TÉCNICO LISBOA

Emparelhamento de Dados Censitários

Lufialuiso Sampaio Velho

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Prof. Mário Jorge Costa Gaspar da Silva
Prof. Pável Pereira Calado

Júri

Presidente: Prof. Daniel Jorge Viegas Gonçalves
Orientador: Prof. Mário Jorge Costa Gaspar da Silva
Vogal: Prof. Pedro Manuel Moreira Vaz Antunes de Sousa

Outubro 2017

Agradecimentos

Ao Deus de toda a criação, Alfa e Ômega, meu Paracleto agradeço pelo dom da vida, Saúde, Paz e Motivação desde o início do MEIC até a data presente. Nada seria possível sem o Senhor (Salmos 90:1-2).
À Mimosa Velho, minha amada Esposa, conselheira e companheira de longa data e aos meus filhos: Ntemos e Dino Velho. Obrigado pelo amor que sempre demonstraram, pela paciência e constante fé de que chegaríamos até aqui.
Aos meus Pais, Sogros, Irmãos, Sobrinhos e Tios,
Aos meus Orientadores: Prof. Mário J. Gaspar da Silva e Prof. Pável Calado pelo vosso constante apoio e direcção durante a elaboração deste trabalho. Tenho-vos como inspiração para a minha carreira científica.

Ao INE pela disponibilização dos seus recursos durante a elaboração deste trabalho, e um especial obrigado para a equipa do Gabinete dos Censos 2021: Eng.ª Anabela Delgado, Doutora Sandra Lagarto, Dr.ª Paula Paulino, João Capelo e Paula Nabais pela vossa constante atenção e disponibilidade em ajudar-nos em todas as nossas dúvidas e inquietações ao longo deste trabalho,

Ao Prof. Doutor Mateus Padoca Calado e sua família, pelo vosso apoio imensurável, força, confiança e motivação durante este longo e valioso percurso. Que Deus esteja sempre contigo e abençoe a sua família, pois foste um grande impulsionador desta formação. A sua perseverança ajudou-me a ter esperança e prosseguir. O nosso muito Obrigado!
Aos guerreiros e companheiros Dr. Amândio de Jesus Almada e Dr. João José da Costa. Sim, mais do que companheiros, hoje sois como meus irmãos. Passamos momentos alegres e difíceis e em silêncio caminhamos sempre para o alvo. Confio que concluirão a vossa formação com êxito.

Aos Professores Vangajala Soki, Augusta Martins, Suzanete Nunes da Costa, Orlando da Mata, Samuel Vitorino, Agatângelo Eduardo, e Inês Massukinini pelo vosso grande apoio, muitos de forma indirecta mas reconheço a vossa prestação para a concretização deste propósito.

Ao Ministério Cristo é a Resposta, presidido pelo Pastor Oenes Andrade: a minha família em Lisboa. Nem que pudesse escrever um livro, não caberia o quanto significam para mim. Amo-vos e sempre serão a minha família.

Aos meus colegas da UAN e IST: MSc. Rui Menezes da Silva, MSc. Dikiefu Fabiano, MSc. Vicente Lopes, Dr. Emanuel Tunga, Dr. Aureliano Francisco, MSc. Dizando Norton, Dr. Edson Novais, Mário Katala, Dr. Domingos da Conceição (XP), Drª Cândida John, MSc. Glória Gonçalves, MSc. João Eduardo, MSc. Nuno Roboredo, MSc. Frederico Felisberto, MSc. Lilian Gomes, Francisco Calisto, MSc. Artur Arêde, Eduardo Benjamim Melo. Muito obrigado por tudo!

Abstract

A feasibility study is under way to enable Statistics Portugal to obtain part of the census information through administrative data sources. The process becomes complex because there is not a personal unique number, inconsistencies in the data and anonymised/pseudonimized data by determination of the Data Protection Authority (CNPD). This work presents an approach based on matching available data using Machine Learning methods. With the developed system, it was possible, for example, to detect 244,903 new matches between records of the databases of the Civil Population Register (Citizen's Card) and Tax Authority (IRS), representing an increase of 64.94%, and 47,836 new matches, an increase of 19.21%, with the Social Security database considering records not matched by exact methods. The obtained results support the feasibility of the methodology and software developed for pairing the administrative data that are now available at Statistics Portugal.

Keywords

Record Linkage, Data Quality, Machine Learning, Census

Resumo

Está em curso um estudo de viabilidade para que Portugal, através do Instituto Nacional de Estatística (INE), possa obter parte da informação censitária através de fontes de dados administrativos. O processo torna-se complexo devido ao facto de não haver um número único do cidadão, inconsistências nos dados, e dados anonimizados/pseudonimizados por determinação da Comissão Nacional de Protecção de Dados (CNPD). Esta dissertação apresenta uma abordagem baseada no emparelhamento dos registos disponibilizados recorrendo a métodos de aprendizagem automática (probabilísticos). Com o sistema desenvolvido, foi possível, a título de exemplo detectar 244.903 novos emparelhamentos entre registos das bases dados de Identificação Civil (Cartão do Cidadão) e Autoridade Tributária (Imposto Sobre o Rendimento das Pessoas Singulares (IRS)), representando um acréscimo de 64,94%, e 47.836 novos emparelhamentos, um acréscimo de 19,21%, com a base de dados da Segurança Social, relativamente aos registos não emparelhados por métodos exactos. Os resultados obtidos sustentam a viabilidade da metodologia e do software desenvolvido para o emparelhamento dos dados administrativos que são hoje disponibilizados ao INE.

Palavras Chave

Integração de Dados, Qualidade de Dados, Aprendizagem Automática, Censos

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

AT	Autoridade Tributária
AR	Autorização de Residência
ACSS	Administração Central do Sistema de Saúde
BDIC	Base de Dados da Identificação Civil
BI	Bilhete de Identidade
BKV	Blocking Key Values
BPMN	Business Process Model and Notation
BPR	Base da População Residente
BS	Blocking Schema
CC	Cartão de Cidadão
CGA	Caixa Geral de Aposentações
CNPD	Comissão Nacional de Protecção de Dados
CSV	Comma Separated values
CTP	Census Transformation Programme
DGEEC	Direcção Geral de Estatísticas da Educação e Ciência
EDUC	Fonte de Dados da Educação e Ciência
GEP	Gabinete de Estratégia e Planeamento do Ministério da Solidariedade e Segurança Social
HESA	Higher Education Statistics Agency
IRN	Instituto dos Registos e do Notariado

INE	Instituto Nacional de Estatística
IISS	Instituto de Informática da Segurança Social
IEFP	Instituto do Emprego e Formação Profissional
IRS	Imposto Sobre o Rendimento das Pessoas Singulares
LSOA	Lower Super Output Area Mid-Year Population Estimates
MPI	Módulo de Produção da Informação
NIC	Número de Identificação Civil
NISS	Número de Identificação da Segurança Social
NIF	Número de Identificação Fiscal
ONS	Office for National Statistics
PCS	Population Coverage Survey
PR	NHS Patient Register
SEF	Serviços de Estrangeiros e Fronteiras
SPD	Statistical Population Dataset
SGBD	Sistema de Gestão de Base de Dados
SQL	Structured Query Language
SRE	Statistical Research Environment
SSH	Secure Shell

1

Introdução

Conteúdo

Os Censos representam a fonte de grande parte da informação estatística sociodemográfica da população e do parque habitacional em todo o mundo [INE,Gabinete dos Censos 2021, 2016b]. Constituem o único meio que fornece informações precisas acerca do número da população residente num determinado país, a dimensão das famílias, características da população, assim como as condições e tipos de habitação em que elas vivem [Office for National Statistics, 2016b].
Na maioria dos países do mundo quer em África, Europa, Ásia, ou Américas, este levantamento é realizado em cada 10 anos. Por exemplo: Argélia, Portugal, China, Estados Unidos da América e Equador [1].
Considerando os custos financeiros, a carga sobre os cidadãos para responder aos inquéritos e a frequência requerida para estes estudos, vários países, tais como a Alemanha, Polônia, Reino Unido, Espanha, Áustria e Itália procuram migrar do modelo tradicional, baseado em inquéritos porta a porta, para novos modelos censitários, baseados unicamente em ficheiros administrativos ou ainda modelos híbridos que combinam os dois.
Está em curso um estudo de viabilidade para que Portugal nos Censos 2021 migre para um modelo combinado, baseado em fontes de dados administrativos complementado por inquéritos. Actualmente, estão disponíveis no INE várias fontes de dados administrativas, provenientes de nove organismos da administração pública:

- Instituto dos Registos e do Notariado (IRN),
- Autoridade Tributária (AT),
- Caixa Geral de Aposentações (CGA),
- Instituto de Informática da Segurança Social (IISS),
- Direcção Geral de Estatísticas da Educação e Ciência (DGEEC),
- Serviços de Estrangeiros e Fronteiras (SEF),
- Gabinete de Estratégia e Planeamento do Ministério da Solidariedade e Segurança Social (GEP),
- Instituto do Emprego e Formação Profissional (IEFP),
- Administração Central do Sistema de Saúde (ACSS).

O INE tem como objectivo integrar as fontes de dados referidas para a criação de uma BPR, que consiste numa relação em que cada registo nela contida corresponde a um cidadão residente em Portugal num determinado ano.

[1] https://unstats.un.org/unsd/demographic/sources/census/censusdates.htm

Este processo torna-se complexo, atendendo vários motivos:

- Não existe um identificador pessoal único que seja comum entre as diferentes fontes de dados administrativas;
- Baixa taxa de preenchimento de alguns atributos demográficos;
- Falta de um padrão único para a representação da informação;
- Um aumento exponencial do número de comparações entre pares de registos em função do número de fontes de dados consideradas e quadrático no número de registos.

Os motivos acima referidos, acrescidos às variações na formatação dos dados, abreviações, omissões, erros de digitação ou ainda a utilização de *stop words*, limitam o emparelhamento com métodos exatos, uma vez que potenciais emparelhamentos seriam excluídos devido a diferenças entre pares de atributos como as indicadas.
Nestes casos, é comum o uso de medidas de similaridade para comparar os registos complementados com métodos de classificação por regras ou probabilísticos para determinar se pares de registos de fonte distintas se referem ou não à mesma entidade no mundo real.

1.1 Caracterização das Fontes de Dados

Os dados disponíveis no INE carregados a partir das várias fontes fornecedoras de informação, encontram-se armazenados num Sistema de Gestão de Base de Dados (SGBD) Oracle. O processamento e o armazenamento dos resultados finais do tratamento da informação serão realizados nesta plataforma.
Os dados das várias fontes encontram-se codificados de forma a proteger a privacidade dos cidadãos. A protecção da privacidade, é resultante da deliberação nº 929/2014 da CNPD [CNPD, 2014], que estabelece o seguinte:

1. Os identificadores numéricos tais como: o Número de Identificação Civil (NIC), Número de Identificação Fiscal (NIF), Número de Identificação da Segurança Social (NISS) e Autorização de Residência (AR), devem ser cifrados com um hash SHA256 antes do envio ao INE;
2. O nome de cada indivíduo é representado pelas 3 primeiras letras do 1º nome e as 3 últimas letras do último nome.
3. Quanto ao endereço do indivíduo, são disponibilizados unicamente o Código Postal (7 dígitos) e a Localidade.

Por outro lado, o INE realizou algumas transformações sobre as fontes de dados disponíveis [INE,Gabinete dos Censos 2021, 2015], tais como:

1. O atributo sexo encontra-se representado por valores numéricos (1 e 2), acompanhado de um atributo que designa cada valor (1 para Masculino e 2 para Feminino);

2. A data de nascimento encontra-se representada em formato numérico. Por exemplo,2016/04/29 é representado por 20160429;

3. O código postal encontra-se representado por dois atributos, nomeadamente um para os 4 primeiros dígitos do código postal e o outro para os 3 últimos dígitos do código postal.

4. Foi acrescentado às fontes de dados um atributo ANO, com intuíto de identificar claramente o ano que os respectivos dados correspondem.

1.2 Objectivos

A presente dissertação tem enquadramento dentro do estudo de viabilidade de um novo modelo censitário para Portugal que permita vir a substituir os inquéritos directos pela obtenção de dados a partir das fontes. Identificam-se os seguintes objectivos:

- Propor uma metodologia para emparelhamento dos registos das diferentes fontes de dados Administrativos disponibilizados ao INE, tendo em conta que as mesmas, em geral, não possuem um identificador pessoal único comum;

- Propor um método para a avaliação da qualidade dos emparelhamentos realizados entre os registos provenientes das diversas fontes ao longo das várias etapas da metodologia preconizada.

1.3 Contribuições

As contribuições que resultaram do desenvolvimento deste trabalho incluem:

- Um gerador de pares de registos candidatos a emparelhamento de custo computacional tratável;

- Um avaliador de emparelhamentos, cujo o intuito é medir a qualidade da classificação e a reavaliação da qualidade dos emparelhamentos realizados pelo INE durante a construção dos protótipos da BPR nas versões 2011 e 2015.

1.4 Metodologia

O presente trabalho foi realizado com base nas seguintes etapas:

1. Revisão da literatura e o estudo de emparelhamentos realizados em instituições similares, como por exemplo a ONS (congénere do INE no Reino Unido);
2. Familiarização do ambiente de trabalho e a revisão do processo de emparelhamento exacto realizado pelo INE;
3. Elaboração da arquitectura da solução e a proposta de metodologia de emparelhamento dos dados;
4. Implementação dos módulos do processo de emparelhamento;
5. Avaliação dos emparelhamentos realizados (e a comparação com os resultados anteriores do INE obtidos por métodos exactos);
6. Documentação.

O desenvolvimento do trabalho em referência realizou-se em conjunto com o estudante Rui Menezes da Silva, no qual a minha dissertação e contribuições incidem em maior detalhe na elaboração da metodologia usada para o emparelhamento das fontes de dados e na avaliação da qualidade dos dados em cada uma das etapas do processo de emparelhamento. As técnicas de aprendizagem automática para o emparelhamento dos registos são abordadas em maior detalhe na dissertação do meu colega [Silva, Rui, 2017].

1.5 Organização da Dissertação

A presente dissertação está organizada pelo seguinte modo:

1. No Capítulo 2, abordo primeiramente os fundamentos teóricos que servirão de suporte para a compreensão deste trabalho, e de seguida faço referência sobre alguns sistemas similares que considerei relevantes e que solucionam problemas de emparelhamento de registos;
2. No Capítulo 3, abordo a sobre a metodologia implementada para o Emparelhamento das Fontes de dados administrativas disponibilizadas ao INE para o Censo;
3. No Capítulo 4, abordo sobre a metodologia implementada para a avaliação da qualidade dos Emparelhamentos produzidos e a Validação dos Emparelhamentos realizados pelo INE;
4. No Capítulo 5, apresento as conclusões do trabalho realizado.

2

Métodos de Emparelhamento de Dados e Sistemas Similares

Conteúdo

Neste capítulo abordo numa visão geral os conceitos e técnicas que servirão de suporte para a compreensão do presente trabalho. Seguidamente abordarei alguns trabalhos de maior relevância realizados e em curso, focados na resolução de problemas de emparelhamento de dados (*data matching*).

2.1 Métodos de Emparelhamento de Dados

Nesta secção abordo as etapas envolvidas no processo de emparelhamento de registos e algumas técnicas para avaliação da qualidade dos emparelhamentos baseadas em métricas de qualidade de dados.
As subsecções seguintes abordam as etapas de um processo de emparelhamento de registos com base na seguinte ordem:

- Na subsecção 2.1.1 abordo o pré-processamento de dados (limpeza e normalização de dados), cujo o objectivo é garantir que as duas relações a serem integradas estejam no mesmo formato em termos de estrutura e conteúdo;
- Na subsecção 2.1.2 abordo algumas técnicas de indexação ou *Blocking*(em português colocar em blocos) cujo o objectivo é garantir a redução da complexidade do número de comparações entre pares de registos candidatos a emparelhamento;
- Na subsecção 2.1.3 abordo as medidas de similaridade para comparação de pares de atributos, que podem ser *Strings* ou numéricos, considerando o facto de que o emparelhamento de registos inicia num nível mais abaixo que corresponde ao emparelhamento de pares de atributos dos registos, para posteriormente comparar os registos e decidir se os registos referem-se a verdadeiros emparelhamentos ou não;
- Na subsecção 2.1.4 abordo os algoritmos de aprendizagem automática usados para a classificação de registos, assim como as técnicas existentes para a validação da qualidade do Classificador;
- Na subsecção 2.1.5 abordo as técnicas existentes para a fusão de registos, cujo o objectivo é fundir os registos emparelhados, mantendo um única representação da entidade no mundo real;
- Na subsecção 2.1.6 abordo as métricas de avaliação da qualidade dos dados, cujo o objectivo é garantir que os dados a serem emparelhados obedecem a qualidade exigida, e no final do processo de emparelhamento, garantir que o resultado obtido corresponde ao esperado.

2.1.1 Limpeza e Normalização dos Dados

Os dados a serem emparelhados, podem conter atributos não peenchidos, erros, inconsistências ou ainda variações em termos de formato e estrutura. Estes problemas são resolvidos numa etapa inicial do processo de emparelhamento de fontes de dados designado por limpeza e normalização ou pré-processamento de dados.
[Christen, 2012] considera três passos fundamentais nesta etapa:

1. Remoção de caracteres tais como vírgulas, pontos e outros caracteres especiais também designados por *stop words*;

2. Correção da variação do conteúdo dos atributos, por exemplo abreviações do tipo Av. ao invés de Avenida no caso de ruas, assim como erros ortográficos;

3. Segmentação do conteúdo dos atributos, como é o caso de atributos que possuem vários pedaços de informação representando o endereço de um indivíduo, por exemplo o nome da rua, edifício e o código postal, quando na outra fonte estes encontram-se separados; e a correcção de valores nos atributos, como é o caso dos códigos postais ou ainda números telefónicos que variam de região em região.

