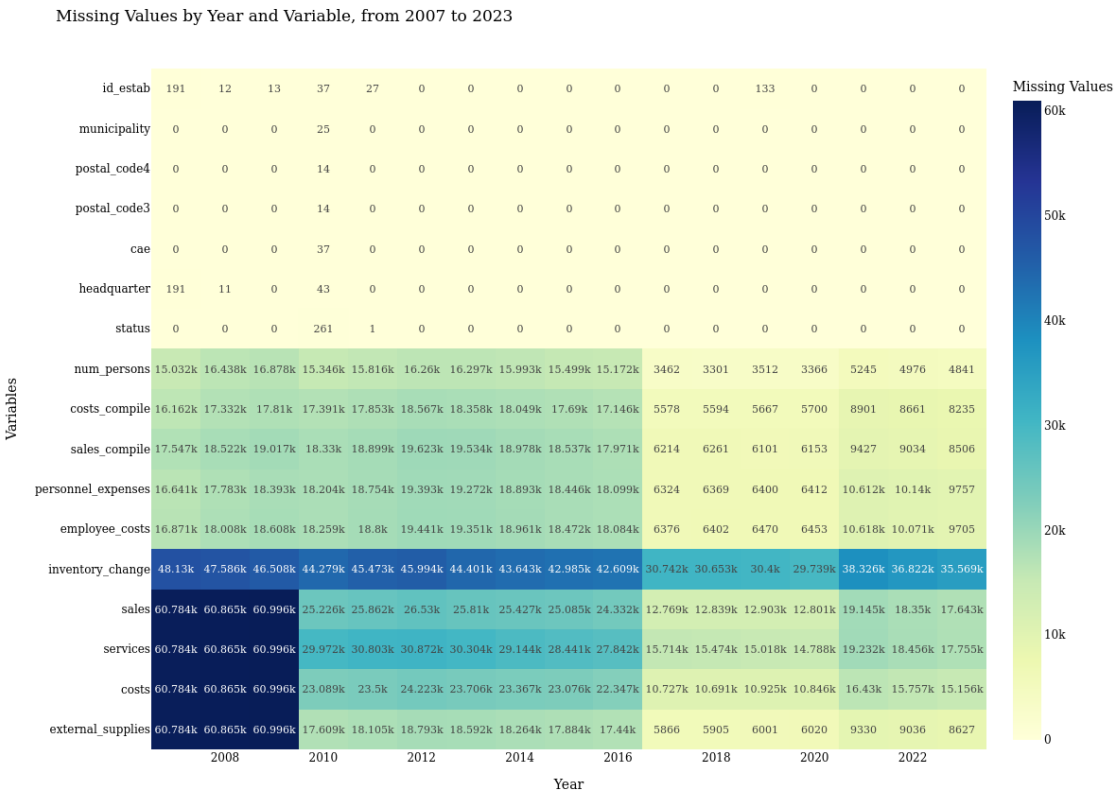
[3] Earlier years, particularly from 2007 to 2016, show a high concentration of missing values, as reflected by the more intense blue shading. However, there is a pronounced reduction in missing data beginning in 2017, suggesting a significant enhancement in data reporting practices. Although a moderate increase in missing values is visible between 2021 and 2023, it remains notably lower than the levels observed prior to 2017. The reduction in missing values over time can be attributed to several factors, including improvements in data collection and reporting processes, increased awareness of the importance of complete and accurate data for analysis and decision making.

However, financial variables are those with the highest proportion of missing values, which may create identification issues for establishments that share identical location characteristics and cae classification. Since these variables are important for distinguishing between such establishments, their absence makes it more difficult to determine whether records correspond to the same entity (duplicates) or to different ones (but share the same location characteristics and cae classification).

The free-text variables related to street address and locality were also analyzed for inconsistencies and errors, as they often contain typographical mistakes, spelling variations, and inconsistent formatting. These issues can negatively affect the accuracy of the matching pro-

[3]Note that some variables present zero values that represent missing information rather than a true zero, and are therefore not accounted for in the missing value count displayed in the figure.

Figure 3.1: Missing Data Matrix, from 2007 to 2023.



cess. To address this, a preprocessing stage was implemented to reduce such inconsistencies.

The preprocessing steps included converting all text to lowercase, removing extra whitespace, punctuation, and other artifacts, and applying standardization procedures. Standardization was applied only to the street address fields, using a manually created mapping of common abbreviations and frequent misspellings. This mapping was developed by the BPLIM team, based on the most recurring patterns identified in the dataset.

Given the large volume of records, this process was computationally demanding. To improve efficiency, a parallel text processing approach (from the *multiprocessing* library) was adopted, in which the dataset was divided into smaller chunks and processed simultaneously across multiple CPU cores. This significantly reduced preprocessing time, particularly for large datasets.

The standardization procedure was carried out in three stages. The first stage focused on the first word of each record and included a mapping of 606 corrections. The second stage targeted the second word, with 530 corrections applied. Finally, the third stage addressed corrections throughout the entire record, with a mapping of 35 additional adjustments.

The mapping applied to the first word focuses primarily on the standardization of street types and related location descriptors. This includes common designations such as ”rua” (street), which may appear in multiple forms (e.g., ”r u a”, ”r rua”, ”r do r”, ”r ua”), as well as other street types such as ”avenida”, ”travessa” (lane), and ”estrada” (road). It also covers open spaces and public areas, including ”largo” (plaza) (e.g., ”lr”, ”lrg”, ”larg”, ”lg”), ”praceta”

(small square), and ”pátio” (courtyard). In addition, it standardizes terms related to residential or rural areas (e.g., ”bairro” (neighborhood), ”quinta” (farm)), commercial or industrial zones (e.g., ”centro comercial” (shopping center), ”zona industrial” (industrial zone)), buildings and structures (e.g., ”edifício” (building), ”lote” (lot)), and public facilities (e.g., ”mercado” (market), ”estação” (station)). The mapping also includes titles and prefixes such as ”doutor/a” (doctor), ”engenheiro/a” (engineer), and ”santo/a” (saint), as well as infrastructure-related terms like ”estrada nacional” (national road) and ”ponte” (bridge).

The mapping of the second word focuses on the standardization of commonly occurring terms that complement address information. These include professional and academic titles (e.g., ”doutor/a” (doctor), ”professor/a” (professor)), military ranks (e.g., ”general” (general), ”tenente” (lieutenant)), religious titles (e.g., ”padre” (father), ”arcebispo” (archbishop)), and nobility or honorifics (e.g., ”dom” (lord), ”dona” (lady), ”visconde” (viscount)). It also addresses diplomatic titles (e.g., ”embaixador” (ambassador)), institutional and organizational terms (e.g., ”industrial” (industrial), ”municipal” (municipal), ”sociedade” (society), ”fundação” (foundation)), and types of facilities (e.g., ”hospital” (hospital), ”parque” (park)). Additionally, it standardizes geographical and descriptive terms (e.g., ”república” (republic), ”bombeiros” (firefighters)), acronyms and institutional names, and Portuguese proper names containing special characters (e.g., ”António”, ”João”, ”Açores”, ”Guiné”).

Finally, the mapping applied throughout the entire record focuses on general abbreviations and positional descriptors. This includes directional indicators (e.g., ”esquerdo” (left), ”frente” (front)), floor-level expressions (e.g., ”rés do chão” (ground floor)), and structural elements (e.g., ”bloco” (block), ”lote” (lot)). It also reinforces the standardization of certain street types and titles already addressed in the previous steps, ensuring consistency across the full address string.

Table 3.3 presents a fictitious example of the corrections made in the street address field, including the original and standardized versions of the addresses. This examples illustrate the types of inconsistencies that were addressed during the preprocessing step.

Table 3.3: Standardization examples of the street address field.

Raw address	**Standardized address**
Ru a Nsa Sra, Porto, primeiro esq lt 2	rua nossa senhora, porto, primeiro esquerdo lote 2

For Portuguese establishments, the postal code components (postal_code4 and postal_code3) were concatenated into a single seven digit variable. In cases where the last three digits were missing, zeros were added at the beginning of the postal_code3 to ensure a complete postal code. Furthermore, municipality fields were parsed to separate numeric identifiers from municipality names.

Together, these preprocessing steps produced two cleaned and internally consistent datasets ready for the record linkage pipeline that will be described in the following chapter.

Chapter 4

Model Implementation

This chapter presents the two record linkage approaches developed to assign stable establishment level identifiers across IES reporting periods. Section 4.1 introduces the mathematical formulation of the problem. Section 4.2 describes the empirical pipeline, a three-step sequential approach. Section 4.3 presents the semantic pipeline, which replaces the second and third steps of the empirical pipeline with a fine-tuned BERT-large Cross-Encoder.

4.1 String Matching Pipeline Model

The string matching approach can be represented as a mathematical model that captures the essential components and processes involved in the pipeline. First, a concise definition: Let $e_{i,t}$ denote the observed characteristics of establishment i in year t. The objective is to estimate a latent identifier $z_{i,t}$ such that observations corresponding to the same real establishment across years receive the same identifier despite reporting inconsistencies, duplicate records, and temporal instability. The problem is formulated as a constrained record linkage optimization problem based on similarity functions computed over spatial, categorical, and financial characteristics.

Definition of the notation used in the model:

- t denotes the years, where $t = 1, 2, ..., T$.
- s denotes the window size of the previous year with the historical years, where $s < t$.
- i denotes an establishment observation in year t, where $i = 1, 2, ..., N_t$ and N_t is the number of observations in year t.
- F_t denotes the set of firms in year t, $F_t = \{f\}_t$.
- $E_{f,t}$ denotes the set of establishments belonging to firm f in year t, $E_{f,t} = \{e\}_f^t$.
- $e_{i,t}$: denotes the characteristics of establishment i in year t.
- $z_{i,t}$: denotes the latent identifier for establishment i in year t.

Each observed establishment record is characterized by p features, which can be represented as a vector $e_{i,t} = (e_{i,t}^{(1)}, e_{i,t}^{(2)}, ..., e_{i,t}^{(p)})$. The goal is to assign a latent identifier $z_{i,t}$ to each establishment such that records corresponding to the same real establishment across different years receive the same identifier.

$$z_{i,t} \in Z$$

where Z is the set of all possible latent identifiers. So if $z_{i,t} = z_{j,s}$, then $e_{i,t}$ and $e_{j,s}$ represent the same real establishment, based on a set of characteristics. The goal is to estimate a mapping such that $z_{i,t} = z_{j,s}$ if and only if $e_{i,t}$ and $e_{j,s}$ represent the same real establishment.

The similarity function can be defined as follows:

$$S(e_{i,t}, e_{j,s}) = \sum_{k=1}^{p} w_k * s_k(x_{i,t}^{(k)}, x_{j,s}^{(k)})$$

where s_k is a similarity function for the k-th feature, and w_k is a weight that reflects the importance of that feature in determining whether two records match.

The classification rule is based on two parameters, θ_{upper} and θ_{lower} that will be used to define a match: $0 < \theta_{\text{lower}} < \theta_{\text{upper}} \leq 1$. Then, the classification rule can be defined as follows:

- **Match if** $S(e_{i,t}, e_{j,s}) \geq \theta_{\text{upper}}$
- **Non-match if** $S(e_{i,t}, e_{j,s}) < \theta_{\text{lower}}$
- **Partial Match if** $\theta_{\text{lower}} \leq S(e_{i,t}, e_{j,s}) < \theta_{\text{upper}}$

ID assignment is challenging because multiple establishment records within the same firm can be similar to each other, increasing the risk that distinct establishments are incorrectly assigned the same identifier. This can be formulated as an optimization problem, where the goal is to minimize the total matching cost: higher similarity corresponds to lower cost, and lower similarity corresponds to higher cost. The optimization problem can be defined as follows:

$$\min_{x} \sum_{i} \sum_{j} C_{i,t} * X_{i,j}$$

where i is the establishment in year t and j is the establishment in previous years s, and $C_{i,j} = 1 - S(e_{i,t}, e_{j,s})$ is the cost of assigning the same identifier to records $e_{i,t}$ and $e_{j,s}$, and $X_{i,j}$ is a binary variable that equals 1 if records $e_{i,t}$ and $e_{j,s}$ are assigned the same identifier, and 0 otherwise. However, there are some constraints that need to be added to the optimization problem to ensure that the same identifier is assigned to records that represent the same real establishment:

- $\sum_j X_{i,j} \leq 1$ for all i, which ensures that one establishment cannot correspond to multiple historical records.
- $\sum_i X_{i,j} \leq 1$ for all j, which ensures that each historical record is assigned to at most one establishment in year t.

Duplication detection of the establishment can be formulated as:

$$D(e_{i,t}, e_{j,t}) = 1$$

if

$$x_{i,t}^{(k)} = x_{j,t}^{(k)}$$

for all k, where k corresponds to the features used to define a unique establishment. This means that if two records have the same address and main economic activity, they are considered possible duplicates. This does not mean that they are the same establishment, but it is a strong indication that they might be, and to determine that, the features related with the financial characteristics can be used to determine if they are the same establishment or not.

Lastly, the historical search can be formulated as follows: For each establishment record $e_{i,t}$ in year t, that was not matched to any establishment in $t-1$, define the candidate set as:

$$H_{i,j} = U_{s=t-5}^{(t-1)} E_{f,s}$$

For each candidate record $e_{j,s}$ in the candidate set, the similarity score is computed $S(e_{i,t}, e_{j,s})$, and the classification rule is applied to determine if they match or not. If a match is found, the same identifier is assigned to both records.

4.2 Record Linkage Empirical Pipeline Implementation

The implementation of the model involves several steps and the following sections describe each of these steps in detail.

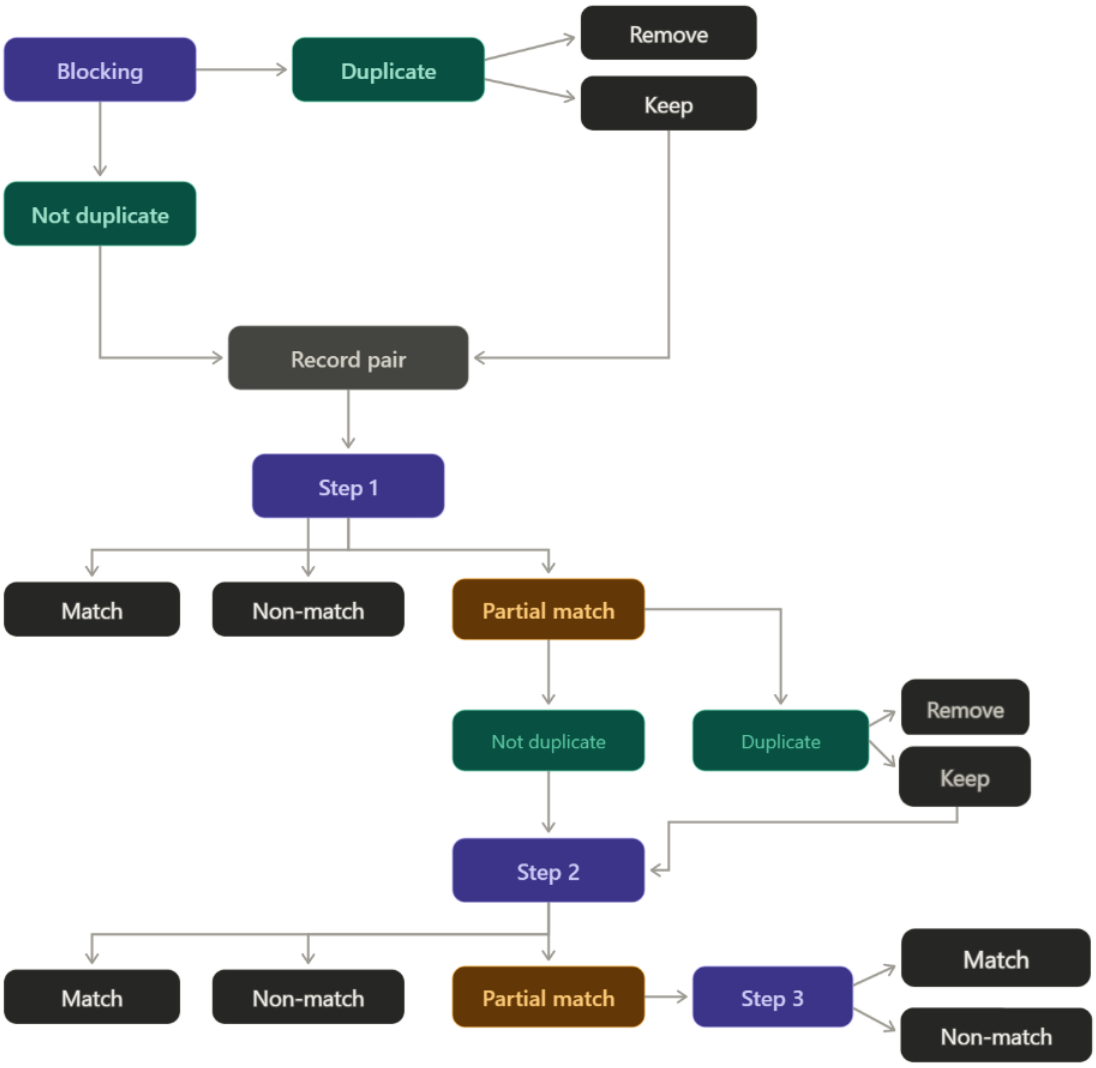
The algorithm processes the data through a sequential yearly pipeline, where each year constitutes a separate stage whose output feeds into the next. While each year is processed as a self-contained unit, the stages are not fully independent: results from previous years inform the processing of the current year, particularly during the historical search step. Basically, the algorithm is divided into three main steps that are applied sequentially for each year, represented in Figure 4.1.

4.2.1 Blocking

Blocking aims to group records in blocks that share some common attributes. This block will be used to calculate the similarity scores between records within the same block, which is computationally more efficient than calculating similarity scores between all possible pairs of records. The blocking key is chosen based on the features that are most likely to indicate a potential match.

In the dataset, the most suitable key to perform blocking is the *nipc* variable, which is a unique identifier for firms. Based on that, all the comparisons are made between records that share the same *nipc*. Given that the reporting of the *nipc* variable is mandatory, it allows to uniquely and unambiguously identify a firm in the data.

Figure 4.1: Record linkage pipeline using probabilistic matching.



In this way the *nipc* is a powerful blocking key, because it ensures that all records that belong to the same firm are grouped together, and it also ensures that records that belong to different firms are not grouped together. This is a very efficient way to perform blocking, because it reduces the number of comparisons that need to be made, and it also increases the accuracy of the matching process, because it ensures that only records that are likely to match are compared.

4.2.2 Identification of Possible Duplicated Records

Before performing the Record Pair Comparison phase, a classification step was conducted to identify whether a firm had multiple establishments sharing the same address and same main economic activity. In such cases, these establishments were flagged as potential duplicates. Subsequently, additional analysis and metrics were applied to determine whether they were true duplicates. Only records classified as non-duplicates were then used in the Record Pair

Comparison phase. This step is essential, as it ensures that duplicate records are identified and removed prior to matching. In this way, it prevents situations where multiple records represent the same establishment, which could otherwise lead to incorrect matches and unreliable results.

However, it is possible that more than one establishment shares exactly same address, defined as the combination of *street_address*, *postal_code*, *locality*, *municipality*, and *parish*. The main economic activity is also included *cae_estab*, because two or more establishments at the same location may have a different economic activity (for example, a firm with a factory and store located in the same address) and in this case is not a duplicate.

Quantitative variables used to distinguish

The flagged potential duplicate establishments must be resolved through a multi-phase process. The first phase focuses on the available quantitative variables, including *num_persons*, *costs_compile*, *sales_compile*, *personnel_expenses*, *employee_costs*, *inventory_change*, *costs*, *external_supplies*, *sales*, and *services*.

As a preprocessing step, all missing values in these variables are replaced with zero. Additionally, values greater than zero but smaller than one are also set to zero. This adjustment addresses specific cases where establishments report negligible values (e.g., 0.01), which can interfere with duplicate detection. Such values are considered implausible for annual activity and would otherwise prevent correct identification of duplicates due to strict inequality comparisons.

To distinguish true duplicates from establishments that merely share the same location and main economic activity, the quantitative variables are compared across records. Establishments are considered distinct if all quantitative variables differ and are at least equal to 1. Conversely, records are treated as true duplicates if all quantitative variables are equal, or if one record contains valid (non-zero) values while the others contain only zeros. In such cases, only the record with valid information is retained, and the remaining records are removed.