Um aspecto importante a ter em conta antes da execução desta etapa, consiste em realizar uma cópia das fontes de dados a serem utilizadas e aplicar sobre estas cópias as operações requeridas, para garantir que os dados originais não sejam sobrepostos [Christen, 2012].

2.1.2 Indexação

Suponha duas relações A e B com 1.000 registos para cada uma. Caso se pretenda obter os pares candidatos a emparelhamento destas relações, seriam necessárias no total 1.000 x 1.000 =1.000.000 comparações. Num problema de emparelhamento que envolve milhões de registos como é o caso do presente trabalho, estas comparações teriam elevado custo de computação, o que e se tornaria impraticável.
A indexação, *sorting* (em português, organizar) ou *blocking* (em português, colocar em blocos), surge devido a necessidade de reduzir o elevado número de computações entre os vários pares de registos durante a fase de comparação.
Assim sendo, o principal propósito é criar uma estrutura de dados baseada em índice, que seja capaz de agrupar em blocos ou em *cluster* todos os valores similares de acordo a um certo critério designado por *blocking key* (em português, chave de bloqueamento) [Christen, 2012], que é definida a partir de um atributo ou a concatenação de vários atributos, ou ainda a partir de pequenos pedaços dos atributos, que podem ser codificadas com funções *hash* ou outras para garantir maior performance.

Para a definição de uma chave de bloqueamento, são considerados alguns aspectos tais como: a qualidade dos dados nos atributos (menor percentagem de valores nulos ou vazios), a frequência dos valores dos atributos, e o *trade-off* entre o número de chaves a serem geradas versus o tamanho de valores em cada bloco.
Dentre as várias técnicas de *Blocking* existentes, a seguir faço menção de duas, nomeadamente a *Standard Blocking* que é o método tradicional de *Blocking* e a *Sorted Neighbourhood*.

2.1.2.A Standard Blocking

O *Standard Blocking* é a abordagem tradicional de indexação, baseia-se no agrupamento dos registos em vários blocos de acordo com uma chave de bloqueamento definida, também designado por *critério de Blocking*. Isto é, se considerarmos um problema de emparelhamento, o processo de agrupamento em blocos é realizado de modo independente para cada relação, e no final são gerados os pares candidatos mediante a combinação dos vários pares de blocos cujo valor da chave de bloqueamento (Blocking Key Values (BKV)) seja igual .
No caso de ser aplicada uma função *hash* sobre a chave de bloqueamento, este código *hash* definirá o *Bucket* onde um determinado registo se encontra e a comparação entre os pares de registos das fontes de dados serão feitas apenas em caso dos dois registos terem o mesmo *hash*.
Uma abordagem referida por [Christen, 2012], consiste na construção de uma estrutura de dados baseada em índice invertido, onde a chave de bloqueamento representa a chave de índice e todos os identificadores dos registos que estejam no mesmo bloco, são inseridos na mesma lista de índices invertidos, tal como ilustra a figura 2.1.

Identificadores	Sobrenome	BKVs (Código Soundex)
R1	Smith	S530
R2	Miller	M460
R3	Peters	P362
R4	Myler	M460
R5	Smyth	S530
R6	Millar	M460
R7	Smyth	S530
R8	Miller	M460

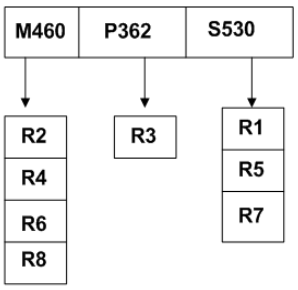

Figura 2.1: Bloqueamento baseado em índice invertido. Adaptado de [Christen, 2012]

A abordagem baseada na aplicação de função *hash* sobre a chave de bloqueamento, apresenta limitações para o caso de emparelhamentos aproximados, uma vez que o *hash* de dois registos aproximados resulta em uma codificação diferente. Uma das alternativas para estes casos é o uso de algoritmos como

o *Soundex*, *NYSIIS* ou o *Metaphone* em atributos com maior discriminação, como é o caso dos nomes e apelidos para facilitar a comparação dos campos similares [Elmagarmid et al., 2007].

2.1.2.B Sorted Neighbourhood Blocking

Este método foi desenvolvido por [Hernández and Stolfo, 1995], é baseado inicialmente na combinação de duas ou mais relações em uma lista sequencial com N registos, distinguidos por uma *flag* que indica a relação de proveniência. De seguida são aplicados à lista três passos do método *Sorted Neighbourhood Blocking*:

- **Criação das chaves:** baseia-se na computação da chave de cada registo, criada com base em atributos relevantes ou porções dos atributos do registo;
- **Ordenação dos dados:** é feita alfabeticamente com base na chave de bloqueamento criada no passo anterior, com o intuito de garantir que os dados equivalentes estejam muito próximos uns dos outros na lista ordenada;
- **Geração de pares candidatos:** neste passo considera-se uma janela com tamanho fixo $w > 1$, a qual é movida sobre a lista sequencial dos registos, e os pares candidatos são gerados a partir dos registos que se encontram na janela num dado passo. É necessário garantir que em cada par gerado, exista um registo de cada uma das relações.

Considerando w como o tamanho da janela, os novos registos a serem inseridos na janela formarão pares com os w-1 registos anteriores, conforme ilustra a Figura 2.2. Entretanto, cada par único será comparado uma única vez na etapa de comparação.

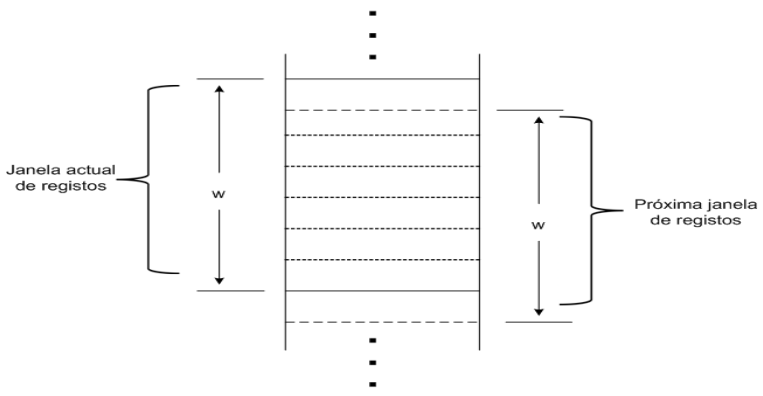

Figura 2.2: Movimentação da Janela durante a geração de candidatos. Adaptado de [Hernández and Stolfo, 1995].

[Christen, 2011] refere que nesta abordagem, o número de posições da janela é denotado por $(n_a + n_b - w + 1)$, onde na e nb representam os números de registos contidos nas relações a e b.
O número de pares candidatos gerados na primeira posição da janela é equivalente a $\frac{n_a \times n_b}{(na+nb)^2} \times w^2$, enquanto que para as restantes posições, são gerados $\frac{na \times nb}{(na+nb)^2} \times 2(n_a+n_b-w)(w-1)$ pares candidatos.
O número total de pares candidatos únicos é estimado pela seguinte equação:

$$PC = \frac{n_a \times n_b}{(n_a + n_b)^2}(w^2 + 2(n_a + n_b - w)(w - 1)) \tag{2.1}$$

Durante a execução de cada posição da janela sobre a lista sequencial, vários potenciais candidatos a emparelhamento podem não estar próximos um do outro considerando os casos em que o tamanho da janela seja insuficiente para cobrir todos os candidatos.
Uma abordagem referida em [Christen, 2011] consiste na utilização de uma estrutura de dados baseada em índice invertido, onde as chaves de índice referem-se aos valores das chaves de bloqueamento ordenados alfabeticamente. A janela é movida sobre os valores das chaves de bloqueamento e os pares candidatos são formados para todos os registos contidos nas listas de índice correspondentes.
Tal como referido anteriormente, nesta abordagem os pares candidatos únicos serão igualmente comparados uma única vez na etapa de comparação.
O número de posições da janela é denotado por $(b - w + 1)$, onde b representa o número de valores da chave de bloqueamento. Considerando uma distribuição uniforme das chaves de bloqueamento, cada lista de índice invertido conterá $\frac{n_a}{b} + \frac{n_b}{b}$ identificadores de registos.
O número de pares de registos candidatos é estimado pela seguinte equação:

$$PC = \frac{n_a \times n_b}{b^2}(w^2 + (b - w)(2w - 1)) \tag{2.2}$$

2.1.3 Métricas de Emparelhamento de Strings

Nesta subsecção, farei menção de algumas técnicas de comparação de atributos ou registos baseadas em medidas de similaridade, partindo do pressuposto de que os atributos ou registos a serem comparados referem-se a um par de Strings. Estas métricas são comumente designadas como Métricas de *String Matching*.
[Doan et al., 2012] define o problema de *String Matching* (emparelhamento de cadeias de caracteres em português) usando dois conjuntos de Strings X e Y, em que se pretende encontrar todos os pares de Strings (x, y), onde $x \in X$ e $y \in Y$, de tal modo que x e y representam a mesma entidade no mundo real. Estes pares são denominados por emparelhamentos [Doan et al., 2012].
Para resolver um problema de emparelhamentos, é necessário considerar dois grandes desafios : exatidão e a escalabilidade [Doan et al., 2012, Christen, 2012].

O desafio de atingir a maior exatidão possível surge devido ao facto de que as *Strings* que representam a mesma entidade no mundo real, em algumas vezes podem diferir em algum aspecto, nomeadamente erros de escrita ou abreviações. Como exemplo, consideremos duas *Strings* Conceição e Conceicao que se referem a nomes e representam a mesma entidade no mundo real, mas diferem-se pelo modo de escrita.

Tanto o problema do exemplo como qualquer outro *String Matching* no geral é solucionado aplicado medidas de similaridade entre os pares de *Strings*, cujo o resultado é um valor no intervalo $[0, 1]$. Similaridades com valores mais próximos de 1, garantem maior probabilidade de serem emparelhamentos. Para alguns casos, é usado $s(x, y) \geq t$, onde $s(x, y)$ representa a medida de similaridade entre x e y, e t representa o *threshold* (limiar em português) cujo o intuito é definir o limite inferior de classificação para os pares de Strings que garantem maior probabilidade de serem a mesma entidade no mundo real. Neste caso seriam considerados como matches os pares cuja similaridade é igual ou superior a t. O desafio da escalabilidade foi abordado na Secção 2.1.2. A seguir, farei menção de três medidas de similaridade, nomeadamente a de Jaccard, distância de Levenshtein e Jaro-Winkler.

1. **Distância de Levenshtein:** é denotada por uma função $d(x, y)$ que representa o custo mínimo para transformar uma *String* x para outra y [Rieck, 2011]. A transformação é realizada atribuindo um custo unitário para cada operação de edição efectuada, nomeadamente a inserção, eliminação e a substituição de um caracter por outro [Doan et al., 2012]. Nesta métrica, quanto menor for a distância de edição entre as duas Strings maior é a similaridade entre as mesmas. A distância de edição pode ser usada para capturar erros de escrita cometidos durante a inserção dos dados, tais como a inserção de caracteres adicionais, troca da posição entre os caracteres, ou ainda nos casos da falta de acentuação de caracteres.

   A função $d(x, y)$ pode ser convertida em uma função de similaridade $s(x, y)$ denotada por:

   onde, $d(x, y)$ representa o custo mínimo das operações de edição, e pode ser calculada usando a programação dinâmica; e o $max(|x|, |y|)$ representa o valor máximo dentre os comprimentos das Strings x e y respectivamente.

   O algoritmo da programação dinâmica é baseada numa matriz, onde a célula $d[i, j]$ na linha i $(0 \leq i \leq |x|)$ e coluna j $(0 \leq j \leq |y|)$, corresponde ao número de operações de edição necessárias para converter os primeiros i caracteres da String x para os primeiros j caracteres da String y. O algoritmo é executado recursivamente mediante a uma equação de recorrência denotada por:

$$d(x, y) = \text{min } d(i-1, j-1), \text{ se } x_i = y_j. \text{ / cópia}$$
$$d(x, y) = \text{min } d(i-1, j-1) + 1, \text{ se } x_i \neq y_j. \text{ / Substituição do caractere de } x \text{ pelo caractere de } y$$

$d(x, y)$= min $d(i-1, j)+1$. / Elimina-se o caractere x_i

$d(x, y)$= min $d(i, j-1)+1$, / inserção do caractere y_i na nova string

2. **Jaro-Winkler:** é uma família de medidas de similaridade criada especificamente para a comparação de pequenas Strings tais como nome, apelido e nome de ruas [Winkler, 1994].
   A medida de Jaro é calculada do seguinte modo: considere x, y como duas Strings, e i, j como as posições de cada caractere de x e y respectivamente.

   - Encontre os caracteres comuns entre x_i e y_j tal que $x_i = y_j$ e $|i-j| <= min|x|, |y|/2$ em caso dos mesmos se encontrarem em posições muito próximas entre i e j.
     O número de caracteres comuns é denotado por c.
   - Compare os caracteres comuns na posição i de x, e j de y. Caso não sejam iguais, então existe uma transposição. O número de transposições é denotado por t.
   - Calcule a medida de Jaro considerando a fórmula:

   $$jaro(x, y) = \frac{1}{3}(\frac{c}{|x|} + \frac{c}{|y|} + (c - \frac{t}{2})/c) \tag{2.3}$$

   Para os casos em que o valor resultante do cálculo da similaridade entre as Strings aplicando a métrica de *Jaro* seja muito abaixo do esperado mas possuem prefixos comuns, diferindo apenas nos caracteres finais ou medianos das Strings, surge a extensão da métrica de Jaro denominada *Jaro-Winkler*, que introduz dois parâmetros: o comprimento do prefixo comum mais longo entre as Strings denotado por $PL(0 \leq PL \leq 4)$, e o peso atribuído ao prefixo, denotado por PW, que tem o valor padrão de 0.1.
   A fórmula de jaro-winkler é expressa por:

   $$jaro - winkler(x, y) = (1 - PL * PW) * jaro(x, y) + PL * PW \tag{2.4}$$

3. **Jaccard:** considere A_x e B_y como dois conjuntos contendo vários pedaços denomonados *tokens*, gerados a partir das Strings x e y.
   A similaridade de Jaccard entre as Strings x e y é denotada por:

   $$J(x, y) = \frac{|B_x \cap B_y|}{B_x \cup B_y|}, \tag{2.5}$$

   onde, $|B_x \cap B_y|$ corresponde ao número de tokens comuns entre as duas *S*trings; e $|B_x \cup B_y|$ corresponde ao conjunto de $tokens$ que formam as *Strings* x e y.

Consideremos o exemplo referido por [Doan et al., 2012]: Seja x=dave, y=dav e $O(x, y) = |B_x \cap B_y|$; e sobre cada uma delas são extraídos todos os conjuntos de 2-gramas ou *tokens*, que formam os conjuntos $B_x = \{\#d, da, av, ve, e\#\}$ e $B_y = \{\#d, da, av, v\#\}$.
Assim sendo, $O(x, y) = 3$, $|B_x \cup B_y| = 6$ e a medida de *Jaccard* resultará em $J(x, y) = \frac{3}{6}$.

2.1.4 Classificação de Registos

[Doan et al., 2012] define o problema de classificação considerando duas relações X e Y com esquemas idênticos, onde cada registo em X e Y descrevem propriedades de uma entidade (por exemplo, indivíduo, livro ou filme). Considera-se que $x \in X$ emparelha com $y \in Y$ se os mesmos referirem a mesma entidade no mundo real. O principal propósito da Classificação consiste em encontrar todos os emparelhamentos entre as relações X e Y.
Enquadrando os aspectos abordados nas secções anteriores, importa referir que a Classificação permite Classificar em várias categorias ou *Classes* (duas ou mais) os pares candidatos gerados por técnicas de indexação e detalhadamente comparadas usando métricas de similaridade [Christen, 2012].
A classificação de registos é um processo realizado em duas etapas [Han and Kamber, 2006]:

1. É feita a construção do Classificador ou modelo mediante a aprendizagem ou treino de um conjunto de registos também designados por exemplos, amostras ou objectos selecionados a partir da base de dados em análise. Estes exemplos possuem uma etiqueta associada que representa a classe ou categoria em que pertencem. Por exemplo: 1 (*Match*) ou 0 (*Not-match*).
   Dado o facto de cada exemplo estar associado uma etiqueta que representa a classe, esta etapa é igualmente designada por *aprendizagem supervisionada*. Isto significa que a aprendizagem do classificador é supervisionada. Para além deste método de aprendizagem, existe o método não supervisionado (*clustering*) onde a etiqueta da classe não é conhecida e o número ou conjunto de classes a serem treinadas não é conhecido atempadamente. Neste caso, são usadas técnicas como por exemplo o *Clustering* para permitir agrupar os tuplos que partilham alguma semelhança.