Resolving the Remaining Possible Duplicates

In this stage, it is required that the *num_persons* variable be equal across records, or that only one record contains a non-zero (or non-missing) value. Additionally, a reduced set of quantitative variables is considered:*costs_compile*, *sales_compile*, *personnel_expenses*, *employee_costs*, *inventory_change*. This restriction is motivated by the fact that, following Annex R aggregation, several disaggregated variables are newly introduced and often contain identical or missing values. Therefore, only the aggregated variables are used, as they are more reliable.

Within this phase, several cases are addressed. If a record contains only a single valid value among the selected variables, it is interpreted as a correction or update of a reporting error. In such cases, the record is removed, and its valid value is incorporated into the corresponding field of the more complete record.

For duplicate candidates with identical values across all selected variables, only one record is

retained and the others are removed. In situations where duplicate records contain complementary information, meaning each record has valid values in different variables and missing values in others, the records are merged. This occurs when, taken together, the records fully populate all variables without conflict. For example, consider Record A: (1111, nan, 235, nan, 58742) and Record B: (nan, 2887, nan, 3698, nan). These records are combined into a single record: (1111, 2887, 235, 3698, 58742), and the redundant record is removed. After resolving all true duplicates, only one consolidated record remains for each establishment. The next step is to assign identifiers by computing similarity scores using the previously defined similarity function.

Record Pair Comparison and ID attribution

To reduce computational complexity, comparisons are restricted to records within the same *municipality*. Additionally, records that have already been assigned identifiers are excluded from the comparison set to prevent inconsistencies. An identifier is assigned only if the similarity score is greater than or equal to the upper threshold, selecting the match that maximizes the similarity score when multiple candidates exist within the same firm. If no such match is found, the establishment is classified as new and assigned a new identifier.

For each such group of undifferentiated establishments, the algorithm computes similarity scores (based on the function used on Subsection 4.2.3.1) between current year records and all unmatched records from the previous year. Only the top four candidates with similarity scores equal to or above to the upper threshold are retained. Among these, a subset of tied candidates is identified, defined as those with similarity scores greater than or equal to 0.984 across all features and whose identifiers have not yet been assigned. This near one threshold is used instead of strict equality to account for minor data inconsistencies. The assignment process then follows three prioritized cases:

1. **Common identifier case.** When the reported establishment identifier (*id_estab*) is consistent across years for establishments sharing the same address, the algorithm exploits this correspondence to preserve previously assigned identifiers. Specifically, if a current year establishment has an *id_estab* that matches one of the tied previous year candidates, it inherits the identifier assigned in the prior year. This ensures that establishments which are genuinely distinct, but very difficult to distinguish, retain a stable identity over time, even when their address is shared with other establishments.

2. **Financial similarity case.** If neither of the previous conditions applies, the assignment is based on *num_persons* and normalized financial features: *costs_compile*, *sales_compile*, *personnel_expenses*, *employee_costs*. A weighted mean squared error Mean Squared Error (MSE) matrix is computed for all candidate pairs, where *num_persons* is assigned a weight of 0.4 and each of the remaining variables a weight of 0.2. Matches are determined greedily by selecting the pair with the lowest weighted MSE, ensuring that each establishment is matched at most once. The higher weight assigned to *num_persons* reflects its relative stability over time and lower incidence of missing values.

3. **New identifier case.** Any current year establishment that remains unmatched after

the previous steps is classified as a new establishment and assigned a new sequential identifier.

4.2.3 Record Pair Comparison

The algorithm performs the linkage in three steps. The methodology is based on a progressive filtering approach, beginning with a more restrictive phase and evolving into increasingly broader phases, in order to minimize the occurrence of matching errors. The process must follow a temporal order: records from year t can only be linked to records from year $t-1$ or earlier, never to records from future years. Ideally, the linkage for year t should be completed before attempting the linkage for year $t+1$. This reflects a methodological choice: since the data is derived from annual extractions, the algorithm assumes that the earliest available record for an establishment is correct and builds its longitudinal history forward from there. An alternative design could instead proceed backward from the most recent year, but the forward-building approach was adopted to ensure that the resulting dataset reflects the natural progression of an establishment over time.

4.2.3.1 Step 1: Matches Identification

The first step of the algorithm is designed to be more restrictive, meaning that it uses a more stringent set of criteria to determine whether two records match, as well as stringent thresholds. This step is intended to identify the most obvious matches, where the records are very similar and there is a high degree of confidence that they represent the same establishment.

Features Selection

The criteria used in this step are based on the features that are most likely to indicate a potential match. Therefore, the features used to calculate the similarity scores are the ones related to the location of the establishment and its main activity:

- $street\_address\_new$: corresponds to the *street_address* with the preprocessing step described in the previous chapter.
- $locality\_new$: corresponds to the *locality* feature subjected to the preprocessing described in the previous chapter.
- $postal\_code\_new$: corresponds to the aggregation of the *postal_code4* and $postal\_code3$ features preprocessed.
- $municipality\_id$: corresponds to the numeric identifier of the *municipality* feature.
- $parish$.
- cae: the code of the main activity was not used in the transition year in which the classification changes: 2008.

Blocking

The process begins by defining the year to be processed. Based on this year, two datasets are created: one containing the records from the previous year and another containing historical records from the previous five years. Then, for each firm *nipc* in the current year, the algorithm checks whether it was present in the previous year. Depending on this result, the firm follows one of two paths.

Similarity measures used in each feature selected

For firms present in the previous year, a similarity score is computed between each current year establishment and the establishments from the previous year using the features described above. The final similarity score is obtained as the mean of the individual similarity scores calculated for each feature.

Based on experimental results, the feature *street_address_new* is designed to maximize the similarity score by leveraging two metrics from the RapidFuzz library: token sort ratio and token set ratio. These methods first tokenize the input strings and then either sort the tokens alphabetically or consider only the shared and unique tokens between the strings. The similarity is then computed using the Levenshtein distance, and the higher of the two scores is selected. Since these metrics ignore token order, they are particularly well-suited for address comparison, where word order may vary but the semantic content remains consistent. For example, the strings "Shopping Center, Rua das Camélias, 45" and "Rua das Camélias, Shopping Center, 45" would be identified as highly similar using token-based approaches, whereas a direct string comparison would fail to capture this similarity.

The feature *locality_new* is computed using the normalized Jaro-Winkler similarity. This metric is appropriate because locality names are typically short and contain few words. Additionally, Jaro-Winkler assigns greater importance to matching prefixes, which often carry the most relevant information in locality names.

For the features *postal_code_new*, *municipality_id*, *parish*, and *cae*, the similarity is computed using the Ratcliff-Obershelp similarity. These attributes are generally short and numeric, making this metric suitable due to its simplicity. It measures similarity as the proportion of matching characters relative to the total number of characters in both strings.

The final similarity score is calculated as the arithmetic mean of all feature level similarities, assigning equal weight to each feature. Experimental evaluation showed that this uniform weighting scheme provides the best overall performance (see Section 5.2).

Record Pair Comparison

To compare records from the current year with those from the previous years, each establishment in the current dataset is compared against all establishments from the previous years within the same firm. For each record, the candidate with the highest similarity score is selected as the best potential match.

If the similarity score of this best candidate exceeds a predefined upper threshold, the pair is considered a candidate match. In cases where the similarity score equals 1 (i.e., an exact match), the corresponding record from the previous year is removed from further consideration to prevent it from being matched with multiple current records.

However, the method allows multiple current year records to be matched with the same previous year record, provided that their similarity scores exceed the upper threshold. This design choice accommodates situations where multiple establishments are highly similar to a single record from the previous year, and it is not possible to confidently determine a unique match. Such cases are classified as multiple high-confidence matches, and their resolution is addressed in a subsequent step.

For all candidate matches with similarity scores above the upper threshold, the identifier of the matched record from the previous year is assigned to the corresponding record in the current year. If no candidate exceeds the threshold, no identifier is assigned.

Firms Present in the Previous Year

To address cases where multiple current year records are matched to the same record from the previous year, a dedicated function was developed. First, the candidate establishments are sorted in descending order according to their similarity scores. Four possible scenarios are then considered:

- **Case 1.** If the highest similarity score is exactly equal to 1, the corresponding record is assigned the previous year identifier. The remaining records are then evaluated through a historical matching process. If a match is found in the historical data, the corresponding historical identifier is assigned; otherwise, the records are classified as new establishments and assigned new sequential identifiers. When multiple records fall into this category, identifiers are assigned incrementally, ensuring no duplication.
- **Case 2.** If all similarity scores exceed the upper threshold but none equals 1, all establishments are considered highly similar, and it is not possible to determine a unique match. In this case, all records are classified as partial matches and are handled in a later stage of the algorithm.
- **Case 3.** If only one establishment has a similarity score above the upper threshold, that record is assigned the identifier of the corresponding previous year record. Any additional records with similarity scores in the interval$[\text{upper_threshold} - 0.05, \text{upper_threshold}[$ are treated as partial matches and deferred to a later stage. The remaining records are processed using historical data: if a match is found, the historical identifier is assigned; otherwise, they are classified as new establishments and assigned new sequential identifiers.
- **Case 4.** If all similarity scores are below the upper threshold, but some exceed a predefined lower threshold, those records are classified as partial matches and handled in a later stage. The remaining records are then processed through historical matching; if a match is found, the corresponding historical identifier is assigned, otherwise they are classified as new establishments and assigned new sequential identifiers.

For cases where there is no conflict involving multiple matches to the same previous year record, the assignment follows a straightforward rule based on the similarity score. If the similarity score exceeds the upper threshold, the identifier of the matched record from the previous year is directly assigned. If the similarity score falls between the upper and lower thresholds, the record is classified as a partial match and deferred to a later stage of the algorithm for further resolution. If the similarity score is below the lower threshold, the record before being considered a non-match is compared with the historical data, and from it could be considered as a match or non-match. In this context, a new establishment refers to a non-match classification, meaning that the record from the current year cannot be matched to any record from either the previous year or earlier historical data.

The historical search can be interpreted as a second stage matching process for records that were not successfully matched in the initial comparison. This step is designed to capture matches that may have been missed due to reporting inconsistencies or temporal variability, such as missing records. In this phase, only establishments that remain unassigned from previous steps are considered, ensuring that each establishment identifier is assigned to at most one record.

As a result, the number of comparisons is reduced, improving computational efficiency. The historical search uses the same similarity metrics and classification rules as the initial matching step. However, instead of comparing current year records with those from the previous year, the comparison is performed against the five most recent years available records.

The decision rules remain consistent: if the similarity score exceeds the upper threshold, the historical identifier is assigned (match); if the score falls between the upper and lower thresholds, the record is classified as a partial match and handled later; and if the score is below the lower threshold, the record is classified as a new establishment and assigned a new sequential identifier.

Firm Not Present in the Previous Year

When the *nipc* is not found in the previous year, the corresponding establishments follow one of two possible paths: they are either matched through the historical search or classified as new establishments.

The algorithm first checks whether the *nipc* exists in the historical dataset. If it is not present, all establishments associated with that firm are classified as new and assigned new sequential identifiers. If one of these establishments is identified as the *headquarter*, it is assigned identifier 1, in accordance with Annex R, which specifies that the headquarter must get the identifier 1 of the firm. The remaining establishments are then assigned subsequent identifiers starting from 2. If no headquarter is identified, the assignment begins directly at identifier 2.

If the *nipc* is present in the historical dataset, the algorithm computes the similarity score for each establishment using the previously defined similarity function, comparing it with records from the most recent year in which that *nipc* appears in the historical data. Three scenarios may then occur:

- **Scenario 1.** If the similarity score is equal to or greater than the upper threshold, the

identifier of the matched historical record is assigned.

- **Scenario 2.** If the similarity score is below the upper threshold but equal to or greater than the lower threshold, the establishment is classified as a partial match.

- **Scenario 3.** If the similarity score is below the lower threshold, an additional historical search is performed. In this step, records that have already been assigned an identifier are excluded from consideration. If a match is found with a similarity score equal to or greater than the upper threshold, the corresponding identifier is assigned; otherwise, the establishment is classified as new and assigned a new sequential identifier.

4.2.3.2 Step 2: Partial Matches from Step 1

This stage attempts to resolve the partial matches identified in the previous step (Step 4.2.3.1), by adopting a similar, but more comprehensive approach. The main distinction lies in the treatment of the *locality* attribute. Based on inspection of the data, this variable was found to be highly inconsistent: it represents a relatively small geographic area, is recorded as free text, and is therefore prone to vary across reporting periods. For example, the same locality may be described differently in consecutive years, or even incorporated into the *street_address* field. Due to this subjectivity and instability, the *locality* attribute is excluded from the matching criteria in this step.

Identification and Resolution of Potential Duplicate Records

In addition, potential duplicates are re-evaluated under a stricter definition that excludes *locality* from the set of variables selected in Step 1, and are then assessed using the same resolution procedure described in the previous step (Section 4.2.2), including the handling of true duplicates and establishments that share the same location and main economic activity but represent distinct entities. In this phase, only previous records that have not been assigned an identifier are considered, ensuring consistency with earlier stages of the algorithm.

Records Pairs Comparison

After resolving potential duplicates, the remaining partial matches are addressed using a new similarity function specifically designed for this stage. This function considers once again all the features used in the previous step with the exception that *street_address* and *locality* are combined into one feature, the *partial_address*.

For each pair of current and previous year establishments, a similarity score is computed for each feature. The Ratcliff-Obershelp similarity is used for all features except *partial_address*. For the *partial_address* attribute, similarity is defined as the maximum of three measures: token sort ratio, token set ratio (both from the RapidFuzz library), and TF-IDF cosine similarity. The TF-IDF vectorizer is fitted on the union of street addresses, combining the partial addresses from the current year (treated as the document set) with all available records from previous and historical data with a ID not matched within the same firm (the corpus). This

ensures that term weights reflect the full vocabulary across both cross-sections. The vectorizer incorporates unigrams, with tokenization based on whitespace-separated address components. This hybrid approach leverages the strengths of both TF-IDF and fuzzy matching techniques. TF-IDF captures the relative importance of terms within the corpus, downweighting common and uninformative tokens frequently found in Portuguese addresses (e.g., "Rua", "Avenida") while emphasizing more distinctive elements such as street names or locality identifiers. In parallel, fuzzy matching methods account for abbreviation inconsistencies and typographical variations, which are common in administrative data.

The decision to define similarity as the maximum of the three measures is motivated by robustness to different types of variation in address data. Rather than averaging the scores, which could dilute strong signals when one method performs particularly well, the maximum operator preserves the highest-confidence match across the three metrics.

The final similarity score is calculated as a weighted mean, using the following features: *partial_address*, *postal_code4*, *municipality_id*, *parish*, and *cae*. In 2008, when the classification of the economic activity changes, *cae* is not used.

Records Classification

Candidate pairs within each firm block are evaluated through pairwise similarity computation, yielding a similarity matrix $S \in R^{m,n}$, where m denotes the number of establishments in year t and n the number of establishment in $t-1$ and historical years available.

An optimal one-to-one assignment is obtained using the Hungarian algorithm (Kuhn, 1955), which identifies the matching that maximizes the total similarity across all pairs. This global optimization approach is preferred over greedy methods (because in greedy methods each establishment is sequentially matched to its highest scoring candidate) as it avoids suboptimal allocations in which a locally optimal match precludes a better overall assignment. In particular, greedy strategies may assign a highly similar pair early on, even when one of the establishments involved represents the only viable match for another record.

From a computational perspective, the Hungarian algorithm has time complexity $O(k^3)$, where $k = max(m, n)$. While this may be computationally demanding for large matrices, it is only applied within firm blocks, which reduces k.

When the number of establishments in the current year exceeds that of the previous years $(m > n)$, unmatched records are assigned a similarity score of -1, indicating that they require further processing.

Following the assignment, a duplicate resolution step is performed. If multiple current year establishments are assigned to the same previous year identifier, only the match with the highest similarity score is retained, and the remaining records are reassigned a score of -1. A match is confirmed only if the similarity score is greater than or equal to the upper threshold.

For establishments that do not meet this threshold, a second comparison is conducted using the same similarity function, but now including historical records (excluding identifiers that have already been assigned). If the similarity score exceeds the upper threshold, the establish-

ment is matched; if it falls below the lower threshold, the establishment is classified as new and assigned a new sequential identifier; otherwise, it remains classified as a partial match.

Finally, for the remaining partial matches, an additional rule is applied. If the similarity score is less than or equal to $0.05 + lower\_threshold$, and the reported establishment identifier *id_estab*, has not been previously assigned, the establishment is classified as new and assigned a new sequential identifier based on the maximum identifier already assigned within the firm. All other cases remain classified as partial matches for further analysis.

4.2.3.3 Step 3: Partial Matches from Step 2

The final step is deterministic and aims to resolve all remaining partial matches from Step 2, ensuring that every establishment is definitively classified as either a match or a non-match and assigned an identifier.

At this stage, the set of identifiers already assigned to establishments in year t is computed, along with the identifiers from the previous year and historical data that remain available. If no identifiers are available for a given firm, all remaining partial matches are classified as new establishments (non-matches) and assigned new sequential identifiers.

To resolve the remaining establishments, a new similarity function is applied using the same features as in Step 1. The main difference in this step lies in the similarity computation for the feature that can generate more controversy, the *street_address*. For *street_address_new*, Jaro-Winkler is used as the primary metric, replacing the combination of token-based and Jaccard similarities used in previous steps. For the *locality* feature, the similarity score is defined as the maximum of the token sort ratio, token set ratio, and Ratcliff-Obershelp similarity. The remaining features continue to use the Ratcliff-Obershelp similarity.

The overall similarity score is computed as the mean of the individual feature similarities. Matching is then performed using a greedy one-to-one assignment strategy. If the similarity score is greater than or equal to the upper threshold, the pair is considered a match and the corresponding identifier is assigned. Otherwise, the similarity score is set to -1, indicating a non-match, and the establishment is assigned a new sequential identifier.

4.3 Record Linkage using Semantic Analysis Implementation

This approach extends the previous matching strategy, but its goal is to recover the remaining partial matches by using an Natural Language Processing (NLP)-based method that captures the semantic meaning of words from their context in a sentence (Devlin, Chang, Lee, & Toutanova, 2019). Instead of comparing strings only, it performs a bidirectional contextual analysis and therefore understands semantics rather than relying solely on syntax. This makes it more effective in cases involving synonyms, such as shopping center and shopping mall, and in many situations where partial matches are ambiguous or present contradictory signals.

This approach combines deterministic matching and embedding-based matching, using the

preprocessing treatment used in the section 3. From the previous approach (4.2), it reuses the following steps: the blocking technique (4.2.1) with NIPC as the blocking key, the identification of possible duplicates (4.2.2), and the same resolution process described in Step 1 (4.2.3.1), preserving the matched and non-matched records as solved records. In this way, the easy matches are resolved using a syntactic approach, while the difficult matches are handled using a semantic approach. The remaining partial matches from Step 1 are then processed by a BERT model and classified as either a match or a non-match according to the defined threshold. Using both techniques helps avoid wasting too many resources on cases that can be solved trivially. The model operates in a closed environment, without access to the internet or any external resources, relying exclusively on the representations learned during fine-tuning. Using both techniques helps avoid wasting too many resources on cases that can be solved trivially.

4.3.1 Solving Partial Matches using Deep Learning

This record linkage approach is a binary text classification problem that determines if two records refer to the same physical establishment. Here, the pipeline is divided into three stages: preprocessing (including text representation and training pair construction with hard negatives), cross-encoder fine-tuning, and threshold based inference with optimal assignment.

4.3.1.1 Preprocessing

To fine-tune the cross-encoder, the first step is to create a labeled training dataset. The dataset consists of positive pairs *(label=1)*, meaning that both records refer to the same establishment, and negative pairs *(label=0)*, which corresponds to the most similar incorrect candidate, that is, the one with the highest similarity score among the non-matching records. In this way, the negative examples are genuinely confusable pairs that the fine-tuned model must learn to reject. For establishments that do not appear in any previous year, the record is considered a new establishment and is assigned *label=0*.

The features used to build the pairs combine geographic identifiers, economic activity identifiers, and financial indicators, allowing the model to access both structural and economic information. The features used are *street_address*, *locality*, *postal_code*, *municipality*, *parish*, *cae*, *headquarter*, *num_persons*, *cost_compile*, *sales_compile*, *personnel_change*, *employee_costs*, *inventory_changes*, and *status*.