2. O modelo treinado é usado para a classificação de registos cuja classe é desconhecida. Antes disso, é estimada a exactidão do modelo. Para tal, são usados um conjunto de registos etiquetados para a etapa de teste. Estes registos diferem-se dos usados para o treino, afim de evitar que a estimativa seja otimista, e tenha uma baixa performance quando usado em registos não classificados (*overfiting*). Sendo assim, o recomendado é seleccionar outros registos de forma aleatória a partir da base de dados em análise.
   A exactidão do classificador num dado conjunto de teste corresponde a percentagem dos tuplos que foram corretamente classificados pelo Classificador. A etiqueta da classe em cada exemplo de teste é comparada com a previsão atribuída pelo classificador para o referido exemplo. Caso

a exactidão for considerada aceitável, o classificador pode ser usado para classificar os registos cuja *classe* é desconhecida.

Na literatura sobre Record Linkage, Data Mining e Machine Learning [Doan et al., 2012, Han and Kamber, 2006, Christen, 2012, Hastie et al., 2009] são abordadas várias técnicas usadas para a classificação de registos, como por exemplo as Árvores de Decisão, Naive Bayes, Belief Networks, Regressão Logística e Support Vector Machine.

2.1.5 Fusão de Dados

Considere duas relações A e B com um esquema comum, contendo os seguintes atributos: **BI, Nome, Sexo, Data_nasc e Morada**, cada uma com um único registo. A Fusão de Dados ocorre quando deseja-se integrar as duas relações, com o intuito de manter uma única representação para cada entidade no mundo real, tal como é apresentado na Tabela 2.1.

Tabela 2.1: Exemplo de Fusão de dados

Relação	BI	Nome	Sexo	Data_nasc	Morada
A	25564	Ana Celeste	F	21-04-1976	Barcarena
B	25564	Ana Celeste	F	21-04-1976	Barcarena
Fusão	25564	Ana Celeste	F	21-04-1976	Barcarena

Regra geral, quando os dados provêm de fontes heterogêneas, podem surgir dois tipos de inconsistências durante o processo de fusão de dados [Bleiholder and Naumann, 2006]:

1. **Inconsistências ao nível dos esquemas:** ocorrem quando as relações não possuem os mesmos atributos, atributos com a mesma semântica mas com nomes diferentes ou ainda casos em que os dados são armazenados em diferentes estruturas.

2. **Inconsistências ao nível dos dados:** ocorrem quando existem conflitos entre dois ou mais valores usados para descrever a mesma propriedade de uma entidade no mundo real.

No caso presente, as insconsistências ao nível dos esquemas não representa um problema, uma vez que os esquemas das fontes distintas vêm com as variáveis requeridas pelo INE. Nesta secção, farei menção dos conflitos que ocorrem ao nível dos dados e as estratégias existentes para o tratamento dos mesmos. Dentre os conflitos destacam-se dois tipos: **contradições e incertezas**.
Considere novamente o exemplo da Tabela 2.1, supondo algumas alterações nos atributos BI, Sexo e Morada, tal como é apresentado na Tabela 2.2.

Tabela 2.2: Exemplo de conflitos na fusão de dados

Relação	BI	Nome	Sexo	Data_nasc	Morada
A	25564	Ana Celeste	F	21-04-1976	Barcarena
B	25664	Ana Celeste	NULL	21-04-1976	Areeiro
Fusão	?	Ana Celeste	?	21-04-1976	?

Estamos perante uma contradição quando existe um conflito entre dois ou mais valores diferentes, usados para descrever uma mesma propriedade de uma entidade [Bleiholder and Naumann, 2006]. Como exemplo, é apresentado na Tabela 2.2 a contradição existente entre os números de BI e a morada da Cidadã Ana Celeste.

Estamos perante uma incerteza, quando existe um conflito entre um valor não nulo e um ou mais valores nulos, ambos usados para descrever a mesma propriedade de uma entidade [Bleiholder and Naumann, 2006]. Como exemplo, é apresentado na Tabela 2.2 a incerteza existente entre os valores do par de atributo Sexo da Cidadã Ana Celeste.

Para resolver estes tipos de conflito são aplicadas estratégias de tratamento de conflitos, classificadas em três grupos [Bleiholder and Naumann, 2006] tal como ilustra a Fig 2.3.

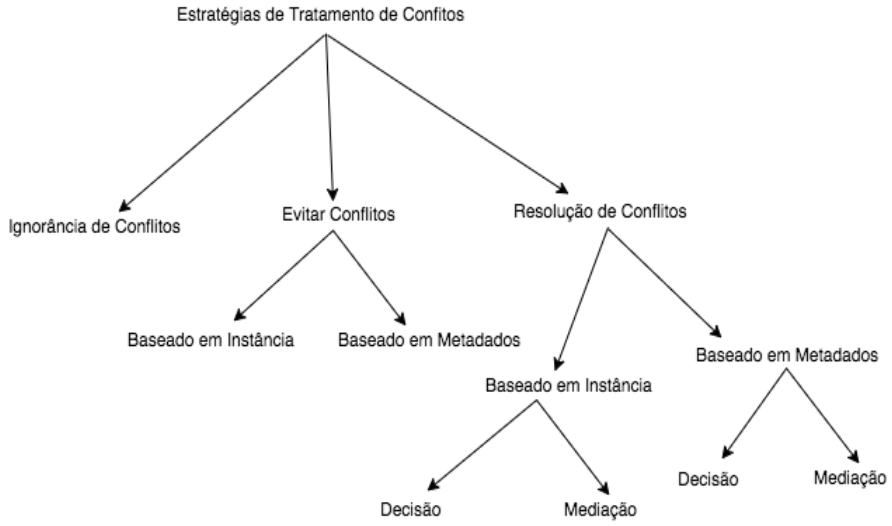

Figura 2.3: Classificação das estratégias para Fusão de Dados

A seguir, farei menção de forma resumida as estratégias existentes em cada um dos grupos:

1. **Ignorar conflitos:** inclui estratégias em que a resolução dos conflitos é dependente de uma acção humana ou de um software específico. Neste grupo, destacam-se duas estratégias [Bleiholder and Naumann, 2006]:

(a) **Pass it on:** consiste em transferir todos os valores em conflito para um utilizador ou um software específico para que estes decidam o modo de tratamento dos mesmos.

(b) **Consider all possibilities:** fornece ao utilizador todas as combinações de valores possíveis que um atributo pode tomar dando-o a liberdade de escolha.

2. **Evitar conflitos:** inclui estratégias que não resolvem os conflitos de forma individual, mas ainda assim tratam das inconsistências nos dados. Para a tomada de decisão, as estratégias deste grupo encontram-se subdivididas em dois subgrupos: As estratégias baseadas em instâncias e as baseadas em metadados.
   A seguir abordo cada uma das estratégias:

   (a) **Take the Information:** é uma estratégia baseada em instâncias, tem por objectivo resolver o conflito da incerteza excluindo os valores nulos do conjunto de resultados.

   (b) **No Gossiping:** é uma estratégia baseada em instâncias, tem por objectivo retornar como conjunto de resultados apenas os registos consistentes, deixando de parte os inconsistentes.

   (c) **Trust your friends:** é uma estratégia em que a tomada de decisão é baseada nos metadados das fontes de dados envolvidas no processo de fusão. Uma vez tomada a decisão de que fonte de dados usar, as acções são refletidas para todos os valores não levando em conta se existem ou não conflitos entre os mesmos.

3. **Resolução de conflitos:** inclui estratégias consideram todos os dados e os metadados das fontes de dados antes de decidir o método de resolução de conflitos a ser aplicado. Em contraste com os grupos de estratégias referidos em 1 e 2, esta subdivide-se ainda em dois subgrupos [Bleiholder and Naumann, 2006], nomeadamente: Decidir e Mediar.
   As estratégias de decisão dependem unicamente do valor dos dados ou dos metadados existentes a fim de escolher o valor adequado para a resolução do conflito. Este subgrupo inclui as seguintes estratégias:

   (a) **Cry with the wolves:** baseia-se na escolha do valor mais frequente, ou seja o que é reportado pela maioria das fontes de dados [Cecchin et al., 2010].

   (b) **Roll the dice:** considera todos os valores em conflito e escolhe um de forma aleatória.

   Por outro lado, as estratégias de mediação podem escolher valores que não estejam necessariamente entre os conflituosos, mas podem ser incluídos valores que não existiam anteriormente [Bleiholder and Naumann, 2006]. Este grupo inclui a seguinte estratégia:

   (a) **Meet in the middle:** esta estratégia segue o princípio do compromisso, não preferindo um valor dentre todos os existentes, mas trata de inventar um valor aproximado aos valores

presentes. Outro princípio que pode ser usado baseia-se no cálculo da média dos valores [Bleiholder and Naumann, 2006].

A estratégia **Keep up to date** é referida em [Bleiholder and Naumann, 2006] como uma estratégia de decisão baseada em Metadados, uma vez que utiliza o valor mais recente adicionado com uma informação de timestamp para garantir a sua recência. A informação de recência pode ser armazenada nas próprias tabelas como um atributo separado ou pode ser obtida com base na informação de proveniência.

2.1.6 Qualidade de Dados

Os processos realizados numa data warehouse exigem uma elevada qualidade dos dados nela contidos. O contrário leva a conclusões erradas destes processos, o que resultarão em grandes perdas ao nível de tomadas de decisão [Kulkarni and Bakal, 2014].
Os dados que provêm de fontes externas normalmente contêm erros ortográficos, campos em falta ou ainda diferentes formatos quer ao nível dos esquemas como ao nível da representação dos dados. A solução para este problema parte pelo investimento de tempo e dinheiro em processos de limpeza de dados a fim de garantir algum nível de qualidade.
[Kulkarni and Bakal, 2014] define a qualidade de dados como o grau em que os dados atendem às necessidades específicas de clientes específicos, que contém várias dimensões.
A qualidade dos dados é avaliada com base em diferentes aspectos denominados dimensões. A seguir faço referência de algumas dimensões a serem usadas neste trabalho:

- **Exatidão:** avalia o nível de proximidade que existe entre dois valores (v e v´) considerados como correctos no mundo real [Batini and Scannapieco, 2016]. Considere os seguintes elementos: VP (verdadeiros positivos), VN (verdadeiros negativos), FP (Falsos positivos) e FN(Falsos negativos). A exactidão é medida com base na seguinte expressão:

$$Exactidao = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.6}$$

Nesta dimensão, podem ainda ser consideradas outras medidas, nomeadamente:

1. **Precisão:** considera entre todos os pares de registos classificados, quantos o modelo emparelhou correctamente. É denotada pela fórmula:

$$Precisao(R) = \frac{VP}{VP + FP} \tag{2.7}$$

Onde,

- VP : representam os verdadeiros emparelhamentos (refere-se aos registos que o classificador classificou correctamente como emparelhados)
- FP : representam os falsos emparelhamentos (refere-se aos registos que o classificador classificou incorretamente como emparelhados)

2. **Abrangência:** considera entre todos os registos que deveriam ter sido emparelhados, quantos o modelo conseguiu de facto emparelhar correctamente. É denotada pela fórmula:

$$Abrangencia(R) = \frac{VP}{VP + FN} \tag{2.8}$$

Onde,

- FN : representam os falsos não emparelhados (refere-se aos registos que o classificador classificou incorretamente como não emparelhados)

- **Completude:** avalia a taxa de preenchimento de um determinado atributo x numa relação R.
  No âmbito deste trabalho,considere $CP_k(x)$ como a completude de um atributo x num dado registo k, onde k corresponde a ordem do registo no intervalo de $[1..n]$. $CP_k(x)$ toma o valor 1 (caso o atributo esteja preenchido) ou 0 (caso contrário).
  A completude total da relação referente ao atributo x é denotada por $CP_R(x) = \frac{\sum_{k=1}^{n} CP_k(x)}{N}$, onde N corresponde ao total de registos observados na relação.

- **Conformidade:** avalia a taxa de conformidade no cumprimento dos padrões estipulados para os atributos dos registos nas fontes de dados.
  No âmbito deste trabalho,considere $CP_k(x)$ como a conformidade de um atributo x num dado registo k, onde k corresponde a ordem do registo no intervalo de $[1..n]$. $C_k(x)$ toma o valor 1 (caso satisfaça os padrões estabelecidos) ou 0 (caso contrário).
  A conformidade total da relação referente ao atributo x é denotada por $C_R(x) = \frac{\sum_{k=1}^{n} C_k(x)}{N}$, onde N corresponde ao total de registos observados na relação.

- **Inconsistência:** avalia taxa de violação de regras semânticas definidas sobre itens de dados, onde os itens referem-se a pares de atributos de uma relação R.
  No âmbito deste trabalho, a inconsistência é medida com base nas incertezas e contradições existentes na relação. Considere $IN_k(x, y)$ como a incerteza de um par de atributos x e y num dado registo k, onde k corresponde a ordem do registo no intervalo de $[1..n]$. $IN_k(x)$ toma o valor 1 (em caso de incerteza) ou 0 (caso contrário).
  O total de incertezas da relação referente ao par de atributo x e y é denotada por $IN_R(x, y) = \frac{\sum_{k=1}^{n} IN_k(x)}{N}$, onde N corresponde ao total de registos observados na relação. A taxa de contradição é medida de modo similar.

2.2 Sistemas Similares

Nesta secção abordo alguns trabalhos realizados e em curso no domínio de emparelhamento de registos. Farei menção dos mesmos subdividindo-os com base nas diferentes etapas de um processo de emparelhamento comum.

Inicialmente, importa referir que a criação de uma óptima chave de bloqueamento representa um dos grandes desafios na etapa de Bloqueamento. Para a criação das chaves, são utilizados métodos aplicados aos atributos ou ainda a partes de atributos, o que se designa por Blocking Schema (BS) ou esquema de bloqueamento, e é representado por $BS = \{metodo, atributo\}$.

Uma das técnicas usadas para obter uma qualidade considerável de pares de registos candidatos consiste em combinar os resultados de várias execuções da técnica *sorted neighborhood blocking* (Multi-pass approach) [Hernández and Stolfo, 1995] com uma chave de bloqueamento diferente em cada execução.

[Michelson and Knoblock, 2006] aborda uma alternativa baseada em *machine learning* para a obtenção de um esquema de bloqueamento efectivo baseando-se na aprendizagem automática de vários esquemas, o que se designa por regras; e o esquema efectivo resulta da disjunção das várias regras obtidas. A abordagem usada nesta alternativa consiste na modificação do algoritmo *Sequential Covering* (SCA) [Riddle, 1997], que ao invés de retornar a disjunção de todas as regras aprendidas, mantém apenas as regras mais específicas deixando de parte as menos específicas dentro do conjunto de regras anteriormente aprendidas.

Como exemplo, considere BS como esquema de bloqueamento, e duas regras constituídas por conjunções: $(\{3-prim\_letras\_nome, nome\} \wedge \{match\_anterior, data\_nascimento\})$ e $(\{3prim\_letras\_nome, nome\} \wedge \{match\_anterior, cod\_fregesia\})$. Supondo que o treino da regra 2ª inclua todos os tuplos obtidos pela 1ª regra acrescido a pelo menos 3 tuplos, a segunda regra considerar-se-á mais específica do que a primeira. Se não houver mais nenhum melhoramento nas regras a serem adicionadas, o esquema de bloqueamento final é definido por: $BS = (\{3-prim\_letras\_nome, nome\} \wedge \{match\_anterior, cod\_fregesia\})$

Mas, caso seja criada uma regra cujo os resultados sejam diferentes, quase que disjuntos, o esquema de bloqueamento corresponderá a união (disjunção) dos mesmos.

[Michelson and Knoblock, 2006] avaliam a qualidade do esquema efectivo baseando-se na aprendizagem de regras que minimizem os falsos positivos e cobrem um número suficiente de verdadeiros positivos. As métricas usadas são as seguintes:

Taxa de Redução: avalia o quanto um esquema de bloqueamento reduz o número de pares candidatos, denotada por: $TR = 1 - \frac{C}{N}$, onde C representa o número de candidatos e N representa o produto cartesiano das relações consideradas.

Completude do Par: avalia a cobertura dos verdadeiros emparelhamentos com base na quantidade dos verdadeiros positivos encontrados no conjunto dos pares candidatos e a quantidade de positivos

extraídos do produto cartesiano, denotada por: $CP = \frac{S_m}{N_m}$

Para avaliação do método, foi feita uma comparação com as abordagens dos esquemas de bloqueamento usadas por Marlin System [Bilenko and Mooney, 2003], Hybrid-Field Matcher (HFM) [Minton et al., 2005] e Winkler [Winkler, 2005]. O referido experimento aponta que a abordagem de [Michelson and Knoblock, 2006] apresenta uma taxa de redução (RR) bastante superior na ordem dos 99,26% e 99,86% contra 55,35% e 47,92% e uma completude do par (CP) a uma taxa de 98,16% e 99,92% contra 100% e 99,97%. Isto significa que os métodos usados por Marlin System e Hybrid-Field Matcher procuram maximizar a Completude do Par mas geram um elevado número de pares candidatos.

Por outro lado, as cinco melhores conjunções consideradas por Winkler para os dados dos Censos denominados "best five" blocking schema [Winkler, 2005], resultou numa taxa de redução superior comparado ao esquema de bloqueamento proposto por [Michelson and Knoblock, 2006] na ordem de 99,52% contra 98,12% e uma completude do par inferior na ordem de 99,16% contra 99,85%.

Na vertente dos Censos, são notáveis os trabalhos realizados por [Winkler, 2005], [Feigenbaum, 2016], [Winkler, 2006], [Winkler, 1995] e dos que se encontram em curso importa referir alguns tais como a Nova Zelândia [March et al., 2015], Noruega [Harald, 2000], Canadá [Trépanier et al., 2013] e o Reino Unido, o qual faremos referência mais detalhadamente.

A Office for National Statistics (ONS) é a entidade responsável pelos estudos estatísticos no Reino Unido. A mesma tem disponível no âmbito dos Censos cinco principais fontes de dados administrativas [Office for National Statistics, 2015] e outras mencionadas em [Office for National Statistics, 2016c], cujo o principal objectivo consiste na integração destas fontes que resultará na Statistical Population Dataset (SPD) (base de dados similar a BPR), que é complementada pela Population Coverage Survey (PCS) para ajustamentos e cobertura de erros para permitir a produção de estimativas da população residente na Inglaterra e Gales.

Uma vez que as referidas fontes de dados não possuem um identificador comum, o emparelhamento é feito com base nos atributos quase-identificadores dos indivíduos [Office for National Statistics, 2014a]. A abordagem para a produção das estimativas da população é apresentada em [Skinner et al., 2013] e comporta os seguintes elementos: duas ou mais fontes de dados à entrada, um processo de emparelhamento, um processo de aplicação de regras aos registos emparelhados e como resultado obtém-se a SPD.

Paralelamente a estes elementos, existe uma componente de avaliação da qualidade que interage com as fontes de dados, e os processos de emparelhamento e de aplicação das regras. As estimativas são obtidas integrando os resultados da SPD com os da PCS.

A abordagem usada para o emparelhamento das fontes é constituída por quatro etapas, nomeadamente pré-processamento, importação dos dados para o Statistical Research Environment (SRE), emparelha-

mento por regras e o emparelhamento Probabilístico.
Os registos a emparelhar são primeiramente encriptados com um hash SHA-256 na fase de pré processamento, para garantir que os dados carregados para o SRE estejam completamente anonimizados para posteriores emparelhamentos com métodos determinísticos ou probabilístico.
Aplicando esta política, torna-se difícil o emparelhamento exacto dos registos similares, uma vez que o hash dos mesmos nunca seria o mesmo. A alternativa aplicada pelos autores consiste na criação de várias chaves de emparelhamento construídas por vários pedaços de informação a fim de serem obtidas chaves únicas que permitam o emparelhamento um para um. Um exemplo da chave de emparelhamento refere-se a: 3 primeiras letras do nome e do apelido combinado com a data de nascimento, sexo e o código postal do distrito de residência do indivíduo.

A alternativa das chaves de emparelhamento capturam em média 95% dos emparelhamentos disponíveis.Uma das formas usadas para tentar reduzir o número de falsos positivos consiste em emparelhar apenas os registos cuja a chave é única nas duas fontes de dados.
Em caso de haverem múltiplos registos com a mesma chave, o emparelhamento não é efectuado. Estes são considerados como residuais e passam para a etapa da classificação por método probabilístico.
Ao considerar a aplicação de métodos probabilísticos, surge a limitação de serem aplicadas medidas de similaridade ao par de Strings encriptados. A alternativa usada para este caso consiste na criação de tabelas de similaridade na etapa de pré-processamento para cada par de atributos nome, apelido e data de nascimento de cada indivíduo.
Os potenciais candidatos são gerados com base no resultado do total de similaridade obtido nas três tabelas de comparação. Uma vez calculadas as similaridades, os registos são encriptados e importados para o SRE.

O método probabilístico escolhido para classificar os potenciais candidatos é o método supervisionado de regressão logística por parecer uma das maneiras de desenvolver uma classificação binária com uma única pontuação para um dado threshold. O método serviu primeiramente para calcular os pesos de cada variável a emparelhar e de seguida para classificar-los com base num threshold de 0.5 [Office for National Statistics, 2014a].
As variáveis de previsão para o modelo foram: as pontuações do nome e do apelido, os códigos de concordância da data de nascimento, sexo e código postal, os pesos do nome, apelido, a quantidade de nomes (combinação do nome e apelido) e a distância entre os centróides do Lower Super Output Area Mid-Year Population Estimates (LSOA) (fonte de dados que contém a estimativa da população por idade e sexo, residentes numa determinada área da Inglaterra e Gales num total da população de 1.000 a 3.000 pessoas e uma média de 1.600 pessoas) [Office for National Statistics, 2014b, Office for

National Statistics, 2016b]

A [Office for National Statistics, 2014a] refere que esta técnica foi aplicada a uma série de conjuntos de dados de teste com resultados consistentes alcançados. Nos conjuntos de dados de treino, o modelo foi capaz de prever uma correspondência em emparelhamentos com a decisão clerical em aproximadamente 97% dos casos. Quando as equações do modelo são executadas em uma amostra independente de pares, que são manualmente emparelhados por comparação, o nível de concordância é geralmente superior a 95%.

No início do Beyond 2011 agora designado por Census Transformation Programme (CTP) (Projecto criado com a finalidade de coordenar os Censos 2021 no Reino Unido) [Office for National Statistics, 2016a], foi realizado um exercício de comparação para comparar 10.000 registros de estudantes do conjunto de dados do Higher Education Statistics Agency (HESA) com o NHS Patient Register (PR). O objectivo deste exercício consistiu em avaliar se a abordagem da integração de dados anonimizados era ou não viável. Segundo [Office for National Statistics, 2014a], os resultados foram encorajadores, concluindo assim que era viável o emparelhamento de fontes anonimizadas.

3

Abordagem para o Emparelhamento de Registos do INE

Conteúdo

Neste capítulo, apresento a abordagem usada para o emparelhamento de registos do INE, que recorre a métodos de aprendizagem automática (probabilísticos). É realizado por meio de um conjunto de *scripts* de software que constituem o *Módulo de Produção da Informação (MPI)*.

O MPI tem por objectivo produzir indícios de residência da população por meio do emparelhamento sucessivo dos vários pares de fontes de dados disponíveis no INE. Produz como resultado a *Matriz Base da População*, uma relação que contém o registo dos indícios de residência respeitantes a cada indivíduo potencialmente residente em Portugal em cada ano. É a partir desta *Matriz* que o INE estabelece a sua *Base de População Residente* (BPR).

A Figura 3.1 ilustra, na notação Business Process Model and Notation (BPMN)[1], o processo de geração e actualização da BPR, as principais actividades envolvidas e resultado final do processo. Sem perda de generalidade, descrevo neste capítulo o processo usado para o emparelhamento dos registos das fontes BDIC (relação que contém o registo de todos os cidadãos de nacionalidade Portuguesa e Brasileira com estatuto de porto seguro com morada de residência em Portugal ou morada desconhecida) e AT (relação referente aos cidadãos que apresentam a declaração de IRS), provenientes do IRN e AT, respectivamente. Para outros pares de fontes, o processo é similar.

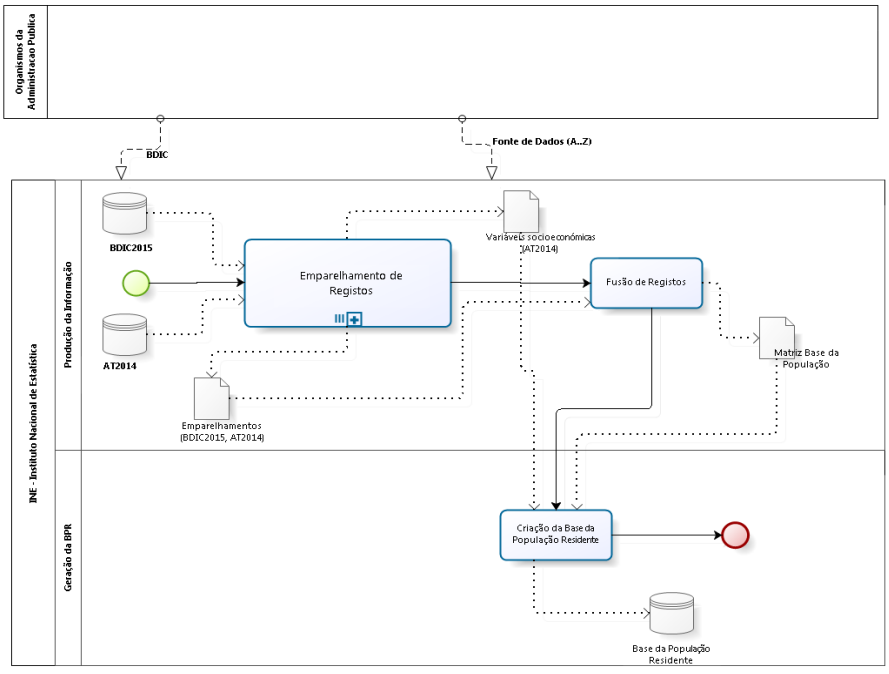

Figura 3.1: Processo de Geração da BPR com as Fontes de Dados BDIC e AT

[1] http://www.omg.org/spec/BPMN/2.0.2/

O MPI é constituído por dois sub-módulos que correspondem a duas actividades que se executam sequencialmente: o *Emparelhamento de Registos* e a *Fusão de Registos*.
A actividade Emparelhamento de registos funciona como sintetizado na Tabela 3.1.

Tabela 3.1: Síntese da actividade de emparelhamento de registos

Actividade	Emparelhamento de Registos
Entrada	1. Par de relações (por ex: BDIC , AT)
Saída	1. Relação com os emparelhamentos produzidos, cujo esquema possui os pares de atributos demográficos e geográficos de cada emparelhamento.
Funcionalidade	Preencher a relação de saída com os emparelhamentos efectuados.

Primeiramente, os dados das fontes são carregados para o SGBD Oracle do INE e armazenados em tabelas com o esquema originalmente recebido, e que é variável de fonte para fonte.
Para simplificação da gestão de informação, é também útil dispor dos dados de todos os anos de residência normalizados e armazenados em tabelas, uma para cada fonte. Por exemplo, no caso da BDIC e AT, os registos de todos os anos são armazenados em tabelas normalizadas com o mesmo nome.
Foi definido um esquema normalizado para processar com os mesmos *scripts* todas as fontes de informação. Para a criação do esquema, criámos um *script* em *Python* (*esquema_basedados.py*) que produz um *script* em Structured Query Language (SQL), a fim de ser executado no SGBD e criar todas as tabelas necessárias à execução do processo de emparelhamento de todas as fontes.

Para o emparelhamento dos pares de relações de um dado ano, é feita uma selecção dos registos sobre estas relações com base no ano de residência. Por exemplo, para o cenário em curso foram emparelhados os registos da relação BDIC do ano 2015 com os registos da AT do ano 2014.
Uma vez realizados os emparelhamentos das fontes BDIC e AT, surge a necessidade de manter um único registo associado a cada indivíduo no mundo real. Esta actividade é denominada, tal como o módulo correspondente, *Fusão de Registos*. Tem como resultado uma relação que permitirá adicionar novos registos à *Matriz Base da População* ou atualizá-la com novos identificadores pessoais. O módulo de *Fusão de Registos* não foi implementado no âmbito deste trabalho, uma vez que seria trivial, dado que o INE já estabeleceu uma estratégia própria para a Fusão de registos, baseada na selecção com base numa hierarquia estabelecida de confiabilidade das fontes.

Aos registos presentes na *Matriz Base da População* serão aplicados regras de residência estabelecidas pelo INE a fim de decidir se os mesmos serão adicionados ou não para a BPR [INE,Gabinete dos Censos 2021, 2016a]. Este processo encontra-se ficticiamente representado na Figura 3.1 como

Geração da BPR, um módulo de software que permite executar a actividade de *Criação da Base da População Residente*.

A Figura 3.2 representa, na notação BPMN, as várias sub-actividades da actividade de Emparelhamento de Registos, assim como a relação entre as mesmas durante o emparelhamento de registos. As secções seguintes descrevem cada uma das actividades, assim como o modo como cada uma foi implementada no âmbito deste trabalho:

1. Limpeza e Normalização,
2. Blocking,
3. Comparação de registos,
4. Treino,
5. Classificação de Emparelhamentos

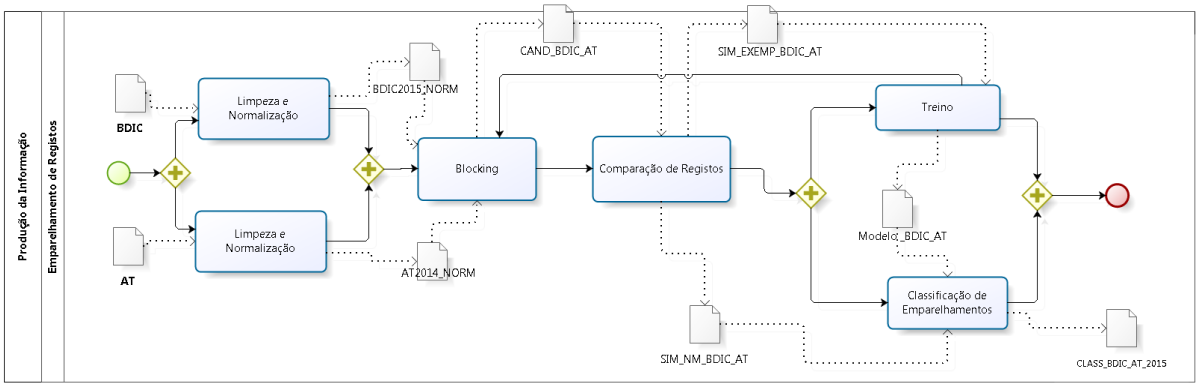

Figura 3.2: Processo de Emparelhamento de Registos das Fontes de Dados BDIC e AT

3.1 Limpeza e Normalização

Esta actividade, constitui a primeira etapa do processo de emparelhamento de registos a ser executada; ocorre logo após a importação dos ficheiros para o SGBD Oracle. O seu funcionamento encontra-se sintetizado na Tabela 3.2.

Tabela 3.2: Síntese da actividade de Limpeza e Normalização

Actividade	Limpeza e Normalização
Entrada	1. Relação com dados obtidos da fonte (por ex: BDIC) 2. Etiqueta a atribuir para identificação de proveniência (por exemplo, "BDIC2015" aos dados da BDIC do ano 2015)
Saída	1. Dados preenchidos na relação normalizada correspondente. Por exemplo: BDIC $\rightarrow$ BDIC_NORM
Funcionalidade	Garantir dados de anos sucessivos para cada fonte limpos e armazenados num esquema comum, com informação de proveniência associada
Scripts usados	1. *preprocessamento.py* 2. *normalizacao.py*

A execução da limpeza e normalização é feita através do *Script* (1) de carregamento dos dados da relação original para a relação com os dados normalizados. Atendendendo ao facto do INE ter realizado anteriormente algumas transformações nos dados recebidos das fontes, tal como referido no Capítulo 1, o desenvolvimento do *script* foi simplificado:

1. A vertente de Limpeza, consiste em detectar e remover o carácter *til* nos dados contidos no atributo RESID_LOCAL_POSTAL com auxílio de instruções em SQL. Esta tarefa é realizada devido ao facto de nas etapas posteriores do processo, a informação a processar ser convertida em de ficheiros Comma Separated values (CSV). Observou-se que o carácter vírgula que é o separador padrão dos ficheiros CSV ocorre com frequência nos dados deste atributo nas relações AT2014(IRS) e IISS2015, e a sua remoção simplifica o tratamento dos dados nas etapas subsequentes.

2. Na vertente de Normalização, foi identificado um conjunto de atributos demográficos e geográficos comuns entre as fontes, e de seguida, com o auxílio do *script* em *Python* (2) que, dado o nome e o ano do carregamento de uma fonte de dados, gera um *script* SQL que permite realizar as seguintes transformações:

   (a) Repartir o atributo Nome do indivíduo em dois distintos (NOME_3PRI e NOME_3ULT);

   (b) Repartir o atributo Data de nascimento em três distintos (A_NASC, M_NASC, D_NASC);

(c) Uniformizar os nomes dos atributos comuns entre as diferentes fontes de dados.

(d) Normalizar os dados nos atributos identificadores da naturalidade e residência do indivíduo de cada fonte de dados, substituindo strings vazias por NULL e valores desconhecidos como por exemplo ZZ0000 e 00 para NULL.

O esquema da relação resultante desta actividade é comum a todas as fontes, e inclui unicamente os atributos demográficos e geográficos de cada indivíduo, tal como ilustra a Tabela 3.3. Os nomes do esquema das relações são padronizados no *script esquema_basedados.py*, referindo o nome da fonte ou fontes e a actividade em que são preenchidos. Por exemplo, para o carregamento dos dados referentes a 2015 da BDIC e 2014 da AT em forma normalizada, seriam criadas pelo *script* as relações BDIC_NORM e AT_NORM respectivamente, obedecendo o padrão de nomes $< FONTE >$_NORM.

Tabela 3.3: Esquema normalizado das relações

Relação	$< FONTE >$_NORM
Atributo	**Descrição**
ID	Identificador pessoal do Indivíduo (NIC, NIF)
PROV	Proveniência. Por exemplo: "BDIC2015"
NOME_3PRI	Três primeiras letras do 1º nome
NOME_3ULT	Três últimas letras do último nome
SEXO	Sexo
A_NASC	Ano de nascimento
M_NASC	Mês de nascimento
D_NASC	Dia de nascimento
EST_CIVIL	Estado Civil
NAT_DT	Distrito de Naturalidade
NAT_MN	Município de Naturalidade
NAT_FR	Freguesia de Naturalidade
NAC_ISO	País de nacionalidade no indivíduo no Padrão ISO 3166-2
RESID_DT	Distrito de Residência
RESID_MN	Município de Residência
RESID_FR	Freguesia de Residência
RESID_CP4	Quatro primeiros dígitos do Código Postal
RESID_CP3	Três últimos dígitos do Código Postal
RESID_LOCAL_POSTAL	Local Postal
RESID_ISO	País de residência do indivíduo no Padrão ISO 3166-2

No esquema normalizado, o atributo *ID* corresponde ao identificador pessoal usado em cada fonte de informação, isto é, ao NIC para a tabela BDIC_NORM e NIF para AT_NORM. Para registar a proveniência dos dados carregados nas tabelas com os dados normalizados, o atributo PROV é preenchido com um valor que permite identificar o ficheiro de dados de onde foi carregada a informação, sendo no cenário em consideração "BDIC2015" e "AT2014", para os dados da BDIC de 2015 e da AT referentes ao IRS de 2014, respectivamente.