Each establishment is then represented as a single structured string by concatenating these features in a consistent format.[1] The representation is as follows:

Street: value | Locality: value | Municipality: value | Cae: value | Postal Code: value | ... | status: value

To create the labeled dataset, a bi-encoder was used to generate the positive and negative pairs. The labeled training pairs are constructed from the ground-truth panel, where, for each establishment in year t, two types of pairs are generated under the same blocking strategy: positive pairs *(label=1)* and hard-negative pairs *(label=0)*. Hard-negative pairs are critical

[1] If a variable has a negative value, it is omitted from the representation.

to training quality because they help the model learn the rejection boundary. By selecting the most similar incorrect candidate, the model is forced to learn fine-grained distinctions between establishments that share the same firm, location, or sector.

These pairs are produced using a bi-encoder that scores candidate similarity with a lightweight ensemble of three pre-trained models. Each establishment is encoded independently into a fixed-size embedding vector of 2,816 dimensions, and similarity is measured using cosine similarity between the vectors. The candidate records are those that belong to the same NIPC but have a different establishment ID and fall within the interval from year $t-1$ to year $t-5$.

A bi-encoder encodes each record independently into a dense embedding vector. Matching is then performed by computing cosine similarity between the two vectors. The three transformer models are available on Hugging Face (Hugging Face, n.d.), and they were pre-trained on diverse tasks, which makes them suitable for fine-tuning on specific problems. The Portuguese-language models are:

1. **BERT-large** (Souza, Nogueira, & Lotufo, 2020), embedding size of 1024. The weight is 0.45.

2. **E5-large** (L. Wang et al., 2024), embedding size of 1024. The weight is 0.35.

3. **MPNet** (Reimers & Gurevych, 2019b), embedding size of 768. The weight is 0.20.

The three models were used to leverage complementary semantic representations and improve robustness. The ensemble reduces dependence on a single embedding space and helps capture subtle differences in noisy or ambiguous records.

BERT is loaded as an AutoModel, whereas E5 and MPNet are loaded as sentence-transformer models. BERT alone only converts words into tokens and therefore does not directly produce sentence embeddings. For this reason, the input sentences are first tokenized with padding and pooling. Padding ensures that all sequences in a batch have the same length by adding [PAD] tokens, while pooling averages the token embeddings across the sequence dimension, producing a single representative vector for each sentence. After that, L2 normalization is applied to the BERT embeddings to scale the output vectors to unit Euclidean length. This places the BERT embeddings on the same hypersphere as the sentence-transformer outputs when they are concatenated; otherwise, the BERT embeddings would contribute disproportionately to the final dot product, even when they are less semantically informative.

By contrast, sentence-transformer models perform tokenization and embedding generation automatically, so the resulting sentence embeddings can be used directly.

As a result, similar records have vectors pointing in similar directions, while dissimilar records have vectors pointing in different directions. To create a positive pair, the most recent candidate belonging to the same firm and having the same establishment ID is selected. Negative pairs are created based on cosine similarity scores: among the candidates from the same firm but with a different establishment ID, the record with the highest score is selected as the most similar incorrect match. A fictitious example of these training pairs is shown in Figure 4.1.

Table 4.1: Pairs of Positive and Highly Negative Matches generated by the Bi-Encoder (illustrative).

NIPC		Record	Label	Similarity
A	Record in year T	Street: rua das virtudes, lote 2, primeiro esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1234 | ... | Status: Active	1	0.998
	Candidate Record in historical	Street: rua des virtudes, lote 2, 1 esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1234 | ... | Status: Active		
A	Record in year T	Street: rua das virtudes, lote 2, primeiro esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1234 | ... | Status: Active	0	0.923
	Candidate Record in historical	Street: rua das virtudes, lote 5, 3 esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1232 | ... | Status: Active		

4.3.1.2 Cross-Encoder Fine-Tuning

To fine-tune the cross-encoder as a binary classifier capable of distinguishing between matching and non-matching records in a sentence similarity task, the negative pairs generated by the bi-encoder are combined with the positive pairs formed from the ground truth.

The cross-encoder receives both sentences simultaneously as input to the Transformer network. In other words, both establishments, one from year t and one from the historical record, are fed jointly into the model as a sentence pair, allowing the network to model the interaction between them during the encoding process and to learn their relationship directly. For this reason, it does not compute sentence embeddings independently.

In this setup, both records are tokenized as a single sequence (fictitious example): *[CLS] Street: rua das virtudes, lote 2, primeiro esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1234 | ... | Status: Active [SEP] Street: rua des virtudes, lote 2, 1 esquerdo | Locality: São Bento | Municipality: Porto | CAE: 1234 | ... | Status: Active [SEP]*

This representation allows direct comparison between the two records, enabling the Transformer self-attention mechanism to operate across both inputs simultaneously. As a result, the model can detect subtle differences, such as a change in door number or a variation in postal code, which the bi-encoder may fail to capture because it encodes each record independently.

The model used in this stage was the BERT-large model mentioned in the previous section. To fine-tune it, the ground truth data was split into a training set (70%), a validation set (15%), and a test set (15%). This split was performed at the firm level, meaning that each NIPC appears in only one of the three sets. The training set contains 20,309 firms, while both the validation and test sets contain 4,352 firms each. In total, there are 923,176 available pairs. This strategy helps prevent overfitting, since the model has no opportunity to memorize firm-specific patterns in the training set and reproduce them in validation or test data. The training set is used to learn the model parameters, the validation set is used to select the probability threshold and the best checkpoint during training, and the test set is used to

evaluate performance using the F1-score.

The BERT-large model was trained using the hyperparameters shown in Figure 4.2.

Table 4.2: Parameters used in the fine-tune of the cross-encoder.

Hyperparameter	**Value**
Epochs	3
Batch Size	16
Learning rate	$2e^{-5}$
Warmup ratio	0.1
Gradient Clipping	1.0
Weight decay	0.01
Eval steps	2000
Compute Metrics	F1
Save total limit	2

The loss function is the categorical cross-entropy. For each pair, the model outputs two logits, one per class, which are converted to probabilities via the softmax function. The loss for a single example is then defined as

$$\mathcal{L} = -\log(p_y),$$

where p_y is the probability assigned to the correct class y (as given by the ground truth label). The total loss for a batch is the average of this penalty across all examples in the batch.

The model is evaluated on the validation set every 2,000 steps. At the end of training, the checkpoint that achieved the highest validation F1 is restored as the final model, rather than simply using the weights from the last training step.

At every evaluation step, the model produces predictions on the validation set. Logits are converted to probabilities via softmax and thresholded at 0.5 to obtain binary predictions, from which accuracy, precision, recall, and F1 are computed via the resulting confusion matrix. This 0.5 threshold is used only for monitoring training progress; the threshold used in production is selected separately, by sweeping a range of thresholds over the validation predictions and choosing the value that maximizes F1.

The cross-entropy loss quantifies how far the model's predictions are from the ground truth. During training, backpropagation computes the gradient of this loss with respect to every model parameter, and the weights are updated accordingly. This process is repeated for every batch of pairs, across epochs full passes through the training set, gradually reducing the loss as training progresses.

4.3.1.3 Threshold Optimal Selection

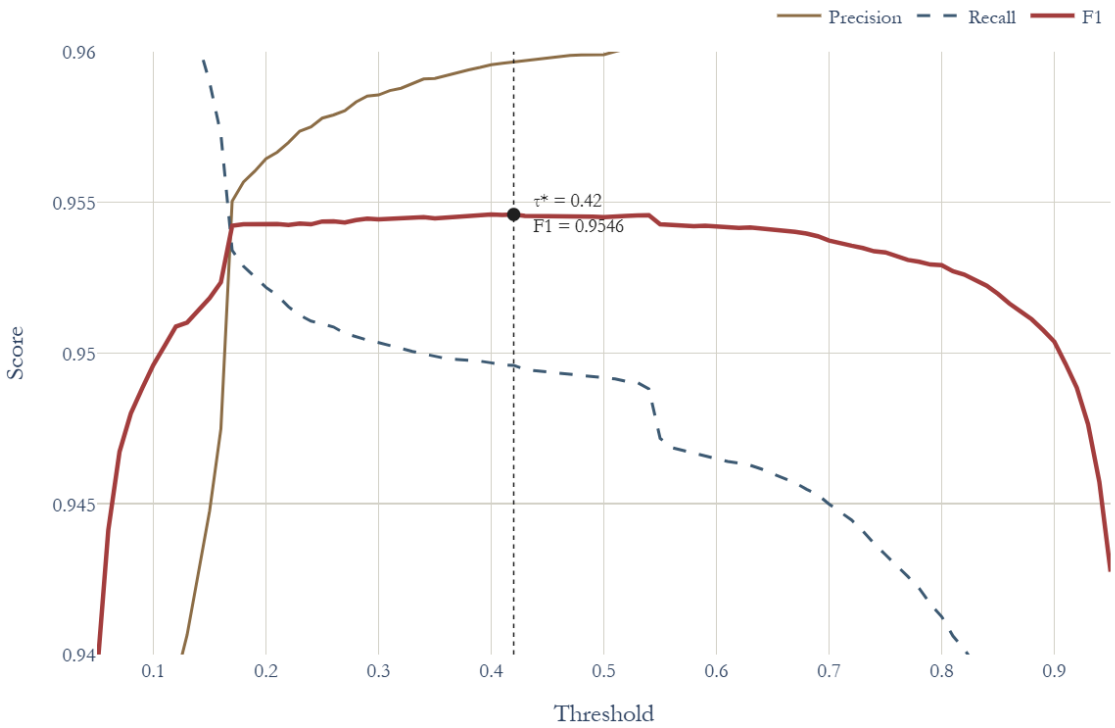
The fine-tuned model does not directly assign the labels match or non-match to each pair. Instead, it produces a probability score, and the objective of this section is to determine the

optimal threshold that separates matches from non-matches.

To select the optimal threshold without biasing the final performance, threshold selection and final evaluation are performed on two disjoint sets. First, a fine grid of candidate thresholds, starting at 0.05 and increasing by 0.01 up to 0.95, is evaluated on the *validation* set. The threshold that maximizes the chosen objective, F1, is then retained. Second, this fixed threshold is applied to the *test* set.

Figure 4.2 shows the precision-recall trade-off across the threshold grid for the BERT model on the validation set. Recall decreases monotonically as the threshold increases, from 0.98 at threshold 0.05 to 0.91 at threshold 0.95, while precision increases monotonically in the opposite direction, from 0.91 to 0.98 over the same range. The optimal threshold corresponds to the value that maximizes F1, which is 0.96 at a threshold of 0.42. At this level, precision is 0.96 and recall is 0.95. The threshold sweep also reveals a flat F1 plateau for threshold values in the interval $[0.38, 0.54]$, where F1 varies by less than 0.0002. This indicates that the model produces well separated probability scores with very few uncertain cases near the decision boundary.

Figure 4.2: Precision, recall and F1 vs. decision threshold (BERT).



The results on the *test* set, composed of 4,352 firms (corresponding to 148,971 establishments), are highly satisfactory and indicate that the selected threshold is appropriate for this dataset. The model attains an accuracy of 94.26%, an F1-score of 95.46%, a precision of 93.49%, and a recall of 95.29%.

4.3.1.4 Establishment ID Assignment

To finalize the process, a pair of records is considered a match if its probability is greater than or equal to 0.42; otherwise, it is considered a non-match.

For pairs classified as matches, the establishment receives the same identifier as the candidate record. However, to prevent multiple current year records from being assigned the same historical candidate identifier, the same algorithm used in Step 2 of the empirical pipeline (4.2.3.2), the Hungarian algorithm, is applied.

For pairs classified as non-matches, the records are treated as new establishments, and therefore the identifier assigned is the next available number after the highest one already used.

The methodological choices underlying each step of the empirical pipeline are empirically justified in the following chapter, which presents the experimental results and compares the performance of the two approaches against the ground truth.

Chapter 5

Results

This chapter explains the methodological choices underlying the final record linkage pipeline, the naive experiments that contributed to refining it, and evaluates its performance by comparing its results with the ground truth. The ground truth consists of the establishments annotated by BPLIM for the 2007 to 2016 period. It then extends the analysis to the unsupervised dataset, assessing how the pipeline performs in the later period.

Section 5.1 introduces the confusion matrix framework and the evaluation metrics used throughout the chapter. Section 5.2 presents the baseline naive heuristic experiment, in which thresholds were set *ad hoc*, establishing a reference point for subsequent improvements. Section 5.3 investigates the impact of different similarity metric choices on matching performance. Section 5.4 describes the Bayesian threshold optimization procedure and justifies the final threshold configuration adopted. Section 5.5 compares the performance of both pipelines against the reported establishment identifier, demonstrating the advantage of the proposed approach. Section 5.6 provides a comprehensive evaluation against the ground truth, analyzing establishment entry, exit, survival, re-entry, and gap distributions. Finally, Section 5.7 extends the analysis to the unsupervised period, assessing the pipeline's behaviour on unlabelled data and examining the full longitudinal dynamics from 2007 to 2023.

5.1 Confusion Matrix Interpretation

A confusion matrix is a two-dimensional table used to evaluate the performance of the model on a classification problem. It compares the actual values with the predicted values (Yang & Berdine, 2024). In this study, it is used to assess whether an establishment corresponds to an already exiting establishment or represents a new one, allowing a comparison between the ground truth and the model prediction.

To evaluate model performance, a confusion matrix is used and the metrics presented in Section 2.3.4 will be calculated based on its results.

The confusion matrix includes four possible outcomes:

- True Positive: the model predicts that the record matches an exiting establishment, and

the ground truth confirms that match.

- False Negative: the model predicts that the record is a new establishment, but the ground truth indicates that it matches an exiting establishment.

- False Positive: the model predicts that the record matches an exiting establishment, but the ground truth indicates that it is a new establishment.

- True Negative: the model predicts that the record is a new establishment, and the ground truth confirms it.

5.2 Heuristic initialization approach

The development of the empirical pipeline presented in Section 4.2 was preceded by an initial set of experiments, in which several assumptions were introduced to evaluate the quality of the approach. The goal was to examine how the algorithm behaved across the three steps, assess the quality of the matching, and test the similarity functions implemented in a preliminary experiment. For this purpose, the similarity thresholds were set *ad hoc*.

In this way, a baseline experiment was carried out. The main differences between this preliminary approach and the final algorithm were the following:

- In the first similarity function, only one similarity metric was used for all variables, namely Ratcliff-Obershelp similarity. The initial thresholds were set to 0.9 (upper threshold) to consider a match and 0.6 (lower threshold) to consider a non-match. Values between 0.6 and 0.9 were treated as partial matches.

- In the second similarity function, the selected similarity features included *partial_address* as a joint word representation of *street_address* and *locality*, while *parish* was not included.

- In the second similarity function, the similarity of *street_address* was calculated using the Jaccard similarity coefficient, while the remaining variables were evaluated using Ratcliff-Obershelp similarity. It also used a weighted average, but the initial weights were, 0.25, 0.30, 0.10, 0.20, and 0.15 (*street_address*, *postal_code4*, *new_postal_code*, *municipality*, and *cae*, respectively). The upper threshold was set to 0.85 and the lower threshold to 0.6. For cases classified as partial matches and with a reported establishment ID identified as a new ID, a threshold of 0.7 was used to determine a non-match.

- In the third similarity function, the similarity of *street_address* was computed using cosine similarity, while the remaining variables were again evaluated using Ratcliff-Obershelp similarity. The threshold used to consider a match was 0.86, while values below this threshold were considered non-matches.

- Finally, this approach always considers the economic activity, *cae*, including in the year of classification change.

To evaluate this initial *ad hoc* experiment several measures were considered: the number of partial matches remaining after the first two steps, the confusion matrix, and the corresponding metrics, including recall, precision, F1-score, and accuracy. Accuracy was calculated separately for each year.

Table 5.1: Number of new establishments, partial matches remaining and accuracy obtained for the *ad hoc* experiment.

Year	New Establishment	Step 1 partial	Step 2 partial	Accuracy (%)
2007	51,001	0	0	100.00
2008	11,849	13,110	2	88.70
2009	6,893	6,217	1,919	88.70
2010	6,333	7,536	2,413	88.04
2011	5,670	4,819	1,581	88.16
2012	5,102	5,579	1,650	87.95
2013	4,659	6,481	1,404	87.53
2014	4,401	5,129	1,439	87.18
2015	4,666	4,660	1,318	87.28
2016	4,511	4,851	1,323	87.33

Table 5.2: Confusion matrix and evaluation metrics on *ad hoc* experiment.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	64,535	11,252
	Negative	5,788	24,459

Metric	Value
Precision	91.77%
Recall	85.15%
F1-score	88.34%

Before analysing these results, it is important to note three structural breaks in the data. In 2008, the feature *cae* underwent a classification change; in 2010, Annex R form was reformulated; and in 2013, Portugal implemented a territorial reorganisation of the *parish* boundaries. As a consequence, similarity scores in these transition years may be slightly lower, since two records referring to the same establishment may differ in one or more affected features due to administrative reclassification rather than an actual change in the establishment.

Given that the threshold used in this experiment is 0.9, any record for which at least one of the six features differs substantially may fall below this threshold and be classified as a partial match for further resolution.

As shown in Table 5.1, the years 2008, 2010, and 2013 present the highest number of partial matches in Step 1, which is consistent with the structural changes described above.

The per-year accuracy values are relatively stable, averaging approximately 87.87%. However, Table 5.2 shows that the number of true negatives is considerably high, indicating that the

algorithm tends to assign new identifiers more often than necessary, even in cases where the same identifier should be preserved. This behaviour reflects the conservative nature of the high threshold.

The false positive rate is relatively low, representing approximately 5.46% of all classified records and roughly half the number of false negatives. In this context, this is desirable, since incorrectly assigning the same identifier to two different establishments is more problematic than temporarily classifying a true match as uncertain, as it can lead to false merges and propagate errors throughout the longitudinal panel.

According to the results in Table 5.2, precision reaches a high value, confirming that when the model assigns a match, it does so reliably. Recall, however, is lower, at 85.15%, which reflects the conservative threshold: a non-negligible number of true continuations are not captured in Step 1 and are instead forwarded to the partial match stage.

The F1-score, which combines precision and recall into a single metric, is approximately 88.37%. This value reflects the trade-off between the two measures: precision is favoured by the strict threshold, while recall is reduced by it.

5.3 Impact of Similarity Measures Choices on the Matching Performance

This section aims to evaluate how the model behaves when different similarity metrics are used in the similarity functions. To this end, three experiments were conducted using the years 2008, 2009, and 2010, with the analysis focused mainly on the number of partial matches and the accuracy.

To assess the effect of changing the similarity metrics, three experiments were performed, with the results depicted in Table 5.3. These experiments followed the same code structure described in the previous section, with the only differences being the similarity metrics and thresholds used in Step 1.

Table 5.3: Number of partial matches and accuracy across experiments, for years 2008 to 2010.

	Experiment 1			Experiment 2			Experiment 3		
Year	Step 1 partial	Step 2 partial	Accuracy (%)	Step 1 partial	Step 2 partial	Accuracy (%)	Step 1 partial	Step 2 partial	Accuracy (%)
2008	19,627	4	88.50	8,941	2	89.94	6,195	2	91.86
2009	6,251	1,861	88.50	3,901	1,460	89.71	3,752	1,342	91.23
2010	7,688	2,365	87.88	4,966	2,006	88.86	4,530	1,887	90.16

In Experiment 1, the similarity of *street_address_new* was calculated using token sort ratio, *locality* was compared using Jaro-Winkler, and the remaining variables were evaluated using a ratio based metric with a 20.00% penalty when an exact match was not obtained. The results indicate that applying a penalty to the categorical variables was less effective, since the number of partial matches in Step 1 increased and the accuracy became slightly lower. However, the number of partial matches remaining after Step 2 decreased when compared with the initial

approach in Table 5.1.

Experiment 2 used the same similarity metrics as Experiment 1, but the upper threshold in Step 1 was reduced to 0.85. As a result, the number of partial matches in Step 1 decreased on average by nearly 34.00% relative to the Table 5.1, while accuracy increased on average by approximately 1.00%.

Finally, Experiment 3 changed the similarity metrics but retained a restrictive upper threshold of 0.9, equal to the initial value. In this experiment, *street_address_new* was compared using Levenshtein distance, *locality* remained based on Jaro-Winkler, and the other categorical variables were evaluated using Ratcliff-Obershelp similarity. This configuration produced the lowest number of partial matches in Step 1 (reducing this number around 44.00% compared to Table 5.1) and increased the accuracy by an average of 2.60% compared with *ad-hoc* results.

These results suggest two main conclusions. First, applying a hard penalty to categorical variables is not ideal; instead, it is preferable to use similarity metrics that tolerate minor discrepancies while remaining sensitive to token order and string length. This is particularly relevant for the *parish* variable, which is encoded as a six-digit integer in which the first two digits identify the district, the third and fourth identify the municipality, and the last two identify the parish itself. Since the district and municipality fields in Annex R are implemented as cascading dependent selectors, any recording error is likely to affect only the last two digits. In such cases, a similarity metric such as Ratcliff-Obershelp similarity is more appropriate, because it rewards partial agreement, whereas a binary penalty would ignore this information entirely. Second, using a very conservative threshold in Step 1 does not appear to be optimal, as the initial approach already showed.

5.4 Bayesian Threshold Optimization

The fixed thresholds used in the previous sections adopt a conservative strategy. To improve the flexibility of the record linkage pipeline, an optimization procedure was implemented to determine the best combination of thresholds for Steps 1, 2, and 3. These threshold values were treated as hyperparameters and optimized on the ground truth years using a weighted objective function.

The optimization process was based on the previous heuristic approach, with minor modifications in the similarity functions used at each step. In Step 1, the similarity metric for the variables *street_address* and *locality* was changed to token sort ratio, implemented with RapidFuzz, and Jaro-Winkler, respectively. In Step 2, the weighted average similarity was computed using the variables *partial_address*, *postal_cod4*, *postal_code_new*, *municipality_id*, and *cae*, with the same weights defined in the *ad-hoc* experiment. In Step 3, the only modification was the use of cosine similarity to compute *street_address_new* similarity.

The optimization was done used Optuna, an open-source hyperparameter optimization framework. Optuna searches over the defined threshold space, evaluates each trial according to a predefined objective function, and identifies the configuration that yields the best performance. The optimization is guided by a Tree-structured Parzen Estimator (TPE) sampler, which uses information from previous trials to propose more promising parameter combi-

nations. In addition, unpromising trials can be pruned to reduce computation time (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

Optuna was used to calibrate the five threshold parameters of the three-step record linkage pipeline: the upper and lower thresholds of Steps 1 and 2, and the threshold of Step 3. The objective was to balance two competing goals: maximizing yearly accuracy and minimizing the number of partial matches per year. The objective function was defined as: $max(0.7 * accuracy - 0.3 * (\frac{partial\_matches}{total\_establishments})$ For each year, the search was conducted over 30 trials. The weights were set *ad-hoc* to achieve a satisfactory balance between performance and flexibility, thereby making the algorithm less restrictive.

5.4.1 Experiment Regarding the Year 2008

From the first experiment, the thresholds obtained through the optimization procedure for the year 2008 are presented in Table 5.4.

Table 5.4: Threshold resulting from the hyperparameterization, regarding the year 2008.

Year	Lower Step 1	Upper Step 1	Lower Step 2	Upper Step 2	Threshold Step 3
2008	0.7413	0.8035	0.6584	0.7519	0.7719

Table 5.5: Number of partial matches and accuracy obtained, with thresholds optimized for 2008.

Year	Step 1 partial	Step 2 partial	Accuracy (%)
2007	0	0	100.00
2008	3,133	8	96.51
2009	2,092	180	95.60
2010	2,522	369	94.78
2011	1,782	301	94.42
2012	2,325	385	93.96
2013	2,503	310	93.38
2014	1,991	258	92.81
2015	1,493	227	92.68
2016	1,553	219	92.52

Comparing Tables 5.1 and 5.5, it is evident that the number of partial matches produced in Step 1 and Step 2 decreased on average by $(34.6 \pm 3.4)\%$ and $(17.6 \pm 4.0)\%$, respectively. In Step 2 the only exception was the year 2008, for which the number of partial matches increased by 75.00%, because increase from 2 to 8 (however, this increase is not substantial in absolute terms). Overall, the hyperparameter optimization was effective and led to a significant improvement in model performance, better accuracy and a reduce number of partial

Table 5.6: Confusion matrix and evaluation metrics, for threshold optimized for 2008.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	70,184	5,523
	Negative	5,900	24,426

Metric	**Value**
Precision	92.25%
Recall	92.71%
F1-score	89.23%

matches), this suggest that the thresholds previously defined a priori were too conservative and did not fit the data as well.

According to the confusion matrix in Table 5.6, the number of false negatives decreased substantially, while the number of true positives increased considerably. Although the number of false positives increased by 1.90%, this small rise is acceptable given that the remaining metrics improved overall. These results indicate that the model fits the data well.

5.4.2 Experiment Regarding a Combination of Years

The experiment regarding the combination of years was conducted under the same conditions as the previous experiments. It is important to note that the combination of years does not mean that data from different years were merged into a single dataset. Instead, each year was processed independently and in isolation, as shown in Table 5.7. To obtain the final threshold values, the results from each year were combined and summarized using the average and the median, as presented in Table 5.8.

Table 5.7: Threshold resulting from the hyperparameterization, regarding the years of 2008, 2010, 2012, 2013, 2014, and 2016.

Year	**Lower Step 1**	**Upper1 Step 1**	**Lower2 Step 2**	**Upper Step 2**	**Threshold Step 3**
2008	0.682237	0.824791	0.696306	0.862179	0.854152
2010	0.605789	0.857380	0.626010	0.881226	0.873196
2012	0.630075	0.895978	0.610163	0.889357	0.889026
2013	0.665165	0.918491	0.605392	0.827004	0.808367
2014	0.616116	0.868635	0.614009	0.860230	0.824104
2016	0.662095	0.903801	0.677212	0.790132	0.815887

The lower threshold of Step 1 (Table 5.7) exhibits greater variability, particularly in the transition years associated with changes in economic activity and parish classifications. The highest values are observed in more recent years (e.g., 2013 and 2016), suggesting that the model requires a higher minimum confidence level to classify observations, which may indicate improvements in data quality and reporting consistency over time. The upper threshold of Step

1 generally shows an increasing trend from 2008 to 2013, reaching its maximum value in 2013 (0.918), although some fluctuations occur afterwards. This behaviour suggests that the model becomes more selective when identifying highly reliable observations. In contrast, the lower threshold of Step 2 remains relatively stable throughout the study period, with only minor deviations. The lowest values occur in the middle years, while the first (2008) and last (2016) years display slightly higher thresholds. The upper threshold of Step 2 and the final threshold of Step 3 do not show a clear increasing or decreasing pattern, but they remain within a relatively narrow range, indicating a stable calibration process across years.

Overall, the distance between the lower and upper thresholds tends to decrease in the more recent years. This reduction suggests that the model's uncertainty region becomes narrower, meaning that classifications can be made with greater confidence and consistency. In other words, the model appears to become more robust over time, likely benefiting from improvements in data quality, reporting practices, and the standardization of administrative classifications.

Table 5.8: Threshold resulting from the combination of years (2008, 2010, 2012, 2013, 2014 and 2016).

-	**Average Thresholds**	**Median Thresholds**
Lower Step 1	0.6435795	0.6460850
Upper Step 1	0.8781793	0.8823065
Lower Step 2	0.6376413	0.6200095
Upper Step 2	0.8516880	0.8612045
Upper Step 3	0.8441223	0.8391280

The results respecting the number of partial matches, confusion matrix and evaluation metrics from the average threshold approach are in Appendix A.3, A.3, and A.4, respectively. And the results from the computation of the median are in Appendix A.5, A.6, and A.7, respectively.

The average consistently outperforms the median across all metrics and years, and the gap widens slightly in later years. Given that the threshold distributions across the six years are not strongly skewed (the values in Table 5.8 are fairly symmetric), this is somewhat counterintuitive, because the median should be more robust to outliers. However, this was not verified because the upper threshold in Step 1 of 2013 (around 0.9185) is a mild outlier that pulls the average up in a way that happens to fit the data better.

Comparing the results across the three experiments (2008 optimized thresholds, average combined thresholds, and median combined thresholds), in all the experiments the accuracy declines monotonically from 2008 to 2016, indicating that the administrative data becomes harder to link over time.

However, there is a significant difference in the number of partial matches resulting from the Step 2. Because in the experiments using the average and the median thresholds those values increase drastically (having a maximum of 2,024 and 2,243, respectively) compared to a maximum of 385 in the experiment of 2008. This also happens in the partial resulting from the Step 1. Suggesting that combined thresholds (median or average) are more permissive in

the ambiguous cases.

The 2008 optimized model (Table 5.6) also has a more balanced error profile (having 5,523 cases classified as False Negatives and 5,900 cases classified as False Positives), while both combined threshold variants shift toward more false negatives (8,456 and 8,754 FN) and slightly fewer false positives (5,865 and 5,845). So, the combined thresholds approaches are more conservative, because they miss more true links, have almost the same cases of false negatives that can distort establishment continuity estimates downward. For these reasons, the thresholds of the 2008 optimized model will be used in the final algorithm (Table 5.4).

5.5 Record Linkage Empirical and Semantic Pipeline Identifier vs Reported Establishment Identifier

Before presenting the main results, several alternative experiments were conducted on the empirical pipeline. These included using a weighted mean in the similarity function of Step 1, where each variable was weighted according to the minimum number of characters shared between the two records; removing the *parish* feature in the transition year 2013, which led to slightly worse results, with an increase of 1.65% in false negatives and 0.1% in false positives; and modifying the weights in the similarity function of Step 2, including treating *postal_code_new* as an independent variable. However, these alternatives consistently produced worse performance than the final configuration. So the similarity score implemented in the final version of the Step 4.2.3.2 consider the following variables: *partial_address*, *postal_code4*, *municipality_id*, *parish*, and *cae*, where the weights associated are 0.25, 0.20, 0.20, 0.15, and 0.20, respectively. In 2008, when the classification of the main economic activity changes, *cae* is not used and the weights are: 0.30, 0.25, 0.25, and 0.20, respectively.

Annex R has the variable *id_estab* (see Table 5.9), which is the institutional variable that reports the establishment ID, whose main goal is to exclusively identify the same establishment of the same firm yearly. However, Table 5.9 highlights the limitations of relying solely on the reported establishment identifier, because its accuracy decreases substantially over time, falling from 84.4% in 2008 to 73.5% in 2016, a decline of nearly 11 Percentage Points (pp) over the study period. This suggests that the original identifier does not allow establishments to be consistently tracked over time.

In contrast, the developed empirical algorithm consistently maintains higher accuracy throughout the period, ranging from 97.64% in 2008 to 93.09% in 2016. Although some decline is also observed, likely due to increasing data complexity and changes in administrative classifications, the gap relative to the reported establishment identifier widens over time and reaches approximately 19.6 pp in 2016. First, as time passes and reports are submitted further from the reference period, there is a greater likelihood that the report is completed by a different person, which may lead to inconsistencies in how establishment names and attributes are recorded, the same establishment may be described using a different designation across years, making identifier based tracking less reliable. Second, as the observation window extends, the natural turnover of establishments increases: more establishments close and new ones begin activity, reducing the persistence of the panel over time and making longitudinal

matching inherently more challenging. This comparison underscores the value of a dedicated record linkage pipeline, since the institutional identifier alone is not sufficiently reliable for longitudinal analysis and could introduce systematic bias into downstream results.

Comparing the results from the empirical pipeline and the semantic pipeline, in terms of accuracy both models have similar results. The semantic pipeline underperforms the empirical pipeline by around 0.17 pp in terms of accuracy.

Note that the semantic pipeline did not receive the same optimization effort and refinement as the empirical pipeline. Additionally, the semantic pipeline generates 1,886 establishments (373 firms) with duplicate IDs, meaning that, within the same year, more than one establishment shares the same combination of *nipc*, *establishment_id*, and *ano*. The empirical pipeline, by contrast, does not exhibit any repeated IDs for the same firm.

Table 5.9: Performance comparison between the accuracy of the reported establishment identifier vs the developed algorithm, based on the ground truth for multi establishments.

		Reported ID			Empirical ID			Semantic ID		
Year	**Total GT**	**Total**	**Correct**	**Acc. (%)**	**Total**	**Correct**	**Acc. (%)**	**Total**	**Correct**	**Acc. (%)**
2007	51,526	51,372	51,372	100.00	51,444	51,444	100.00	51,444	51,444	100.00
2008	50,132	50,128	42,366	84.52	50,057	48,875	97.64	50,061	48,545	96.97
2009	49,560	49,552	40,542	81.82	49,467	47,794	96.62	49,459	47,413	95.86
2010	46,241	46,193	35,472	76.79	46,187	44,218	95.74	46,181	43,761	94.76
2011	47,411	47,402	36,393	76.78	47,341	45,091	95.25	47,335	44,675	94.38
2012	45,824	45,824	34,423	75.12	45,752	43,318	94.68	45,752	42,888	93.74
2013	44,043	44,043	32,724	74.30	43,973	41,395	94.14	43,969	40,907	93.04
2014	42,898	42,898	32,028	74.66	42,824	40,058	93.54	42,822	39,531	92.32
2015	42,328	42,328	31,488	74.39	42,264	39,437	93.31	42,263	38,921	92.09
2016	42,407	42,407	31,173	73.51	42,338	39,412	93.09	42,338	38,896	91.87
Total	462,370	462,147	367,981	–	461,647	441,042	–	461,624	436,981	–

Note: The total number of reported identifiers is lower than the ground truth total, due to missing values in the reported identifier field.

Table 5.10 provides a direct comparison between the performance of the reported identifier and the algorithm based identifier (empirical and semantic) in classification terms. The most notable difference concerns the false negative counts: the reported establishment identifier fails to recover 23,988 true establishment links, compared with only 4,777 and 3,300 for the empirical and semantic algorithms, a reduction of approximately 80.00%. In a longitudinal dataset, missed links translate into artificial establishment entry and exit, making this level of error highly consequential for analyses of establishment survival.

The metric comparison reinforces this conclusion. While both approaches achieve similar precision (90.60% versus 92.38% and 92.07%), the reported establishment identifier shows substantially lower recall (68.56% versus 93.69% and 91.96%), indicating that it frequently updates the identifier rather than preserving continuity across records.

Both algorithms achieve a more balanced trade-off between precision and recall compared to the reported establishment identifier, suggesting that neither substantially over-links nor under-links establishments. This balance is especially important in record linkage, because

false positives incorrectly merge distinct establishments, while false negatives disrupt true longitudinal continuity. Nevertheless, the semantic pipeline struggles to detect repeated establishments but performs reasonably well in identifying new ones, resulting in a higher number of false negatives, which is only partially offset by a slightly higher number of false positives than in the empirical pipeline.

Table 5.10: Confusion matrix and evaluation metrics for the reported identifier and algorithm based identifiers.

		Reported Identifier		Empirical Reported		Semantic Reported	
		Pred. Positive	**Pred. Negative**	**Pred. Positive**	**Pred. Negative**	**Pred. Positive**	**Pred. Negative**
Actual	**Positive**	52,318	23,988	70,972	4,777	69,543	6,079
	Negative	5,429	24,262	5,855	24,431	5,992	24,422
Metrics (%)	**Precision**	90.60		92.38		92.07	
	Recall	68.56		93.69		91.96	
	F1-score	78.06		93.03		92.01	

Overall, the results suggest that the developed empirical pipeline provides a considerably more reliable basis for longitudinal establishment tracking than the administrative identifier originally intended for that purpose, and that it balances false positives and false negatives more effectively than the semantic pipeline.

5.6 Record Linkage Empirical and Semantic Pipeline Compared with Ground Truth

The supervised dataset (covering reports from 2007 to 2016) includes both single establishment and multi-establishment firms. A firm is classified as multi-establishment if in a given year it has at least two establishments, even if it has one establishment in some of the years. The following analysis focuses exclusively on multi-establishment firms. Under this restriction, the raw sample contains 462,679 establishments over the ten-year period.

In the ground truth data, 158 establishments were identified as duplicates and removed, resulting in a final sample of 462,370 establishments. In contrast, the record linkage empirical pipeline identified 896 duplicated establishments, yielding a final sample of 461,147 establishments, and record linkage semantic pipeline identified 861 duplicated establishments (this number differ because on the semantic pipeline it was not applied the second duplicates search not concerning the locality feature), yielding a final sample of 461,624 establishments. Appendix A.8 reports the number of establishments removed per year.

Table 5.11 presents the yearly distribution of new establishments, matched establishments (i.e., establishments with high similarity scores relative to those reported in previous years), and partial matches, which are resolved in subsequent steps of the pipeline. Excluding the first year, when all establishments are, by construction, classified as new, in the empirical pipeline approximately $(86.09 \pm 1.32)\%$ of the establishments are matched in the first step. Additionally, $(4.58 \pm 0.73)\%$ are classified as partial matches in Step 1, and $(0.39 \pm 0.12)\%$ in

Step 2. These results indicate that Step 1 successfully captures the vast majority of matches, with only a small fraction of establishments requiring further disambiguation.

By contrast, the semantic pipeline produces, on average, three fewer partial matches than the empirical pipeline. In 2011, however, this difference increases to 33 fewer partial matches. At the same time, the number of non-matches in the semantic pipeline increases by approximately 58.45% relative to the empirical pipeline, while the number of matches decreases by 32.89%. Overall, these results indicate that the semantic pipeline adopts a more conservative matching strategy, leading to fewer matches and more non-matches than the empirical pipeline.

Table 5.11: Number of new establishment and partial match remain obtained for the final pipelines.

	Empirical Pipeline									Semantic Pipeline					
	Step 1			Step 2			Step 3			Step 1			BERT		
Year	New establish.	Matches	Partial matches	New establish.	Matches	Partial matches	New establish.	Matches	Total	New establish.	Matches	Partial matches	New establish.	Matches	Total
2007	51,507	0	0	0	0	0	0	0	51,507	51,507	0	0	0	0	51,507
2008	6,228	41,486	2,386	829	1,547	6	3	3	50,096	6,228	41,486	2,386	1,315	1,071	50,100
2009	5,357	42,081	2,052	710	1,223	117	74	43	49,488	5,351	42,086	2053	1,169	874	49,490
2010	4,789	39,033	2,370	558	1,552	259	188	71	46,191	4,781	39,038	2373	1,329	1,037	46,192
2011	4,638	40,991	1,714	448	1,050	215	132	83	47,342	4,569	41,027	1747	1,045	695	47,343
2012	3,794	39,537	2,424	557	1,614	251	146	105	45,753	3,757	39,570	2,428	1,264	1,162	45,755
2013	3,305	37,865	2,807	625	1,937	244	153	91	43,976	3,293	37,869	2,815	1,277	1,533	43,977
2014	3,275	37,586	1,967	479	1,308	178	117	61	42,826	3,267	37,599	1,962	1,051	907	42,828
2015	3,740	37,039	1,487	393	962	132	72	60	42,266	3,709	37,066	1,491	849	641	42,266
2016	3,517	37,247	1,574	380	1,032	162	95	67	42,338	3,497	37,269	1,572	896	676	42,338
Total	90,150	352,865	18,781	4,979	12,225	1,564	980	584	461,783	38,452	404,517	18,827	10,195	8,596	461,796

Figure 5.1 presents the comparison of the number of opening and closing establishments by year, as well as the net change. The record linkage empirical pipeline shows the largest divergence from the ground truth in 2008 and 2009 in terms of newly identified establishments, overestimating them by 6.00% and 2.70%, respectively. In contrast, in 2012 the empirical pipeline underestimates the number of new establishments by 3.60%. In the remaining years, the empirical pipeline identifies a number of new entries that is very close to the ground truth.

However, the semantic pipeline systematically overestimates the number of both opening and closing establishments relative to the ground truth, by an average of 12.00% and 7.40%, respectively. For closing establishments, the comparison is limited to 2014, because from that point onward the semantic pipeline underestimates the number of closures, by about 0.50% in 2014 and 10.80% in 2015.