Existem fontes de dados que possuem mais que um identificador de cada indivíduo no seu esquema

original. Neste caso um deles é escolhido para ser registado no atributo *ID*, e os restantes são incluídos como atributos auxiliares que, não fazendo parte do esquema normalizado, são incluídas na relação normalizada.
Por exemplo, nos dados da Segurança Social (com os identificadores NISS, NIC e NIF), o *ID* corresponde ao NISS e os atributos NIF e NIC são adicionados ao esquema normalizado como atributos auxiliares.

3.2 Blocking

No âmbito deste trabalho, o *Blocking* foi criado com o intuito de realizar duas funções distintas:

1. gerar emparelhamentos candidatos para cada par de fontes de dados a emparelhar, com intuito de garantir a redução do número de comparações entre os pares de registos das referidas fontes. Por exemplo, a BDIC (dados de 2015) e a AT (dados de 2014) possuem 11.825.786 e 9.370.879 registos respectivamente. Sem a aplicação de técnicas de indexação como é o caso do Blocking, seria necessário efectuar biliões de comparações, correspondentes ao produto cartesiano das relações. Essas comparações na sua maioria não produzirão emparelhamentos, e por outro lado o custo da computação seria bastante elevado;

2. gerar exemplos para treino do modelo de classificação usado na actividade subsequente.
   Os exemplos podem ser de dois tipos:

   (a) *Positivos:* correspondem a registos com pares etiquetados como verdadeiros emparelhados.

   (b) *Negativos:* correspondem aos registos antecipadamente etiquetados como verdadeiros não emparelhados.

A Tabela 3.4 sintetiza o funcionamento desta actividade.

Tabela 3.4: Síntese da actividade de Blocking

Actividade	Limpeza e Normalização
Entrada	1. Par de relações normalizadas. Por ex. BDIC_NORM e AT_NORM ilustradas na Fig. 3.2. Estas relações são obtidas considerando uma condição de selecção de proveniência (por exemplo a proveniência da BDIC="BDIC2015" e da AT = "AT2014")
Saída	1. Relação com emparelhamentos candidatos e para treino (por exemplo, a relação CAND_BDIC_AT da Fig.3.2, cujo esquema é dado na Tabela 3.5)
Funcionalidade	1. Gerar os candidatos dos registos a emparelhar; 2. Gerar um conjunto de exemplos positivos e negativos para o treino do modelo de classificação.
Scripts usados	1. *blocking.py* 2. *blocking.sql*

O Blocking é realizado com base na abordagem tradicional referida no Capítulo 2. Optou-se por esta técnica numa fase inicial pelo facto de ser simples de implementar, tendo-se verificado posteriormente que é possível obter bons resultados na maioria das fontes emparelhadas.

O critério de blocking é especificado pelo utilizador, que indica quais os atributos de cada relação de entrada a concatenar para formar a chave de blocking a produzir na saída.

Para os emparelhamentos realizados no INE, foi usado como critério de Blocking a concatenação dos seguintes atributos: *3 primeiras letras do 1º nome e a data de nascimento* para as fontes de dados emparelhadas com a BDIC, e os atributos país de naturalidade e a data de nascimento do indivíduo para as fontes emparelhadas com o SEF. A escolha deste critério, baseou-se numa análise prévia da qualidade dos dados nas relações cujos pares de registos são previamente conhecidos como verdadeiros emparelhamentos. Esta relação denotada por TEMP_$< FONTEA_FONTEB >$, é preenchida com base no cruzamento do par de relações a emparelhar, por meio dos seus identificadores. Por exemplo, o par BDIC e IEFP, são cruzados por meio de um identificador comum que é o NIC. Existem casos como por exemplo o emparelhamento BDIC e AT, em que será necessário uma terceira relação, neste caso a IISS devido ao facto desta possuir os identificadores auxiliares NIC e NIF para além do seu próprio ID, que é o NISS.

A execução do blocking é realizada com auxílio de um script em *Python* (1) que recebe como parâmetros o nome das fontes de dados a emparelhar, os respectivos anos do carregamento (ano de residência a que se referem os dados) e o critério de Blocking a ser usado. Este *script* em *Python* gera outro *script* em SQL (2) que permite realizar o Blocking e preencher a relação final da actividade, que no caso da BDIC e AT é denotada por CAND_BDIC_AT.

Para melhor compreensão, consideremos a actividade de Blocking para o par BDIC e AT. O *script* (2), primeiramente efectua o preenchimento da relação TEMP_BDIC_AT com base no critério anteriormente referido. Esta relação possui para além dos atributos demográficos e geográficos, dois atributos (BKV_BDIC e BKV_AT) que representam as chaves de Blocking dos registos na relação TEMP_BDIC_AT. Seguidamente, são preenchidas duas relações BDIC_NM_NORM e AT_NM_NORM que correspondem aos registos que não foram possíveis de emparelhar usando o critério aplicado no preenchimento da relação TEMP_BDIC_AT. Cada uma das relações possui igualmente o atributo BKV que é preenchido com base no critério de Blocking definido pelo utilizador.
A relação de saída do Blocking, que no caso do cenário em curso é denotada por CAND_BDIC_AT é preenchida pelos seguintes registos diferenciados por um atributo denominado *CLASSE*:

1. Registos gerados pelo cruzamento da relação TEMP_BDIC_AT consigo mesma, usando como critério os atributos BKV_BDIC e BKV_AT. Os registos que correspondem a verdadeiros emparelhamentos são preenchidos com o valor *um* (1) no atributo CLASSE, os restantes são preenchidos pelo valor *zero* (0);

2. Registos gerados pelo cruzamento das relações BDIC_NM_NORM e AT_NM_NORM usando como critério o atributo BKV definidos nos seus esquemas. Estes registos são preenchidos com o valor (-1) no atributo CLASSE.

Os registos em CAND_BDIC_AT com a CLASSE=1 e CLASSE=0 correspondem aos exemplos Positivos e Negativos para o treino do modelo de Classificação; e os registos com a CLASSE= -1 correspondem aos registos a serem emparelhados em actividade subsequente.
Apresento na Tabela 3.4, o esquema da relação de saída do Blocking para o emparelhamento BDIC e AT (CAND_BDIC_AT). O atributo *RECNUM* corresponde a ordem do registo na relação. A Estratégia aplicada para a BDIC e AT é similar para os restantes pares de fontes.

Tabela 3.5: Esquema da relação Candidatos BDIC2015 e AT2014

Relação	CAND_BDIC_AT
Atributos	
RECNUM	
ID_BDIC	ID_AT
NOME_3PRI_BDIC	NOME_3PRI_AT
NOME_3ULT_BDIC	NOME_3ULT_AT
SEXO_BDIC	SEXO_AT
A_NASC_BDIC	A_NASC_AT
M_NASC_BDIC	M_NASC_AT
D_NASC_BDIC	D_NASC_AT
EST_CIVIL_BDIC	EST_CIVIL_AT
NAT_DT_BDIC	NAT_DT_AT
NAT_MN_BDIC	NAT_MN_AT
NAT_FR_BDIC	NAT_FR_AT
NAC_ISO_BDIC	NAC_ISO_AT
RESID_DT_BDIC	RESID_DT_AT
RESID_MN_BDIC	RESID_MN_AT
RESID_FR_BDIC	RESID_FR_AT
RESID_CP4_BDIC	RESID_CP4_AT
RESID_CP3_BDIC	RESID_CP3_AT
RESID_LOCAL_POSTAL_BDIC	RESID_LOCAL_POSTAL_AT
RESID_ISO_BDIC	RESID_ISO_AT
CLASSE	

3.3 Comparação de Registos

Esta actividade realiza a comparação entre os pares de registos das relações a emparelhar, produzidos na actividade de *Blocking*, usando métricas de similaridades como por exemplo as abordadas no Capítulo 2.

A comparação de registos é antecedida pela selecção de um conjunto de atributos (*features* em *Machine Learning*), considerados relevantes para o treino do modelo e de para a classificação dos registos, as quais podem variar de relação para relação. Esta variação deve-se ao facto de cada par de relações a emparelhar possuir um conjunto de atributos relevantes que não são comuns a todos os pares de relações. Por exemplo, os atributos que identificam a naturalidade do indivíduo são relevantes para o treino e para a classificação dos pares BDIC-AT mas não o são para o caso BDIC-CGA, pelo facto de na fonte CGA estes atributos não estarem preenchidos na referida relação.

A Tabela 3.6 apresenta uma síntese do funcionamento desta actividade.

Tabela 3.6: Síntese da actividade Comparação de Registos

Actividade	Comparação de Registos
Entrada	1. Ficheiro CSV contendo os registos a processar. Por exemplo: CAND_BDIC_AT da Fig.3.2
Saída	1. Ficheiro CSV contendo os vectores de similaridades entre os vários pares de atributos. Por exemplo: SIM_EXEMP_BDIC_AT e SIM_NM_BDIC_AT da Fig.3.2
Funcionalidade	1. Calcular as similaridades entre os pares de atributos contidos no ficheiro CSV de entrada usando as métricas de similaridade estabelecidas.
Script usado	1. *similarities.py*

Para esta actividade, foi desenvolvido um módulo em *Python* (1) que recebe como entrada um ficheiro CSV com os atributos a comparar. Este ficheiro de entrada é obtido a partir da exportação dos dados provenientes da relação de saída da actividade de *Blocking* (CAND_BDIC_AT) para 2 ficheiros:

1. Ficheiro CSV com os registos a serem usados para o treino do modelo de Classificação. Isto é, os registos *CLASSE=1* e *CLASSE=0* na relação de entrada;

2. Ficheiro CSV com os registos com a *CLASSE= -1* (desconhecida), ou seja os por emparelhar.

Dependendo da etapa que se pretende executar (treino ou classificação), o *script* selecciona o ficheiro adequado, calcula as similaridades entre cada par de atributos a comparar e produz um ficheiro CSV contendo os vectores de similaridades. A métrica usada para o cálculo das similaridades é a distância de edição (distância de *Levenshtein*).

3.4 Treino

Esta actividade consiste no treino de um modelo de classificação, com base num conjunto de exemplos *Positivos* e *Negativos*. Este modelo é utilizado como suporte para decidir que *CLASSE* atribuir aos pares de registos não etiquetados ou seja o conjunto de registos por emparelhar.

A Tabela 3.7 apresenta em síntese o funcionamento desta actividade.

Tabela 3.7: Síntese da actividade de Treino

Actividade	Treino
Entrada	1. Ficheiro CSV contendo as similaridades dos exemplos para o treino do modelo de classificação. Por exemplo: SIM_POS_BDIC_AT da Fig.3.2
Saída	1. Ficheiro no formato PKL contendo o modelo de classificação treinado. Por exemplo: Modelo_BDIC_AT da Fig.3.2
Funcionalidade	1. Treinar e gerar um modelo de classificação com base nos exemplos fornecidos.
Script usado	1. *train_model.py*

O modelo do Classificador de emparelhamentos foi treinado usando a técnica de regressão logística [Alpaydin, 2004, Hastie et al., 2009], com base em registos de exemplos gerados segundo a estratégia referida na actividade de Blocking (secção 3.2). A escolha desta técnica deveu-se ao facto de ser simples de implementar e de rápida execução.
Para tal, foi desenvolvido um módulo em *Python* (1) que executa a actividade sintetizada na Tabela 3.7, para cada fonte de dados a ser emparelhada.

3.5 Classificação dos Emparelhamentos

Esta actividade tem por objectivo prever a *CLASSE* em que cada par de registos não emparelhado se encontra. Esta decisão é tomada com base no modelo de classificação treinado, conforme descrito na etapa de treino (secção 3.4).
A Tabela 3.8 apresenta em síntese o funcionamento desta actividade.

Tabela 3.8: Síntese da actividade de Classificação de Emparelhamentos

Actividade	Classificação de Emparelhamentos
Entrada	1. Ficheiro CSV contendo os vectores de similaridades dos registos não classificados (Registos com a Classe= -1). Por exemplo: SIM_NM_BDIC_AT ilustrado na Fig.3.2 2. Modelo de Classificação. Por exemplo: Modelo_BDIC_AT
Saída	1. Ficheiro CSV com os resultados do emparelhamento. Por exemplo: o ficheiro CLASS_BDIC_AT_2015 ilustrado da Fig.3.2
Funcionalidade	1. Classificar os registos do ficheiro de entrada com base no modelo de Classificação de entrada, produzindo o ficheiro de saída com o atributo adicional *CLASSE* preenchido.
Script usado	1. *match_records.py*

Para o caso dos emparelhamentos do INE, foi desenvolvido um módulo em *Python* (1) que gera um ficheiro de saída que possui 4 campos no seu esquema:

1. **recnum:** valor numérico que representa a ordem do registo no ficheiro
2. **non-match:** representa a probabilidade do registo não ser um match [0 , 1]
3. **match:** representa a probabilidade do registo ser um match [0 , 1]
4. **predicted:** representa a CLASSE atribuída pelo classificador (0 ou 1).

Uma vez obtido o resultado da classificação, o ficheiro de saída é importado para o SGBD e armazenado numa relação cujo esquema é formado pelos 4 atributos, referidos acima. Para o caso do emparelhamento entre a BDIC e AT, denominamo-la por CLASS_BDIC_AT.

A associação do resultado da classificação aos indivíduos correspondentes é feita mediante o cruzamento da relação dos candidatos não emparelhados com as respectivas classificações.
Para o cenário em curso, consideram-se as relações CAND_BDIC_AT (referida na secção 3.2) selecionando apenas os registos com a *CLASSE = -1* e CLASS_BDIC_AT, cruzadas por meio do atributo *RECNUM* que gerará a relação CLASS_BDIC_AT_2015.

3.6 Ambiente de Execução do MPI

O MPI é constituído por um conjunto de *Scripts* de software nas linguagens *Python* e SQL. A Fig.3.3 ilustra o ambiente de execução do MPI, disponível no INE. É composto por um servidor de bases de dados relacional, um servidor aplicacional para execução de código *Python*, e um repositório de software com controlo de versões.

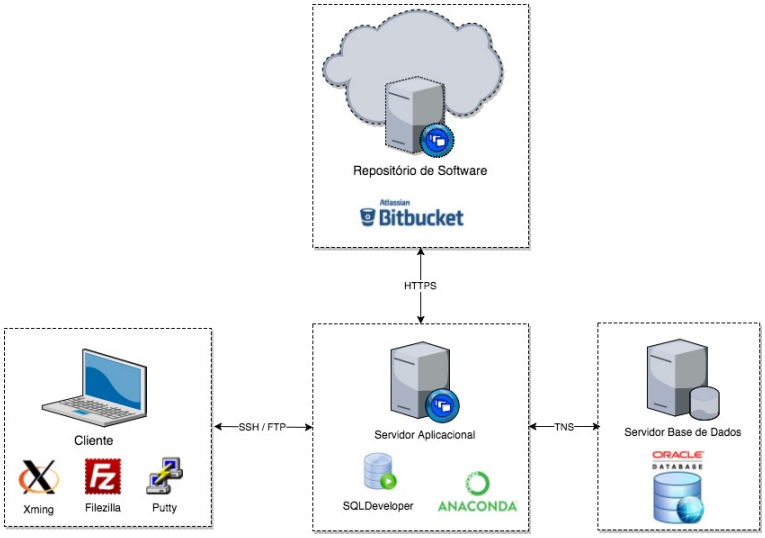

Figura 3.3: Ambiente Computacional para o MPI

O acesso ao Servidor Aplicacional é feito através de um terminal Cliente configurado com o ambiente Windows (Windows 7), no qual se encontram instaladas as seguintes ferramentas:

1. **Xming:** servidor de janelas X Window Systems, para o sistema operativo Windows. Foi utilizado para permitir executar remotamente, a partir do servidor aplicacional (Linux), a aplicação Cliente do SGBD Oracle (*SQLDeveloper*);

2. **Filezilla:** ferramenta para a transferência de ficheiros do servidor aplicacional para o terminal Cliente e vice-versa;

3. **Putty:** terminal de ligação segura (Secure Shell (SSH)) ao Servidor Aplicacional.

O Servidor Aplicacional é um ambiente *Linux*, onde estão instalados o *Anaconda* (ambiente de programação para o *Python*) e o *SQLDeveloper* para acesso ao SGBD Oracle.
No servidor de Base de Dados Oracle, encontram-se armazenadas as tabelas criadas a partir dos ficheiros de dados provenientes das várias instituições da Administração Pública, e todas outras tabelas criadas para o funcionamento do MPI. O código do MPI encontra-se armazenado no repositório com controlo de versões *Bitbucket*[2].