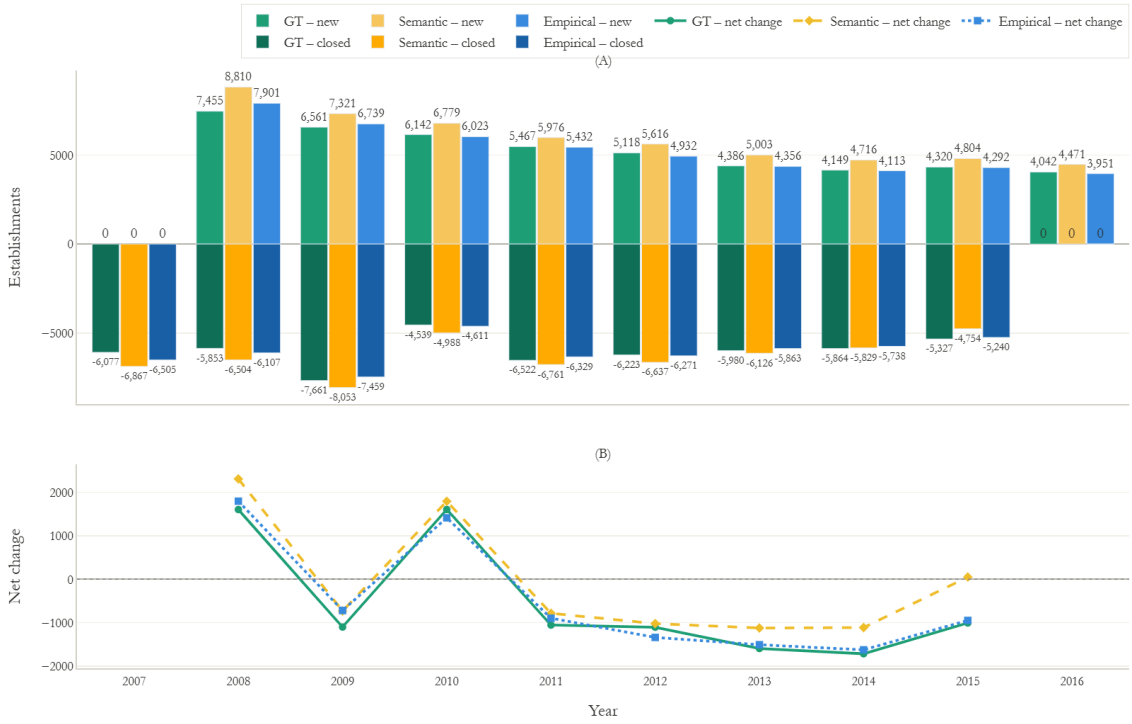
Regarding closed establishments, the largest discrepancies occur in 2007 and 2008, where the empirical pipeline reports higher closure rates, exceeding the ground truth by 7.00% and 4.30%, respectively. It is important to note that these differences are associated with the higher number of duplicated establishments dropped in those years, more than double compared to other years (see Appendix A.8). This discrepancy propagates over time, affecting subsequent yearly comparisons.

As shown in Figure 5.1 B, which depicts the net change, the empirical and ground truth series closely track each other, capturing the same peaks (2008 and 2010) as well as the sustained negative trend after 2010 until 2014. The largest divergence occurs in 2008, when the empirical pipeline slightly overestimates the ground truth. This is consistent with the over counting

of both new establishments (+446) and closed establishments (−254), as illustrated in Figure 5.1 A.

As expected, the net change estimated by the semantic pipeline consistently exceeds the ground truth, while still following the same overall pattern, with similar peaks and troughs across the years. As well, always exceed the empirical pipeline with the exception of 2009.

Figure 5.1: (A) New & closed establishments, (B) Net change, from 2007 to 2016 (ground truth (GT) vs Empirical Pipeline vs Semantic Pipeline).



To assess how well the pipeline tracks establishments over time, the difference in survival rates between the ground truth and the pipelines identifiers is computed. A positive difference indicates that the ground truth survival rate is higher than that of the pipeline, implying that the pipeline prematurely classifies some establishments as closed. Conversely, a negative difference indicates that the pipeline overestimates survival, continuing to track establishments that have, in reality, exited.

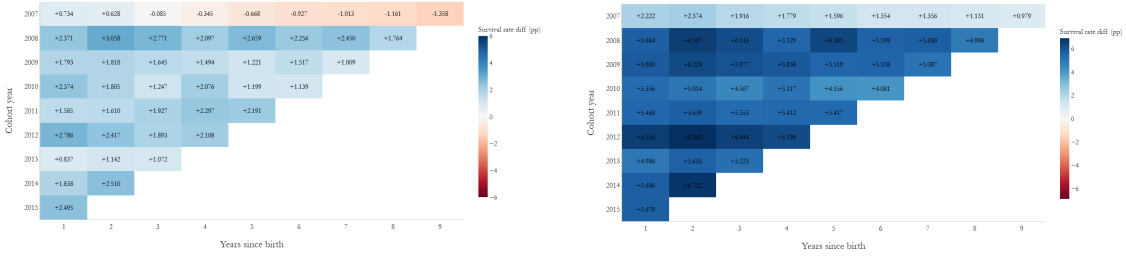
Figure 5.2 shows that the survival rate difference is predominantly positive across cohorts and years. This indicates that the pipeline systematically under counts surviving establishments, slightly over attributing exits. An exception is observed for the first cohort (2007) on the empirical pipeline, which exhibits a negative and increasing drift over time, suggesting that the pipeline overestimates survival for these establishments. One possible explanation is that changes in reporting practices over time reduce the effectiveness of the similarity functions, making it more difficult to correctly match establishments across years. The 2008 cohort, for the empirical pipeline, displays the largest positive deviations, with a peak difference of +3.06 pp in its second year (2010), and remains elevated throughout its observed lifespan. This pattern is likely related to propagation effects from the initial year: inaccuracies in 2007 may introduce noise into the baseline, which then affects the tracking of subsequent cohorts.

For cohorts from 2009 to 2015, the average survival rate difference is approximately $(1.75 \pm 0.52)\%$. This relatively stable and modest deviation suggests that empirical pipeline errors do not accumulate significantly over time across cohorts.

The semantic pipeline displays a markedly different pattern from its empirical counterpart. The survival rate difference is uniformly positive across all cohorts and years, and the negative drift observed for the 2007 cohort under the empirical pipeline is absent here. However, the magnitude of the deviation is substantially larger, typically ranging between $+5.00$ and $+6.90$ pp. This indicates that the semantic pipeline systematically under counts surviving establishments to a much greater extent, more frequently classifying establishments as closed when, in reality, they continued to operate.

Across cohorts, the deviation does not exhibit a clear accumulating trend within each cohort's lifetime; instead, it stabilizes around a persistently high level, generally between $+5.00$ and $+6.50$ pp. This suggests that the larger error associated with the BERT pipeline reflects a systematic bias, likely originating from the matching threshold or the partial match resolution step, rather than an error that compounds over time.

Figure 5.2: Survival rate difference (ground truth − pipeline) by cohort × years since birth.



(a) Empirical pipeline **(b)** Semantic pipeline

Table 5.3 presents the comparison between the ground truth and the pipelines across four dimensions: first entries, re-entries; exits (last appearance); and re-exits.

In panel (A), the empirical pipeline overestimates the number of new establishments, particularly in the first two years. This can be problematic because the pipeline cannot assign identifiers from prior years, it creates new ones. After this initial period, both series exhibit very similar dynamics and nearly converge, with a noticeable divergence again in 2012, when the pipeline over creates establishments relative to the ground truth.

In panel (B), which reports re-entries, the empirical and ground truth series are nearly identical throughout the period. The consistently small differences indicate that the pipeline is highly effective at identifying re-entry events, although it shows a slight tendency to underestimate them. This underestimation becomes more pronounced in the final years of the sample, which is consistent with the use of a five year historical window.

Panel (C) shows the last observed appearance of establishments. From 2009 onward, empirical and ground truth series follow a very similar pattern; however, the pipeline consistently underestimates the true values. In contrast, in 2007 and 2008 the pipeline substantially

[0]The survival rate is computed as the ratio of active establishments to the cohort size, calculated separately for the ground truth and the pipeline. The values reported correspond to the difference between the two rates.

overestimates last appearances (by 459 and 256 establishments, respectively), again reflecting initialization issues in the early years.

Finally, panel (D) presents re-exits, in which the largest discrepancies are observed between the empirical and the ground truth series. From 2010 onward, the pipeline systematically under counts re-exits. This suggests that it fails to detect some re-entry and re–exit cycles, likely because certain re-entries are misclassified as continuous spells. Since re-exit identification depends on correctly linking re-entry, subsequent survival, and eventual exit, any matching error at the re-entry stage propagates forward, leading to an underestimation of re-exits.

While the empirical pipeline's errors are concentrated at the 2007 and 2008 and track ground truth closely after that, the semantic pipeline shows a systematic, persistent inflation across entries, re-entries, exits, and re-exits throughout the whole 2007 to 2016 window. This is consistent with semantic pipeline being more prone to fragmenting true establishment trajectories, splitting a single continuous establishment history into multiple shorter extra entries and exits rather than correctly chaining them, which would inflate all four flow categories simultaneously rather than just the boundary years.

Figure 5.3: Comparison between (A) First entry (B) Re-entry (C) Final Exit (D) Re-exit establishments yearly.

To explore the unique identifier attributions between both pipelines and the ground truth, the Stata panelstat package (BPLIM, n.d) was used. This tool enables the comparison of two alternative identifier variables, evaluating whether they coincide across years or diverge, that is, whether the same physical establishment is assigned a consistent unique identifier over time, or whether a single establishment receives multiple distinct identifiers across different years

(fragmentation), and conversely, whether multiple distinct establishments are collapsed under the same identifier (merging). Both fragmentation and merging represent failure modes of the linkage pipeline and can introduce systematic bias into longitudinal analyses that rely on establishment level identifiers for tracking entry, exit, and continuity.

The comparison of ID attribution between the pipelines and the ground truth at the establishment level (Table 5.12) shows that in 81.62% of establishments the empirical pipeline assigns a single ID consistent with the ground truth, whereas the semantic pipeline achieves this in only 77.56% of cases.

The main source of error arises in cases where both approaches result in many-to-many ID assignments, which correspond to false positives. These errors are primarily driven by similarities in address related attributes. For example, establishments located on the same street but differing only in door or store numbers are often incorrectly linked. Additional discrepancies arise when address information is recorded differently across years (e.g., abbreviations versus extended forms not converging in the preprocessing step), or when administrative attributes such as parish or CAE codes are different.

Another important source of error is the five year backward search window used by the pipeline. If an establishment is inactive for more than five years and later reappears, the pipeline fails to link it to its historical record and instead assigns a new ID. This can also lead to ID collisions when identifiers previously used in the distant past are reassigned and change in the address happen. Furthermore, when multiple establishments share identical address and CAE characteristics, the pipeline may be unable to distinguish them.

There are cases where the ground truth assigns a new identifier due to minor address changes (e.g., a different door number), while the pipeline still classifies the record as a match because the similarity score exceeds the threshold. In addition, when a firm transitions from a single establishment structure to a multi-establishment one, the ground truth may update the identifier following address changes, whereas the pipeline, which by design is restricted to multi-establishment firms, does not revise previously assigned identifiers. Furthermore, since the pipeline only processes firm-years in which the firm operates as a multi-establishment entity, years in which the firm had a single establishment are entirely excluded from the linkage procedure. Consequently, any location changes that occurred during the single establishment period are not captured by the pipeline, which retains the previously assigned identifier regardless of any address changes during those unobserved years.

Importantly, when multiple ground truth IDs are mapped to a single pipeline ID, this indicates that the pipeline tends to merge distinct establishments rather than split a single one. This behavior is consistent with the undercounting of re-exits observed in Figure 5.3 (D). When two establishments are merged into a single entity, re-entry and re-exit cycles become unobservable because the interruption in activity is effectively removed. Overall, this suggests that the pipeline's linking strategy is conservative, favoring under-segmentation over over-segmentation of the establishment population.

Finally, Figure 5.4 compares the distribution of gaps (periods in which an establishment disappears from the panel and later reappears) between the ground truth and the two pipelines. Note that on the ground truth there is 93.33% of establishment with zero gaps, on the empirical pipeline 93.68 %, and in the semantic pipeline 92.20%.

Table 5.12: Comparison of attribution between the pipelines and the ground truth at the establishment level.

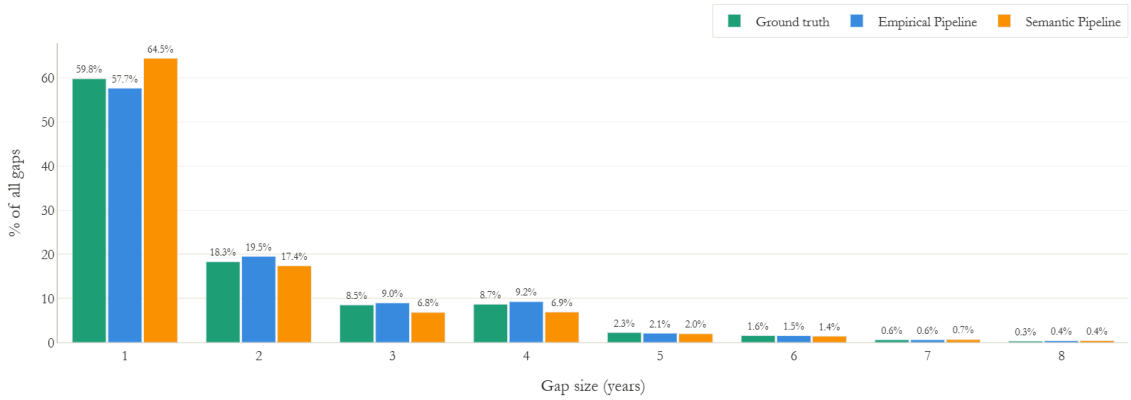
GT:Pipeline	Nº of Establishments (Empirical)	Percentage of Empirical (%)	Nº of Establishments (Semantic)	Percentage of Semantic (%)	Meaning
1:1	88,420	81.62	88,403	77.56	Perfect match, where one GT ID maps to one RL pipeline ID.
1:multiples	2,189	2.02	1,170	1.06	GT only have one ID, but the pipeline RL splits it into at least 2 ID's. (over segmentation)
multiples:1	4,980	4.60	10,772	9.45	GT has multiple establishments, but the pipeline merges into one (under segmentation)
multiples:multiples	12,626	11.66	13,638	11.96	Both have multiple values, that create ambiguity (False Positives)
Total	108,215	100.00	113,983	100.00	Number of unique establishments.

The empirical pipeline reproduces the distribution almost perfectly, suggesting that it captures the temporal structure of establishment interruptions with high accuracy. The distribution is strongly right-skewed: gaps of one year account for around 58.00% of all cases, and the share declines sharply up to four year gaps. Beyond that point, the frequency of gaps is close to zero, indicating that most re-entering establishments return relatively quickly.

The semantic pipeline exhibits a different pattern from the empirical pipeline when it comes to gap structure. While the overall distribution remains strongly right-skewed, as in the ground truth, the semantic pipeline over represents 1 year gaps (64.50% versus 59.80% in the ground truth) and correspondingly under represents gaps of 2 to 4 years, with differences of up to 2 pp. Gaps of 5 years or more are essentially unaffected, indicating that the distortion is concentrated in the short-gap range.

This behaviour aligns with the fragmentation pattern observed in the entry and exit flow analysis (Figure 5.3), where the semantic pipeline systematically overestimates (re-)entries (re-)exits. A plausible explanation is that the pipeline occasionally fails to match an establishment to its correct identifier in a single year despite its presence in adjacent years, generating a spurious one-year exit-reentry pair instead of a continuous spell. Unlike the empirical pipeline, whose deviations concentrate in the cold-start cohort, the semantic pipeline's bias is spread across the whole panel and larger in magnitude, pointing to a persistent tendency to under-link establishments across consecutive years.

Figure 5.4: Gap size distribution (%), comparison between the pipeline and the ground truth.)



5.7 Record Linkage Empirical and Semantic Pipeline - Unsupervised Dataset

This section aims to evaluate the behavior of the record linkage pipelines on unsupervised data (2017 to 2023), and subsequently to assess its performance over the full period from 2007 to 2023. As in the previous section, the analysis focuses exclusively on multi-establishment firms.

During the unsupervised period, a total of 552 and 494 establishments were identified as duplicates and removed (see Appendix A.4 for the yearly distribution) in empirical and semantic pipeline, respectivatly. Based on the data processed in the previous section, 21,300 firms were already active at the beginning of the unsupervised period for the empirical pipeline and 21,547 for the semantic. Over the subsequent seven years, the total number of firms increases to 33,081.

Table 5.13 shows that the empirical pipeline performs consistently across the 2017 to 2023 period. The majority of establishments are solved in Step 1, where high-confidence matches account for approximately $(84.92 \pm 7.01)\%$ of total processed units in most years. The number of new establishments identified in 2021 exhibits a marked increase, but this reflects the broader expansion of the sample that year rather than a deterioration in pipeline performance, the Step 1 match rate remains stable at around 70.00%, and the share of partial matches does not increase disproportionately. Steps 2 and 3 continue to solve a small residual share of cases, with Step 3 accounting for fewer than 150 establishments per year, confirming that the Step 2 solver operates on a narrow and well-contained gray zone. Comparing the two periods (Tables 5.11 and 5.13), the annual totals in the 2017 to 2023 sample are lower than in the previous period, with a declining trend through 2020 followed by a recovery from 2021 onward, stabilising at approximately 46,500 to 47,400 establishments per year by 2022 and 2023. The recovery observed from 2021 is consistent with aggregate national trends, according to Instituto Nacional de Estatística (2023), the number of new firm births in Portugal increased.

Regarding the semantic pipeline, BERT resolves a comparable total volume of gray zone cases to the empirical pipeline's Steps 2 and 3 combined (9,523 versus 9,480 establishments), but with a markedly different internal composition: BERT classifies a substantially higher share of these cases as new establishments rather than matches, particularly in 2021 (1,550 new establishments against 945 matches, compared to 1,131 new against 1,217 matches for the empirical Step 2 solver). This is consistent with the fragmentation behaviour identified in the entry and exit flow and gap distribution analyses, where it favours the creation of new identifiers over the confirmation of existing matches when classification confidence is not high. The multi-establishment counts reported in Table A.4 reinforce this pattern: while the share of establishments flagged as *Only multi-est.* is nearly identical across pipelines, the broader multi-establishment counts are substantially higher under the semantic pipeline.

Figure 5.5 displays the annual number of establishments classified by the pipelines as entries and exits, along with the implied net change. Notably, this figure also includes first time observations of establishments that initially appear as single establishment firms but later become part of multi-establishment firms.

Table 5.13: Number of new establishment, partial match remain.

	Empirical Pipeline									Semantic Pipeline					
	Step 1			Step 2			Step 3			Step 1			BERT		
Year	New establish.	Matches	Partial matches	New establish.	Matches	Partial matches	New establish.	Matches	Total	New establish.	Matches	Partial matches	New establish.	Matches	Total
2017	4,405	33,733	1,361	235	1,024	100	45	55	39,497	4,392	33,737	1,370	590	775	39,499
2018	3,245	35,075	1,069	261	685	122	49	73	39,388	3,241	35,076	1,072	556	515	39,389
2019	3,270	34,710	1,033	252	699	82	25	57	39,013	3,257	34,711	1,045	534	511	39,013
2020	2,809	34,668	1,120	229	760	129	52	75	38,593	2,805	34,679	1,113	509	602	38,597
2021	9,123	37,820	2,489	1,131	1,217	139	54	83	49,428	9,118	37,816	2,498	1550	945	49,432
2022	3,934	42,416	1,084	293	701	89	22	67	47,433	3,920	42,425	1,089	566	523	47,434
2023	3,701	41,465	1,336	288	851	197	107	90	46,502	3,689	41,465	1,348	644	703	46,502
Total	30,487	294,597	9,492	2,689	5,937	858	354	500	299,854	30,422	259,909	9,535	4,949	4,574	299,866

A pronounced structural break is observed between 2016 and 2017. In 2017, entries sharply increase to 11,178, while exits fall to 3,709 (in empirical pipeline), resulting in an unusually large positive net change. This pattern likely reflects a data construction artifact rather than a genuine economic shift. Specifically, the supervised (2007 to 2016) and unsupervised (2017 to 2023) datasets were processed separately when identifying multi-establishment firms. As a result, firms that operated as single establishment entities during the supervised period but transitioned into multi-establishment firms after 2016 are only captured starting in 2017. Consequently, their true start dates are not recorded in the earlier period, artificially inflating entries and suppressing exits in 2017.

After 2017, the series stabilizes: entries become more consistent over time, while exits show a gradual upward trend. An additional spike in entries is observed in 2021, with 9,942 establishments classified as new. This may reflect post-pandemic dynamics, such as re-entries, firm restructuring, or increased sensitivity of the pipeline to changes in establishment characteristics.

Overall, the empirical pipeline produces plausible establishment turnover dynamics across most of the period, with the main exception being the structural discontinuity at the 2016 to 2017 boundary, which leads to an overestimation of new establishments in 2017.

The semantic pipeline displays the same structural break at the 2016-2017 boundary, with entries rising to 11,488 and exits falling to 3,966 in 2017, confirming that this discontinuity originates from the underlying data construction process rather than from a specific matching methodology. However, the semantic pipeline consistently records higher entries and exits than the empirical pipeline in nearly every year of the sample, including the 2021 spike (10,292 against 9,942 new establishments). This is consistent with the fragmentation behaviour identified. Despite this gross inflation in entries and exits, the implied net change for the semantic pipeline remains very close to that of the empirical pipeline throughout the period, suggesting that the additional fragmentation affects entry and exit counts roughly symmetrically without materially distorting the overall turnover signal.