[2]https://bitbucket.org/

3.7 Sumário

Neste capítulo foram detalhadas cada uma das cinco componentes da solução de emparelhamento dos dados do INE desenvolvida:

1. Limpeza e Normalização,
2. *Blocking*,
3. Comparação de registos,
4. Treino e
5. Classificação de emparelhamentos.

Todas estas componentes são executadas com a intervenção do utilizador, por meio de *scripts* em *Python* e SQL. O código dos vários *scripts* encontra-se disponível no repositório *Bitbucket* em `https://bitbucket.org/pavel_calado/ine2016`

A solução é geral, não sendo necessário um desenvolvimento adicional para emparelhar uma nova fonte fora das existentes. Alguma parametrização será porém necessária em função das características das novas fontes de dados que poderão surgir, ou ainda em caso de substituição dos algoritmos usados em cada etapa do processo.
O ambiente de execução usa exclusivamente ferramentas abertas e gratuitas, além do SGBD Oracle que é já usado no INE. Desde que seja usado o *SQL Standard*, pode ser usado um outro SGBD relacional.
A solução descrita neste trabalho usa a *Regressão Logística* como técnica de Classificação, sendo portanto adequada ao emprego na análise exploratória de dados. Porém, outras técnicas de aprendizagem podem ser aplicadas sobre estes dados.
Para a comparação dos pares de atributos foi usada a métrica de distância de edição, e para o *Blocking* a técnica do Blocking Tradicional usando o seguinte critério: *3 primeiras letras do 1º nome, Ano de nascimento, Mês de nascimento e Dia de nascimento*.
O uso destas técnicas, em conjunto permitiram o emparelhamento de 8 pares de fontes de dados, onde 5 pares correspondem a emparelhamentos com a BDIC e 3 com o SEF.

No capítulo seguinte, abordarei os resultados obtidos na aplicação da metodologia descrita neste trabalho, usando como base as métricas de qualidade descritas no Capítulo 2.

4

Monitorização e Resultados do Emparelhamento de Registos

Conteúdo

Neste capítulo apresento as funcionalidades do Processo de Monitorização, que permitem avaliar a qualidade do Processo de Produção da Informação ao processar as fontes de dados do INE.
A Figura 4.1 ilustra a interação entre as actividades do processo de Produção da Informação com as correspondentes actividades de monitorização. A cada etapa do Processo de Produção da Informação sucede uma do Processo de Monitorização. As etapas da Produção da Informação posteriores à de Limpeza e Normalização requererão do operador uma aprovação prévia da qualidade dos dados das etapas anteriores. Por exemplo, a execução da actividade de Blocking requer uma validação prévia da qualidade dos dados na etapa de Limpeza e Normalização.

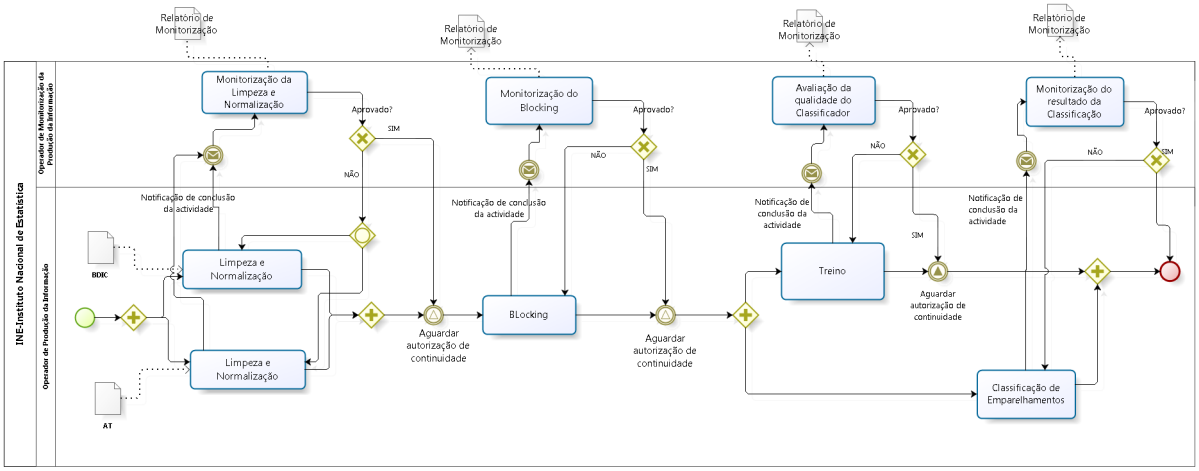

Figura 4.1: Processo de Monitorização da Produção da Informação das Fontes de Dados BDIC e AT

Para este efeito, foi desenvolvido um módulo (*monitorizacao.py*) que consulta os dados produzidos pelo Processo de Produção da Informação, armazenados em tabelas e no sistema de ficheiros do Servidor Aplicacional, calcula as métricas de qualidade de dados e fornece os relatórios de monitorização em cada uma das etapas. As referidas etapas são apresentadas ao utilizador a partir de uma interface em *Python*, que contém as seguintes opções:

1. Limpeza e Normalização,
2. Blocking,
3. Treino e
4. Classificação de Emparelhamentos.

A actividade de Comparação de registos abordada no Capítulo 3, não se encontra referenciada no processo de monitorização uma vez que a avaliação da qualidade nesta etapa pode ser feita a partir dos resultados de monitorização do Blocking ou ainda das etapas posteriores.

Nas secções seguintes, abordarei em maior detalhe as métricas e os resultados obtidos após a execução de cada uma das actividades de monitorização.

4.1 Limpeza e Normalização

A avaliação da qualidade dos dados nesta etapa, é realizada com base nas seguintes métricas:

1. Completude, conforme definida no Capítulo 2,
2. *Null Count*: fornece a quantidade de registos cujo o valor no atributo em análise é indefinido (NULL),
3. Unicidade: fornece a quantidade de valores distintos tomados pelo atributo em análise.

A Tabela 4.1 sintetiza o funcionamento desta actividade. Tendo em conta que esta etapa é influente para o sucesso das etapas posteriores da Produção da Informação, importa que seja verificado que o nível de qualidade nos dados a serem usados para os novos emparelhamentos se mantém ou melhora. Para tal, pode ser feita a comparação das métricas obtidas com os dados de anos / carregamentos anteriores com o carregamento a ser processado. Como exemplo, apresento na Figura 4.2 as medidas de qualidade obtidas nesta etapa, considerando as fontes BDIC do ano de 2014 e BDIC do ano de 2015.

Tabela 4.1: Síntese da actividade de Monitorização da Limpeza e Normalização

Actividade	Monitorização da Limpeza e Normalização
Entrada	1. Relação com dados obtidos da fonte (por ex: BDIC) 2. Especificação da restrição de proveniência (por exemplo, os dados da "BDIC2015").
Saída	1. Relatório de Monitorização com os valores das métricas escolhidas para esta etapa
Funcionalidade	Fornecer o relatório das métricas de Completude, *Null Count* e Unicidade para cada fonte de dados indicada na Entrada.

ATRIBUTO	COMPLETUDE (%)	NULL COUNT (Registos)	VALORES DISTINTOS (Registos)
PROV	100.0	0	1
ID	100.0	0	11884913
NOME_3PRI	100.0	0	4213
NOME_3ULT	99.99893	127	4505
SEXO	100.0	0	2
A_NASC	99.99978	26	148
M_NASC	99.99955	54	14
D_NASC	99.99955	54	33
EST_CIVIL	100.0	0	6
NAT_DT	100.0	0	33
NAT_MN	100.0	0	363
NAT_FR	100.0	0	3718
NAC_ISO	100.0	0	3
RESID_DT	100.0	0	31
RESID_MN	100.0	0	312
RESID_FR	100.0	0	3380
RESID_CP4	72.28231	3294223	646
RESID_CP3	72.28231	3294223	982
RESID_LOCAL_POSTAL	72.28230	3294224	6910
RESID_ISO	100.0	0	3

a) BDIC (2014)

ATRIBUTO	COMPLETUDE (%)	NULL COUNT (Registos)	VALORES DISTINTOS (Registos)
PROV	100.0	0	1
ID	100.0	0	11825786
NOME_3PRI	100.0	0	4251
NOME_3ULT	99.99894	125	4543
SEXO	100.0	0	2
A_NASC	100.0	0	148
M_NASC	99.99954	54	14
D_NASC	99.99954	54	33
EST_CIVIL	100.0	0	6
NAT_DT	100.0	0	33
NAT_MN	100.0	0	361
NAT_FR	100.0	0	3712
NAC_ISO	99.99996	5	3
RESID_DT	100.0	0	31
RESID_MN	100.0	0	310
RESID_FR	100.0	0	3373
RESID_CP4	76.85867	2736644	642
RESID_CP3	76.85867	2736644	981
RESID_LOCAL_POSTAL	76.85865	2736646	5363
RESID_ISO	100.0	0	3

b) BDIC (2015)

Figura 4.2: Avaliação da Qualidade de Dados da BDIC

Para a análise dos resultados na Figura 4.2 é possível observar alguns aspectos como:

1. A variação no número total de registos nas duas versões (coluna “valores distintos referente ao atributo ID”);
2. As taxas de Completude nos seguintes atributos: último nome, data de nascimento, Estado Civil, Nacionalidade e a informação de residência (CP4,CP3, Local Postal).

Quanto ao ponto 1, observa-se que a versão da BDIC (2015), possui uma quantidade de registos inferior a versão BDIC (2014) em cerca de 0.497%. Isto pode explicar-se devido ao facto da taxa de natalidade ser inferior a taxa de mortalidade durante este período.

Quanto ao ponto 2, é possível observar o seguinte:

1. Apesar da redução do número de registos da BDIC (2015) relativamente à BDIC (2014), ainda assim observa-se alguma quantidade residual de indivíduos sem informação no atributo último nome. Não tendo uma completude a 100%, acrescentando a possibilidade de mudança de apelidos especialmente para os indivíduos do sexo feminino, este atributo não fornece garantias de ser usado como parte do critério de *Blocking*.

2. Para a data de nascimento, observa-se uma taxa de preenchimento abaixo de 100% nos atributos ano, mês e dia, onde cerca de 54 indivíduos possuem a data de nascimento desconhecida. Para este caso, o presente atributo toma o valores NULL ou 0, razão pela qual existem dois valores distintos em excesso em cada atributo que constitui a data de nascimento do indivíduo.

3. Quanto ao Estado Civil, a fonte de dados BDIC possui seis estados: Solteiro, Casado, Divorciado, Separado, Desconhecido e Viúvo;

4. Quanto à nacionalidade, este atributo encontra-se preenchido pelos seguintes valores: PT, BR e NULL quando o país de nacionalidade é desconhecido.

5. Para a informação de residência, as taxas de completude e o total de registos nulos permanecem inalterados devido a redução de 0.497% dos registos, apesar de ter um valor superior de completude na versão BDIC (2015) comparado a versão BDIC (2014).

4.2 Blocking

A qualidade dos dados nesta etapa é medida com base na taxa de redução com que um determinado critério de *Blocking* é capaz de afectar o número de comparações entre cada par de relações a emparelhar.
A Tabela 4.2 sintetiza o funcionamento desta actividade.

Tabela 4.2: Síntese da actividade de Monitorização do Blocking

Actividade	Monitorização do Blocking
Entrada	1. Relação com os pares positivos, negativos e pares candidatos (por ex: CAND_BDIC_AT) 2. Especificação do tipo de pares de registos a analisar (por exemplo, pares de registos de CLASSE=0, CLASSE=1 ou CLASSE= -1).
Saída	1. Relatório de Monitorização com os valores das métricas escolhidas para esta etapa
Funcionalidade	Fornecer o relatório sobre a quantidade de registos existentes em cada fonte de entrada, pares positivos, negativos e candidatos gerados e a Taxa de redução obtida após a aplicação do Blocking.

A execução desta atividade, tem como resultado o relatório como o ilustrado na Figura 4.3. Neste, são referidas as informações gerais sobre as relações BDIC (2015) e AT (2014), o conjunto de pares positivos e negativos onde são extraídos os exemplos para o treino do modelo de classificação, o total de registos candidatos para o emparelhamento BDIC-AT e as taxas de redução obtidas. Para o caso dos pares positivos e negativos, a taxa de redução é comum uma vez que os dois tipos de pares provêm de uma única relação. A informação abaixo permite-nos igualmente analisar que estratégias podem ser aplicadas para garantir o equilíbrio do número de exemplos a usar em cada Classe para o treino do modelo de Classificação.

FONTE DE DADOS	ANO	TOTAL DE REGISTOS
BDIC	2015	11825786
AT	2014	9307121

METRICAS DE QUALIDADE DE DADOS

CATEGORIA ANALISADA	TOTAL DE REGISTOS	TAXA DE REDUCAO COM O BLOCKING (%)
PARES POSITIVOS	4880997	99.99999
PARES NEGATIVOS	3842312	
PARES CANDIDATOS	61038705	99.99994

Figura 4.3: Qualidade de Dados do Blocking (BDIC-AT)

4.3 Treino

Nesta etapa, a qualidade é medida com base na exactidão do modelo de Classificação. A técnica usada para o efeito é a 2- *fold Cross Validation* [Han and Kamber, 2006] aplicada ao conjunto de exemplos (positivos e negativos) extraídos do par de relações a emparelhar. Uma vez feita a *2- fold Cross Validation*, são medidos os valores percentuais da precisão (precision) e da abrangência (recall) referentes a cada *Classe* dos exemplos usados.

A medida de exactidão do modelo ditará se o mesmo pode ou não ser usado para a classificação de novos registos. Para o caso deste trabalho, considero como satisfatório os modelos com o nível de abrangência com um erro de 3%. Este valor foi estabelecido com base na subestimação da taxa de cobertura verificada nos Inquéritos de Qualidade (IQ) dos Censos porta a porta de 2011, cujo o valor corresponde com uma média de 2.5% para todo o país [INE, 2012].

Tendo em conta que as fontes de dados não possuem as variáveis com uma taxa de completude a 100% na maioria dos casos, como por exemplo o caso da Fonte de Dados da Educação e Ciência (EDUC) (Educação e Ciência), escolho estabelecer um valor não muito superior mas equiparado aos Censos 2011 como *baseline* para a qualidade do modelo de Classificação pretendido, embora ter passado um certo tempo até a data presente.

A Tabela 4.3 apresenta de forma sintetizada o funcionamento desta actividade.

Tabela 4.3: Síntese da actividade de Monitorização do Treino

Actividade	Avaliação da qualidade do Classificador
Entrada	1. Ficheiro CSV com os vectores de similaridades referentes aos exemplos de um par de fonte de dados (por ex: SIM_EXEMP_BDIC_AT)
Saída	1. Relatório de Monitorização com os valores das métricas escolhidas para esta etapa
Funcionalidade	Avaliar a precisão do Classificador usando uma técnica de avaliação de qualidade, por exemplo *k-fold Cross Validation* e gerar o relatório das métricas de *Precisão* e *Abrangência* para cada par de emparelhamentos indicados na entrada.

A Tabela 4.4 apresenta as informações gerais para o treino e validação de cada modelo de Classificação usado no INE. Nesta Tabela, encontram-se referidas a quantidade de atributos e o número de registos (exemplos) de *CLASSE=1* e *CLASSE=0* usados para a geração do modelo de Classificação. Para cada par de emparelhados, foi gerado um modelo diferente, devido às características que cada conjunto de dados apresenta, e pelo número de atributos comuns e preenchidos em cada par (conforme abordado na Secção 3.3).

Os exemplos usados para o treino do modelo *model_bdic_at* referente às fontes BDIC e AT, encontram-se indisponíveis devido o facto de ter sido o primeiro emparelhamento realizado, e não ter sido possível registar os referidos dados. No caso do emparelhamento das fontes SEF e EDUC (Fonte de Dados da Eduçação), foi usado o modelo *model_sef_iiss* pelo facto de ter sido obtido um número bastante reduzido de exemplos positivos.

Tabela 4.4: Informações Gerais para o Treino do Modelo de Classificação

Fonte de Dados	Modelo	Atributos	Exemplos Positivos	Exemplos Negativos
BDIC (2015)	model_bdic_at	13		
AT (2014)				
BDIC (2015)	model_bdic_iiss	17	44.912	955.088
IISS (2015)				
BDIC (2015)	model_bdic_educ	12	294.514	1.705.486
EDUC (2015)				
BDIC (2015)	model_bdic_iefp	14	174.973	325.027
IEFP (2015)				
BDIC (2015)	model_bdic_cga	11	76.368	423.632
CGA (2015)				
SEF (2015)	model_sef_at	9	13.669	10.860
AT (2014)				
SEF (2015)	model_sef_iiss	14	121.916	160.066
IISS (2015)				
EDUC (2015)				

Considerando a abordagem para a geração de exemplos referida na Secção 3.2, observam-se na Tabela 4.4 algumas discrepâncias referentes a quantidade de exemplos positivos *versus* negativos para cada emparelhamento realizado. Isto deve-se ao facto de existirem blocos com um elevado número de exemplos negativos, sendo que em cada um possui apenas um único exemplo positivo. Esta abordagem de geração de negativos poderia ser melhorada escolhendo no máximo dois a três exemplos negativos para cada exemplo positivo, cuja similaridade se aproxima mais ao exemplo positivo.

Com base nos dados acima referidos, apresento na Tabela 4.5 os resultados obtidos na avaliação da exactidão de cada modelo de classificação treinado. Os modelos de Classificação apresentam valores óptimos, que vão de acordo com o que se pretende, com excepção aos modelos gerados com a EDUC e com a Caixa Geral de Aposentações(CGA) que possuem uma abrangência abaixo dos 97%.