The larger discrepancies observed for the semantic pipeline in the early years of the sample may be partly explained by structural changes in the administrative classifications underlying the IES Annex R data. Three relevant events occurred during this period: a revision of the CAE in 2008, a structural change to Annex R itself in 2010, and a reorganisation of civil parishes in 2013, in which several parishes were merged into existing ones. Although the BERT model was fine-tuned on data spanning the supervised period, the train/validation/test

split was constructed at the NIPC level rather than by year, so the model may not have observed a sufficient number of examples capturing these specific transitions. When such classification codes or form structures change abruptly between consecutive years, which increasing the likelihood that the establishment is misclassified as a new entry or exit rather than as a continuation.

Figure 5.5: New & closed establishments and its net change, from 2007 to 2023.

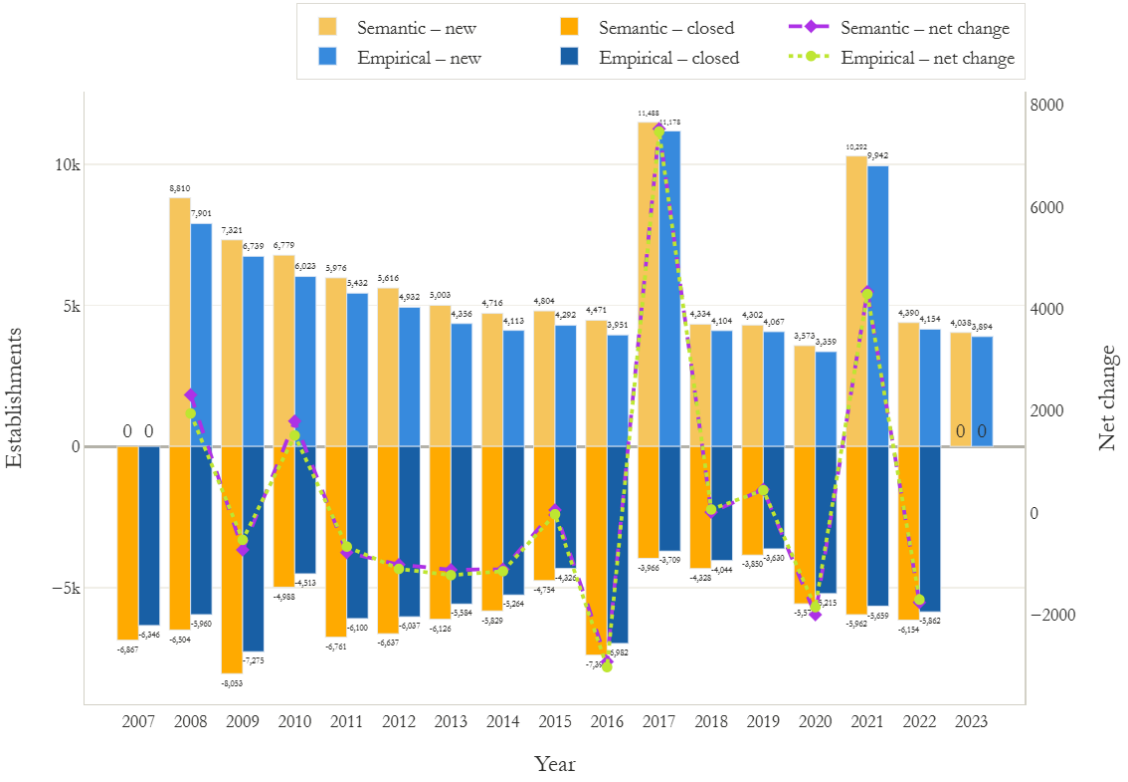


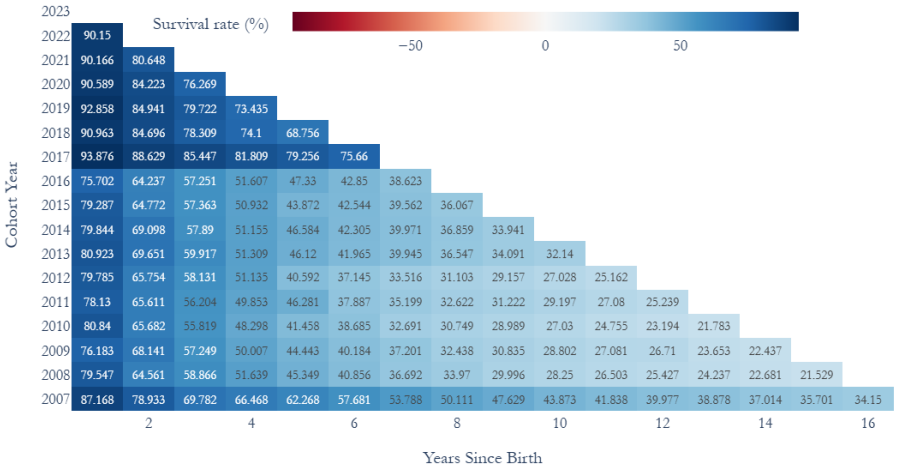
Figure 5.6a presents cohort survival rates for establishments identified by the empirical pipeline over the full 2007 to 2023 period. For most cohorts, year-one survival rates range between 75.00% and 93.00%, then decline gradually to around 40.00 to 50.00% by year five. The figure also reveals a clear structural break at the 2016 to 2017 boundary, visible as a new block of higher survival rates at the top of the matrix.

The 2016 cohort shows unusually low survival, falling from 75.70% in year one to 38.60% by year six. In contrast, the 2017 cohort exhibits the opposite pattern, with year one survival of 93.90% and consistently higher survival than adjacent cohorts throughout its observed life. These two deviations are consistent with the transition between the supervised and unsupervised datasets, which hampers the pipeline's ability to link establishments across the boundary. From 2018 onward, survival trajectories return to a more stable pattern, suggesting that the pipeline reaches a steadier state.

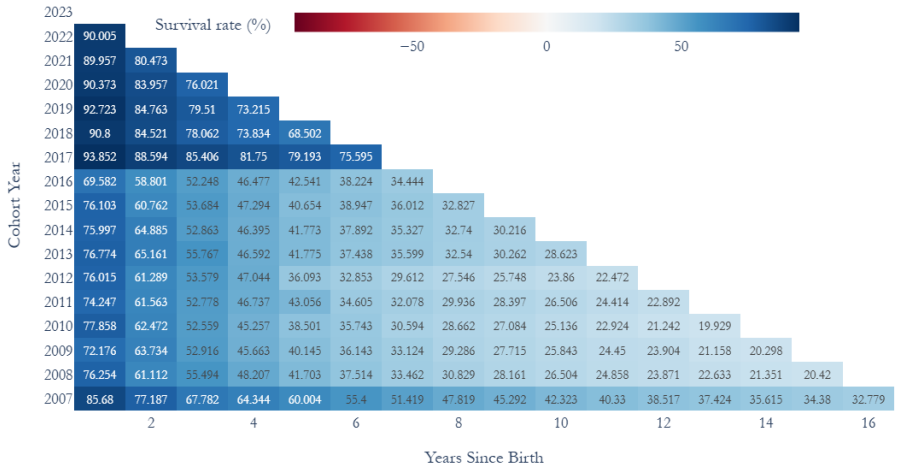
Both pipelines exhibit the boundary in 2016 to 2017. The near identical timing and magnitude of this break across both pipelines reinforces the conclusion that it originates from the transition between the supervised and unsupervised datasets, rather than from the specific matching methodology employed.

Outside the boundary region, survival rates for the semantic pipeline are systematically lower than for the empirical pipeline across nearly all pre-2017 cohorts, consistent with the positive bias identified in the diverging survival rate difference heatmap (Figure 5.6b). From 2017 onward, the two pipelines converge closely, with year one survival differences typically below 0.3 percentage points, suggesting that the semantic pipeline's fragmentation bias is concentrated in the earlier, structurally disrupted period rather than reflecting a persistent weakness of the method itself.

Figure 5.6: Survival rate difference by cohort x years since birth across the pipelines.



(a) Empirical pipeline



(b) Semantic pipeline

Chapter 6

Conclusion

The main goals of this dissertation are to develop and evaluate a record linkage framework designed to construct a stable, unique identifier for each establishment using Annex R data (IES). The motivation for this work arises from a fundamental limitation of the existing data: the reported establishment identifier proved insufficiently reliable for longitudinal analysis, with accuracy declining from 84.52% in 2008 to 73.51% in 2016.

To achieve that purpose, two pipelines were developed and evaluated against a semi-automated annotation ground truth covering the years 2007 to 2016, and subsequently applied to unsupervised data from 2017 to 2023. The empirical pipeline operates in three sequential steps (similar to a waterfall structure). Step 1 applies a deterministic and fuzzy similarity function over spatial and categorical features, using string metrics tailored to the specific characteristics of each variable. Records classified as clear matches or non-matches are resolved at this stage, while ambiguous cases are passed forward as partial matches. Step 2 applies a more comprehensive similarity function that uses a consistently reported subset of variables, introduces a hybrid TF-IDF and fuzzy-matching approach for composite address comparison, and uses the Hungarian algorithm to obtain a globally optimal one-to-one assignment. Step 3 provides a final deterministic resolution for any remaining partial matches resulting from Step 2, ensuring that every establishment receives an identifier. The pipeline uses Bayesian threshold optimization via Optuna, calibrating five threshold hyperparameters against a weighted objective that balances accuracy and the volume of partial matches.

The semantic pipeline replaces Steps 2 and 3 with a fine-tuned Cross-Encoder (using a BERT-large model), trained on hard-negative pairs generated by a Bi-Encoder ensemble of three Portuguese-language Transformer models. The Cross-Encoder classifies partial matches from Step 1 as either matches or non-matches based on probability scores, with the optimal threshold of 0.42 selected by sweeping the validation set. While the Cross-Encoder achieved an accuracy on the test set of 94.26% and an F1-score of 95.76%, it produced 1,886 establishments with duplicate identifiers within the same year, a problem not observed in the empirical pipeline.

Overall, the empirical pipeline performs marginally better than the semantic pipeline across both year on year identifier accuracy and the confusion matrix metrics, achieving slightly higher precision (92.38% versus 92.07%), recall (93.69% versus 91.96%), and F1-score (93.03%

versus 92.01%), as well as consistently higher accuracy in every cohort year. However, this advantage should be interpreted with caution, given that the empirical pipeline underwent considerably more extensive testing, threshold tuning, and refinement than the semantic pipeline, which may account for at least part of the observed performance gap. Moreover, the empirical pipeline is also substantially more efficient computationally, matching the full panel from 2007 to 2023 takes approximately 10 hours on CPU alone, compared to approximately 29 hours for the semantic pipeline's Step 1 and Cross-Encoder classification on GPU.

Across diagnostic analyses, both pipelines reproduced consistent and plausible establishment dynamics. The empirical pipeline closely tracked the ground truth, with survival rate deviations of only $(1.75 \pm 0.52)\%$ across 2009 to 2015 cohorts, whereas the semantic pipeline showed a systematic positive bias of $+5$ to $+6.90$ pp, reflecting its tendency to fragment true establishment trajectories. Both pipelines also exhibited a structural discontinuity at the 2016 to 2017 boundary, due to the separate processing of the supervised and unsupervised datasets: firms with only a single establishment up to 2017 were not considered multi-establishment in the supervised dataset, so if such firms acquired more than one establishment after 2016, this growth would only be captured starting in 2017, artificially inflating the count of new establishments in that year.

This dissertation makes a tangible contribution to the production of high-quality statistical microdata for the Portuguese economy. By developing an algorithm to automatically harmonize the identifier, this work enables the construction of a more consistent establishment level panel dataset.

6.1 Limitations and Future Work

Restriction to multi-establishment firms. By design, the pipeline applies only to firms operating two or more establishments in a given year. Firms that remain single establishment throughout the observation period fall outside the scope of the linkage algorithm. Moreover, the transition of a firm from single to multi-establishment status is not handled within the supervised period, contributing to the structural discontinuity at the dataset boundary.

Five year backward search window. The historical search is limited to the five most recent years preceding the year under analysis. Consequently, if a firm is absent from the data for more than five consecutive years and later reappears (which is a rare event in the data), its identifier sequence restarts from 1, creating a potential conflict with the original identifier.

Address based false positives. The main source of false positives in the empirical pipeline arises from establishments with near identical addresses that differ only in fine-grained details, such as door numbers or floors. In such cases, the similarity score exceeds the matching threshold despite the establishments being genuinely distinct. This is a structural limitation of string based address matching, and the resolution strategy for establishments sharing the same address and CAE code, yet representing distinct entities, warrants further refinement.

Sensitivity to structural administrative changes. Three events introduced non-trivial noise into the data: the 2008 revision of CAE codes, the 2010 restructuring of Annex R, and the 2013 territorial reorganisation of civil parishes. Although the empirical pipeline incorporates specific

handling for the CAE transition year, residual mismatches in affected years are unavoidable. The BERT model may be particularly sensitive to these changes, since its train/validation/test split was stratified by firm rather than by year, potentially limiting its exposure to transition year patterns during training.

Semantic pipeline duplicate identifiers. The semantic pipeline generates 1,886 duplicate identifiers across the full period (the same identifier appears more than once per year). This is a critical limitation for panel data construction and would require an additional deduplication step before the semantic identifiers can be used in longitudinal analysis.

Accuracy decline over time. There is evidence of a gradual decline in accuracy in both pipelines, suggesting that static thresholds calibrated early in the period may become too permissive as data quality improves over time; periodic threshold updates after 2011 could help sustain accuracy.

Year-stratified training for the BERT model. Re-splitting the training data by year rather than by firm would ensure the model observes examples from each structural transition period during training. This could reduce the fragmentation bias observed for the semantic pipeline in transition years and improve its temporal generalization.

Nevertheless, the framework is fully parameterized and designed for adaptability: the restriction to multi-establishment firms, the five year historical window, the number of years processed, and all decision thresholds are configurable, allowing the pipeline to be readily extended to single establishment firms, and longer time spans.

References

Agapito, G., & Cauteruccio, F. (2025). Entity resolution and data integration in bioinformatics. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (2nd ed., Vol. 1, pp. 116–129). Academic Press. Retrieved from https://doi.org/10.1016/B978-0-323-95502-7.00045-3 doi: 10.1016/B978-0-323-95502-7.00045-3

Agostinho, A., & Miguel, A. (2017, July). How to increase quality in the central banks statistical business process? the experience of banco de portugal. In *Papers presented by the statistics department in national and international fora* (pp. 11–16). Banco de Portugal. Retrieved from https://www.bportugal.pt/sites/default/files/anexos/pdf-boletim/suplemento_1_2017_en_2.pdf (Supplement to the Statistical Bulletin 1 | 2017, pp. 11–16. Accessed: December 23, 2025)

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2623–2631). ACM. doi: 10.1145/3292500.3330701

Andoni, A., Indyk, P., Laarhoven, T., Razenshteyn, I., & Schmidt, L. (2015). Practical and optimal lsh for angular distance. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 1225–1233). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2015/hash/61df6d338-Abstract.html

Assembleia da República. (2009, July 13). *Decreto-lei n.º 158/2009: Sistema de normalização contabilística.* Retrieved from https://diariodarepublica.pt/dr/legislacao-consolidada/decreto-lei/2009-34517175 (Diário da República: 1.ª série, n.º 133/2009. Accessed: December 19, 2025)

Assembleia da República. (2013, January 28). *Lei n.º 11-a/2013 de 28 de janeiro – reorganização administrativa do território das freguesias.* Retrieved from https://diariodarepublica.pt/dr/detalhe/lei/11-a-2013-373798 (Diário da República, 1.ª série, N.º 19. Accessed: May 2026)

Assembleia da República. (2025, February 28). *Lei n.º 25-a/2025 de 28 de fevereiro – orçamento do estado para 2025.* Retrieved from https://diariodarepublica.pt/dr/detalhe/lei/25-a-2025-910933580 (Diário da República, 1.ª série, N.º 41. Accessed: May 2026)

Ather, H. (2026). LLM-assisted record linkage: A framework for official statistics. *Statistical Journal of the IAOS*, *42*(1), 211–217. doi: 10.1177/18747655261422068

Autoridade Tributária e Aduaneira. (2025). *Informação empresarial simplificada / declaração anual de informação contabilística e fiscal (ies/da).* Retrieved from https://info.portaldasfinancas.gov.pt/pt/apoio_contribuinte/modelos_formularios/

decl_anual_inf_contabilistica_fiscal/Documents/IES_DANUAL_R.pdf (Modelo Oficial. Accessed: December 15, 2025)

Autoridade Tributária e Aduaneira, Instituto dos Registos e do Notariado, Instituto Nacional de Estatística, & Banco de Portugal. (2015). *Informação empresarial simplificada - ies/declaração anual: Perguntas & respostas.* Instituto dos Registos e do Notariado. Retrieved from https://irn.justica.gov.pt/Portals/33/Contas%20anuais/IES-PF-PerguntasFrequentes.pdf (Accessed: December 5, 2025)

Babu, B. V., & Santoshi, K. J. (2014). Unsupervised detection of duplicates in user query results using blocking. *International Journal of Computer Science and Information Technologies*, *5*(3), 3514–3520.

Banco de Portugal Microdata Research Laboratory (BPLIM). (2025). *Establishment data - annex r.* Banco de Portugal - BPLIM. Dataset.

Barlaug, N., & Gulla, J. A. (2021). Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data*, *15*(3), Article 53. Retrieved from https://doi.org/10.1145/3442200 doi: 10.1145/3442200

Beheshti, M., Gondara, L., & Zachary, I. (2025). Leveraging language models for automated patient record linkage. *Studies in Health Technology and Informatics*, *329*, 1632–1633. Retrieved from https://doi.org/10.3233/SHTI250438 doi: 10.3233/SHTI250438

Bilenko, M., Kamath, B., & Richardson, M. (2006). Adaptive blocking: Learning to scale up record linkage. In *Proceedings of the sixth ieee international conference on data mining (icdm-06)* (pp. 87–96). Retrieved from https://doi.org/10.1109/ICDM.2006.13 doi: 10.1109/ICDM.2006.13

Borkar, V., Deshmukh, K., & Sarawagi, S. (2001). Automatic segmentation of text into structured records. *SIGMOD Record*, *30*(2), 175–186. Retrieved from https://doi.org/10.1145/376284.375682 doi: 10.1145/376284.375682

BPLIM. (n.d). *panelstat: Descriptive statistics for panel data sets.* https://github.com/BPLIM/Tools/tree/master/ados/General/panelstat. (Stata package. BPLIM. Accessed: 2026-06-02)

Celko, J. (2005). Character data types in sql. In *Joe celko's sql for smarties: Advanced sql programming* (3rd ed., pp. 169–184). Morgan Kaufmann. Retrieved from https://datubaze.wordpress.com/wp-content/uploads/2020/01/celkos-sql-for-smarties-2005.pdf

Christen, P. (2012a). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer. Retrieved from https://doi.org/10.1007/978-3-642-31164-2 doi: 10.1007/978-3-642-31164-2

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, *24*(9), 1537–1555. Retrieved from https://doi.org/10.1109/TKDE.2011.127 doi: 10.1109/TKDE.2011.127

Christen, P., Churches, T., & Zhu, J. X. (2002). Probabilistic name and address cleaning and standardisation. In *Proceedings of the australasian data mining workshop (adm 2002).* Canberra, Australia. Retrieved from https://users.cecs.anu.edu.au/~Peter.Christen/publications/adm2002-cleaning.pdf

Churches, T., Christie, J., Lim, K., & Zelinsky, S. (2002). Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, *2*, Article 9. Retrieved from https://doi.org/10.1186/1472-6947-2-9 doi: 10.1186/1472-6947-2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of*

the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers) (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1810.04805 doi: 10.18653/v1/N19-1423

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., … Reizenstein, J. (2024). *The faiss library.* Retrieved from https://doi.org/10.48550/arXiv.2401.08281 (arXiv preprint arXiv:2401.08281) doi: 10.48550/arXiv.2401.08281

Elfeky, M. G., Verykios, V. S., & Elmagarmid, A. K. (2002). Tailor: A record linkage toolbox. In *Proceedings of the 18th international conference on data engineering* (pp. 17–28). Retrieved from https://doi.org/10.1109/ICDE.2002.994694 doi: 10.1109/ICDE.2002.994694

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. Retrieved from https://doi.org/10.1080/01621459.1969.10501049 doi: 10.1080/01621459.1969.10501049

Franke, M., Christen, V., Christen, P., Rohde, F., & Rahm, E. (2024). (privately) estimating linkage quality for record linkage. In *Proceedings of the 27th international conference on extending database technology (edbt)* (pp. 294–306). Retrieved from https://doi.org/10.48786/edbt.2024.26 doi: 10.48786/edbt.2024.26

Franke, M., Sehili, Z., Gladbach, M., & Rahm, E. (2018). Post-processing methods for high quality privacy-preserving record linkage. In J. Garcia-Alfaro, J. Herrera-Joancomartí, G. Livraga, & R. Rios (Eds.), *Data privacy management, cryptocurrencies and blockchain technology* (pp. 263–278). Springer. Retrieved from https://doi.org/10.1007/978-3-030-00305-0_18 doi: 10.1007/978-3-030-00305-0_18

Gomatam, S., Carter, R. L., Ariet, M., & Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Statistics in Medicine*, *21*(10), 1485–1496. Retrieved from https://doi.org/10.1002/sim.1147 doi: 10.1002/sim.1147

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). *Record linkage: Current practice and future directions* (Tech. Rep. No. Technical Report No. 03/83). CSIRO Mathematical and Information Sciences. Retrieved from https://www.researchgate.net/publication/2479819_Record_Linkage_Current_Practice_and_Future_Directions

Han, J., Pei, J., & Tong, H. (2023). Data, measurements, and data preprocessing. In *Data mining: Concepts and techniques* (4th ed., pp. 23–84). Morgan Kaufmann / Elsevier. Retrieved from https://doi.org/10.1016/B978-0-12-811760-6.00012-6 doi: 10.1016/B978-0-12-811760-6.00012-6

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, *2*(1), 9–37. Retrieved from https://doi.org/10.1023/A:1009761603038 doi: 10.1023/A:1009761603038

Holmes, D., & McCabe, M. C. (2002). Improving precision and recall for soundex retrieval. In *Proceedings of the international conference on information technology: Coding and computing* (pp. 22–26). Retrieved from https://doi.org/10.1109/ITCC.2002.1000354 doi: 10.1109/ITCC.2002.1000354

Hugging Face. (n.d.). *Hugging face – the AI community building the future.* https://huggingface.co. (Accessed: 2026-06-23)

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual acm symposium on theory of computing (stoc '98)* (pp. 604–613). Retrieved from https://doi.org/10.1145/276698.276876

doi: 10.1145/276698.276876

Instituto Nacional de Estatística. (2023). *Empresas em Portugal — demografia das empresas 2021*. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=586619963&DESTAQUESmodo=2. (Destaque INE, 8 de março de 2023. Accessed: 2026-06-30)

Instituto Nacional de Estatística. (n.d.). *Sistema integrado de metainformação: Categoria*. Retrieved from https://smi.ine.pt/Versao (Integrated Metadata System: Category. Accessed: December 2025)

Instituto Nacional de Estatística. (1994a). *Empresa (conceito) [concept 4456]*. Retrieved from https://smi.ine.pt/Conceito/Detalhes/4456?modal=1 (Sistema de Metainformação – INE. Accessed: December 5, 2025)

Instituto Nacional de Estatística. (1994b). *Estabelecimento (conceito) [concept 3464]*. Retrieved from https://smi.ine.pt/Conceito/Detalhes/3464?modal=1 (Sistema de Metainformação – INE. Accessed: December 5, 2025)

Karapiperis, D., Tjortjis, C., & Verykios, V. S. (2025). Lsblock: A hybrid blocking system combining lexical and semantic similarity search for record linkage. In P. K. Chrysanthis, K. Stefanidis, Z. Zhang, & K. Nørvåg (Eds.), *Advances in databases and information systems: 29th european conference, adbis 2025* (pp. 131–146). Springer. Retrieved from https://doi.org/10.1007/978-3-032-05281-0_9 doi: 10.1007/978-3-032-05281-0_9

Kaufman, A. R., & Klevs, A. (2021). Adaptive fuzzy string matching: How to merge datasets with only one (messy) identifying field. *Political Analysis*, *30*(4), 590–596. Retrieved from https://doi.org/10.1017/pan.2021.38 doi: 10.1017/pan.2021.38

Knuth, D. E. (1998). *The art of computer programming: Vol. 3. sorting and searching* (2nd ed.). Addison-Wesley.

Konda, P., Das, S., Suganthan, G. C. P., Doan, A., Ardalan, A., Ballard, J. R., … Raghavendra, V. (2016). Magellan: Toward building entity matching management systems. *Proceedings of the VLDB Endowment*, *9*(12), 1197–1208. Retrieved from https://doi.org/10.14778/2994509.2994535 doi: 10.14778/2994509.2994535

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83–97.

Lait, A. J., & Randell, B. (1996). *An assessment of name matching algorithms* (Tech. Rep. No. Technical Report Series No. 550). University of Newcastle upon Tyne. Retrieved from http://homepages.cs.ncl.ac.uk/brian.randell/Genealogy/NameMatching.pdf

Marcelino, D. (2015). *Soundexbr: Phonetic-coding for portuguese*. Retrieved from https://doi.org/10.32614/CRAN.package.SoundexBR (R package (Version 1.2), Comprehensive R Archive Network (CRAN)) doi: 10.32614/CRAN.package.SoundexBR

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., … Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 international conference on management of data* (pp. 19–34). Retrieved from https://doi.org/10.1145/3183713.3196926 doi: 10.1145/3183713.3196926

Munk, M., & Feitelson, D. G. (2022). *When are names similar or the same? introducing the code names matcher library*. Retrieved from https://doi.org/10.48550/arXiv.2209.03198 (arXiv preprint) doi: 10.48550/arXiv.2209.03198

O'Hare, K., Jurek-Loughrey, A., & De Campos, C. (2021). High-value token-blocking: Efficient blocking method for record linkage. *ACM Transactions on Knowledge Discovery from Data*, *15*(5), Article 79, 1–23. Retrieved from https://doi.org/10.1145/3450527 doi:

10.1145/3450527

Ordem dos Contabilistas Certificados. (2025, April 17). *Ies - manual de preenchimento da declaração ies e taxonomias - 2025*. Retrieved from https://www.occ.pt/pt-pt/publicacoes/ies-manual-de-preenchimento-da-declaracao-ies-e-taxonomias-2025 (Accessed: December 5, 2025)

Ratcliff, J. W., & Obershelp, J. A. (1988). Pattern matching: The gestalt approach. *Dr. Dobb's Journal*(141), 46–51.

Reimers, N., & Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3982–3992). Retrieved from https://doi.org/10.18653/v1/D19-1410 doi: 10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2019b, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1908.10084

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th brazilian conference on intelligent systems, BRACIS, rio grande do sul, brazil, october 20-23 (to appear)*.

Statistics Canada. (2021, September 2). *Record linkage*. Retrieved from https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/5214780-eng.htm (Statistics: Power from Data! Accessed: December 2025)

Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In J. Domingo-Ferrer (Ed.), *Privacy in statistical databases* (pp. 253–268). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-11257-2_20 doi: 10.1007/978-3-319-11257-2_20

Tokle, J., & Bender, S. (2021). Record linkage. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science: Data science methods and tools for research and practice* (2nd ed., chap. 3). Chapman and Hall/CRC. Retrieved from https://textbook.coleridgeinitiative.org/chap-link.html

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wang, M., Haberland, V., Yeo, A., Martin, A., Howroyd, J., & Bishop, J. M. (2016). A probabilistic address parser using conditional random fields and stochastic regular grammar. In *2016 ieee 16th international conference on data mining workshops (icdmw)* (pp. 225–232). Retrieved from https://doi.org/10.1109/ICDMW.2016.0039 doi: 10.1109/ICDMW.2016.0039

Wikipedia contributors. (2025, January 27). *Postal codes in portugal*. Retrieved from https://en.wikipedia.org/wiki/Postal_codes_in_Portugal (In Wikipedia. Accessed: December 2025)

Winkler, W. E. (1990). *String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage*. U.S. Bureau of the Census. Retrieved from https://eric.ed.gov/?id=ED325505

Winkler, W. E. (1993). *Matching and record linkage* (Tech. Rep. Nos. Research Report Series, Statistics #RR93/08). U.S. Bureau of the Census. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/1993/adrm/rr93-8.pdf

Winkler, W. E. (2006). *Overview of record linkage and current research directions* (Tech. Rep. Nos. Research Report Series, Statistics #2006-2). U.S. Census Bureau. Retrieved from https://www.census.gov/library/working-papers/2006/adrm/rrs2006-02.html

Yang, S., & Berdine, G. (2024). Confusion matrix. *The Southwest Journal of Medicine*, *12*(53), 75–79. doi: 10.12746/swrccc.v12i53.1391

Zobel, J., & Dart, P. (1996). Phonetic string matching: Lessons from information retrieval. In *Proceedings of the 19th annual international acm sigir conference on research and development in information retrieval* (pp. 166–172). Retrieved from https://doi.org/10.1145/243199.243258 doi: 10.1145/243199.243258

Appendix A

Appendix

A.1 Annex R

MODELO EM VIGOR A PARTIR DE JANEIRO DE 2022

ANTES DE PREENCHER ESTE ANEXO LEIA COM ATENÇÃO AS INSTRUÇÕES QUE O ACOMPANHAM

MUITO IMPORTANTE

IES

DECLARAÇÃO ANUAL

IES – INFORMAÇÃO EMPRESARIAL SIMPLIFICADA

(ENTIDADES RESIDENTES QUE EXERCEM, A TÍTULO PRINCIPAL, ATIVIDADE COMERCIAL, INDUSTRIAL OU AGRÍCOLA, ENTIDADES NÃO RESIDENTES COM ESTABELECIMENTO ESTÁVEL E EIRL)

01	Nº DE IDENTIFICAÇÃO FISCAL (NIPC)	02	EXERCÍCIO
1		1	

IE

ANEXO R

03 NÚMERO DE ESTABELECIMENTOS DA ENTIDADE

EM TERRITÓRIO NACIONAL: 1 ____ FORA DO TERRITÓRIO NACIONAL: 2 ____

04 ESTABELECIMENTOS LOCALIZADOS NO TERRITÓRIO NACIONAL – Exercícios de 2009 e anteriores

MORADA: 1 _____

CÓDIGO POSTAL: 2 ____ - ____ 3 LOCALIDADE: _____

DISTRITO: 4 ____ CONCELHO: 5 ____ FREGUESIA: 6 ____

FAX: 7 _____ TELEFONE: 8 _____

E-MAIL: 9 _____ SITUAÇÃO PERANTE A ATIVIDADE: 10 ____

ATIVIDADE PRINCIPAL: 11 _____

CAE Rev. 3 12 _____ Nº DE ORDEM DO ESTABELECIMENTO: 13 ____ ESTABELECIMENTO SEDE: SIM 14 ☐ NÃO 15 ☐

CONTAS POC			
-	Número médio de pessoas ao serviço durante o ano	R101	
61 + 62	Custo das mercadorias vendidas e das matérias consumidas e fornecimentos e serviços externos	R102	. . ,
64	Custos com o pessoal	R103	. . ,
641 + 642	Remunerações	R104	. . ,
71 + 72	Vendas e prestações de serviços	R105	. . ,
vp	Variação da produção	R106	. . ,
42 + 441/6 + 448	Aumentos de imobilizado corpóreo	R107	. . ,
422	Dos quais: Aumentos em edifícios e outras construções	R108	. . ,

04-A ESTABELECIMENTOS DA ENTIDADE - Exercícios de 2010 e seguintes

CARACTERIZAÇÃO

1 PAÍS: 1 ☐☐☐

2 TIPO DE LOCALIZAÇÃO: 16 ☐☐

MORADA: 1 _____

CÓDIGO POSTAL: 2 ☐☐☐☐ - ☐☐☐ 3 LOCALIDADE: _____

DISTRITO: 4 ☐☐ CONCELHO: 5 ☐☐ FREGUESIA: 6 ☐☐

FAX: 7 ☐☐☐☐☐☐☐☐☐ TELEFONE: 8 ☐☐☐☐☐☐☐☐☐

E-MAIL: 9 _____ SITUAÇÃO PERANTE A ATIVIDADE: 10 ☐☐

DATA DE INÍCIO DE EXPLORAÇÃO: 17 ANO ☐☐☐☐ MÊS ☐☐ DIA ☐☐

ATIVIDADE PRINCIPAL: 11 _____

CAE Rev. 3 12 ☐☐☐☐☐ Nº DE ORDEM DO ESTABELECIMENTO: 13 ☐☐☐☐ ESTABELECIMENTO SEDE: SIM 14 ☐ NÃO 15 ☐

INSÍGNIA: 18 _____

ÁREA TOTAL DO ESTABELECIMENTO m^2: 19 ☐☐☐☐☐

ÁREA DE EXPOSIÇÃO E VENDA m^2 20 ☐☐☐☐☐

ÁREA DE ARMAZENAGEM m^2 21 ☐☐☐☐☐

ÁREA DE PRESTAÇÃO DE SERVIÇOS m^2 22 ☐☐☐☐☐

RESTANTE ÁREA m^2 23 ☐☐☐☐☐

INFORMAÇÃO ECONÓMICA

CONTAS SNC			
-	Número médio de pessoas ao serviço durante o ano	R201	
71	Vendas	R202	. . ,
72	Prestações de serviços	R203	. . ,
73	Variações nos inventários da produção	R204	. . ,
61	Custo das mercadorias vendidas e das matérias consumidas	R205	. . ,
62	Fornecimentos e serviços externos	R206	. . ,
63	Gastos com o pessoal	R207	. . ,
631 + 632	Remunerações	R208	. . ,
31	Compras	R209	. . ,
37 ...	Aquisições em ativos biológicos	R210	. . ,
43 ...	Aquisições em ativos fixos tangíveis	R211	. . ,
432 ...	Das quais: Em edifícios e outras construções	R212	. . ,
42 ...	Aquisições em propriedades de investimento	R213	. . ,
422 ...	Das quais: Em edifícios e outras construções	R214	. . ,
	Capitais próprios ou equiparados	R215	. . ,

Outras informações:

05 ESTABELECIMENTOS LOCALIZADOS FORA DO TERRITÓRIO NACIONAL - Exercícios 2009 e anteriores

CONTAS POC			
-	Número médio de pessoas ao serviço durante o ano	R109	
61 + 62	Custo das mercadorias vendidas e das matérias consumidas e Fornecimentos e serviços externos	R110	. . ,
64	Custos com o pessoal	R111	. . ,
641 + 642	Remunerações	R112	. . ,
71 + 72	Vendas e Prestações de serviços	R113	. . ,
vp	Variação da produção	R114	. . ,
42 + 441/6 +448	Aumentos de imobilizado corpóreo	R115	. . ,
422	Dos quais: Aumentos em edificios e outras construções	R116	. . ,

INFORMAÇÃO ESTATÍSTICA (IE) – INFORMAÇÃO EMPRESARIAL SIMPLIFICADA

ESTABELECIMENTOS DA ENTIDADE

Sujeitos passivos residentes que exercem, a título principal, atividade de natureza comercial, industrial ou agrícola, entidades não residentes com estabelecimento estável e Estabelecimentos Individuais de Responsabilidade Limitada (EIRL)

INSTRUÇÕES PARA O PREENCHIMENTO DO *ANEXO R* À IES / DECLARAÇÃO ANUAL

INDICAÇÕES GERAIS

No âmbito da Informação Empresarial Simplificada (IES), criada pelo Decreto-Lei n.º 8/2007, de 17 de janeiro, o **Anexo R** deve ser apresentado **:**

1) CONJUNTAMENTE com o **Anexo A** pelas entidades residentes que exercem, a título principal, uma atividade de natureza comercial, industrial ou agrícola, ou por entidades não residentes com estabelecimento estável;

2) CONJUNTAMENTE com o **Anexo I** pelos Estabelecimentos Individuais de Responsabilidade Limitada (EIRL).

Com a submissão conjunta e por via eletrónica dos referidos Anexos, considera-se disponibilizada a informação necessária ao cumprimento das seguintes obrigações legais compreendidas na IES:

- entrega da declaração anual de informação contabilística e fiscal (alínea c) do n.º 1 do artigo 117.º e artigo 121.º do CIRC e n.º 1 do artigo 113.º do CIRS);
- registo da prestação de contas junto das conservatórias do registo comercial (n.º 1 do artigo 15.º do Código do Registo Comercial);
- prestação de informação de natureza estatística ao Instituto Nacional de Estatística (n.º 1 do artigo 4.º da Lei do Sistema Estatístico Nacional);
- prestação de informação relativa a dados contabilísticos anuais para fins estatísticos ao Banco de Portugal (artigo 13.º da Lei Orgânica do Banco de Portugal);
- prestação de informação de natureza estatística à Direção-Geral das Atividades Económicas (DGAE), para os efeitos previstos no regime jurídico de acesso e exercício de atividades de comércio, serviços e restauração, aprovado em anexo ao Decreto-Lei n.º 10/2015, de 16 de janeiro;
- confirmação da informação sobre o beneficiário efetivo, nos termos previstos em legislação especial (artigo 15.º da Lei n.º 89/2017, de 21 de agosto).

Estas obrigações legais são exclusivamente cumpridas através da entrega da IES (n.º 3 do artigo 2.º do Decreto-Lei n.º 8/2007, de 17 de janeiro).

Para os **exercícios de 2009 e anteriores**, deve preencher o quadro 04 e a informação a constar deste quadro deve ser desagregada por estabelecimento, devendo ser preenchidos tantos quadros quantos os estabelecimentos que a entidade possui/explora em território nacional. Se possuir/explorar estabelecimentos fora do território nacional, deverá preencher o quadro 05.

Para os **exercícios de 2010 e seguintes**, deve preencher o quadro 04-A e a informação a constar deste quadro deve ser desagregada por estabelecimento, devendo ser preenchidos tantos quadros quantos os estabelecimentos que a entidade possui/explora, quer em território nacional, quer fora deste.

O somatório dos valores atribuídos aos vários estabelecimentos, localizados no território nacional e/ou fora do território nacional, deve corresponder aos valores da entidade.

Estabelecimento – corresponde a uma entidade ou parte desta (fábrica, oficina, mina, armazém, loja, escritório, entreposto, sucursal, filial, agência, etc.) situada num local topograficamente identificado. Nesse local ou a partir dele exercem-se atividades económicas para as quais, regra geral, uma ou várias pessoas trabalham (eventualmente a tempo parcial), por conta de uma mesma entidade. A sede da entidade deve ser considerada como um estabelecimento.

Nota: não devem ser criados estabelecimentos de natureza virtual para enquadramento de vendas à distância (vendas *online*, por *e-mail*, telefónicas, por correspondência ou similares). As vendas sem contacto presencial devem ser afetas a um estabelecimento físico ou à sede.

Nos casos em que a entidade possui/explora apenas um estabelecimento coincidente com a sede da entidade, só devem ser preenchidos os campos 1 a 15 do Quadro 04 deste anexo (exercícios de 2009 e anteriores) ou os campos 1 a 23 do Quadro 04-A (períodos de 2010 e seguintes).

Quadro 01 – N.º de Identificação Fiscal (NIPC/NIPC)

Caso se trate de pessoa coletiva, inscrever o número de identificação de pessoa coletiva ou equiparada (NIF/NIPC) atribuído pelo Ministério da Justiça e constante do respetivo cartão de empresa ou pessoa coletiva.

Caso se trate de um EIRL, deve inscrever o número de identificação fiscal (NIF) que lhe tiver sido atribuído.

Quadro 02 – Exercício/Período

Indicar o exercício a que respeitam os rendimentos e demais variáveis económicas.

Tendo-se adotado um período de tributação diferente do ano civil, deve ser indicado o ano em que se integre o primeiro dia do referido período.

Quadro 03 – Número de estabelecimentos da entidade

No campo 1 indicar o número de estabelecimentos que a entidade possui/explora **em território nacional**, incluindo a sede, mesmo que nestes não seja exercida atividade produtiva.

No campo 2 indicar o número de estabelecimentos que a entidade possui/explora **fora do território nacional**, mesmo que nestes não seja exercida atividade produtiva.

Quadro 04 – Estabelecimentos localizados no território nacional – Exercícios de 2009 e anteriores

Este quadro deve ser preenchido isoladamente **para cada um** dos estabelecimentos indicados no campo 1 do Quadro 03.

No campo 1 indicar a morada, no campo 2 o código postal e no campo 3 a localidade do estabelecimento.