Quanto ao comportamento dos modelos face aos exemplos fornecidos, é possível verificar por exemplo, no caso do modelo *model_bdic_iiss* que o classificador classificou 99% dos exemplos positivos como pares de Classe=1 (*Precisão*), sendo que apenas 98% desta classificação correspondem efectivamente a pares de Classe=1 (*Abrangência*), o que significa que o classificador errou 1% dos casos. E, para os exemplos negativos do mesmo modelo, notam-se valores percentuais de 100% para os casos de *Precisão* e *Abrangência*.

Tabela 4.5: Avaliação da Qualidade por Modelo de Classificação

Modelo	Precisão (%)		Abrangência (%)	
	Exemplos Positivos	Exemplos Negativos	Exemplos Positivos	Exemplos Negativos
model_bdic_at	98	97	97	98
model_bdic_iiss	99	100	98	100
model_bdic_educ	93	99	94	99
model_bdic_iefp	99	99	98	99
model_bdic_cga	88	99	95	98
model_sef_at	97	97	97	98
model_sef_iiss	98	98	97	98

Resultados similares são verificados nos restantes modelos, com excepção ao *model_bdic_cga* (1) onde os valores de Precisão e Abrangência dos exemplos de Classe=1 possuem uma discrepância de 7%. Isso deve-se ao facto de haver um grande desequilíbrio entre a quantidade de exemplos positivos versus negativos usados para o treino do modelo de Classificação.
Para melhorar a Precisão, treinou-se um outro modelo *model_bdic_cga* (2) contendo 900.000 exemplos ao invés de 500.000 do modelo anterior (1), dentre os quais 138.279 registos de Classe=1 e 761.721 registos de Classe=0. Com isto, obteve-se:

1. Para os exemplos de Classe=1, obteve-se uma Precisão de 92% e Abrangência de 93%;
2. Para os exemplos de Classe=0, obteve-se uma Precisão de 99% e Abrangência de 98%.

Em termos dos resultados das métricas em referência, os valores da qualidade aumentaram significativamente no caso da precisão mas a referida melhoria não refletiu nos resultados dos emparelhamentos. Isto é, o número de emparelhamentos obtidos e os que foi possível verificados na *Matriz* (emparelhamentos equivalentes obtidos por métodos exactos) é inferior aos obtidos com modelo *model_bdic_cga* (1) na ordem dos 1,04%, supondo-se que o elevado número de exemplos negativos face aos exemplos positivos, possa ter tornado o modelo enviesado.

4.4 Classificação de Emparelhamentos

Nesta etapa, a qualidade é medida com base nas contradições e incertezas existentes no resultado dos emparelhamentos realizados.

A Tabela 4.6 sintetiza o funcionamento desta actividade.

Tabela 4.6: Síntese da actividade de Monitorização da Classificação de Emparelhamentos

Actividade	Monitorização do resultado da Classificação
Entrada	1. Relação com o resultado os emparelhamentos realizados (por ex: CLASS_BDIC_AT_2015)
Saída	1. Relatório de Monitorização com os valores das métricas escolhidas para esta etapa
Funcionalidade	Medir as inconsistências (contradições e incertezas) existentes nos emparelhamentos realizados.

Uma vez obtida a relação dos emparelhamentos como descrito na Secção 3.5, é necessário realizar algumas operações de tratamento dos dados. Por exemplo, na relação CLASS_BDIC_AT_2015 há que:

1. Verificar e extrair da relação os registos que tenham sido previamente emparelhados pelo INE por métodos exactos. Esta extração é realizada devido ao facto desses emparelhamentos já terem sido previamente realizados e não são considerados unicamente como produto da aplicação do modelo probabilístico às fontes de dados. Os referidos emparelhamentos realizados pelo INE encontram-se armazenados numa relação denominada Matriz.
   O esquema da Matriz possui um conjunto de atributos, dentre os quais estão incluídos alguns que representam a presença do registo em uma ou mais fontes de dados. No INE, estes atributos são denominados por *flags* (por exemplo *BDIC2015*, *IRS2014*, *CGA2015*). Estas *flags* tomam o valor NULL (caso o registo não se encontra na referida fonte) ou diferente de NULL (caso tenha sido emparelhado por identificador numérico, igualdade de atributos ou por outra regra).
   Considerando o emparelhamento BDIC e AT, o processo de verificação é realizado por meio de uma consulta em SQL que permite verificar se um determinado registo contido na relação CLASS_BDIC_AT_2015 com o NIC e NIF e as *flags BDIC2015* e *IRS2014* activas (diferente de NULL) encontram-se na Matriz. No final desta operação, são inseridos na relação CLASS_N_MATRIZ_BDIC_AT_2015 apenas os registos que não cumprem as condições acima referidas.

2. Partindo da relação CLASS_N_MATRIZ_BDIC_AT_2015, surge o seguinte problema: considerando que os emparelhamentos foram realizados por métodos probabilísticos, um registo na BDIC pode emparelhar com um ou mais registos na AT, caso as suas características sejam bastante simila-

res.

Para minimizar o impacto deste problema nos resultados que se pretende, aplicou-se a regra das *probabilidades máximas* que consiste no seguinte: Dado um conjunto de emparelhamentos com os ID X e Y (neste caso NIC e NIF), efectua-se um agrupamento dos registos por NIC e extraí-se o conjunto de registos que tenham a probabilidade máxima de emparelhamento em cada grupo de NIC. Denotemos a relação contendo este conjunto de registos por CLASS_PROB_MAX_NIC. Seguidamente, a partir da última relação são extraídos unicamente os registos cujo a probabilidade de emparelhamento com o NIF seja a máxima neste conjunto. Os registos obtidos por meio destas operações são inseridos na relação final denominada CLASS_PROB_MAX_BDIC_AT_2015. A referida relação contém registos de dois tipos:

(a) Emparelhamentos que carecem de uma revisão administrativa (*Clerical Review*), por possuírem probabilidades máximas de emparelhamentos iguais, os quais devem ser decididos por um especialista;

(b) Emparelhamentos únicos que servirão para actualizar a *Matriz* do INE, resultando na activação de novas *flags* para cada registo e com isto adicionar novos registos à BPR.

4.4.1 Resultados e Qualidade dos Emparelhamentos

A Tabela 4.7 apresenta a estatística dos registos a emparelhar por fonte de dados.
Nesta tabela, importa salientar três casos particulares:

1. O par de fontes de dados BDIC e AT foi emparelhado com auxílio de uma terceira fonte de dados que é a IISS, conforme referido no final da Secção 3.2.

2. Para os pares de fontes de dados SEF e IISS, SEF e EDUC foi necessário realizar a operação de união em SQL do resultado do cruzamento realizado por NISS e o resultado do cruzamento realizado por NIF. Isso deve-se ao facto de cada cruzamento em separado ter resultado numa quantidade reduzida de registos que seria insuficiente para o treino do modelo de Classificação.

3. O número de registos presentes na EDUC (2015) refere-se a um subconjunto de cidadãos com o tipo de documento de identificação Bilhete de Identidade (BI)/Cartão de Cidadão (CC) e Cédula de nascimento e os da fonte IEFP (2015) referem-se aos cidadãos cujo documento de identificação é BI/CC, ambos para o emparelhamento com a BDIC (2015).

Tabela 4.7: Estatíticas dos registos a emparelhar por par de Fontes

Fonte de Dados	Nº registos	Identificador de ligação(a)	Registos emparelhados (b)	Registos por emparelhar(c)
BDIC (2015)	11.825.786	NIC	4.892.526	6.933.267
AT (2014)	9.370.879	NIF		4.414.595
BDIC (2015)	11.825.786	NIC	5.542.658	6.283.141
IISS (2015)	6.927.720			1.385.062
BDIC (2015)	11.825.786	NIC	1.595.050	10.230.736
EDUC (2015)	1.680.018			84.968
BDIC (2015)	11.825.786	NIC	622.576	11.203.211
IEFP (2015)	686.198			63.622
BDIC (2015)	11.825.786	NIC	810.843	11.014.943
CGA (2015)	1.032.133			209.642
SEF (2015)	383.764	NISS e NIF	118.155	253.742
IISS (2015)	6.927.720			624.118
SEF (2015)	383.764	NIF	163.237	220.315
AT (2014)	9.370.879			9.023.088
SEF (2015)	383.764	NISS e NIF	7885	375.872
EDUC (2015)	87.017			79.132

(a) Atributo que permitiu cruzar cada par de fontes de dados afim de preencher a relação TEMP_$< FONTEA_FONTEB >$ abordada na Secção 3.2
(b) Corresponde ao total de registos emparelhados usando o identificador de ligação referido em (a).
(c) Corresponde ao total de registos não emparelhados por identificador numérico (total de registos a emparelhar por fonte de dados)

A Tabela 4.8 apresenta o resultado obtido após o emparelhamento dos registos usando o modelo probabilístico.

Tabela 4.8: Emparelhamentos do Método Probabilístico por par de Fontes

Fonte de Dados	Registos por emparelhar (a)	Registos emparelhados (b)	Registos encontrados na Matriz (c)	Novos (d)
BDIC (2015)	6.933.267	3.631.740	3.262.651 (84,88%)	244.903
AT (2014)	4.414.595			
BDIC (2015)	6.283.141	4.667.315	582.237 (68,4%)	47.836
IISS (2015)	1.385.062			
BDIC (2015)	10.230.736	61.943	8.224 (20,17%)	51.138
EDUC (2015)	84.968			
BDIC (2015)	11.203.211	74.468	55.260 (61,68%)	11.974
IEFP (2015)	63.622			
BDIC (2015)	11.014.943	300.823	168.158 (93,18%)	60.545
CGA (2015)	209.642			
SEF (2015)	253.742	32.823	2.249 (8,71%)	30.120
IISS (2015)	624.118			
SEF (2015)	220.315	66.028	7.990 (35,72%)	52.177
AT (2014)	9.023.088			
SEF (2015)	375.872	23.381	2.691 (30,02%)	12.796
EDUC (2015)	79.132			

(a) Total de registos a emparelhar por par de fontes
(b) Total de registos emparelhados pelo Modelo Probabilístico por par de fontes de dados. Deste total de emparelhamentos, são aplicadas às operações referidas na Secção 4.4.
(c) Total de emparelhamentos comuns entre os métodos probabilístico e exactos, acompanhado do total percentual que o mesmo representa.
(d) Total de registos acrescentados pelo Modelo Probabilístico: refere-se a novos emparelhamentos encontrados por par de fontes de dados.

FONTE DE DADOS	ANO	TOTAL DE REGISTOS
CLASS_PROB_MAX_BDIC_AT	2015	244903

METRICAS DE QUALIDADE DE DADOS

ATRIBUTO	CONTRADICOES (Registos)	CONTRADICOES (%)	INCERTEZAS (Registos)	INCERTEZAS (%)
NOME_3PRI	0	0.0	0	0.0
NOME_3ULT	2527	1.03184	1	0.00041
SEXO	1380	0.56349	0	0.0
A_NASC	0	0.0	0	0.0
M_NASC	0	0.0	0	0.0
D_NASC	0	0.0	0	0.0
NAT_DT	399	0.16292	1790	0.73090
NAT_MN	664	0.27113	1790	0.73090
NAT_FR	1433	0.58513	1790	0.73090
NAC_ISO	2664	1.08778	0	0.0
RESID_DT	0	0.0	244903	100.0
RESID_MN	0	0.0	244903	100.0
RESID_FR	0	0.0	244903	100.0
RESID_CP4	3625	1.48018	61206	24.99194
RESID_CP3	5163	2.10818	61206	24.99194
RESID_LOCAL_POSTAL	81191	33.15231	62091	25.35330
RESID_ISO	7050	2.87869	0	0.0

Figura 4.4: Avaliação da Qualidade do Emparelhamento BDIC-AT

Na Figura 4.4 é ilustrado em maior detalhe o resultado dos novos emparelhametos das fontes BDIC e AT baseando-se nas medidas de qualidade de dados da etapa de Classificação (contradição e incerteza).

Dentre os vários pares de atributos que podem ser comparados, farei referência dos que possuem uma taxa elevada relativamente aos outros nas duas métricas:

1. Grau de contradições entre os pares de atributos:
   Quanto às contradições, importa referir as observadas no atributo RESID_LOCAL_POSTAL por possuírem um valor superior comparado as demais. Estas são maioritariamente problemas de inserção de dados. Por exemplo:
   (a) Na BDIC: São João dos Montes e na AT: São João Montes
   (b) Na BDIC: Duas Igrejas PNF e na AT: Duas Igrejas
2. Quanto às incertezas, o maior grau encontra-se nos atributos de residência. Dos quais foi possível observar:
   (a) Do total existente nos atributos RESID_DT, RESID_MN e RESID_FR, consistem em emparelhamentos onde o valor destes atributos na fonte AT é indefinido;
   (b) Um total de 61.206 registos cujo o valor dos atributos RESID_CP4 e RESID_CP3 encontram-se indefinidos na AT
   (c) Um total de 62.091 registos cujo o valor do atributo RESID_LOCAL_POSTAL encontra-se indefinido na BDIC

Os emparelhamentos referidos na Figura 4.4, foram validados por um especialista do INE usando os atributos de ligação NIC e NIF, cruzados com as fontes de dados IISS 2016, ACSS 2015 e CGA 2015, tendo concluído o seguinte:

1. No cruzamento das relações CLASS_PROB_MAX_BDIC_AT_2015 e IISS (2016) foi possível validar 10.497 registos,
2. Com a CGA (2015) foram validados 205 registos,
3. Com o IEFP (2015) foram validados 528 registos e
4. Com a ACSS (2015) foram validados 337 registos.

4.5 Validação dos Emparelhamentos Realizados pelo INE

A Tabela 4.9 apresenta os resultados obtidos no processo de validação e avaliação de emparelhamentos. Inicialmente, foram observados na Matriz os vários pares de emparelhamentos nela existentes com base nas flags abordadas na Secção 4.4, tendo sido obtido o total de registos refletidos em (a).

Tabela 4.9: Estatísticas da validação dos emparelhamentos realizados pelo INE

Fonte de Dados	Nº registos	Emparelhamentos na Matriz (a)	Registos validados (b)	Registos validados (%) (c)
BDIC (2015)	11.825.786	8.736.100	8.155.177	93,35
AT (2014)	9.370.879			
BDIC (2015)	11.825.786	6.393.870	6.124.895	95,79
IISS (2015)	6.927.720			
BDIC (2015)	11.825.786	1.635.819	1.603.274	98,01
EDUC (2015)	1.680.018			
BDIC (2015)	11.825.786	712.163	677.836	95,17
IEFP (2015)	686.198			
BDIC (2015)	11.825.786	991.296	979.001	98,75
CGA (2015)	1.032.133			
SEF (2015)	383.764	143.950	120.404	83,64
IISS (2015)	6.927.720			
SEF (2015)	383.764	185.605	171.227	92,25
AT (2014)	9.186.325			
SEF (2015)	383.764	16.848	10.576	62,77
EDUC (2015)	87.017			

(a) Refere-se ao total de emparelhamentos por par de fontes realizados pelo INE, verificados com base na ativação das flags (consultar a Secção 4.4).
(b) Refere-se ao total de registos validados pela nossa abordagem por par de fontes
(c) Total percentual de registos validados.

A validação dos registos refletidos em (a) é feita com base na seguinte estratégia:

1. Consideram-se como correctos os emparelhamentos obtidos pelo cruzamento do par de relações a emparelhar usando um identificador comum.

2. Os registos obtidos pelo emparelhamento probabilístico e encontrados na Matriz usando o método abordado no ponto 1 da Secção 4.4 são igualmente considerados correctos.

O total de registos referidos em (b) representa o somatório destes dois conjuntos de emparelhamentos, os quais podemos considerar como sendo verificados.
Com base nesta premissa, é possível observar que as percentagens de validação e avaliação encontram-se entre os 62,77 à 98,75%. Comparado com os resultados obtidos pelo INE, observa-se igualmente que o número de emparelhamentos por eles encontrados é superior. Isto deve-se ao facto dos seus emparelhamentos terem sido obtidos aproveitando emparelhamentos realizados entre outros pares de fontes de dados.
O emparelhamento SEF-EDUC possui um valor abaixo dos demais, o que supõe-se que é devido ao

facto de ter sido usado o modelo SEF-IISS a fim de realizar os emparelhamentos deste par, tal como abordado na Secção 4.3.

4.6 Sumário

Neste capítulo, foi apresentada a metodologia proposta para a monitorização da qualidade dos dados das fontes disponíveis no INE, assim como os emparelhamentos produzidos pelo software do processo desenvolvido segundo a metodologia que possui quatro actividades principais:

1. Limpeza e Normalização,
2. Blocking,
3. Treino e
4. Classificação de Emparelhamentos.

A actividade de Limpeza e Normalização é crucial para decidir que critérios utilizar na etapa de Blocking e de modo geral avaliar o estado dos dados a serem usados para emparelhamento. É monitorada com base nas métricas de Completude, Null Count e valores distintos. O nível de qualidade nesta etapa pode ser medido analisando as variações das referidas métricas relativamente a versões dos anos anteriores de uma determinada fonte de dados, permitindo determinar se houve ou não melhorias comparado a actual versão.
A actividade de Blocking é monitorada com base na taxa de redução do número de registos a emparelhar por critérios de similaridade face ao total de combinações possíveis. De acordo com o critério de Blocking usado, foi possível obter em média três registos candidatos em cada bloco, o que torna o custo computacional tratável.
A qualidade da etapa de treino foi avaliada com base na técnica 2- fold Cross Validation aplicada ao conjunto de exemplos (positivos e negativos) usados para treinar o modelo. Foram testadas as técnicas 5 e 10 - *fold Cross Validation* usando o mesmo número de exemplos comparado a validação com a 2- *fold Cross Validation* e verificou-se que os resultados das métricas de precisão e abrangência permaneceram inalterados, com execpção de gastarem mais CPU comparado ao gasto pelo 2- *fold Cross Validation*.
Quanto à qualidade dos modelos de Classificação, de modo geral encontram-se dentro do nível pretendido (mínimo de 3% de erro de abrangência), 0.5% acima do erro de cobertura verificado em todo o País nos Censos 2011, com excepção dos modelos gerados com a EDUC e com a CGA que possuem uma taxa de abrangência abaixo dos 97%. Quanto a generalização do modelo aos dados não emparelhados (Classe= -1), serão necessárias optimizações a fim de melhorar os resultados comuns entre os

pares de fontes (93,18% melhor caso e no pior caso 20,17%).

O nível de qualidade de dados nas fontes em geral é satisfatório em algumas fontes de dados como por exemplo a BDIC, AT e IISS, mas em algumas como por exemplo a EDUC existem vários atributos do indivíduo com uma taxa de completude muito inferior.

5

Conclusões e Trabalhos Futuros

Conteúdo

O trabalho realizado no âmbito desta dissertação teve como principais propósitos propor:

1. Uma metodologia para o emparelhamento de registos das diferentes fontes de dados disponíveis no Instituto Nacional de Estatística (INE) recorrendo a métodos probabilísticos;
2. Um método para a avaliação da qualidade dos emparelhamentos realizados entre os registos provenientes das diversas fontes ao longo das várias etapas da metodologia preconizada.

A metodologia de emparelhamento de registos e o método de avaliação da qualidade estão concretizados numa plataforma disponível no INE numa uma solução constituída por dois processos:

- Processo de Produção da Informação
- Processo de Monitorização.

Ambos os processos são constituídos por um conjunto de etapas que funcionam em paralelo:

- Limpeza e Normalização,
- Blocking,
- Comparação de Registos,
- Treino
- Classificação de Emparelhamentos.

Cada uma destas componentes possui interfaces bem definidas para permitir aplicar, substituir ou combinar algoritmos de forma independente.

Foram emparelhados oito pares de fontes de informação disponibilizadas ao INE, dentre os quais cinco referem-se a emparelhamentos realizados com a Base de Dados da Identificação Civil (BDIC) e três pares referem-se a emparelhamentos realizados com a Base de Dados do Serviço de Estrangeiros e Fronteiras (SEF).
A realização desta dissertação permite concluir que a qualidade dos dados das fontes é boa na generalidade das fontes, com excepção a algumas em que a completude é bastante reduzida em vários atributos (em particular a da Caixa Geral de Aposentações (CGA) e Educação e Ciência (EDUC)).
A qualidade dos modelos de classificação produzidos, recorrendo a regressão logística, com abrangência com erro máximo de 3%, está em geral dentro do que seria necessário atingir considerando como *baseline* o erro de cobertura do inquérito de qualidade dos Censos 2011 (2,5%). Observa-se uma excepção ao nível de abrangência nos modelos treinados com a fontes da EDUC e CGA que possuem um erro superior aos 3% observados nos restantes emparelhamentos.

Dos novos emparelhamentos encontrados, espera-se que venham a permitir acrescentar um número substancial de ligações à Base de População Residente do INE (BPR), na ordem de 64,94% dos 401.829 registos não emparelhados com a Autoridade Tributária (AT) e 19,21% dos 248.953 registos não emparelhados com o Instituto de Informática da Segurança Social (IISS), aplicando os critérios antes usados pelo INE com o método exacto. O total de registos não emparelhados referidos nos dois casos encontram-se referenciados no Relatório sobre a Metodologia de actualização da Base de População Residente - Construção da BPR 2015 (QUAR 2016) [INE,Gabinete dos Censos 2021, 2016a].

Os emparelhamentos produzidos com o processo descrito nesta dissertação foram também usados na validação dos emparelhamentos antes realizados pelo INE recorrendo a métodos exactos. A validação permite concluir que:

- **Para os emparelhamentos realizados com a BDIC:** foram validados no melhor caso 98,75% do total de emparelhamentos feitos pelo INE por métodos exactos (emparelhamento entre as fontes BDIC - CGA) e no pior caso foram validados 93,35% (fontes BDIC - AT). Sendo a AT uma fonte transversal das demais, tais como a IISS, EDUC, Instituto do Emprego e Formação Profissional (IEFP) e CGA, muitos dos emparelhamentos validados por estes pares podem ser aproveitados para melhorar o total percentual dos emparelhamentos entre a BDIC e a AT, refletindo-se igualmente para os casos dos restantes pares de emparelhamentos.
- **Para os emparelhamentos realizados com o SEF:** foram validados no melhor caso 92,25% do total de emparelhamentos feitos INE por métodos exactos (refere-se ao emparelhamento entre as fontes SEF - AT) e no pior caso foram validados 62,77% (refere-se ao emparelhamento entre as fontes SEF - EDUC).

De entre as contribuições resultantes do meu trabalho, foi possível produzir um método de emparelhamento probabilístico que, em resultado da técnica de blocking aplicada, permite seleccionar com grande exactidão os potenciais candidatos ao emparelhamento. A abordagem permite obter em média cerca de 3 candidatos similares em cada bloco. Apesar de ter um custo computacional superior (há que analisar o triplo dos pares de registos comparativamente ao método exacto) ainda assim mantém-se tratável e com a vantagem de garantir maior número de emparelhamentos correctos em relação ao método exacto, sem degradar significativamente o erro.

A utilização do avaliador da qualidade dos emparelhamentos produzidos, permite aferir quão bons ou maus foram os métodos usados para o emparelhamento de registos. A título de exemplo, a avaliação dos novos emparelhamentos produzidos entre o par de fontes BDIC e AT com base nas inconsistências entre os pares de atributos demográficos mais significativos do indivíduo permite concluir que, do total de 244.903 novos emparelhamentos, podem ser excluídos não mais do que 1.087% destes registos caso o atributo que identifica a nacionalidade do indivíduo for considerado como de maior relevância

em termos de identificação do indivíduo ou uma percentagem menor caso forem considerados os outros atributos, diferentes dos de residência, uma vez que as residências nem sempre são fixas.
Com base nos resultados obtidos, é possível sustentar a viabilidade do uso da metodologia e o software de emparelhamento probabilístico para as fontes de dados administrativas disponíveis no INE. Importa igualmente referir que o presente trabalho é na primeira versão do software, sendo assim serão necessárias melhorias significativas na metodologia, assim como toda a componente de software, afim de permitir ao INE construir a sua BPR de forma rápida e eficaz.

5.1 Trabalhos Futuros

Esta dissertação constitui uma primeira abordagem para avaliação da viabilidade do recurso a métodos probabilísticos no emparelhamento de dados administrativos para fins censitários. Há muitas melhorias e desenvolvimentos adicionais a considerar antes de o sistema desenvolvido vir a ser usado em produção. Desde logo, poderiam ser consideradas as melhorias ao processo abaixo descritas.

Implementação de uma estratégia que permita a combinação de múltiplos critérios de Blocking a fim de melhorar a actividade de geração de candidatos. O uso de múltiplos critérios, permitirá reduzir a exclusão de registos do processo de emparelhamento, causado por alguma inconsistência nos valores do(s) atributo(s) escolhidos para realizar o Blocking quando é aplicado um único critério.

Melhoramento do método de geração de exemplos negativos para o treino do modelo de Classificação. Poderiam ser escolhidos como exemplos negativos os dois ou três exemplos em cada bloco que melhor se assemelham ao exemplo positivo conhecido em cada bloco. Esta estratégia permitiria equilibrar a quantidade de exemplos positivos e negativos para cada modelo a treinar, uma vez que podem existir potencialmente blocos com múltiplos exemplos negativos para cada exemplo positivo, levando a que uma classe seja fortemente maioritária e enviese o modelo treinado.

Disponibilizar uma interface gráfica com um *Dashboard* para a invocação das etapas de processamento/análise e visualização das estatísticas dos dados e emparelhamentos realizados. A solução existente funciona em linha de comandos e não permite monitorar os dados de forma interactiva e amigável em todas as etapas da metodologia. A interface poderia ainda permitir a revisão administrativa (*Clerical Review*) de forma interactiva e fácil para decidir os emparelhamentos correctos dentro de um conjunto (por exemplo, emparelhamentos duplicados), onde os modelos de aprendizagem não têm forma de decidir. Esta interface deveria igualmente permitir a inserção dos registos classificados como

emparelhamentos numa relação temporária ou definitiva (por exemplo a *Matriz*);

Re-aproveitar os emparelhamentos entre pares de fontes de dados a fim de aumentar o número de ligações entre as várias fontes de dados. Tendo em conta que os pares de fontes emparelhados possuem identificadores diferentes para o mesmo indivíduo, os emparelhamentos podem ser cruzados a fim de obter um maior número de ligações. Por exemplo, considerando os emparelhamentos BDIC - AT (identificados por NIC e NIF) e BDIC - IISS (identificados por NIC e NISS), o cruzamento destes emparelhamentos permitirá obter os três identificadores fornecidos pelo Cartão de Cidadão Português (NIC, NIF, NISS), o que identifica o cidadão de maneira unívoca.

Os emparelhamentos realizados com as versões dos anos anteriores das fontes de dados poderiam também ser aproveitados de forma a serem usados como indícios para os emparelhamentos nas versões actuais, o que reduzirá o custo computacional e por outro lado reduzirá igualmente o número de falsos negativos considerando o facto de que os dados podem sofrer alterações positivas ou negativas ao longo dos anos.

Finalmente, tendo em conta que a qualidade dos dados em algumas fontes apresenta ainda bastante potencial para ser melhorado (vários atributos não preenchidos), poderão surgir casos em que os registos classificados como não emparelhados por um par de fonte de dados sejam classificados como emparelhados noutros pares. Uma das formas de resolver este problema, passaria por cruzar os registos classificados como não emparelhados num par de fontes com os registos classificados como emparelhados noutro outro par. Esta medida permitiria reduzir o erro de abrangência na etapa de classificação.

Bibliografia

[Alpaydin, 2004] Alpaydin, E. (2004). *Introduction to Machine Learning*, volume 53. MIT Press, London, Engand.

[Batini and Scannapieco, 2016] Batini, C. and Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer International Publishing.

[Bilenko and Mooney, 2003] Bilenko, M. and Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 39–48.

[Bleiholder and Naumann, 2006] Bleiholder, J. and Naumann, F. (2006). Conflict Handling Strategies in an Integrated Information System. *Proceedings of the IJCAI Workshop on Information on the Web*, (197):1–13.

[Cecchin et al., 2010] Cecchin, F., de Aguiar Ciferri, C., and Hara, C. (2010). XML data fusion. In *International Conference on Data Warehousing and Knowledge Discovery*, page 297–308.

[Christen, 2011] Christen, P. (2011). A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24:1–20.

[Christen, 2012] Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage,Entity Resolution and Duplicate Detection*. Springer-Verlag Berlin Heidelberg.

[CNPD, 2014] CNPD (2014). Deliberação da CNPD nº 929/2014 para os Censos 2021. Technical report, Comissão Nacional de Protecção de Dados (CNPD), Lisboa. URL: https://goo.gl/HDiFrD (acesso em 17 Out. 2016).

[Doan et al., 2012] Doan, A., Halevy, A., and Ives, Z. (2012). *Principles of Data Integration*. Elsevier, Inc.

[Elmagarmid et al., 2007] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate Record Detection : A Survey. *IEEE Transactions on knowledge and data engineering*, 19(1):1–16.

[Feigenbaum, 2016] Feigenbaum, J. J. (2016). A Machine Learning Approach to Census Record Linking. URL: https://goo.gl/NNYx36. pages 1–34.

[Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*, volume 12. Morgan Kaufmann, 2nd edition.

[Harald, 2000] Harald, U. (2000). Population and Housing Censuses in Norway Towards a Register Based Solution. Technical Report 3, Statistical Office of the European Communities (EUROSTAT), Geneva. URL: https://goo.gl/JYTX3Q (Acesso em 25 Jan. 2017).

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, volume 1. Springer, second edition.

[Hernández and Stolfo, 1995] Hernández, M. A. and Stolfo, S. J. (1995). The Merge/Purge Problem for Large Databases. In *ACM Sigmod Record*, pages 127–138. ACM.

[INE, 2012] INE (2012). Censos 2011, Resultados Definitivos. Technical report, Instituto Nacional de Estatística (INE), Lisboa. URL: https://goo.gl/XuYkvZ (acesso em 27 Set. 2017), isbn = 9789892501857.

[INE,Gabinete dos Censos 2021, 2015] INE,Gabinete dos Censos 2021 (2015). Interligação das diferentes bases de dados provenientes de fontes administrativas , no âmbito do novo modelo censitário para 2021 (Relatório QUAR *). Technical report, Instituto Nacional de Estatística, Lisboa. URL: https://goo.gl/fmo6Z7 (acesso em Nov. 2016).

[INE,Gabinete dos Censos 2021, 2016a] INE,Gabinete dos Censos 2021 (2016a). Metodologia de atualização da Base de População Residente - Construção da BPR 2015. Technical report, Instituto Nacional de Estatística, Lisboa. URL: https://goo.gl/fmo6Z7 (acesso em 02 Nov. 2016).

[INE,Gabinete dos Censos 2021, 2016b] INE,Gabinete dos Censos 2021 (2016b). Novo modelo censitário - Estudo de viabilidade Programa de Trabalho. Technical report, Instituto Nacional de Estatística, Lisboa. URL: https://www.ine.pt/xurl/doc/265780886 (acesso em 10 Out. 2016).

[Kulkarni and Bakal, 2014] Kulkarni, P. S. and Bakal, J. W. (2014). Survey on Data Cleaning. *International Journal of Engineering Science and Innovative Technology (IJESIT)*, 3(4):615–620.

[March et al., 2015] March, B., Bycroft, C., and Nz, S. (2015). Census Transformation : progress in New Zealand. Technical Report March, Stats NZ, URL: https://goo.gl/KMspw2 (Acesso em 02 Out. 2016).

[Michelson and Knoblock, 2006] Michelson, M. and Knoblock, C. A. (2006). Learning Blocking Schemes for Record Linkage $*$. *American Associaton for Artificial Intelligence*, pages 440–445.

[Minton et al., 2005] Minton, S. N., Nanjo, C., Knoblock, C. A., Michalowski, M., and Michelson, M. (2005). A heterogeneous field matching method for record linkage. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 314–321.

[Office for National Statistics, 2014a] Office for National Statistics (2014a). Beyond 2011: Matching Anonymous Data. Technical Report July 2013, Office for National Statistics, URL: https://goo.gl/shbwxw (Acesso em 04 Out. 2016).

[Office for National Statistics, 2014b] Office for National Statistics (2014b). SAPE15DT1 - Lower Super Output Area Mid-Year Population Estimates, formatted, Mid-2013 - Superseded. URL: https://goo.gl/9E5XT4 (Acesso em 25 Jan. 2017).

[Office for National Statistics, 2015] Office for National Statistics (2015). ONS Census Transformation Programme Administrative Data Update. Technical Report October 2015, Office for National Statistics, URL: https://goo.gl/59Sfgc (Acesso em 04 Out. 2016).

[Office for National Statistics, 2016a] Office for National Statistics (2016a). Beyond 2011 (Census Transformation Programme). URL: https://goo.gl/W2enJP (Acesso em 04 Out. 2016).

[Office for National Statistics, 2016b] Office for National Statistics (2016b). Census Transformation Programme. Annual assessment of ONS' progress towards an Administrative Data Census post 2021. May 2016. Technical Report May, URL:https://goo.gl/nNja21 (acesso em 04 Out. 2016).

[Office for National Statistics, 2016c] Office for National Statistics (2016c). Statement of administrative sources.

[Riddle, 1997] Riddle, P. (1997). Learning Sets of Rules Learning Disjunctive Sets of Rules. *Machine Learning*, pages 229–253.

[Rieck, 2011] Rieck, K. (2011). Similarity measures for sequential data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):296–304.

[Silva, Rui, 2017] Silva, Rui (2017). Matching census data records. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa.

[Skinner et al., 2013] Skinner, C., Hollis, J., and Murphy, M. (2013). Beyond 2011 : Independent Review of Methodology. Technical report, Office for National Statistics (ONS), URL: https://goo.gl/M8BSR6 (Acesso em 25 Jan. 2017).

[Trépanier et al., 2013] Trépanier, J., Pignal, J., and Royce, D. (2013). Administrative Data Initiatives at Statistics Canada. In *2013 Federal Committee on Statistical Methodology Research Conference*.

[Winkler, 1994] Winkler, W. E. (1994). Advanced methods for record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

[Winkler, 1995] Winkler, W. E. (1995). Matching and Record Linkage. In *Business survey methods*, volume 1, pages 355–384.

[Winkler, 2005] Winkler, W. E. (2005). Approximate String comparator search strategies for very large administrative lists. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 2.

[Winkler, 2006] Winkler, W. E. (2006). Overview of record linkage and current research directions. Technical Report 2006-2, US. Bureau of the Census, Washington, DC. URL: https://goo.gl/mCzFG4.