No campo 4 indicar o distrito, no campo 5 o concelho e no campo 6 a freguesia do estabelecimento.
No campo 7 indicar o número de fax, no campo 8 o número de telefone e no campo 9 o endereço de e-mail do estabelecimento.

No campo 10 indicar a situação perante a atividade do estabelecimento. Este campo pode assumir os seguintes valores:

(01) aguarda início de atividade;
(02) em atividade;
(03) atividade suspensa, ou
(04) cessou a atividade.

No campo 11 descrever, em texto livre, a *atividade principal do estabelecimento*. Esta corresponde à atividade com maior importância no conjunto das atividades exercidas pelo estabelecimento. O critério para a sua aferição é o valor acrescentado bruto ao custo dos fatores. Na impossibilidade da sua determinação por este critério, considera-se como *atividade principal* a que representa o maior volume de negócios ou, em alternativa, a que ocupa, com carácter de permanência, o maior número de pessoas ao serviço.

No campo 12 indicar o CAE Rev. 3 do estabelecimento, ou seja o código da atividade principal exercida no estabelecimento de acordo com a Classificação Portuguesa das Atividades Económicas (Decreto-Lei n.º 381/2007, de 14 de novembro – CAE Rev. 3).

No campo 13 deve atribuir a cada estabelecimento que a entidade possui/explora em território nacional um **número de ordem.** Ao estabelecimento sede deve ser atribuído o número de ordem igual a um. Para os restantes estabelecimentos o número de ordem a atribuir deve ser 2, 3, 4, … dependendo do número de estabelecimentos que a entidade possui/explora em território nacional.

O N.º de ordem atribuído deve ser mantido em futuras IES. Nestas, se o estabelecimento em causa cessar a atividade, não deve voltar a utilizar o número de ordem que lhe estava atribuído em novos estabelecimentos. A estes deve ser atribuído o número de ordem imediatamente a seguir ao do último número atribuído.

No campo 14 ou 15 deve indicar se o estabelecimento corresponde ou não à sede da entidade.

No campo R101 indicar o número médio de pessoas ao serviço no estabelecimento durante os meses do ano em que o estabelecimento esteve em atividade.

> ***Pessoas ao serviço do estabelecimento*** – deve incluir o pessoal que trabalha no estabelecimento/entidade e que recebe uma remuneração em dinheiro ou em espécie como contrapartida do trabalho prestado

(incluindo sócios), o pessoal que trabalha para o estabelecimento/entidade sem usufruir qualquer tipo de remuneração (ex: sócios trabalhadores, trabalhadores familiares), o pessoal ausente por um período não superior a um mês (ex: doença, férias, formação profissional) e o pessoal de outras entidades que se encontre a trabalhar na entidade, sendo por esta diretamente remunerado. **Não deve incluir** o pessoal a trabalhar no estabelecimento/entidade cuja remuneração é suportada por outra entidade, os prestadores de serviços (profissionais liberais), o pessoal do estabelecimento/ entidade ausente por um período superior a um mês (ex: doença, serviço militar obrigatório, licença sem vencimento) e o pessoal com vínculo à entidade deslocado para outras entidades, sendo nessas diretamente remunerado.

Os restantes campos (R102 a R108) deste quadro correspondem a informação económica que se extrai dos valores registados nas contas/subcontas do POC, aprovado pelo Decreto-Lei n.º 410/89, de 21 de novembro e respetivas alterações aplicáveis. Por este motivo remete-se para o referido diploma todas as indicações quanto ao seu âmbito.

Este quadro é flexível permitindo, assim, utilizar tantos quadros quanto os necessários.

Quadro 04-A – Estabelecimentos da entidade – Períodos de 2010 e seguintes

Este quadro deve ser preenchido apenas para os exercícios de 2010 e seguintes.

Este quadro deve ser preenchido isoladamente **para cada um** dos estabelecimentos indicados nos **campos 1 e 2** do Quadro 03.

No campo 1.1 indicar o país onde se situa o estabelecimento, de acordo com a norma ISO 3166 (parte numérica). No caso do país indicado ser diferente de PORTUGAL, não deve preencher os campos 2.1 a 2.8 deste quadro.

No campo 2.16 deve discriminar-se a localização do estabelecimento, conforme este se situe em:
01 arruamento;
02 zona industrial;
03 centro comercial;
04 *retail park*;
05 *outlet center*;
06 mercado, ou
07 outro.

Em complemento deve atender às seguintes especificações:

Centro Comercial: Conjunto de estabelecimentos de venda a retalho e de serviços, concebidos, realizados e organizados como uma unidade, situados num ou mais edifícios contíguos com pelo menos 500 m^2 de área bruta, devendo estes possuir zonas comuns por onde prioritariamente se fará o acesso às lojas nele implantadas, bem como uma unidade de gestão responsável pela implementação, direção e coordenação dos serviços comuns técnico-comerciais.

Outlet Center: Conjunto de estabelecimentos de venda a retalho e de serviços, onde fabricantes e retalhistas vendem mercadorias na sua maioria com desconto no preço, para escoamento rápido de stocks ou por se tratarem de produtos descontinuados ou com pequenos defeitos.

Retail Park: Conjunto de estabelecimentos de venda a retalho e de serviços, concebidos, realizados e organizados como uma unidade, sendo os seus estabelecimentos de dimensão superior ao habitualmente verificado nos centros comerciais, estando integrados num espaço aberto para a via pública, com acesso direto ao parque de estacionamento ou áreas pedonais.

No campo 2.1 indicar a morada, no campo 2.2 o código postal do estabelecimento e no campo 2.3 a localidade.

No campo 2.4 indicar o distrito, no campo 2.5 o concelho e no campo 2.6 a freguesia do estabelecimento.
No campo 2.7 indicar o número de fax, no campo 2.8 o número de telefone (obrigatório) e no campo 2.9 o endereço de *e-mail* (obrigatório) do estabelecimento correspondente à sede.

No campo 2.10 indicar a situação perante a atividade do estabelecimento. Este campo pode assumir os seguintes valores:
(01) aguarda início de atividade;
(02) em atividade;
(03) atividade suspensa, ou
(04) cessou a atividade.

No campo 2.17 indicar a data de início de exploração (ano/mês/dia). A data de início de exploração corresponde à data em que o estabelecimento iniciou a sua atividade. Este campo é de preenchimento obrigatório quando no campo 2.10 for indicado que o estabelecimento se encontra em (02) em atividade; (03) atividade suspensa ou (04) cessou a atividade.

No campo 2.11 descrever, em texto livre, a *atividade principal do estabelecimento*. Esta corresponde à atividade com maior importância no conjunto das atividades exercidas no estabelecimento. O critério para a sua aferição é o valor acrescentado bruto ao custo dos fatores. Na impossibilidade da sua determinação por este critério, considera-se como *atividade principal* a que representa o maior volume de negócios ou, em alternativa, a que ocupa, com carácter de permanência, o maior número de pessoas ao serviço.

No campo 2.12 indicar o código da atividade principal do estabelecimento de acordo com a Classificação Portuguesa das Atividades Económicas (CAE), o qual deve corresponder à **CAE da Rev. 4**, se a declaração respeitar aos **períodos de 2025 e seguintes**, ou corresponder à CAE da Rev. 3, se a declaração respeitar aos períodos de 2024 e anteriores.

No campo 2.13 deve atribuir a cada estabelecimento que a entidade possui/explora um **número de ordem.** Ao estabelecimento sede deve ser atribuído o número de ordem igual a um. Para os restantes estabelecimentos o número de ordem a atribuir deve ser 2, 3, 4, … dependendo do número de estabelecimentos que a entidade possui/explora.

O Número de ordem atribuído deve ser mantido em futuras IES. Nestas, se o estabelecimento em causa cessar a atividade, não deve voltar a utilizar o número de ordem que lhe estava atribuído em novos estabelecimentos. A estes deve ser atribuído o número de ordem imediatamente a seguir ao do último número atribuído.

Os **números de ordem** atribuídos devem respeitar a numeração definida na **declaração IES do período anterior**, desde que:

- ao estabelecimento sede corresponda o número um, e
- a cada novo estabelecimento seja atribuído um número não utilizado em anos anteriores, devendo ser imediatamente superior aos números em uso e antes usados.

A numeração dos estabelecimentos só poderá ser alterada, face à declaração do período anterior, nos casos em que:

- a sede não corresponda ao número um, ou
- tenha existido anteriormente algum lapso na numeração, como seja o uso do mesmo número para dois estabelecimentos diferentes, em anos diversos.

No campo 2.14 ou 2.15 deve indicar se o estabelecimento corresponde (ou não) à sede da entidade.

No campo 2.18 indicar a insígnia do estabelecimento, ou seja, a designação comercial do estabelecimento. Caso o estabelecimento não opere sob nenhuma insígnia, o campo deverá ser preenchido com a indicação “N.A” ou “não aplicável”.

> ***Insígnia:*** designação figurativa que identifica no mercado um estabelecimento ou um conjunto de estabelecimentos, pertencentes ou não à mesma entidade e que obedecem a um determinado conceito de negócio.

No campo 2.19 deve ser apurada a **área total do estabelecimento** (expressa em metros quadrados) e deve corresponder à soma da área de exposição e venda, da área de armazenagem, da área de prestação de serviços e da restante área que porventura exista no estabelecimento. A área deve ser expressa em números inteiros, sem casas decimais.

O campo 2.20 (área de exposição e venda) deve ser preenchido relativamente a estabelecimentos onde seja exercida uma atividade de comércio por grosso ou a retalho (expressa em metros quadrados).

Em complemento deve atender às seguintes especificações:

> ***Estabelecimento comercial:*** estabelecimento situado num local topograficamente identificado, onde é exercida, exclusivamente ou principalmente, uma ou mais atividades de comércio, com exceção das atividades respeitantes à reparação de bens pessoais e domésticos.
>
> ***Área de exposição e venda:*** área, de um estabelecimento (em m^2) onde é exercida uma atividade comercial, destinada a venda à qual os compradores têm acesso e na qual os produtos se encontram expostos ou são preparados para entrega imediata. Incluem-se as zonas ocupadas pelas caixas de saída e de circulação dos consumidores internas ao estabelecimento, nomeadamente as escadas de ligação entre os vários pisos. Excluem-se as áreas ocupadas pelo armazenamento, pelos escritórios, serviços administrativos e ainda outros espaços não ligados diretamente a exposição e venda.
>
> ***Área de armazenagem:*** área destinada (em m^2) ao depósito de produtos para venda posterior. Não se incluem as áreas de exposição e venda e as áreas ocupadas pelos escritórios, serviços administrativos e outros espaços não ligados diretamente ao armazenamento.
>
> ***Área de prestação de serviços:*** área de um estabelecimento (em m^2), a que o cliente tem acesso, onde é exercida uma atividade de prestação de serviços.

Restante área: corresponde a toda a área (em m^2) não quantificada nos campos 2.20, 2.21 e 2.22.

No campo R201 indicar o número médio de pessoas ao serviço no estabelecimento durante os meses do ano em que o estabelecimento esteve em atividade.

> ***Pessoas ao serviço do estabelecimento*** – deve incluir o pessoal que trabalha no estabelecimento/entidade e que recebe uma remuneração em dinheiro ou em espécie como contrapartida do trabalho prestado (incluindo sócios), o pessoal que trabalha para o estabelecimento/entidade sem usufruir qualquer tipo de remuneração (ex: sócios trabalhadores, trabalhadores familiares), o pessoal ausente por um período não superior a um mês (ex: doença, férias, formação profissional) e o pessoal de outras entidades que se encontre a trabalhar na entidade, sendo por esta diretamente remunerado. **Não deve incluir** o pessoal a trabalhar no estabelecimento/entidade cuja remuneração é suportada por outra entidade, os prestadores de serviços (profissionais liberais), o pessoal do estabelecimento/ entidade ausente por um período superior a um mês (ex: doença, serviço militar obrigatório, licença sem vencimento) e o pessoal com vínculo à entidade deslocado para outras entidades, sendo nessas diretamente remunerado.

Os restantes campos (R202 a R215) deste quadro correspondem a informação económica que se extrai dos valores registados nas contas/subcontas cuja codificação foi publicada pela Portaria n.º 1011/2009, de 9 de setembro, ou pela Portaria n.º 218/2015, de 23 de julho, e que integra o Sistema de Normalização Contabilístico (SNC), aprovado pelo Decreto-Lei n.º 158/2009, de 13 de julho. Por este motivo remete-se para os referidos diplomas todas as indicações quanto ao seu âmbito.

Este quadro é flexível permitindo, assim, utilizar tantos quadros quanto os necessários.

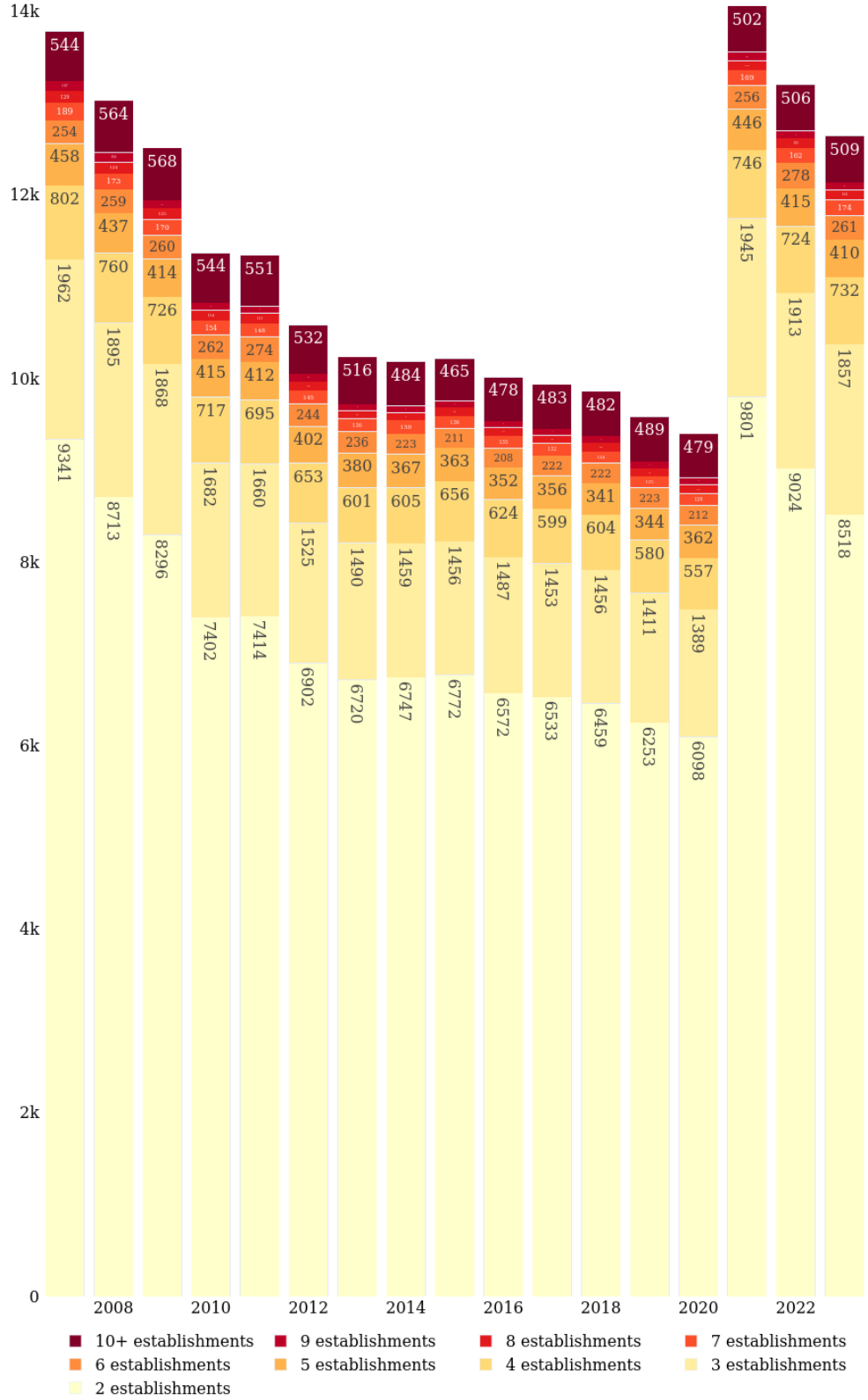
Quadro 05 – Estabelecimentos localizados fora do território nacional – Exercícios de 2009 e anteriores

Os dados individuais dos estabelecimentos que a entidade possui/explora fora do território nacional **devem ser agregados** para efeitos de preenchimento deste quadro, ou seja, este quadro deve ser preenchido **apenas uma vez**, e os valores nele registados devem resultar da agregação dos valores individuais de todos os estabelecimentos localizados fora do território nacional indicados no campo 2 do Quadro 03.

No campo R109 indicar o somatório do número médio de pessoas ao serviço no conjunto dos estabelecimentos localizados fora do território nacional durante o ano (ver instruções do quadro 04, campo R101).

Os restantes campos (R110 a R116) deste quadro, à semelhança do que acontece no quadro 04 para os campos R102 a R108, correspondem a contas do POC, aprovado pelo Decreto-Lei n.º 410/89, de 21 de novembro e respetivas alterações aplicáveis. Por este motivo remete-se para o referido diploma todas as indicações quanto ao seu âmbito.

A.2 Portuguese multi-establishment distribution by level.



A.3 Number of partial matches remain and accuracy obtained, using the average of the combine years (2008, 2010, 2012, 2013,2014 and 2016)

Year	Step 1 partial	Step 2 partial	Accuracy (%)
2007	0	0	100.00
2008	4,953	8	93.54
2009	4,353	1,384	92.79
2010	5,530	2,024	91.69
2011	3,617	1,407	91.64
2012	4,396	1,532	91.19
2013	4,961	1,352	90.60
2014	3,706	1,294	90.12
2015	3,274	1,177	90.03
2016	3,154	1,167	89.98

Confusion Matrix on firm level, resulting from the threshold using the average of the combined years (2008, 2010, 2012, 2013, 2014 and 2016)

		Predicted	
		Positive	**Negative**
Actual	**Positive**	67,262	8,456
	Negative	5,865	24,451

Metric	Value
Precision	91.98%
Recall	88.83%
F1-score	90.38%

A.4 Metrics resulting from the threshold using the average of the combine years (2008, 2010, 2012, 2013,2014 and 2016)

Table A.1: Evaluation metrics.

-	Precision	Recall	F1-score
%	92.98	88.83	90.38

A.5 Number of partial matches remain and accuracy obtained, using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016)

Year	Step 1 partial	Step 2 partial	Accuracy (%)
2007	0	0	100.00
2008	5,138	8	93.14
2009	4,533	1,638	92.45
2010	5,707	2,243	91.37
2011	3,706	1,537	91.35
2012	4,514	1,703	90.92
2013	5,141	1,468	90.34
2014	3,829	1,424	89.88
2015	3,411	1,315	89.77
2016	3,244	1,267	89.69

A.6 Confusion Matrix on firm level, resulting from the threshold using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016)

		Predicted	
		Positive	**Negative**
Actual	**Positive**	66,984	8,754
	Negative	5,845	24,451

A.7 Metrics resulting from the threshold using the median of the combine years (2008, 2010, 2012, 2013,2014 and 2016)

Table A.2: Evaluation metrics.

-	Precision	Recall	F1-score
%	91.97	88.44	90.17

A.8 Duplicated Establishments per year.

Table A.3: Number of duplicated establishments dropped per year.

Year	Ground Truth	Empirical Pipeline	Semantic Pipeline
2007	73	164	163
2008	44	123	118
2009	26	123	118
2010	4	57	54
2011	2	72	68
2012	2	74	69
2013	3	73	68
2014	2	76	72
2015	2	65	64
2016	0	69	67
Total	158	896	861

A.9 Dropped establishment, multi-establishment and establishments that in some year become multi-establishments

Table A.4: Number of dropped establishments, multi-establishments, and establishments that in some year become multi-establishments, Empirical vs. Semantic Pipeline.

	Empirical Pipeline			Semantic Pipeline		
Year	**Dropped**	**Only multi-est.**	**Multi-est.**	**Dropped**	**Only multi-est.**	**Multi-est.**
2017	80	39,456	47,423	77	39,458	57,589
2018	88	39,352	48,116	75	38,354	57,715
2019	82	38,983	48,684	71	38,985	57,794
2020	83	38,563	48,915	73	38,567	57,762
2021	71	49,403	55,442	62	49,405	63,943
2022	70	47,398	54,487	67	47,398	62,747
2023	78	46,464	53,861	69	46,464	61,769
Total	**552**	**299,619**	**356,928**	**494**	**299,631**	**419,319**